Editors D. J. Balding • M. Bishop • C. Cannings

Handbook of
Statistical Genetics

Third Edition

Volume 1

Handbook of Statistical Genetics

Handbook of Statistical Genetics

Third Edition

Volume 1

Handbook of Statistical Genetics

Third Edition

Volume 1

Editors:

D. J. Balding

Imperial College of Science, Technology and Medicine, London, UK

M. Bishop

CNR-ITB, Milan, Italy

C. Cannings

University of Sheffield, UK

Contents

Part 3   EVOLUTIONARY GENETICS   375

12   Adaptive Molecular Evolution   377
Z. Yang

VOLUME 2

26   Inference Under the Coalescent   878
        M. Stephens

36 Population Association   1216

D. Clayton

37 Whole Genome Association   1238

A.P. Morris and L.R. Cardon

38 Family-based Association   1264

F. Dudbridge

Contributors

R.E. Ashcroft
Queen Mary's School of Medicine
and Dentistry
University of London
London, UK

G. Barbujani
Dipartimento di Biologia
ed Evoluzione
Università di Ferrara
Ferrara, Italy

M.A. Beaumont
School of Biological Sciences
University of Reading
Reading, UK

J.M. Bowden
Departments of Health Sciences
and Genetics
University of Leicester
Leicester, UK

J.P. Bollback
Department of Biology
University of Rochester
Rochester, NY
USA

J.F.Y. Brookfield
Institute of Genetics
School of Biology
University of Nottingham
Nottingham, UK

P.R. Burton
Departments of Health Sciences
and Genetics
University of Leicester
Leicester, UK

C. Cannings
Division of Genomic Medicine
University of Sheffield
Sheffield, UK

L.R. Cardon
Wellcome Trust Centre for Human
Genetics
University of Oxford
Oxford, UK

C. Cheng
Department of Biostatistics
St. Jude Children's Research Hospital
Memphis, TN
USA

L. Chikhi
Laboratoire Evolution et Diversité
Biologique
Université Paul Sabatier
Toulouse, France

D. Clayton
Cambridge Institute for Medical
Research
University of Cambridge
Cambridge, UK

M. De Iorio
Division of Epidemiology,
Public Health and Primary Care
Imperial College
London, UK

J. Dicks
Department of Computational
and Systems Biology
John Innes Centre
Norwich, UK

F. Dudbridge
MRC Biostatistics Unit
Institute for Public Health
Cambridge, UK

T.M.D. Ebbels
Division of Surgery, Oncology
Reproductive Biology and Anaesthetics
Imperial College
London, UK

L. Excoffier
Zoological Institute
Department of Biology
University of Berne
Berne, Switzerland

D. Gianola
Department of Animal Sciences
Department of Biostatistics
and Medical Informatics
Department of Dairy Science
University of Wisconsin
Madison, WI
USA

N. Goldman
EMBL-European Bioinformatics Institute
Hinxton, UK

M.D. Hendy
Allan Wilson Center for
Molecular Ecology and Evolution
Massey University
Palmerston North, New Zealand

B.R. Holland
Allan Wilson Center for
Molecular Ecology and Evolution
Massey University
Palmerston North, New Zealand

P. Holmans
Department of Psychological
Medicine
Cardiff University
Cardiff, UK

I. Höschele
Virginia Bioinformatics Institute and
Department of Statistics
Virginia Polytechnic Institute and
State University
Blacksburg, VA, USA

F. Hospital
INRA, Université Paris Sud
Orsay, France

W. Huber
Department of Molecular Genome
Analysis
German Cancer Research Center
Heidelberg, Germany

J.P. Huelsenbeck
Department of Biology
University of Rochester
Rochester, NY
USA

R.C. Jansen
Groningen Bioinformatics Centre
University of Groningen
Groningen, The Netherlands

D.P. Klose
Division of Mathematical Biology
National Institute of Medical
Research
London, UK

S.L. Lauritzen
Department of Statistics
University of Oxford
Oxford, UK

A. Lewin
Division of Epidemiology,
Public Health and Primary Care
Imperial College
London, UK

S. Lin
Department of Statistics
Ohio State University
Columbus, OH
USA

Jun S. Liu
Department of Statistics
Harvard University
Cambridge, MA
USA

T. Logvinenko
Department of Statistics
Harvard University
Cambridge, MA
USA

A.S. Macdonald
Department of Actuarial Mathematics
and Statistics
Heriot-Watt University
Edinburgh, UK

P.M. McKeigue
Conway Institute
University College Dublin
Dublin, Ireland

G. McVean
Department of Statistics
Oxford University
Oxford, UK

L. Moreau
INRA, UMR de Génétique Végétale
Ferme du Moulon
France

A.P. Morris
Wellcome Trust Centre for Human Genetics
University of Oxford
Oxford, UK

C. Neuhauser
Department of Ecology,
Evolution and Behavior
University of Minnesota
Saint Paul, MN
USA

M. Nordborg
Molecular and Computational Biology
University of Southern California
Los Angeles, CA
USA

A. Onar
Department of Biostatistics
St. Jude Children's Research Hospital
Memphis, TN
USA

W.R. Pearson
Department of Biochemistry
University of Virginia
Charlottesville, VA
USA

D. Penny
Allan Wilson Center for
Molecular Ecology and Evolution
Massey University
Palmerston North, New Zealand

S.B. Pounds
Department of Biostatistics
St. Jude Children's Research Hospital
Memphis, TN
USA

S. Richardson
Division of Epidemiology,
Public Health and Primary Care
Imperial College
London, UK

F. Rousset
Laboratoire Génétique
et Environnement
Institut des Sciences de l'Évolution
Montpellier, France

G. Savva
Centre for Environmental
and Preventive Medicine
Wolfson Institute
of Preventive Medicine
London, UK

E.E. Schadt
Rosetta Inpharmatics, LLC
Seattle, WA
USA

N.A. Sheehan
Department of Health Sciences
and Genetics
University of Leicester
Leicester, UK

S.K. Sieberts
Rosetta Inpharmatics, LLC
Seattle, WA
USA

K.D. Siegmund
Department of Preventive Medicine
Keck School of Medicine
University of Southern California
Los Angeles, CA
USA

V. Solovyev
Department of Computer Science
University of London
Surrey, UK

T.P. Speed
Department of Statistics
University of California at Berkeley
Berkeley, CA
USA

and

Genetics and Bioinformatics Group
The Walter & Eliza Hall Institute
of Medical Research
Royal Melbourne Hospital
Melbourne, Australia

D.A. Stephens
Department of Mathematics and
Statistics
McGill University
Montreal, Canada

M. Stephens
Departments of Statistics
and Human Genetics
University of Chicago
Chicago, IL
USA

W.R. Taylor
Division of Mathematical Biology
National Institute of Medical
Research
London, UK

M.D. Teare
Mathematical Modelling and Genetic
Epidemiology
University of Sheffield
Medical School
Sheffield, UK

A. Thomas
Department of Biomedical
Informatics
University of Utah
Salt Lake City, UT
USA

E.A. Thompson
Department of Statistics
University of Washington
Seattle, WA
USA

J.L. Thorne
Departments of Genetics
and Statistics
North Carolina State University
Raleigh, NC
USA

M.D. Tobin
Departments of Health Sciences
and Genetics
University of Leicester
Leicester, UK

M. Vingron
Department of Computational Molecular
Biology
Max-Planck-Institute
for Molecular Genetics
Berlin, Germany

A. von Heydebreck
Department of Computational Molecular
Biology
Max-Planck-Institute
for Molecular Genetics
Berlin, Germany

B. Walsh
Department of Ecology and
Evolutionary Biology
Department of Plant Sciences
Department of Molecular
and Cellular Biology
University of Arizona
Tuscon, AZ
USA

B.S. Weir
Department of Biostatistics
University of Washington
Seattle, WA
USA

J. Whittaker
Department of Epidemiology and
Population Health
London School of Hygiene &
Tropical Medicine
London, UK

T.C. Wood
Department of Biochemistry
University of Virginia
Charlottesville, VA
USA

Z. Yang
Department of Biology
University College London
London, UK

H. Zhao
Department of Epidemiology and
Public Health
Yale University School of Medicine
New Haven, CT
USA

Editor's Preface to the Third Edition

In the four years that have elapsed since the highly successful second edition of the *Handbook of Statistical Genetics*, the field has moved on, in some areas dramatically. This is reflected in the present thorough revision and comprehensive updating: 17 chapters are entirely new, 6 providing a fresh approach to topics that had been covered in the second edition, while 11 new chapters cover areas of recent growth, or important topics not previously addressed. These new topics include microarray data analysis (two new chapters to complement the existing one), eQTL analyses and metabonomics. There are also new chapters on graphical models and on pedigrees and genealogies, admixture mapping and genome-wide association studies, cancer genetics, epigenetics, and genetical aspects of insurance. Of the 26 chapters carried over from the second edition, 21 have been revised, among which 5 very substantial revisions are close to being new chapters.

It will be clear from the topics listed above that we continue to define statistical genetics very broadly. Statistics for us goes beyond mathematical models and techniques, and includes the management and presentation of data, as well as its analysis and interpretation. Genetics includes the search for and study of genes implicated in human health and the economic value of plants and animals, the evolution of genes within natural populations, the evolution of genomes and of species, and the analysis of DNA, RNA, gene expression, protein sequence and structure, and now metabonomics. The latter topics probably fall outside even a liberal definition of 'genetics', but we believe they will be of interest to our readers because of their relevance to studies of gene function and because of the statistical methods being used.

We regard more recent terms, such as 'genomics' and 'transcriptomics' as designating new avenues within genetics, rather than as entirely new fields, and we include their statistical aspects as part of statistical genetics. Similarly, we include much of 'bioinformatics', but we do not systematically survey the available genetic databases or computer software, nor methods and protocols for archiving and annotating genetic data. Some pointers to computer software and other Internet resources are given at the end of relevant chapters.

The 43 chapters are intended to be largely independent, so that to benefit from the handbook it is not necessary to read every chapter, nor read chapters sequentially. This structure necessitates some duplication of material, which we have tried to minimise but could not always eliminate. Alternative approaches to the same topic by different authors can convey benefits. The extensive subject and author indexes allow easy reference to topics covered in different chapters.

For those with minimal genetics background, the glossary of genetic terms (newly updated) should be of assistance, while Wiley's *Biostatistical Genetics and Genetic Epidemiology*, edited by Elston, Olson, and Palmer (2002) provides a more substantial

resource of definitions and explanations of key terms from both genetics and statistics. For those seeking a more substantial introduction to the foundations of modern statistical methods applied in genetics, we suggest *Likelihood, Bayesian, and MCMC methods in Quantitative Genetics* by Sorenson and Gianola (2003).

We thank the many commentators of the first two editions who were generous in their praise. We have tried to take on board many of the constructive criticisms and suggestions. No doubt many more improvements will be possible for future editions and we welcome comments e-mailed to any of the editors. We are grateful to all of our outstanding set of authors for taking the time to write and update their chapters with care, and for meeting their deadlines (and sometimes ours as well). Finally, we would like to express our appreciation to the staff of John Wiley & Sons for initially proposing the project to us, and for their friendly professionalism in the preparation of both editions. In particular, we thank Martine Bernardes-Silva, Layla Harden, and Kathryn Sharples.

<div align="right">

DAVID BALDING
MARTIN BISHOP
CHRIS CANNINGS
August 2007

</div>

Glossary of Terms

GLOSSARY OF GENETIC TERMS:

(prepared by Gurdeep Sagoo, University of Sheffield, UK)

N.B. Some of the definitions below assume that the organism of interest is diploid.

Adenine (A): purine base that forms a pair with thymine in DNA and uracil in RNA.

Admixture: arises when two previously isolated populations begin interbreeding.

Allele: one of the possible forms of a gene at a given locus. Depending on the technology used to type the gene, it may be that not all DNA sequence variants are recognised as distinct alleles.

Allele frequency: often used to mean the relative frequency (i.e. proportion) of an allele in a sample or population.

Allelic association: the non-independence, within a given population, of a gamete's alleles at different loci. Also commonly (and misleadingly) referred to as *linkage disequilibrium.*

Alpha helix: a helical (usually right-handed) arrangement that can be adopted by a polypeptide chain; a common type of protein secondary structure.

Amino acid: the basic building block of proteins. There are 20 naturally occurring amino acids in animals which when linked by peptide bonds form polypeptide chains.

Aneuploid cells: do not have the normal number of chromosomes.

Antisense strand: the DNA strand complementary to the coding strand, determined by the covalent bonding of A with T and C with G.

Ascertainment: the strategy by which individuals are identified, selected, and recruited for participation in a study.

Autosome: A chromosome other than the sex chromosomes. Humans have 22 pairs of autosomes plus 2 sex chromosomes.

Backcross: A linkage study design in which the progeny (F1s) of a cross between two inbred lines are crossed back to one of the inbred parental strains.

Bacterial Artificial Chromosome (BAC): a vector used to clone a large segment of DNA (100−200 Kb) in bacteria resulting in many copies.

Base: (abbreviated term for a purine or pyrimidine in the context of nucleic acids), a cyclic chemical compound containing nitrogen that is linked to either a deoxyribose (DNA) or a ribose (RNA).

Base pair (bp): a pair of bases that occur opposite each other (one in each strand) in double stranded DNA/RNA. In DNA adenine base pairs with thymine and cytosine with guanine. RNA is the same except that uracil takes the place of thymine.

Bayesian: A statistical school of thought that, in contrast with the frequentist school, holds that inferences about any unknown parameter or hypothesis should be encapsulated in a probability distribution, given the observed data. Bayes Theorem allows one to compute the posterior distribution for an unknown from the observed data and its assumed prior distribution.

Beta-sheet: is a (hydrogen-bonded) sheet arrangement which can be adopted by a polypeptide chain; a common type of protein secondary structure.

centiMorgan (cM): measure of genetic distance. Two loci separated by 1 cM have an average of 1 recombination between them every 100 meioses. Because of the variability in recombination rates, genetic distance differs from physical distance, measured in base pairs. Genetic distance differs between male and female meioses; an average over the sexes is usually used.

Centromere: the region where the two sister chromatids join, separating the short (p) arm of the chromosome from the long (q) arm.

Chiasma: the visible structure formed between paired homologous chromosomes (non-sister chromatids) in meiosis.

Chromatid: a single strand of the (duplicated) chromosome, containing a double-stranded DNA molecule.

Chromatin: the material composed of DNA and chromosomal proteins that makes up chromosomes. Comes in two types, euchromatin and heterochromatin.

Chromosome: the self-replicating threadlike structure found in cells. Chromosomes, which at certain stages of meiosis and mitosis consist of two identical sister chromatids, joined at the centromere, and carry the genetic information encoded in the DNA sequence.

cis-Acting: regulatory elements and eQTL whose DNA sequence directly influences transcription. The physical location for cis-acting elements will be in or near the gene or genes they regulate.

Clones: genetically engineered identical cells/sequences.

Co-dominance: both alleles contribute to the phenotype, in contrast with recessive or dominant alleles.

Codon: a nucleotide triplet that encodes an amino acid or a termination signal.

Common disease common variant (CDCV) hypothesis: The hypothesis that many genetic variants underlying complex diseases are common, and hence susceptible to detection using current population association study designs. An alternative possibility is that genetic contributions to the causation of complex diseases arise from many variants, all of which are rare.

complementary DNA (cDNA): DNA that is synthesised from a messenger RNA template using the reverse transcriptase enzyme.

Contig: a group of contiguous overlapping cloned DNA sequences.

Cytosine (C): pyrimidine base that forms a pair with guanosine in DNA.

Degrees of freedom (df): This term is used in different senses both within statistics and in other fields. It can often be interpreted as the number of values that can be defined arbitrarily in the specification of a system; for example, the number of coefficients in a regression model. Frequently it suffices to regard df as a parameter used to define certain probability distributions.

Deoxyribonucleic acid (DNA): polymer made up of deoxyribonucleotides linked together by phosphodiester bonds.

Deoxyribose: the sugar compound found in DNA.

Diploid: has two versions of each autosome, one inherited from the father and one from the mother. Compare with haploid.

Dizygotic twins: twins derived from two separate eggs and sperm. These individuals are genetically equivalent to full sibs.

DNA methylation: the addition of a methyl group to DNA. In mammals this occurs at the C-5 position of cytosine, almost exclusively at CpG dinucleotides.

DNA microarray: small slide or 'chip' used to simultaneously measure the quantity of large numbers of different mRNA gene transcripts present in cell or tissue samples.

Depending on the technology used, measurements may either be absolute or relative to the quantities in a second sample.

Dominant allele: results in the same phenotype irrespective of the other allele at the locus.

Effective population size: The size of a theoretical population that best approximates a given natural population under an assumed model. The criterion for assessing the 'best' approximation can vary, but is often some measure of total genetic variation.

Enzyme: a protein that controls the rate of a biochemical reaction.

Epigenetics: the transmission of information on gene expression to daughter cells at cell division.

Epistasis: the physiological interaction between different genes such that one gene alters the effects of other genes.

Epitope: the part of an antigen that the antibody interacts with.

Eukaryote: organism whose cells include a membrane-bound nucleus. Compare with prokaryote.

Exons: parts of a gene that are transcribed into RNA and remain in the mature RNA product after splicing. An exon may code for a specific part of the final protein.

Expression Quantitative Trait Locus (eQTL): a locus influencing the expression of one or more genes.

Fixation: occurs when a locus which was previously polymorphic in a population becomes monomorphic because all but one allele has been lost through genetic drift.

Frequentist: the name for the school of statistical thought in which support for a hypothesis or parameter value is assessed using the probability of the observed data (or more 'extreme' datasets) given the hypothesis or value. Usually contrasted with Bayesian.

Gamete: a sex cell, sperm in males, egg in females. Two haploid gametes fuse to form a diploid zygote.

Gene: a segment (not necessarily contiguous) of DNA that codes for a protein or functional RNA.

Gene expression: the process by which coding DNA sequences are converted into functional elements in a cell.

Genealogy: the ancestral relationships among a sample of homologous genes drawn from different individuals, which can be represented by a tree. Also sometimes used in

place of pedigree, the ancestral relationships among a set of individuals, which can be represented by a graph.

Genetic drift: the changes in allele frequencies that occur over time due to the randomness inherent in reproductive success.

Genome: all the genetic material of an organism.

Genotype: the (unordered) allele pair(s) carried by an individual at one or more loci. A multilocus genotype is equivalent to the individual's two haplotypes without the phase information.

Guanine (G): purine base that forms a pair with cytosine in DNA.

Haemoglobin: is the red oxygen-carrying pigment of the blood, made up of two pairs of polypeptide chains called globins (2α and 2β subunits).

Haploid: has a single version of each chromosome.

Haplotype: the alleles at different loci on a chromosome. An individual's two haplotypes imply the genotype; the converse is not true, but in the presence of strong linkage disequilibrium haplotypes may be inferred from genotype with few errors.

Hardy-Weinberg disequilibrium: the non-independence within a population of an individual's two alleles at a locus; can arise due to inbreeding or selection for example. Compare with linkage disequilibrium.

Heritability: the proportion of the phenotypic variation in the population that can be attributed to underlying genetic variation.

Heterozygosity: the proportion of individuals in a population that are heterozygotes at a locus. Also sometimes used as short for expected heterozygosity under random mating, which equals the probability that two homologous genes drawn at random from a population are not the same allele.

Heterozygote: a single-locus genotype consisting of two different alleles.

HIV (Human Immunodeficiency Virus): a virus that causes acquired immune deficiency syndrome (AIDS) which destroys the body's ability to fight infection.

Homology: similarities between sequences that arise because of shared evolutionary history (descent from a common ancestral sequence). Homology of different genes within a genome is called paralogous while that between the genomes of different species is called orthologous.

Homozygote: a single-locus genotype consisting of two versions of the same allele.

Hybrid: the offspring of a cross between parents of different genetic types or different species.

Hybridization: the base pairing of a single stranded DNA or RNA sequence, usually labelled, to its complementary sequence.

ibd: identity by descent; two genes are ibd if they have descended without mutation from an ancestral gene.

Inbred lines: derived and maintained by repeated selfing or brother–sister mating, these individuals are homozygous at essentially every locus.

Inbreeding: either the mating of related individuals (e.g. cousins) or a system of mate selection in which mates from the same geographic area or social group for example are preferred. Inbreeding results in an increase in homozygosity and hence an increase in the prevalence of recessive traits.

Intercross: A linkage study design in which the progeny (F1s) of a cross between two inbred lines are crossed or selfed. This design is also sometimes referred to as an *F2 design* because the resulting individuals are known as F2s.

Intron: non-coding DNA sequence separating the exons of a gene. Introns are initially transcribed into messenger RNA but are subsequently spliced out.

Karyotype: the number and structure of an individual's chromosomes.

Kilobase (Kb): 1000 base pairs.

Linkage: two genes are said to be linked if they are located close together on the same chromosome. The alleles at linked genes tend to be co-inherited more often than those at unlinked genes because of the reduced opportunity for an intervening recombination.

Linkage disequilibrium (LD): the non-independence within a population of a gamete's alleles at different loci; can arise due to linkage, population stratification, or selection. The term is misleading and 'gametic phase disequilibrium' is sometimes preferred. Various measures of linkage disequilibrium exist.

Locus (pl. Loci): the position of a gene on a chromosome.

LOD score: a likelihood ratio statistic used to infer whether two loci reside close to one another on a chromosome and are therefore inherited together. A LOD score of 3 or more is generally thought to indicate that the two loci are close together and therefore linked.

Marker gene: a polymorphic gene of known location which can be readily typed; used for example in genetic mapping.

Megabase (Mb): 1000 kilobases = 1,000,000 base pairs.

Meiosis: the process by which (haploid) gametes are formed from (diploid) somatic cells.

messenger RNA (mRNA): the RNA sequence that acts as the template for protein synthesis.

Microarray: see DNA microarray above.

Microsatellite DNA: small stretches of DNA (usually 1–4 bp) tandemly repeated. Microsatellite loci are often highly polymorphic, and alleles can be distinguished by length, making them useful as marker loci.

Mitosis: the process by which a somatic cell is replaced by two daughter somatic cells.

Monomorphic: a locus at which only one allele arises in the sample or population.

Monozygotic twins: genetically identical individuals derived from a single fertilized egg.

Morgan: 100 centiMorgans.

mtDNA: the genetic material of the mitochondria which consists of a circular DNA duplex inherited maternally.

Mutation: a process that changes an allele.

Negative selection: removal of deleterious mutations by natural selection. Also known as *purifying selection*.

Neutral: not subject to selection.

Neutral evolution: evolution of alleles with nearly zero selective coefficient. When $|Ns| << 1$, where N is the population size and s is the selective coefficient, the fate of the allele is mainly determined by random genetic drift rather than natural selection.

Non-Coding RNA (ncRNA): RNA segments that are coded for in the genome, but not translated into protein product. Composed of many classes, the complete range of functions of these molecules has yet to be characterised, but they have been shown to affect the rate of transcription and transcript degradation.

Nonsynonymous substitution: Nucleotide substitution in a protein-coding gene that alters the encoded amino acid.

Nucleoside: a base attached to a sugar, either ribose or deoxyribose.

Nucleotide: the structural units with which DNA and RNA are formed. Nucleotides consist of a base attached to a five-carbon sugar and mono-, di-, or tri-phosphate.

Nucleotide substitution: the replacement of one nucleotide by another during evolution. Substitution is generally considered to be the product of both mutation and selection.

Oligonucleotide: a short sequence of single-stranded DNA or RNA, often used as a probe for detecting the complementary DNA or RNA.

Open Reading Frame (ORF): a long sequence of DNA with an initiation codon at the 5′-end and no termination codon except for one at the 3′-end.

PCR (polymerase chain reaction): a laboratory process by which a specific, short, DNA sequence is amplified many times.

Pedigree: a diagram showing the relationship of each family member and the heredity of a particular trait through several generations of a family.

Penetrance: the probability that a particular phenotype is observed in individuals with a given genotype. Penetrance can vary with environment and the alleles at other loci for example.

Peptide bond: linkages between amino acids occur through a covalent peptide bond joining the C terminal of one amino acid to the N terminal of the next (with loss of a water molecule).

Phase (of linked markers): the relationship (either coupling or repulsion) between alleles at two linked loci. The two alleles at the linked loci are said to be in coupling if they are present on the same physical chromosome or in repulsion if they are present on different parental homologs.

Phenotype: the observed characteristic under study, may be quantitative (i.e. continuous) such as height, or binary (e.g. disease/no disease), or ordered categorical (e.g. mild/moderate/severe).

Pleiotropy: is the effect of a gene on several different traits.

Polygenic: influenced by more than one gene.

Polymorphic: a locus that is not monomorphic. Usually a stricter criterion is imposed: a locus is polymorphic only if no allele has frequency over 99 %.

Polynucleotide: a polymer of either DNA or RNA nucleotides.

Polypeptide: is a long chain of amino acids joined together by peptide bonds.

Polypeptide chain: A series of amino acids linked by peptide bonds. Short chains are sometimes referred to as oligopeptides or simply peptides.

Polytene: refers to the giant chromosomes that are generated by the successive replication of chromosome pairs without the nuclear division, thus several chromosome sets are joined together.

Population stratification (or population structure): Refers to a situation in which the population of interest can be divided into strata such that an individual tends to be more closely related to others within the same stratum than to other individuals.

Positive selection: fixation, by natural selection, of an advantageous allele with a positive selective coefficient. Also known as *Darwinian selection*.

Proband: an individual through whom a family is ascertained, typically by their phenotype.

Prokaryote: organism whose cells have no nucleus.

Promoter: located upstream of the gene, the promoter allows the binding of RNA polymerase which initiates transcription of the gene.

Protein: a large, complex, molecule made up of one or more chains of amino acids.

Pseudogene: a DNA sequence that is either an imperfect, non-functioning, copy of a gene, or a former gene which no longer functions due to mutations.

Purine and Pyrimidine: are particular kinds of nitrogen containing heterocyclic rings.

Purine: adenine or guanine.

Pyrimidine: cytosine, thymine, or uracil.

QTL (Quantitative Trait Locus): a locus influencing a continuously varying phenotype.

Radiation hybrid: a cell line, usually rodent, that has incorporated fragments of foreign chromosomes that have been broken by irradiation. They are used in physical mapping.

Recessive allele: has no effect on phenotype except when present in homozygote form.

Recombination: the formation of new haplotypes by physical exchange between two homologous chromosomes during meiosis.

Restriction enzyme: recognises specific nucleotide sequences in double-stranded DNA and cuts at a specified position with respect to the sequence.

Restriction fragment: a DNA fragment produced by a restriction enzyme.

Restriction site: a 4–8 bp DNA sequence (usually palindromic) that is recognised by a restriction enzyme.

Retrovirus: an RNA virus whose replication depends on a reverse transcriptase function, allowing the formation of a cDNA copy that can be stably inserted into the host chromosome.

Ribonucleic acid (RNA): polymer made up of ribonucleotides that are linked together by phosphodiester bonds.

Ribosome: a cytoplasmic organelle, consisting of RNA and protein, that is involved in the translation of messenger RNA into proteins.

Ribosomal RNA (rRNA): the RNA molecules contained in ribosomes.

Selection: a process such that expected allele frequencies do not remain constant, in contrast with genetic drift. Alleles that convey an advantage to the organism in its current environment tend to become more frequent in the population (positive, or adaptive, selection), while deleterious alleles become less frequent. Under stabilising (or balancing) selection, allele frequencies tend towards a stable, intermediate value.

Sense strand: the DNA strand in the direction of coding.

Sex-linked: a trait influenced by a gene located on a sex (X or Y) chromosome.

Single nucleotide polymorphism (SNP): a polymorphism consisting of a single nucleotide.

Sister chromatids: two chromatids that are copies of the same chromosome. Non-sister chromatids are different but homologous.

Somatic cell: a non-sex cell.

Synonymous substitution: Nucleotide substitution in a protein-coding gene that does not alter the encoded amino acid.

TATA box: a conserved sequence (TATAAAA) found about $25-30$ bp upstream from the start of transcription site in most but not all genes.

Thymine (T): pyrimidine base that forms a pair with adenine in DNA.

trans-Acting: eQTL whose DNA sequence influences gene expression through its gene product. These regulatory elements are often coded for at loci far from or unlinked to the genes they regulate.

Transcription: the synthesis of a single-stranded RNA version of a DNA sequence.

Transition: a mutation that changes either one purine base to the other, or one pyrimidine base to the other.

Translation: the process whereby messenger RNA is 'read' by transfer RNA and its corresponding polypeptide chain synthesized.

Transposon: a genetic element that can move over generations from one genomic location to another.

Transversion: a mutation that changes a purine base to a pyrimidine, or vice-versa.

Uracil (U): pyrimidine base in RNA that takes the place of thymine in DNA, also forming a pair with adenine.

Wild-type: the common, or standard, allele/genotype/phenotype in a population.

Yeast artificial chromosome (YAC): a cloning vector able to carry large (e.g. one megabase) inserts of DNA and replicate in yeast cells.

Zygote: an egg cell that has been fertilized by a sperm cell.

Abbreviations and Acronyms

ABC	Approximate Bayesian Calculation
AD	Alzheimer's Disease
AFLP	Amplified Fragment Length Polymorphism
AGT	Angiotensionogen
AIC	Akaike's Information Criterion
AMOVA	An Analysis of Molecular Variance
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
APM	Affected-Pedigree-Member
ARG	Ancestral Recombination Graph
BAC	Bacterial Artificial Chromosome
BBSRC	Biotechnology and Biological Sciences Research Council
BC	Backcross
BIC	Bayesian Information Criterion
BKYF	Beerli–Kuhner–Yamato–Felsenstein
BLAST	Basic Local Alignment Search Tool
BLUE	Best Linear Unbiased Estimator
BLUP	Best Linear Unbiased Predictor
BMI	Body Mass Index
bp	Base Pairs
CART	Classification and Regression Tree
CASP	Critical Analysis of Structure Prediction
cDNA	Complementary DNA
CEPH	Centre pour l'Etude des Polymorphismes Humains
CGH	Comparative Genomic Hybridization
CHD	Coronary Heart Disease
ChIP	Chromatin Immunoprecipitation
CI	Confidence Interval
CIM	Composite Interval Mapping
CL	Composite Log-Likelihood
CMV	Cytomegalovirus
COGs	Clusters of Orthologous Groups
CTLs	Cytotoxic T Lymphocytes
DAG	Directed Acyclic Graph
df	Degrees of Freedom
DH	Doubled Haploids
DNA	Deoxyribonucleic Acid

EBV	Epstein−Barr Virus
ECHR	European Convention on Human Rights
EC	Extreme Concordant
ECJ	European Court of Justice
ED	Extreme Discordant
EM	Expectation Maximisation
EPD	Eukaryotic Promoter Database
eQTL	Expression Quantitative Trait Loci
EST	Expressed Sequence Tag
FDR	False Discovery Rate
FISH	Fluorescent In Situ Hybridization
FLMs	Finite Locus Models
FM	Fitch−Margoliash Methods
FPM	Finite Polygenic Model
FWER	Family-Wise Error Rate
GA	Genetic Algorithm
GC	Gas Chromatography
GC	Guanine and Cytosine
GEEs	Generalised Estimating Equations
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Model
GNG	Gamma-Normal-Gamma
GO	Gene Ontology
GUI	Graphical User Interface
HA	Haemagglutinin
HBV	Hepatitis B Virus
HMM	Hidden Markov Model
HPD	Highest Probability Density
HSV	Herpes Simplex Virus
HTLV	Human T-Cell Lymphotropic Virus Type I
HVRI	Hypervariable Region
HVRII	Hypervariable Region II
HWE	Hardy−Weinberg Equilibrium
IAM	Infinite-Allele Model
ibd	Identical by Descent
ibs	Identical by State
ICRP	International Commission of Radiological Protection
IID	Independent and Identically Distributed
iis	Identity in State
IS	Importance Sampling
Kb	kilobases
KDEs	Kernel Density Estimators
kNN	k-Nearest Neighbour
LC	Liquid Chromatography
LD	Linkage Disequilibrium
LINEs	Long Interspersed Nuclear Elements
LLR	Log-Likelihood Ratio

LogDet   Logarithm of the Determinant
LOH   Loss of Heterozygosity
LR   Likelihood Ratio
LS   Least-Squares
LTR   Long Terminal Repeat
MAI   Marker-Assisted Introgression
MAS   Marker-Assisted Selection
MC   Monte Carlo
MCMC   Markov Chain Monte Carlo
MH   Metropolis–Hastings
ML   Mapping Methods – Maximum Likelihood
MLE   Maximum Likelihood Estimate
MLR   Maximum Likelihood Ratio
MLR   Multiple Linear Regression
MM   Mismatch
MME   Mixed Model Equations
MP   Maximum Parsimony
MRCA   Most Recent Common Ancestor
mRNA   Messenger Ribonucleic Acid
MS   Mass Spectrometry
MSA   Multiple Sequence Alignment
mtDNA   Mitochondrial DNA
MVN   Multivariate Normal
MY   Million Years
MZ   Monozygous
NcRNA   Noncoding Ribonucleic Acid
NJ   Neighbor-Joining
NMR   Nuclear Magnetic Resonance
NP   Non-Deterministic Polynomial
OLS   Ordinary Least Squares
OR   Odds Ratio
ORF   Open Reading Frame
PAC   Product of Approximate Conditionals
PAM   Partitioning Around Medoids
PCs   Principal Components
PCA   Principal Components Analysis
PCR   Polymerase Chain Reaction
PDB   Protein Data Bank
PDF   Probability Density Function
PI   Paternity Index
PIC   Polymorphism Information Content
PKU   Phenylketonuria
PLS   Partial Least-Squares
PM   Perfect Match
PMLE   Pseudo Maximum Likelihood Estimator
PNNs   Probabilistic Neural Networks
PSA   Population-Specific Alleles

QQ	Quantile–Quantile
QTLs	Quantitative Trait Loci
RAPD	Randomly Amplified Polymorphic DNA
RCTs	Randomized Controlled Trials
REML	Residual Maximum Likelihood
RFLP	Restriction Fragment Length Polymorphism
RH	Radiation Hybrid
RIL	Recombinant Inbred Line
SINE	Small Interspersed Nuclear Elements
SIVagm	SIV from African Green Monkeys
SMM	Stepwise Mutation Model
SNP	Single Nucleotide Polymorphism
SPRT	Sequential Probability Ratio Test
STR	Short Tandem Repeat
STRs	Simple Tandem Repeats
STS	Sequence-Tagged Site
SVM	Support Vector Machine
TDT	Transmission/Disequilibrium Test
TF	Transcription Factor
TPM	Two-Phase Model
TRRD	Transcription Regulatory Regions Database
TSG	Tumour Suppressor Gene
TSS	Transcription Start Site
UA	Ultimate Ancestor
UPGMA	Unweighted Pair-Group Method with Arithmetic Mean
WB	Wilson and Balding
WGA	Whole Genome Association
WPC	Weighted Pairwise Correlation
YAC	Yeast Artificial Chromosome

Part 1
Genomes

1

Chromosome Maps

T.P. Speed

*Department of Statistics, University of California at Berkeley, Berkeley, CA, USA,
Genetics and Bioinformatics Group, The Walter & Eliza Hall Institute of Medical
Research, Royal Melbourne Hospital, Melbourne, Australia*

and

H. Zhao

*Department of Epidemiology and Public Health, Yale University School of Medicine,
New Haven, CT, USA*

Chromosome maps are a natural way of organizing genetic data about chromosomes. Existing
chromosome maps can be broadly divided into four categories: genetic maps, physical maps,
radiation hybrid maps, and gene maps. Although they all make reference to the same biological
entity, namely chromosomes, these maps differ substantially in the types of genetic experiments
conducted and the types of genetic data collected. They further differ in the metrics employed
to define distances and the resolution achievable. Collectively, these maps provide essential tools
to further our understanding of the organization and function of the genome. In this review, we
first describe the biological principles behind each type of chromosome map and then outline the
statistical models and methods that have been developed to construct it. The current state of each
chromosome map is summarized, and links to mapping software are provided for readers interested
in getting hands-on experience with chromosome mapping.

1.1 INTRODUCTION

Chromosome maps are a natural way of organizing genetic data about chromosomes, in
very much the same way that ordinary (cartographic) maps organize geographic data about
continents, countries or cities. Geneticists have long constructed different types of maps
to order genes or markers, breakpoints, deletions, and other features in relation to one
another and to landmarks along chromosomes such as centromeres and telomeres. *Genetic
maps* were the first type of map constructed to position genes along chromosomes, with

the distances between pairs of genes being defined in terms of recombination fractions. Thus genetic maps were unusual for at least two reasons. First, the objects being mapped – genes, later polymorphic markers, collectively described as loci – were frequently abstract, in the sense that data concerning them was only indirectly observed; the genes themselves were never seen. And secondly, the distance was only relative, and defined statistically. Since the rate of recombination varies along chromosomes, genetic map distance is not proportional to actual physical distance, although there are useful average relationships for different organisms.

Physical maps take a number of forms, but common to all is the fact that the objects being mapped are concrete, usually assayable, and the distances are physical, most recently thousands (kb) or millions (Mb) of base pairs, reflecting the fact that chromosomes are long DNA molecules. However, the first physical maps were not exactly of this type. Early examples of physical maps are those based on the salivary gland polytene chromosomes of insects belonging to the order Diptera, such as *Drosophila melanogaster*. In these maps, the positioning is provided by the visible bands. The familiar cytogenetic maps of human (see Figure 1.1) and other mammalian chromosomes created by staining metaphase chromosomes all have a similar character, again with bands providing positioning information.

A slightly confusing aspect is that physical maps refer not only to maps of loci along chromosomes, but also to organized collections of chromosomal segments, such as restriction fragments and more general ordered sets of cloned fragments of a chromosome.

When recombination is used to define distances, the genes or markers must be polymorphic in order to be mappable. By contrast, the physical mapping of loci only requires probes for recognizing specific chromosomal sites or for detecting fragment
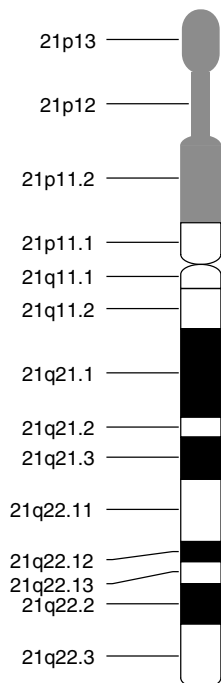
Figure 1.1   Ideogram of human chromosome 21. [Source: `www.gdb.org/hugo/chr21/integratedMaps.html`.]

overlap. The most useful probe in this context is the sequence-tagged site (STS), which is defined by a pair of 20–30 base pairs polymerase chain reaction (PCR) primers that reliably amplify a unique segment in a genome.

Maps of human (and other) chromosomes intermediate in resolution between the genetic and physical maps are the radiation hybrid (RH) maps. These are based on assay data from human–rodent somatic cell hybrids containing small fragments of human chromosomes. They have their own metric, namely the average number of breaks per unit physical distance, reflecting the radiation dose used to fragment the chromosomes.

Next in resolution, although different in character, are *gene maps*. These maps are currently constructed by clustering expressed sequence tags (ESTs), more specifically short DNA sequences in the $3'$-untranslated regions of complementary DNAs (cDNAs), and then locating the clusters on chromosomes using STSs in these regions. Ideally, each point on such a map corresponds to a unique gene, and so gene maps should locate genes on physical maps. Until a genome is completely sequenced, these maps provide the best usable description of the locations of the genes of an organism.

With the exception of the polytene chromosome and cytological maps, all of the kinds of maps we have mentioned make use of statistical methods in their definition, in their construction, or in the assignment of the corresponding map distance. In this chapter, we give a review of some of the statistical models and methods used in chromosome mapping. It will be partial in the sense that we discuss genetic mapping much more fully, in part because it is the most thoroughly developed mapping method, going back nearly a century.

1.2 GENETIC MAPS

1.2.1 Mendel's Two Laws

Modern genetics began with the work of Mendel on garden peas in the 1860s (Mendel, 1866). In his experiments, Mendel studied a number of heritable traits in peas, including seed color. He interpreted his experiments with this trait by postulating the existence of things that we now call genes. He said that two gene variants controlled color in his two lines, *y* (for yellow) and *g* (for green), and that the color gene-pair in the seed determines what color the seed will be. His experiments led him to believe that all cells in the mature plant contain the seed's color gene-pair, with the exception of sex cells, which contain only one of the pair. If the seed's gene-pair is y/g, then half the pollen cells get *y* and half get *g*; similar observations hold for egg cells. Mendel was able to explain his observations with this theory, and it is largely what we believe today. It is often called *Mendel's first law*, the law of segregation. Mendel's first law says that each adult pea plant has a *gene-pair* (say, *y* and *g*) for each character studied, and that the pair *y* and *g* segregate from each other into gametes, so half the gametes will carry *y*, and the other half will carry *g*.

Mendel also considered two or more heritable traits together, for example, seed color and seed shape. Denoting the two variants of seed shape by *s* (for smooth) and *w* (for wrinkled), he first established that, when considered on its own, seed shape inheritance was also explained by supposing that each cell had one gene-pair, in this case one of the pairs s/s, s/w or w/w, and that sex cells had just one of *s* or *w*, effectively chosen at

random from the pair generally present. Mendel then carried out experiments to determine how the two traits, seed shape and seed color, were inherited together. He concluded that each sex cell received one gene from each gene-pair, chosen at random from the available pair, independently for the two gene-pairs. For example, if the mature organism's cells generally possessed gene-pairs y/g and s/w, then its sex cells received ys, yw, gs, or gw with equal frequency $\frac{1}{4}$. Let us see the sense in which this last statement is true. Consider an organism P whose gene-pairs for two traits are y/g and s/w, that is descended from a parent GF that was y/y and s/s, and a parent GM that was g/g and w/w. Then P is ys/gw, getting y and s from GF, and g and w from GM. In a natural sense, y and s were combined together in GF, as were g and w in GM, while y and g (and s and w) were separated at that generation, being present in different individuals. With peas, when y, g, s, and w corresponded to seed color and shape as above, Mendel saw that this togetherness or separateness in the G-generation had no impact on the choice of genes that P passed on to its offspring C: ys, yw, gs, and gw were found to be passed on with equal frequency. Mendel's second law says that during gamete formation, the segregation of one gene-pair is independent of other gene-pairs. When two gene-pairs, say (y, g) and (s, w), segregate, each (haploid) gamete will be equally likely to have genotypes (y, s), (y, w), (g, s), and (g, w).

The above observation, sometimes known as *Mendel's law of independent segregation*, turns out to hold for some, but not all, pairs of genes. The exceptions are the biological basis for genetic mapping. In the early 1900s, deviations from Mendel's second law were observed by Bateson *et al.* (1905) in the sweet pea, and by Morgan (1911) in Drosophila: some genotypes appeared more often than other genotypes, indicating that the gene-pairs were not segregating independently. There are many pairs of traits whose genes do not recombine freely, but tend to stay together, in the sense that the parent P above with composition ys/gw would be more likely to pass on the pairs ys and gw to its offspring C, than the pairs yw and gs. This phenomenon is known as *linkage*: genes that came to P together from the G-generation are preferentially passed on together to offspring in the C-generation. In the most extreme case, C would receive each of P's parental combinations ys and gw with frequency $\frac{1}{2}$, and never receive yw or gs. We would then say that the genes are *completely linked*; no recombining takes place. For a given pair of traits such as seed color and seed shape, with heritable variants (*alleles*) such as y, g, and s, w, we define their recombination fraction to be the frequency with which P's nonparental combinations yw and gs are passed on; with Mendel's examples this fraction was always $\frac{1}{2}$. In the early part of the twentieth century, examples were found where this fraction was noticeably smaller than $\frac{1}{2}$, and to this day, pairs of genes for traits separate into those that freely recombine, and those for which the recombination fraction is less than $\frac{1}{2}$. Using the then much-debated chromosome theory of Mendelian heredity, Morgan explained this nonindependent segregation by supposing the two pairs of genes lie on the same chromosome. A chromosomal exchange between these two genes will result in a recombination between them. Morgan inferred that genes on the same chromosome tend to remain together much more often than if they are on different chromosomes, and called this principle *the third law of heredity*. He also hypothesized that the cross-shaped structure (called *chiasma*) seen during the diplotene phase of meiosis is a manifestation of *crossing-over*. It is now known that crossovers are precise breakage-and-reunion events that are essential for proper segregation, and can promote genetic variation.

1.2.2 Basic Principles in Genetic Mapping

In the following discussion, we make no distinction between *gene*, *marker*, and *locus*, all of which refer to some region on the chromosome. Consider two genes $\mathcal{A}$ (with alleles A and a) and $\mathcal{B}$ (with alleles B and b) and a diploid cell with AB and ab on homologous chromosomes. There are four possible meiotic products, namely, AB, ab, Ab, and aB. The first two are called *parental* types or *nonrecombinants*, because both AB and ab retain the configuration of one of the homologous chromosomes. The other two types, Ab and aB, are called *recombinants*. If two markers are recombined in a meiotic product, then during meiosis an odd number of crossovers must have occurred between the two markers on the strand carrying them. The recombination fraction, r_{AB}, is defined as the proportion of recombinants. It was Sturtevant (1913) who first used the variations in the strength of linkage to determine the sequence in the linear dimension of the chromosome. He argued that if the arrangement of the genes in the chromosome is linear and the recombination frequencies depend on the physical distance between them, then genes can be arranged like dots in a straight line at distances apart proportional to the recombination fraction. For example, for three genes, y (yellow gene), w (white gene), and mi (miniature gene) on the sex chromosome of Drosophila, the observed recombination fraction between y and w was $r_{y,w} = 1.3\,\%$, that between w and mi was $r_{w,mi} = 32.6\,\%$, and that between y and mi was $r_{y,mi} = 33.8\,\%$. Because $r_{y,mi} \approx r_{y,w} + r_{w,mi}$, the white gene can be inferred to lie between the yellow and miniature genes.

For three genes $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ on the same chromosome in the order $\mathcal{A}$–$\mathcal{B}$–$\mathcal{C}$, the additivity among the three recombination fractions, i.e. $r_{AC} = r_{AB} + r_{BC}$, generally holds when the recombination fractions are small (less than $10\,\%$). However, as noted by Sturtevant (1913), the additivity in general does not hold when larger recombination fraction values are involved, and usually $r_{AC} < r_{AB} + r_{BC}$. Deviations from additivity are due to the existence of double crossovers. The next major development in genetic mapping was Haldane's definition Haldane (1919) of the genetic distance between two loci as the average number of crossovers between the loci per meiosis. This gave geneticists an additive distance along chromosomes, albeit one that was rapidly found not to correlate precisely with any apparent physical distance. The unit of genetic distance is the centimorgan (cM). Two markers are 1 cM apart if on average there is one crossover occurring between these two markers on a single strand for every 100 meioses.

Therefore, we have two basic concepts in genetic mapping: the recombination fraction, which can be estimated from data on the offspring of suitable parents; and map distance, which will be based upon the same data, but can only be estimated using a probabilistic model for recombination. With experimental organisms such as the fruit fly, maize, mice, fungi and yeast, establishing linkage and estimating recombination fractions was generally straightforward, because crosses could be planned, and large numbers of offspring examined. With humans, even establishing linkage between a pair of genes was a major achievement in the classical era, and estimating recombination fractions was a challenging statistical problem. Part of the reason for this lies in the longer generation times, and hence the difficulty in obtaining large sets of data, and part lies in the fact that matings are not subject to experimental control, forcing the human geneticist to make use of nonrandomly sampled family or pedigree data. One further complication with human data was the existence of genes with only an indirect relationship between genotype and phenotype, the issue of penetrance. Dominant and recessive traits are instances of what

are termed *incompletely penetrant traits*, and there are many human genetic diseases with quite complex patterns of penetrance, including age and sex dependence.

1.2.3 Meiosis, Chromatid Interference, Chiasma Interference, and Crossover Interference

Before describing statistical methods for genetic mapping in detail, we briefly review the process of meiosis and the genetic concepts relevant to genetic mapping. At the start of meiosis two chromosome sets are present, one coming from each parent in the previous generation. Each chromosome thus has a partner called a *homolog*. During the pachytene and diplotene phases of meiosis, homologous chromosomes pair and each of the paired chromosomes duplicates, resulting in a bundle of four homologous *chromatids*. Chromatids that are copies of the same chromosome are called *sister chromatids*, and those originating from homologous chromosomes are called *nonsister chromatids*. Crossovers take place after the formation of this four-strand structure, with each crossover involving two nonsister chromatids. The number and locations of crossovers vary from chromosome to chromosome for the same meiosis, and from meiosis to meiosis for the same chromosome.

Most genetic mapping efforts have focused on the case where data from only one of the four products of any given meiosis can be observed. Extending terminology from fungal genetics, we call this *single-spore data* in recognition of the fact that in organisms such as *Saccharomyces cerevisiae* (baker's yeast) and *Neurospora crassa* (red bread mold) all four products of a single meiosis can be recovered together in what are known as *tetrads* or *octads*. Genetic studies on these organisms have contributed greatly to our knowledge of many biological mechanisms. Some interesting statistical models that have been developed using tetrad and octad data will be discussed in later sections.

Mather (1933) distinguished two aspects of crossing-over that are relevant to the observed recombination outcome: the distribution of crossover events along the bundle of four chromatids; and the pairs of nonsister chromatids to be involved in crossovers. To distinguish crossover events occurring on the four-strand bundle and crossover events on single strands in the following, we describe crossover events on the four-strand bundle as chiasmata, and those on single strands as crossovers. Chiasma interference refers to nonrandom distribution of chiasmata on the four-strand bundle, whereas crossover interference refers to nonrandom distribution of crossover locations along single strands. Muller (1916) first noted that simultaneous recombinations are not independent, e.g. double recombinations take place at a frequency below that expected under the independence assumption. For example, for the three genes discussed above – yellow, white, and miniature – the expected double recombination frequency is $1.3\% \times 32.6\% = 0.43\%$. However, the observed frequency was only 0.045%. This suggests that the occurrence of one recombination reduces the chance of other recombinations in the nearby region. Crossover interference is seen in almost all organisms, including humans, and the presence of one crossover usually inhibits the formation of crossovers in a nearby region. The biological nature of crossover interference is still not well understood.

With respect to the pairs of nonsister chromatids involved in crossovers, we say there is no chromatid interference (NCI) if any pair of nonsister chromatids are equally likely to be involved in any chiasma, independent of which pairs were involved in other chiasmata.

The observation of crossover interference on the meiotic products (single strands) can be the result of chiasma interference alone, of chromatid interference alone, or of both types of interference. Zhao and Speed (1996) noted that the operation of two types of

interference can lead to no apparent crossover interference, and therefore these two types of interference cannot be separated on the basis of single-strand recombination data. In contrast, tetrad data carries information to distinguish these two types of interference.

1.2.4 Genetic Map Functions

Until the mid-1980s, most linkage mapping was two-point, that is, involved the estimation or testing of a single recombination fraction. For two-point data, we can infer the unobservable genetic distance between two markers from the observable recombination fraction through genetic map functions. Under the assumption of NCI, Mather (1935) showed that, given k (≥ 1) chiasmata between two markers on the four-strand bundle, the probability of observing recombination between these two markers is $\frac{1}{2}$. Therefore, the overall recombination fraction between two markers is $\frac{1}{2}(1 - p_0)$, where p_0 is the probability of having zero chiasmata between these two markers. This is called *Mather's formula*. Assuming chiasmata occurring independently of each other, Haldane derived the now well-known Haldane map function relating recombination fraction and map distance: $r = \frac{1}{2}(1 - e^{-2d})$ with inverse $d = -\frac{1}{2}\log(1 - 2r)$. Nearly 90 years later, this approach has proved to be very satisfactory for a wide variety of organisms. Note that Haldane derived his map function under the two-strand model, i.e. assuming only two strands (the two homologous chromosomes) are involved in the crossover process. Although this assumption is incorrect, we would arrive at the same map function under the four-strand model with no chromatid interference.

In addition to deriving the Haldane map function in his seminal 1919 paper, Haldane also proposed the empirical inverse map function $d = 0.7r + 0.3(-\frac{1}{2}\log(1 - 2r))$ to account for crossover interference in the data then available, and introduced a differential equation method that permitted the construction of a variety of map functions. A variety of other genetic map functions embodying different degrees of crossover interference have been proposed, including by Ludwig (1934); Kosambi (1944); Carter and Falconer (1951); Sturt (1976); Rao *et al.* (1977); Felsenstein (1979); and Karlin and Liberman (1978).

For all these map functions, genetic distance is very close to recombination fraction when the latter is small, and map distances can be (and in the fly group were) estimated without a model – provided the pair of genes were connected by a sequence of closely linked genes – by adding small recombination fractions.

1.2.5 Genetic Mapping for Three Markers

Historically, the first formal linkage analysis involving more than two loci was given by Fisher (1922). He showed how to combine data from a number of two-point analyses in order to obtain efficient estimates of a set of recombination fractions. For three markers $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ in an arbitrary but fixed order, the joint recombination probabilities may be denoted by $\mathbf{p} = (p_{i_1 i_2})$, where the subscript $i_k = 1$ corresponds to recombination across the kth interval, and $i_k = 0$ corresponds to no recombination across the same interval. Therefore, we have four probabilities $\mathbf{p} = (p_{00}, p_{01}, p_{10}, p_{11})$ for three markers, corresponding to the four patterns of recombination or no recombination across $\mathcal{A}$–$\mathcal{B}$ and $\mathcal{B}$–$\mathcal{C}$. Although the data were all two-point, Fisher needed to express the recombination fraction across the union $\mathcal{A}$–$\mathcal{C}$ of two adjacent intervals $\mathcal{A}$–$\mathcal{B}$ and $\mathcal{B}$–$\mathcal{C}$ in terms of their individual recombination fractions. He did so by making the assumption of *complete interference*, that is, by assuming that at most one recombination could occur across any

pair of adjacent intervals. This is equivalent to the following joint distribution:

$$p_{00} = 1 - r_1 - r_2; \quad p_{01} = r_2; \quad p_{10} = r_1; \quad p_{11} = 0,$$

where r_1 and r_2 are the recombination fractions across $\mathcal{A}$–$\mathcal{B}$ and $\mathcal{B}$–$\mathcal{C}$, respectively. This model would not be appropriate for the analysis of three-point data in which double recombinants are observed, but it has been used in modern times with very short intervals. In human linkage analysis one finds almost exclusive use of the extremely tractable Poisson or *no-interference* model, whose joint probabilities for three loci take the form

$$p_{i_1 i_2} = r_1^{i_1}(1 - r_1)^{1 - i_1} r_2^{i_2}(1 - r_2)^{1 - i_2},$$

where, for $i = 1, 2$, the recombination fractions r_i may be expressed in terms of genetic distances d_i by

$$r_i = \tfrac{1}{2}(1 - \mathrm{e}^{-2d_i}).$$

It seems that although this model and its extension to more than three loci fail to fit most data sets of any size, the recombination fractions and locus orderings obtained are generally satisfactory; see Speed *et al.* (1992). Any map function, $r = M(d)$, can be used to analyze three-point data. This is because $r_1 = p_{10} + p_{11} = M(d_1)$, $r_2 = p_{01} + p_{11} = M(d_2)$, and $p_{11} = \tfrac{1}{2}[M(d_1) + M(d_2) - M(d_1 + d_2)]$, and we can derive all the $p_{i_1 i_2}$ from a given map function. Therefore, likelihood functions for the observed data can be constructed and maximum likelihood estimates of genetic distances can be obtained.

For an arbitrary crossover process model, under the assumption of NCI, Speed *et al.* (1992) derived a set of inequality constraints and showed the robustness of the ordering. The order with respect to which these probabilities are defined does not need to be the true one, and if we change it, the probabilities need only be relabeled. For example, if we go from the order $O : \mathcal{A}$–$\mathcal{B}$–$\mathcal{C}$ with probabilities $\mathbf{p}$, to $O' : \mathcal{A}$–$\mathcal{C}$–$\mathcal{B}$ with probabilities $\mathbf{p}'$, then $\mathbf{p}'$ is related to $\mathbf{p}$ as follows:

$$p'_{00} = p_{00}, \quad p'_{10} = p_{10}, \quad p'_{01} = p_{11}, \quad p'_{11} = p_{01}.$$

Three-point phase known crosses (in which allelic combinations across loci are present together on the same chromosome) have been used for decades to order loci in experimental organisms, without any explicit model assumptions. This works because, under very general conditions, the smallest of the four probabilities ($p_{i_1 i_2}$) corresponds to the event of double recombination across two consecutive intervals when the loci are correctly ordered. For example, if the correct order is $O : \mathcal{A}$–$\mathcal{B}$–$\mathcal{C}$, then (assuming no chromatid interference)

$$p_{11} \leq p_{10}, p_{01} \leq p_{00}.$$

If, on the other hand, $O' : \mathcal{A}$–$\mathcal{C}$–$\mathcal{B}$ is the correct order, but we have written our probabilities relative to O, then $p'_{11} = p_{01}$ will be the smallest probability. It follows that, with sufficiently large samples of data, any set of loci can be ordered by inspection, with only a small chance of error. Naturally this is also possible using only the pairwise recombination fractions, but that would take more data to achieve the same level of confidence in the ordering. More generally, it is possible to show that, under the assumption of NCI, a multipoint recombination probability decreases, or at least does not increase, when any nonrecombinant interval is changed to recombinant status; see Speed *et al.* (1992).

1.2.6   Genetic Mapping for Multiple Markers

Although inefficient from the statistical viewpoint, three or more loci can be mapped using only two-point data, since linear maps are determined by pairwise distances. When there are plenty of data, such as with Drosophila, multipoint analyses may be unnecessary. However, in most contexts, data are scarce. In such cases, multipoint linkage analysis can be viewed as an attempt to make more efficient use of recombination data to further the aims of linkage analysis; see Lathrop *et al.* (1984) and Thompson (1984).

Multipoint linkage analyses make fuller use of available data, and can achieve greater precision or power. They are more complex than two-point analyses in several important ways. First, they require the specification of an order for the loci. Second, they require the specification of a joint distribution for all possible recombination patterns: for n loci, there are 2^{n-1} such patterns (including the parental one). Third, from the perspective of parametric statistical inference, joint distributions over recombination patterns corresponding to distinct orderings of the loci define noncomparable statistical models. Most of the difficulties of multipoint linkage analysis stem from these facts, particularly the rate of increase of the number of orders or patterns with the number of loci. When linkage analysis is being done using pedigree data, the size (number of individuals) and complexity (presence of one or more loops) of the pedigrees are additional limiting factors.

At the initial stage of genetic mapping, linkage groups have to be defined. Two markers are in linkage if the recombination fraction between them is less than $\frac{1}{2}$. A linkage group is defined as a set of markers where each marker is linked to at least one other marker in the same set. With enough markers covering the genome, each linkage group will correspond to a chromosome. However, for three markers $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, it is possible that $\mathcal{A}$ and $\mathcal{B}$ are genetically linked, $\mathcal{B}$ and $\mathcal{C}$ are genetically linked, and yet the recombination fraction between $\mathcal{A}$ and $\mathcal{C}$ may be approaching $\frac{1}{2}$ if they are sufficiently far apart from each other on a chromosome. Although linkage groups have been well defined for humans and some experimental organisms, linkage group construction remains the critical first step for many organisms at the early stage of genetic mapping.

To define whether two markers are in linkage is to test whether the recombination fraction between these two markers is less than $\frac{1}{2}$. This hypothesis testing problem can be carried out using the likelihood ratio test; see Ott (1999). The LOD (log-odds) score is often used to assess the evidence for linkage. It differs from the usual likelihood ratio statistic by a constant factor and is defined as

$$\text{LOD} = \log_{10} \frac{L(data|r)}{L\left(data|r = \frac{1}{2}\right)}.$$

A LOD score of 3 has been used as the threshold for linkage testing. The justification of this threshold is discussed by Ott (1999) and Risch (1991).

After linkage groups are defined, the next task is to order genetic markers within each group. The locus ordering problem resembles the traveling salesman problem (TSP) widely discussed in the field of combinatorial optimization, (see Johnson, 1990), in which there are a large number of discrete states, each of which can be assigned a numerical value by a cost or objective function. The calculation of the objective function can depend either on information from pairwise data (e.g. pairwise LOD scores) or on joint genetic information (e.g. multipoint LOD scores, discussed later). For example, Speed *et al.* (1992) showed that, under the assumption of NCI, a given order imposes linear constraints among

multilocus recombination probabilities. Maximum likelihood under these constraints for each order can be used as the objective function.

With n markers, the ideal ordering approach would be to compute the objective function for each of the $\frac{1}{2}n!$ orders, and then to rank the orders, choosing the one with the largest objective function as the best order. With a few markers, all possible orderings can be considered. However, this quickly becomes impossible with many genetic markers. There is no evidence to suggest that a method exists that is generally better than choosing that order which maximizes the likelihood of the data using a suitable recombination model, at least not when the calculation of the likelihoods corresponding to each of the $\frac{1}{2}n!$ distinct orders is possible. The Poisson or no-interference model is the one typically used in this context. Although there does not appear to be a systematic study of this issue, the available evidence suggests that only small gains in the efficiency of ordering loci are to be found by using a more suitable model when interference exists; see Lathrop *et al.* (1984); Bishop and Thompson (1988); Goldgar and Fain (1988); Speed *et al.* (1992); and Goldstein *et al.* (1995) for related results. Different heuristic ordering strategies were reviewed in Weeks (1991), and more recent development can be found in Mester *et al.* (2003); York *et al.* (2005); and Tan and Fu (2006), among others.

In our previous discussion on genetic distance estimation from two-point or three-point genetic data, we described how map functions can be used to estimate genetic distances. However, when there are more than three markers, the multilocus recombination probabilities cannot be uniquely determined from the map function. A crossover process model is needed to derive joint multilocus recombination probabilities. Several point process models (Fisher *et al.*, 1947; Karlin and Liberman, 1979; Risch and Lange, 1979; King and Mortimer, 1990; Fujitani *et al.*, 2002) have been proposed to incorporate crossover interference in modeling the crossover process. The first satisfactory class of recombination models were the chi-square renewal process models discussed by Fisher and his students and colleagues (Fisher *et al.*, 1947). Bailey (1961) gave a good overview of this research. The simplest of these joint probabilities is too complex to be given here, and this is probably the reason why this class of models has not been used with human data until recently (Lin and Speed, 1996; Broman and Weber 2000; Browning, 2003). The chi-square model has been extended to the Poisson-skip model, which has the chi-square model as its special cases and can also incorporate negative crossover interference; see Lange *et al.* (1997). More recently, Stahl and colleagues have proposed that there exist two separate recombinational pathways, one with independent crossovers and one imposing crossover interference. Empirical data seem to be in favor of this two-pathway model for Arabidopsis (Copenhaver *et al.*, 2002) and humans (Housworth and Stahl, 2003). The major alternatives to the chi-square renewal models are due (independently) to Karlin and Liberman (1979) and Risch and Lange (1979), called *count-location* or *generalized no-interference* models, and the model of Goldgar and Fain (1988). For a review and comparison of different stochastic models for recombination, see McPeek and Speed (1995). One approach that does not depend on specific models for recombination was developed by Weinstein (1936) and was recently used to study human meiosis by Lamb *et al.* (1997); Zhao *et al.* (2000); and Li *et al.* (2001). The only assumption employed by this approach is that there is at most one chiasma in each marker interval, which is likely to be satisfied when many markers are studied on a chromosome. Although substantial number of additional parameters are involved, the results from Weinstein's approach can be used to assess the goodness of fit of different crossover process models and to

identify anomalous features of these models. Despite great efforts made to understand the molecular mechanisms leading to crossover interference, surprisingly little are known to date (e.g. Jones and Franklin, 2006).

Liberman and Karlin (1984) proposed to extend genetic map functions to four or more marker cases by embodying the assumption that, for a pair of noncontiguous intervals, the probabilities for joint recombination patterns across these intervals do not depend on the distance between the intervals, something which is not consistent with observations. Those map functions that can be extended to multilocus data through this approach have been (inappropriately) called *multilocus feasible* by Liberman and Karlin (1984). This criterion excludes many functions that were found to fit well to recombination data, such as the Kosambi map function proposed by Kosambi (1944). However, Zhao and Speed (1996) showed that there exist stationary renewal processes that give rise to most map functions in the literature (including the Kosambi map function). Therefore, these map functions are compatible with the analysis of multilocus data via this approach. Moreover, the interevent distributions of the stationary renewal processes corresponding to most map functions can be closely approximated by γ distributions.

We have discussed the cases where recombination or nonrecombination can be unambiguously scored. For human pedigrees, matters are more complicated at many levels. As with two-point linkage analyses, a major complication in multipoint linkage analyses can be the incompleteness of data. For example, there may be missing data due to some individuals not being typed. All data may be available, but phenotype may not determine genotype, as with dominant traits and other types of incomplete penetrance. Genotypes may be known, but haplotypes may not, that is, phase may be unknown. With known genotypes at n loci, there are 2^{n-1} possible haplotypes. While these incompleteness problems can slow down two-point analyses, they can quickly make exact multipoint analyses impossible. On the other hand, multipoint analyses can make use of data that cannot be used in two-point analyses, for example, when only uninformative data are available at a locus intermediate between two fully informative loci; see Lathrop *et al.* (1985) and Ott (1999). In multipoint linkage analysis using pedigree data, the feasibility of an exact analysis will depend on the number of loci, the size and complexity of the pedigrees involved, and the nature and extent of incompleteness in the data.

For pedigrees with simple structures or with a few genetic markers, the likelihood for a pedigree can be calculated exactly. The exact calculations can be divided into two types of algorithms: the Elston–Stewart algorithm (Elston and Stewart, 1971) and the Lander–Green algorithm (Lander and Green, 1987). Consider a pedigree with m individuals, where $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ is the set of observed phenotypes for the pedigree. If G_i is the set of genotypes g_i compatible with the phenotype of person i, then the likelihood of the pedigree can be written as a sum of products:

$$\sum_{g_1 \in G_1} \cdots \sum_{g_m \in G_m} \prod_i P(x_i | g_i) \prod_{k \text{ founders}} P(g_k) \prod_{\{i_1, i_2, i_3\}} P(g_{i_1} | g_{i_2}, g_{i_3}),$$

where $\{i_1, i_2, i_3\}$ is an offspring–parent triad and i refers to the individuals with observed phenotypes. The probability $P(x_i | g_i)$ is the probability of an individual with genotype g_i having phenotype x_i. For codominant genetic markers, the probability is either 1 or 0. The founder probability $P(g_k)$ is a function of population gene allele frequencies. The Elston–Stewart algorithm can be viewed as a method for choosing an order to perform the iterated sum to minimize the total number of additions and multiplications. The number of

calculations in the Elston–Stewart algorithm scales linearly with the number of individuals in the pedigrees but exponentially with the number of markers.

The Lander–Green algorithm works as follows. Let $\mathbf{x}_L = (\mathbf{x}_{L_1}, \mathbf{x}_{L_2}, \ldots, \mathbf{x}_{L_N})$ denote the collection of phenotypes at locus i, and $\mathbf{g}_L = (\mathbf{g}_{L_1}, \mathbf{g}_{L_2}, \ldots, \mathbf{g}_{L_N})$ denote the collection of ordered genotypes at these loci for the individuals. Then the likelihood for the pedigree can be written as

$$\sum_{g_{L_1} \in L_1} \cdots \sum_{g_{L_N} \in L_N} \left[\prod_i P(x_{L_i}|g_{L_i}) \right] P(g_{L_N}|g_{L_{N-1}}, g_{L_{N-2}}, \ldots, g_{L_1}) \ldots P(g_{L_2}|g_{L_1}) P(g_{L_1}).$$

Assuming no crossover interference, then the likelihood is

$$\sum_{g_{L_1} \in L_1} \cdots \sum_{g_{L_N} \in L_N} \left[\prod_i P(x_{L_i}|g_{L_i}) \right] P(g_{L_N}|g_{L_{N-1}}) \ldots P(g_{L_2}|g_{L_1}) P(g_{L_1}).$$

The Lander–Green algorithm can be extended to incorporate the chi-square model in linkage analysis. The Elston–Stewart algorithm is mostly useful for large pedigrees but only a limited number of markers, whereas the Lander–Green algorithm is useful for multiple markers but is limited in the number individuals in each pedigree. This likelihood can be efficiently evaluated using the forward–backward algorithm of the hidden Markov model methodology. In addition, parameter estimates can be obtained using the expectation maximization (EM) algorithm. The number of operations scales linearly with the number of markers but exponentially with the number of individuals in the pedigree.

Both algorithms will fail if we have large pedigrees with many markers typed, and simulation methods to approximate the likelihood have been proposed by Thompson (1994) and Sobel and Lange (1996). In a recent review, Lin (1996) discussed both the sequential imputation approach of Irwin *et al.* (1994) and Markov chain Monte Carlo methods of Lin and Wijsman (1994).

1.2.7   Tetrads

Recall that tetrads and octads refer to the case where all four products of a single meiosis can be recovered together, such as in yeast and bread mold. Octads are generated from tetrads following one mitosis, and octads can usually be represented by tetrads, except when gene conversions occur. If we ignore the possibility of gene conversions, we need make no distinction between tetrads and octads in the following discussion, and refer to both as tetrads. Genetic studies using tetrad data are very valuable in studying the crossovers during meiosis. Compared to single-spore data, tetrad data have several advantages. First, with tetrad data chromatid interference and chiasma interference can be distinguished. Second, when chromatid interference is absent, chiasma interference can be detected with only two markers, whereas at least three markers are needed for single-spore data. Chiasma interference can even be detected with one marker in some studies. Third, the position of the centromere can be inferred. In some organisms, such as *Neurospora crassa*, the tetrads are produced in a linear order corresponding to the meiotic divisions; these are called *ordered tetrads*. In others, such as *Saccharomyces cerevisiae*, the tetrads are produced as a group without order, and are called *unordered tetrads*.

If a cross involves two strains differing with respect to two genes, geneticists distinguish three possible tetrad types: parental ditype with two representatives of each of the two parental types; nonparental ditype, where all four strands show recombined types; and tetratype, where two of the four strands show parental types and the other two strands show recombined types. For tetrad data involving two genetic markers, let P, T, and N denote the proportion of tetrads having parental ditype, tetratype, and nonparental ditype, respectively. The recombination fraction between the two markers can then be estimated by $N + \frac{1}{2}T$. Although the genetic distance can be estimated from this recombination fraction through a genetic map function, there is more information in the raw tetrad data. Under the assumption of NCI, given two chiasmata between two markers, the probabilities of observing parental ditype, tetratype, and nonparental ditype are $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively. If there is a single chiasma between two markers, the resulting tetrads always have tetratype. Therefore, we can estimate the probability of having two chiasmata by $4N$, and the probability of having one chiasma by $T - 2N$. This leads to an estimated distance of $\frac{1}{2}(T + 6N)$ if we assume there are no more than two chiasmata between the two markers. This formula was first proposed by Perkins (1949).

Under the assumption of NCI, Mather (1935) showed that if $k \geq 1$ chiasmata occur between a pair of markers, then the conditional probabilities p_0^k, p_1^k, and p_2^k of observing a tetrad with parental ditype, tetratype, and nonparental ditype, respectively, are

$$p_0^k = \tfrac{1}{3} \left(\tfrac{1}{2} + (-\tfrac{1}{2})^k \right),$$

$$p_1^k = \tfrac{2}{3} \left(1 - (-\tfrac{1}{2})^k \right),$$

$$p_2^k = \tfrac{1}{3} \left(\tfrac{1}{2} + (-\tfrac{1}{2})^k \right).$$

For a given crossover process model, the above relations can be used to relate the probabilities of three tetrad patterns to the genetic distance between two markers. For example, under the Poisson model, $p_0 = \frac{1}{6}(1 + 2e^{-3d} + 3e^{-2d})$, $p_1 = \frac{2}{3}(1 - e^{-3d})$, and $p_2 = \frac{1}{6}(1 + 2e^{-3d} - 3e^{-2d})$, where p_0, p_1, and p_2 are the probabilities of parental ditype, tetratype, and nonparental ditype between two markers, respectively, and d is the genetic distance; see Haldane (1931).

One unique feature of ordered tetrad analysis is that there is information on centromeres. The distance between a single marker and its centromere can be estimated using data from a single marker. For marker $\mathcal{A}$ with alleles A and a inherited from two parents, there are six distinguishable configurations, as illustrated in Table 1.1. Because spindle–centromere attachment during meiosis is random (see Griffiths *et al.*, 1996), types 1 and 6 have equal

Table 1.1 Six distinguishable patterns for marker $\mathcal{A}$. Strands 1 and 2 are attached to one centromere and strands 3 and 4 are attached to the other.

Strand	S_1	S_2	S_3	S_4	S_5	S_6
1	A	A	A	a	a	a
2	A	a	a	A	A	a
3	a	A	a	A	a	A
4	a	a	A	a	A	A

probability because of random spindle–centromere attachment at the first meiotic division, whereas types 2–5 have the same probability because of random spindle–centromere attachment at the second meiotic division. Types 1 and 6 are called first-division segregation (FDS) pattern and types 2–5 are called second-division segregation (SDS) pattern (Griffiths *et al.*, 1996).

The probabilities of FDS and SDS can be related to the genetic distance between $\mathcal{A}$ and its centromere if a chiasma process model is specified. Let $S_{\mathcal{A}} = P(\text{SDS})$; then $S_{\mathcal{A}} = c_1 = 2d$ under the complete interference model, where c_1 denotes the probability of having one chiasma. For the Poisson model, $S_{\mathcal{A}}(d) = 1 - F_{\mathcal{A}}(d) = \frac{2}{3}(1 - e^{-3d})$. Several chiasma models and various map functions derived from these models were studied by Zhao and Speed (1998a). It was found that most map functions proposed in the literature can be well approximated by the map functions under the chi-square model. Centromeres can also be mapped with three markers on three different chromosomes using unordered tetrads, as shown by Whitehouse (1957). For three markers $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, denote the frequencies of SDS for these three loci by x, y, and z. Then the probability of tetratype between $\mathcal{A}$ and $\mathcal{B}$ is $T_{\mathcal{AB}} = x + y - \frac{3}{2}xy$ when $\mathcal{A}$ and $\mathcal{B}$ are on different chromosomes. Similarly, $T_{\mathcal{AC}} = x + z - \frac{3}{2}xz$ and $T_{\mathcal{BC}} = y + z - \frac{3}{2}yz$. These three equations can be used to solve for three unknown parameters. For example,

$$x = \frac{2}{3}\left\{1 \pm \sqrt{\frac{4 - 6T_{\mathcal{AB}} - 6T_{\mathcal{AC}} + 9T_{\mathcal{AB}}T_{\mathcal{AC}}}{4 - 6T_{\mathcal{BC}}}}\right\}.$$

However, this method only has reasonable precision when at least two of the three loci are fairly close to their respective centromeres.

For a cross involving three markers $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ on the same chromosome, if both marker intervals ($\mathcal{A}$–$\mathcal{B}$ and $\mathcal{B}$–$\mathcal{C}$) show tetratypes, there are three types that can be distinguished according to the pattern between $\mathcal{A}$ and $\mathcal{C}$: parental ditype, tetratype, and nonparental ditype (often called two-strand, three-strand, and four-strand doubles). Under the assumption of NCI, the ratio of these three types is $1:2:1$. A significant deviation from the expected ratio can be attributed to the presence of chromatid interference. Geneticists have examined this ratio in different organisms and, overall, found no consistent evidence against the NCI assumption; see Fincham *et al.* (1979). Although most studies on ordered tetrads and unordered tetrads used only three loci for the detection of chromatid interference, some information is lost when only the $1:2:1$ ratio is examined for each pair of marker intervals. Zhao *et al.* (1995b) derived a set of linear equality and inequality constraints on the probabilities of unordered tetrad patterns with an arbitrary number of loci under the assumption of NCI. For example, for two markers, NCI imposes that $p_0 \geq p_2$ and $p_1 \geq 2p_2$. Similar constraints were derived for ordered tetrads by Zhao and Speed (1998a). These constraints can be used to test the presence of chromatid interference without assuming any model for the chiasma process.

To perform genetic mapping using multiple markers simultaneously, we need to be able to evaluate the probability for any multilocus tetrad pattern under a given model for the chiasma process. Both the count-location model (Risch and Lange, 1983) and the chi-square model (Zhao *et al.*, 1995a; Zhao and Speed, 1998a) have been applied to analyze tetrad data. Detailed procedures can be found in these papers.

1.2.8   Half-tetrads

Half-tetrads can arise either from meiosis I or meiosis II nondisjunctions. The first well-studied half-tetrad data were attached-X chromosomes in Drosophila (Beadle and Emerson, 1935). Half-tetrads were also constructed using autosomes in Drosophila (Baldmin and Chovnick, 1967), and have been used in the study of many other organisms, including maize, potatoes, leopard frog, rainbow trout, salmonid fish, catfish, and zebrafish. In mammals, half-tetrads can be studied in the form of uniparental disomy (Robinson *et al.*, 1993), autosomal trisomies (Morton *et al.*, 1990), nondisjunction in ovarian teratomas (Eppig and Eicher, 1983), and PCR analysis of meiosis I products in individual secondary oocytes (Cui *et al.*, 1992).

Genetic mapping using genetic marker data from human nondisjunction data was discussed by Shahar and Morton (1986); Chakravarti and Slaugenhaupt (1987); Chakravarti *et al.* (1989); and Feingold *et al.* (2000). Map distances, as well as LOD scores for these distances, can be calculated from the observed patterns of nonreduction (heterozygous genotype) and reduction (homozygous genotype) of markers along the nondisjoined chromosome pair. Zhao and Speed (1998b) derived the general relationship between multilocus half-tetrad probabilities and multilocus ordered tetrad probabilities. These relationships can be used for likelihood analysis of half-tetrads, and the same methods have been extended to study uniparental disomy and trisomy.

1.2.9   Other Types of Data

Genetic maps can also be constructed using other types of data, including organisms with more than two copies of chromosomes (Bailey, 1961; Wu *et al.*, 2001; Wu and Ma, 2005; Luo *et al.*, 2006), bacterial and bacteriophage (Stahl, 1979), and recombinant inbred strains (Green, 1981). Genetic background and statistical methods for these types of data can be found in these references.

1.2.10   Current State of Genetic Maps

Statistical methods for establishing linkage and estimating recombination fractions in humans were pioneered by Bernstein (1931), and developed intensively by the British school centered around Fisher and Haldane during the 1930s and 1940s. The first human linkage to be established was between the X-linked genes for hemophilia and red-green color blindness by Bell and Haldane (1937); two decades later, Mohr (1954) found linkage between two blood groups on an autosome. Early ways of establishing linkage were based upon what are now known as *score tests*, and a method using sib-pairs, while likelihood methods quickly came into use for estimation. Several methods of correcting for sampling biases were also developed. All of these ideas continue to be important today.

A major limitation in human genetic mapping before the 1980s was the shortage of genetic markers. Markers are Mendelian factors, often but not necessarily genes in the modern sense, which segregate in human populations. For many years, human genetic markers were mainly blood cell antigens and proteins. They provided the basis of human genetic maps, and were a framework within which new genes could be mapped. Despite there being a fair number of known genetic diseases and Mendelian markers such as those just mentioned, the human genetic map was still very sparse in the 1970s. However, it was during this period that the first good algorithms for calculating probabilities over pedigrees were developed, motivated initially by problems in genetic counseling, and

later by the desire to carry out segregation analyses on large pedigrees. Programs based upon these algorithms continue to play a very important role in modern genetic mapping.

Around 1980, the idea of treating DNA sequence differences as genetic markers arose. It was quickly developed, and the present wide availability of what are collectively known as *molecular markers* has revolutionized human genetic mapping, and that of many other organisms. Development of the centre-d'Étude-du-polymorphisme-humain (CEPH) reference families (Dausset *et al.*, 1990) was a critical step in genetic map construction. The first fairly complete human map was published in 1987, and consisted of about 400 restriction fragment length polymorphisms (RFLPs), mapped using DNA from a panel of 21 three-generation families from the CEPH consortium (Donis-Keller *et al.*, 1987). In order to build this map, new multilocus methods for mapping were developed. The mapping of many loci simultaneously was first carried out by Fisher (1922), but it was only following the availability of cheap, fast computers and suitable algorithms that this idea became widely adopted.

At the time the 1987 map was being developed, the PCR was beginning to revolutionize molecular genetics. Several new types of genetic markers have been developed using PCR, with acronyms such as RAPD (random amplified polymorphic DNAs), STRP (short tandem repeat polymorphism), and SSCP (single-strand conformation polymorphism), and the latest genetic maps include several thousand readily assayed markers. It is now possible to carry out genome-wide scans, effectively searching the entire genome for linkage between a trait and markers. Searches of this kind have been remarkably successful in locating genes contributing to a wide range of disease and other phenotypes. They also raise many new statistical questions, especially as interest now focuses on complex and quantitative traits. Such traits are believed to be influenced by a number of genes, as well as the environment, and mapping these genes with available data remains a challenging task.

In recent genetic maps constructed from CEPH pedigrees, more than 8000 STRPs were mapped to the human genome by Broman *et al.* (1998). This map not only provides guidelines for disease gene mapping, but also allows a very detailed comparison between male and female genetic maps (Broman *et al.*, 1998) and the study of crossover interference (Broman and Weber, 2000). More recently, Kong *et al.* (2002) estimated recombination rates across the human genome through 5136 microsatellite markers typed for 146 Icelandic families, with a total of 1257 meiotic events. They detected 'systematic differences in recombination rates between mothers and between gametes from the same mother, suggesting that there is some underlying component determined by both genetic and environmental factors that affects maternal recombination rates'. In Figure 1.2 we show some features of portions of several genetic maps for human chromosome 21. Most recently, there have been extensive studies of single nucleotide polymorphisms (SNPs) in the human genome. The current map contains more than 10 million SNPs (`www.ncbi.nlm.nih.gov/SNP`) and maps for other types of polymorphisms are also being developed, e.g. the INDEL map (Mills *et al.*, 2006).

Genetic maps for pigs, cows, tomatoes, rice, pine trees and many other species of commercial or scientific interest have followed quickly behind the human maps.

1.2.11 Programs for Genetic Mapping

All the programs described here can be found in the website maintained by the Ott group at Rockefeller University, at `http://linkage.rockefeller.edu/soft/list`.
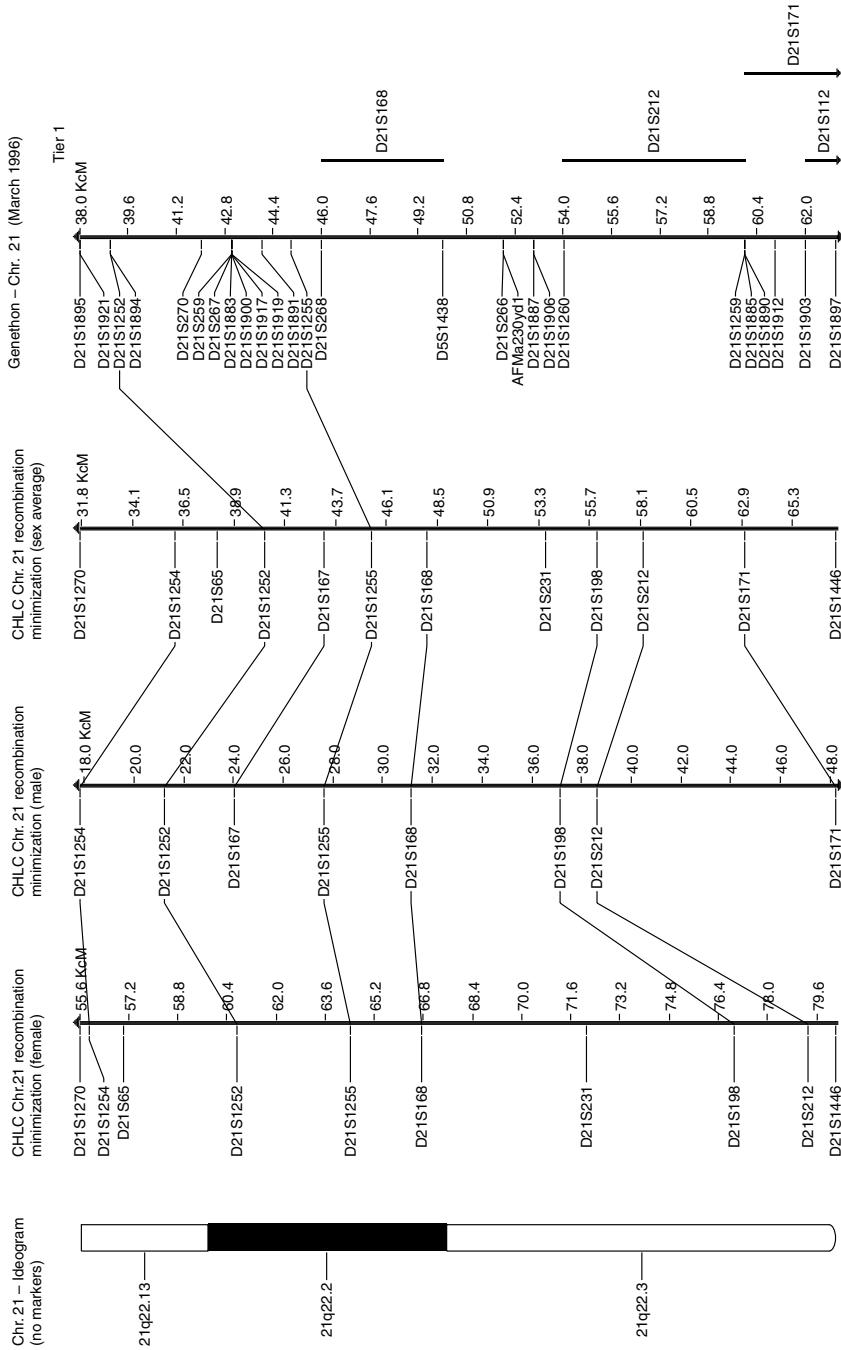
Figure 1.2 Portion of human chromosome 21 genetic map. [Source: `www.gdb.org/hugo/chr21/geneticMaps.html.`]

`html`. As discussed above, algorithms for carrying out multipoint linkage analysis with human (and other) pedigree data are of two kinds: those based upon the Elston and Stewart (1971) approach, using what is known as *peeling*; and those based upon the Lander and Green (1987) hidden Markov model formulation. Each of these classes of algorithms has its strengths and weaknesses, and there are problems that cannot be solved exactly with either of them. The Elston–Stewart approach underlies most of the algorithms discussed in Terwilliger and Ott (1994). For a recent improvement of the implementation of these algorithms, see O'Connell and Weeks (1995). A suite of genetic mapping programs that have gained much popularity uses the basic Lander–Green algorithm in a number of different human linkage problems. These include MAPMAKER for crosses among inbred strains (Lander *et al.*, 1987), analyses with sib-pairs (Kruglyak and Lander, 1995), the analysis of recessive traits with nuclear families (Kruglyak *et al.*, 1995), and multipoint linkage with many markers for general pedigrees of moderate size (Kruglyak *et al.*, 1996). Similar statistical methods underlie the program CRI-MAP, which is most suitable for CEPH-type pedigrees (Green, 1988). MultiMap is another program that assists with map construction (Matise *et al.*, 1994). It consists of framework construction and comprehensive map construction. MultiMap recently incorporated the Gene Mapping System algorithm (Lathrop *et al.*, 1988; Marinov *et al.*, 1999), which is based on identifying and permuting linkage groups within an initial order of all loci. Other programs include Map/Map+ for map integration (Morton *et al.*, 1992), JoinMap for plants (Stam, 1993), and OUTMAP for outbred populations (Ling, 2000).

When exact linkage analysis methods fail because of time or space constraints, Monte Carlo methods may be used. At present these are more research tools than approaches suitable for routine use, but they are developing rapidly, and should become more widely used in the near future. Some of these simulation methods have been implemented in SIMWALK (Sobel and Lange, 1996) and Morgan (Thompson, 1994).

1.3   PHYSICAL MAPS

Physical mapping is the process of determining the locations of 'sites' such as restriction sites (4–8 bp), STSs (20–30 bp), and cloned fragments (kilobase to megabase) on a larger DNA molecule or a chromosome. Among other things, maps of such sites are helpful, if not essential, for cloning genes and for sequencing large stretches of DNA, and have been very widely used in recent years. To quote from an early successful paper in the area, Olson *et al.* (1986, p. 7830):

> a strong case can be made for the value of constructing physical maps of the genomes of intensively studied organisms. We expect the main value of these maps to lie in facilitating the organization of molecular genetic information. Just as conventional cartography provides an indispensable framework for organizing data in fields as disparate as demography and geophysics, it is reasonable to suppose that 'DNA cartography' will prove equally useful in organizing the vast quantities of molecular genetic data that may be expected to accumulate in the coming decades. Furthermore, the principal by-product of these projects – global clone collections that are cross-indexed to the physical maps – could be expected to improve the efficiency of subsequent structural and functional studies of local regions.

The two most common approaches to physical mapping are termed top-down, producing a macrorestriction map, and bottom-up, resulting in a contig map. With either strategy, the maps represent ordered sets of DNA fragments that are generated by cutting genomic DNA.

However, the first physical maps were made from microscope images, and although their construction and interpretation involve no statistics, we discuss them briefly for completeness.

1.3.1 Polytene Chromosomes

Polytene chromosomes are many-stranded chromosomes resulting from repeated chromosomal replication, without the subsequent separation of sister chromatids. Up to 1024 chromatids can be present, giving giant chromosomes visible under a microscope in nondividing cells. Most widely known are those of the salivary glands of insects of the order Diptera, and a classic reference to the polytene chromosomes in the salivary glands of *Drosophila melanogaster* is Bridges (1935).

After appropriate staining, the Drosophila polytene chromosomes have distinctive banding patterns, which have been cataloged, and have proved invaluable for localizing structural alterations such as deletions, and for use with more recent *in situ* hybridization techniques. For further details, please refer to Saura *et al.* (1997) and FlyBase (`http://flybase.bio.indiana.edu`).

1.3.2 Cytogenetic Maps

A closely related class of physical maps are the familiar cytological maps whose human versions are frequently represented in ideogrammatic form (see Figure 1.1). Such maps were originally derived in a variety of organisms by associating mutant phenotypes with chromosomal defects visible by direct microscopic examination. In this way, genes can be physically located on chromosomes, at least to a low level of resolution.

In the late 1960s, staining techniques were discovered, which led quickly to the adoption of banding patterns of human chromosomes now widely used; see, for example, Vogel and Motulsky (1997). This field has evolved greatly in recent years with the advent of fluorescent in situ hybridization (FISH) and multiple coloring of chromosomes; see Trask (1998).

1.3.3 Restriction Maps

A *restriction site* is the location of a sequence, typically 4–6 bp long, where a particular restriction enzyme will cut DNA. Isolated from various bacteria, restriction enzymes recognize short DNA sequences and cut DNA molecules at specific sites in the sequence. Ignoring, for the moment, variations from uniform base composition, restriction enzymes with 4 bp recognition sites will yield pieces – termed *restriction fragments* – on average about 256 bp long, while those with 6 or 8 bp recognition sites will yield pieces of average length 4 or 64 kb, respectively. Since hundreds of different restriction enzymes have been characterized, and they can be used together, DNA can now be cut with them into fragments of many different sizes in many different ways. A restriction map describes the order and distance between restriction sites.

In *top-down mapping*, a chromosome is cut into large DNA fragments using restriction enzymes having rare restriction sites. The fragments are separated by size and assigned to regions by hybridization with genetically or cytogenetically mapped DNA probes. Then

the fragments are assembled into contiguous blocks, resulting in a macrorestriction map. Such fragments may average 1 Mb in size. For a finer map, the ordered fragments may be taken one at a time and dissected with more frequently cutting restriction enzymes.

The simplest way to construct a restriction map is to compare the fragment sizes produced when a DNA molecule is digested with two different restriction enzymes; see Waterman (1995, Chapters 2–4) for a discussion of some of the computational issues here. Restriction maps are easy to generate if there are relatively few cut sites with the enzymes being used, but most enzymes cut frequently, generating a large number of small fragments (from less than 100 bp to 1 kb). Therefore, such mapping is more applicable to small molecules, such as viral and organelle genomes, or to genomic DNA that has already been cloned.

A major advantage of a restriction map (like that in Figure 1.3) is that accurate lengths are known between sets of reference points. We can preserve an overview of the target and we can reach a nearly complete map relatively quickly. The disadvantage of most restriction mapping efforts is that they do not produce the DNA in a convenient form. This approach yields maps with more continuity and fewer gaps between fragments than contig maps, but map resolution is lower and may not be useful in finding particular genes. Currently, this approach allows DNA pieces to be located in regions measuring from about 0.1 Mb to 1 Mb.

1.3.4   Restriction Mapping via Optical Mapping

Optical mapping is a single-molecule approach for the construction of ordered restriction maps developed by Schwartz *et al.* (1993). It uses light microscopy to directly image individual DNA molecules, which are bound to specially derivatized surfaces and then cleaved by restriction enzymes. Cleaved fragments retain their original order, and restriction sites are flagged by small, visible gaps. Optical mapping solves the problem of determining fragment order.

The statistical analysis of optical mapping data is relatively new and quite complex, so we do not attempt to summarize it here. A first solution to the problem can be found in Anantharaman *et al.* (1997). These authors take a Bayesian approach, constructing a prior model for an ordered restriction map and a probability model for restriction map data from single molecules. They then approximate the mode of the posterior distribution of the parameters. Orientation, false cuts, and sizing errors are among the issues to be dealt with. A second, hierarchical Bayes approach to the same problem using reversible-jump Markov chain Monte Carlo can be found in Lee *et al.* (1998). Most recently, sequence assembly methods have been adapted to optical mapping (Valouev *et al.*, 2006). Finally, there have been several successful applications of the method to a whole genome; for example, see Lai *et al.* (1999) and Reslewic *et al.* (2005).

1.3.5   Ordered Clone Maps

Clones – more fully, cloned DNA fragments – are generated by first breaking a large number of identical chromosomes into fragments, either by physical means such as sonication, compression, or irradiation, or by chemical means, typically complete or partial digestion with one or more restriction enzymes. Individual fragments (inserts) of appropriate sizes are then joined to another DNA molecule (the vector) and the result is incorporated into a (host) organism such as *Escherichia coli* or yeast. The average
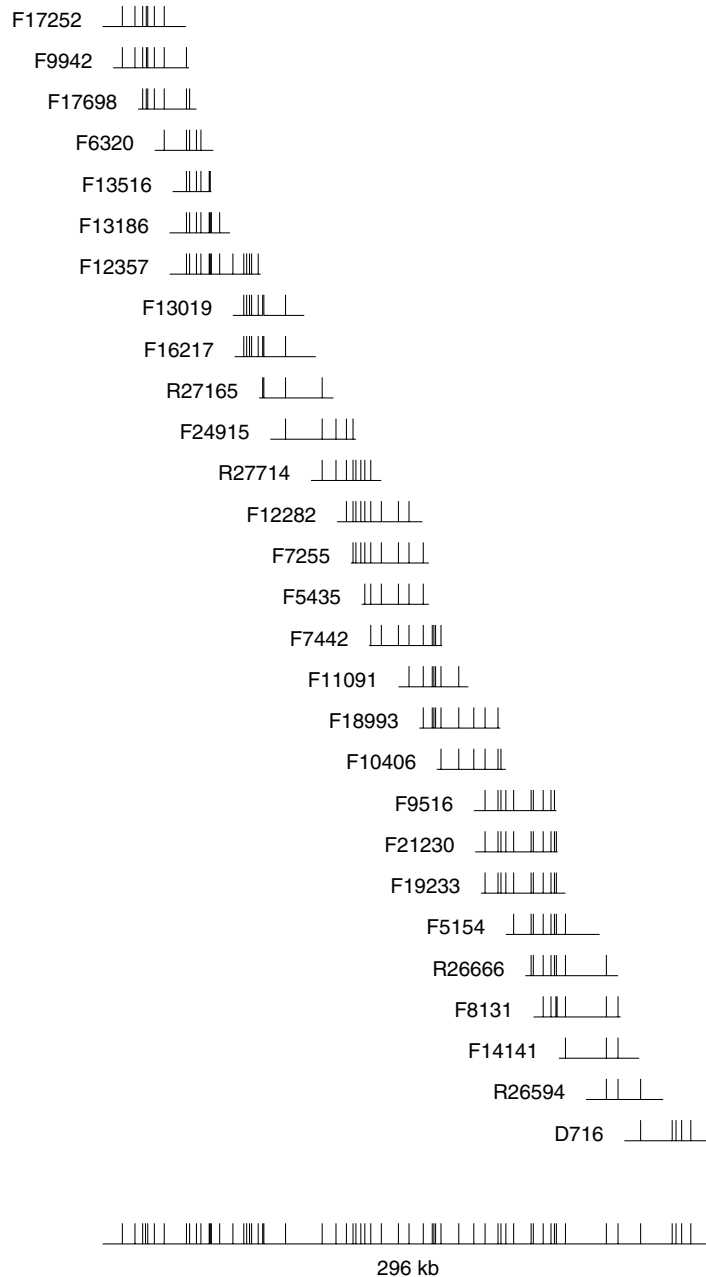
Figure 1.3   Restriction map of cosmid clones. [Source: Lawrence Livermore National Laboratory.]

size of the insert varies widely among different hosts and incorporation methods. Yeast artificial chromosomes (YACs) in yeast may have DNA fragments ranging from 100 kb to 1 Mb. Cosmids in *E. coli* may have fragments ranging from 35 to 45 kb, while the now widely used bacterial artificial chromosomes (BACs) have inserts of sizes in

the range 100–200 kb. The hosts are separated from each other and allowed to grow into colonies, with the fragment in each host being replicated along with the host's DNA during cell divisions. After enough divisions, each host colony can be harvested, resulting in a library of cloned DNA fragments, where each fragment is present in large enough quantities to permit isolation and purification for subsequent biochemical analyses.

The bottom-up approach to physical mapping is usually carried out by breaking up the DNA molecule of interest, cloning selected fragments, and subjecting each clone to one more experiments – restriction digestions, hybridizations or PCR assays with unique or repetitive probes, or sequencing – to obtain what is called a *fingerprint* of the clone. These fingerprints are then used to solve the combinatorial puzzle of inferring the arrangement of clones along the molecule with the help of these data. The ordered fragments form contiguous DNA blocks, which are called *contigs*. Clone ordering usually begins by comparing clones to each other, in order to determine the strength of evidence that any pair of clones overlap, and it is here that statistical ideas enter.

Currently, clone libraries ordered in this way have inserts that vary in size from a few thousand base pairs up to 1 Mb. Contig maps thus consist of a linked library of overlapping clones representing a complete chromosomal segment. An advantage of this approach is the accessibility of these clones to other researchers. While useful for finding genes localized to a small region (under 2 Mb), contig maps can be difficult to extend over large stretches of a chromosome because not all regions are clonable.

The statistical analysis of overlap and the estimation of distances will differ somewhat with different fingerprinting techniques. In a hybridization experiment, the fingerprint will be the list of probes that hybridize to the clone; with restriction digestion the fingerprint is a list of observed fragment sizes resulting from the digestion of the clone, while with STS content mapping (see below) the fingerprint consists of an enumeration of the STSs found to be located on that clone. An example of an STS-based clone map is given in Figure 1.4.

1.3.6   Contig Mapping Using Restriction Fragments

One approach, due by Coulson *et al.* (1986), begins with the calculation, for each pair of clones, of the probability of the observed level of matching of fragment sizes, up to a prescribed tolerance, arising by chance, i.e. when the clones do not in fact overlap. This probability – essentially a *p*-value, but called *the probability of coincidence* – is used for selecting possibly overlapping clone pairs. Clones are then assembled into contigs using a variety of *ad hoc* rules based on these probabilities. For details, we refer to Sulston *et al.* (1988). A modified version of this procedure is embodied in the program FPC (Soderlund *et al.*, 1997), which was widely used in preparing physical maps for human genome sequencing.

An alternative approach involving a likelihood ratio or Bayes posterior odds was initiated by Michiels *et al.* (1987), and then more fully developed by Branscomb *et al.* (1990). We sketch it now, referring the reader to Nelson and Speed (1994a) and Nelson *et al.* (1997) and the references cited there for fuller details of the trinomial model.

For each clone in a library of N clones, we create DNA fingerprint data by restriction digestion, electrophoresis, and sizing. This consists of a list of fragment lengths. For a particular length l, there are four patterns when we compare two clones: (1,1) if a fragment of length l is observed in both clones; (1,0) if a fragment of length l is observed in the first clone but not the second one; (0,1) if a fragment of length l is observed in
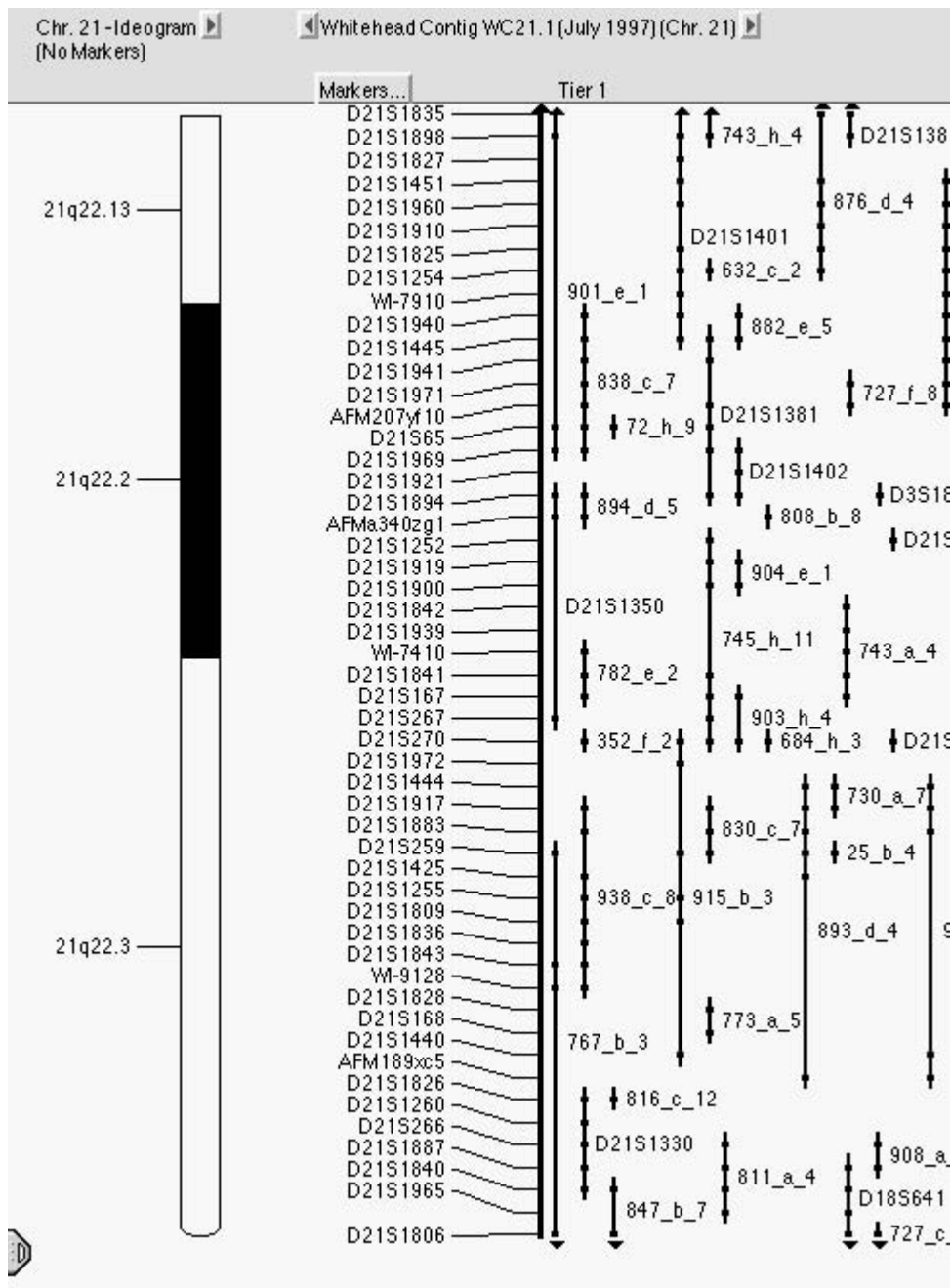
Figure 1.4 STS map of portion of human chromosome 21. [Source: Mapview at `www.gdb.org/hugo/chr21/`.]

the second clone but not the first one; and (0,0) if a fragment of length l is observed in neither clone. The probability of each outcome under a simple randomness model can be approximated by

$$
\begin{aligned}
p_{00} &= q^{L_1 + L_2 - \theta}, \\
p_{01} &= q^{L_1}(1 - q^{L_2 - \theta}), \\
p_{10} &= q^{L_2}(1 - q^{L_1 - \theta}), \\
p_{11} &= 1 - q^{L_1} - q^{L_2} + q^{L_1 + L_2 - \theta},
\end{aligned}
$$

where L_1 and L_2 are the lengths of the two clones, θ is the overlapping amount, $q = \mathrm{e}^{-\lambda_l}$, and λ_l is the intensity for fragments of size l. When all the clones are of the same size, $p_{01} = p_{10}$ and the data can be reduced to a trinomial variable $(n_{00}, n_{01} + n_{10}, n_{11})$. To decide whether two clones overlap, the likelihood ratio test value $L(\theta)/L(\theta = 0)$ can be calculated. Alternatively, with prior information on the θ, posterior odds can be calculated to decide if two clones overlap. The above simple assumptions can be loosened to allow different intensities and errors in fragment size detections. With pairwise similarity measures such as the log posterior odds for overlap, clustering algorithms can be used to build contigs.

1.3.7   Sequence-tagged Site Maps

An STS is defined by two short sequences, each typically $20-25$ bp in length, that have been designed from a region of sequence that appears as a single copy in the human genome. These sequences can act as primers in a PCR assay to score for presence or absence of the site in any DNA sample. One of the aims of the human genome project was to build a high-resolution RH map using STSs as landmarks throughout the human genome. Geneticists would then be able to use the map to isolate genes through nearby landmarks. Sequencers would be able to decide where to prepare clones for the actual sequencing. In addition, the STSs would become part of a common set of markers that can be screened against maps created using different mapping techniques, helping to integrate the efforts of mapping teams worldwide.

For STS content mapping, the data can be summarized as an incidence matrix with N rows corresponding to N clones and M columns corresponding to M STSs. The (i, j)th entry is 1 if the jth STS hits the ith clone, and is 0 otherwise. If there are no errors, the problem can be solved by testing whether this incidence matrix has the consecutive 1s property. An incidence matrix has the consecutive 1s property for rows if its columns can be permuted so as to make all the 1s in each row appear consecutively. Booth and Lueker (1976) described linear-time algorithms for determining if a matrix has the consecutive 1s property, and they provided a compact description of all possible consistent permutations in the form of a PQ-tree. Therefore, the problem is completely solvable in linear time if the data are error-free.

However, real data sets are never error-free, and the consecutive 1s property no longer holds. Before we discuss various algorithms in the literature, most of which use combinatorial approaches to optimize an objective function of orderings, we consider a likelihood-based approach for STS ordering following Yeh (1999). For a pair of STSs (s_i, s_j), distance D apart, we can count the number of clones retaining both STSs (n_{11}),

the first STS but not the second STS (n_{10}), the second STS but not the first STS (n_{01}), and neither STS (n_{00}). Define the set of coretention probabilities for (s_i, s_j) as $\mathbf{p} = (p_{11}, p_{10}, p_{01}, p_{00})$, where p_{11} is the probability that a clone contains both s_i and s_j, p_{10} is the probability that it retains s_i only, p_{01} is the probability that it retains s_j only, and p_{00} is the probability that is retains neither. Assuming clones are random line segments of length L on the chromosome (genome) of size G, the coretention probabilities are

$$p_{11} = l - d,$$
$$p_{10} = d,$$
$$p_{01} = d,$$
$$p_{00} = 1 - l - d,$$

when $d < l$, and

$$p_{11} = 0,$$
$$p_{10} = l,$$
$$p_{01} = l,$$
$$p_{00} = 1 - 2l,$$

when $d \geq l$, where l is the normalized clone length L/G and d is the normalized STS distance D/G. These probabilities allow us to evaluate the likelihood for each ordering of the STSs. The estimated order is the one that maximizes the overall likelihood. Thus the STS ordering problem is equivalent to a TSP with the STSs as vertices and the pairwise likelihoods as the distances. Using arguments similar to those in marker ordering in genetic mapping (Speed *et al.*, 1992); Yeh (1999) showed that this procedure will recover the true order with probability 1 when the number of clones is large. The objective function here is the same as that discussed by Green and Green (1991), which is the first major paper on this topic, and which describes in outline the widely used program SEGMAP. However, that program goes considerably further, including the solution of a linear programming problem to find bounds on the distance between any pair of points in the map (STSs or clone ends).

In addition to obtaining distance estimates among the STSs, this likelihood approach is robust when the error rates are not too high. Mott *et al.* (1993) used simulated annealing to minimize the total discrepancy among adjacent STSs, where the discrepancy between two STSs a and b is defined as

$$d(a, b) = 1 - \frac{\#(\text{clones positive for } a \text{ and } b)}{\#(\text{clones positive for } a \text{ or } b)}.$$

Alizadeh *et al.* (1995) used the TSP algorithms to minimize the total Hamming distance of the clone-probe incidence matrix, which corresponds to the number of gaps of the probe ordering. They also proposed an alternative objective function based on a weighted sum of the number of chimeric clones, the number of false positives, and the number of negatives. Christof *et al.* (1997) formulated the problem as a weighted betweenness problem, assuming the probes are from both ends of all clones. Alizadeh *et al.* (1995) and Nelson *et al.* (1997) described statistical procedures for evaluating overlapping configurations involving more than two clones.

Once STSs are ordered, clones are ordered with respect to the probes by maximizing a measure of fit between the probe data for that clone and the list of ordered probes. Unlike physical maps constructed from restriction maps, the map constructed using STS content mapping would not be tied to a particular set of clones, and thus could be used to order any subsequently generated library.

In the early 1990s, considerable effort was put into the generation of clone contig maps, using STS screening. The major achievement of this phase of physical mapping was the publication of a clone contig map of the entire genome, consisting of 33 000 YACs containing fragments with an average size of 0.9 Mb (Cohen *et al.*, 1993).

The combined STS maps now include positions for almost 7000 simple sequence length polymorphisms that have already been mapped onto the genome by genetic means. As a result, the physical and genetic maps can be directly compared, and the clone contig maps that include STS data can be anchored on to both maps.

1.4   RADIATION HYBRID MAPPING

A radiation hybrid is a rodent cell that contains fragments of chromosomes from a second organism. The technology was first developed in the 1970s by Goss and Harris (1975; 1977) and reintroduced by Cox *et al.* (1990) based on the observation that exposure of human cells to X-rays causes the chromosomes to break up randomly into fragments, and these chromosome fragments can then be propagated if the irradiated cells are fused with nonirradiated hamster or other rodent cells. A routine selection process is used to screen out hamster cells without human chromosome fragments. In its simplest form, a single human chromosome is exposed to a radiation source. For the whole-genome radiation mapping whole genome radiation hybrid (WG-RH), a normal diploid human cell is used as the donor by Walter *et al.* (1994). The WG-RH mapping has the advantage that pieces of many different human chromosomes may be contained in the same hybrid and so a single panel of WG-RHs can be used to map any region of the human genome. Detailed mapping of the entire human genome can be accomplished with fewer than 100 WG-RHs. The resulting panels can be screened for human-specific markers. Data for RH mapping also can be summarized in an incidence matrix like the one for STS content mapping.

Like STS content mapping, the basic premise of RH mapping is that the closer two loci are on the human chromosome, the less likely it is that they will be broken by irradiation. The retention patterns from the various hybrid clones give clues for determining locus order and for estimating the distance between adjacent loci for a given order.

One criterion that quantifies this heuristic is the minimum obligate breaks criterion. For a given order of the loci, we can count the number of changes from 1 to 0 and from 0 to 1. If we sum these changes over all the clones, we get the total number of obligate breaks. The objective is to find the order that minimizes the total number of obligate breaks across all clones. The advantage of the minimum breaks criterion is that it does not depend on any assumptions about how breaks occur and how fragments are retained. Assuming the same retention rate, one can again use arguments similar to those in marker ordering for genetic mapping and show that this criterion is strongly statistically consistent (Lange, 1997, Chapter 11).

1.4.1 Haploid Data

We now turn to probabilistic models for RH mapping. In RH mapping, the distance between sites can be expressed in units (centirays) representing the percentage probability of separation by breakage with a given irradiation dosage. This gives a better measure of physical distance than genetic distance, because the vulnerability to breakage seems to be fairly constant along the whole length of the chromosome. Therefore, in models for RH mapping, the breakage process along the chromosome can be modeled as a Poisson process (Cox *et al.*, 1990). For any two loci, the probability of at least one break θ and the physical distance δ are related by

$$1 - \theta = e^{-\lambda\delta},$$

where the value of λ depends on the irradiation dose. This function is similar to the Haldane map function used in genetic mapping. Note that the parameters δ and λ cannot be separated from the estimation. For RH mapping, in addition to considering breakage, we also need to take retention into account. It is normally assumed that different segments are retained independently; however, different fragments may be allowed to have different retention probabilities.

For two markers, there are four possibilities for a haploid RH: (1,1) when both markers are present; (1,0) when the first marker is present but not the second marker; (0,1) when the second marker is present but not the first one; and (0,0) when neither marker is present. The probabilities for these four patterns are as follows:

$$p_{11} = \theta P_A P_B + (1 - \theta) P_{AB},$$
$$p_{10} = \theta P_A (1 - P_B),$$
$$p_{01} = \theta P_B (1 - P_A),$$
$$p_{00} = \theta (1 - P_A)(1 - P_B) + (1 - \theta)(1 - P_{AB}),$$

where P_A, P_B, and P_{AB} are the probabilities of a hybrid retaining a fragment with marker A only, with marker B only, and with both markers A and B, respectively (Cox *et al.*, 1990). In the general case of many markers, Boehnke *et al.* (1991) derived the probability for a hybrid with any retention pattern.

Note that when $P_A = P_B = P_{AB} = r$ in the above equations, the probabilities reduce to

$$p_{11} = \theta r^2 + (1 - \theta) r,$$
$$p_{10} = \theta r (1 - r),$$
$$p_{01} = \theta r (1 - r),$$
$$p_{00} = \theta (1 - r)^2 + (1 - \theta)(1 - r).$$

Therefore, the probabilities are simply a reparametrization of those we derived when we discussed ordering STSs: simply put $l = 1 - r$ and $d = \theta r (1 - r)$.

1.4.2 Diploid Data

For the WG-RH mapping, two chromosomes instead of one, are involved. For a pair of markers, we have the same four possibilities for an RH. Assuming the same retention rate
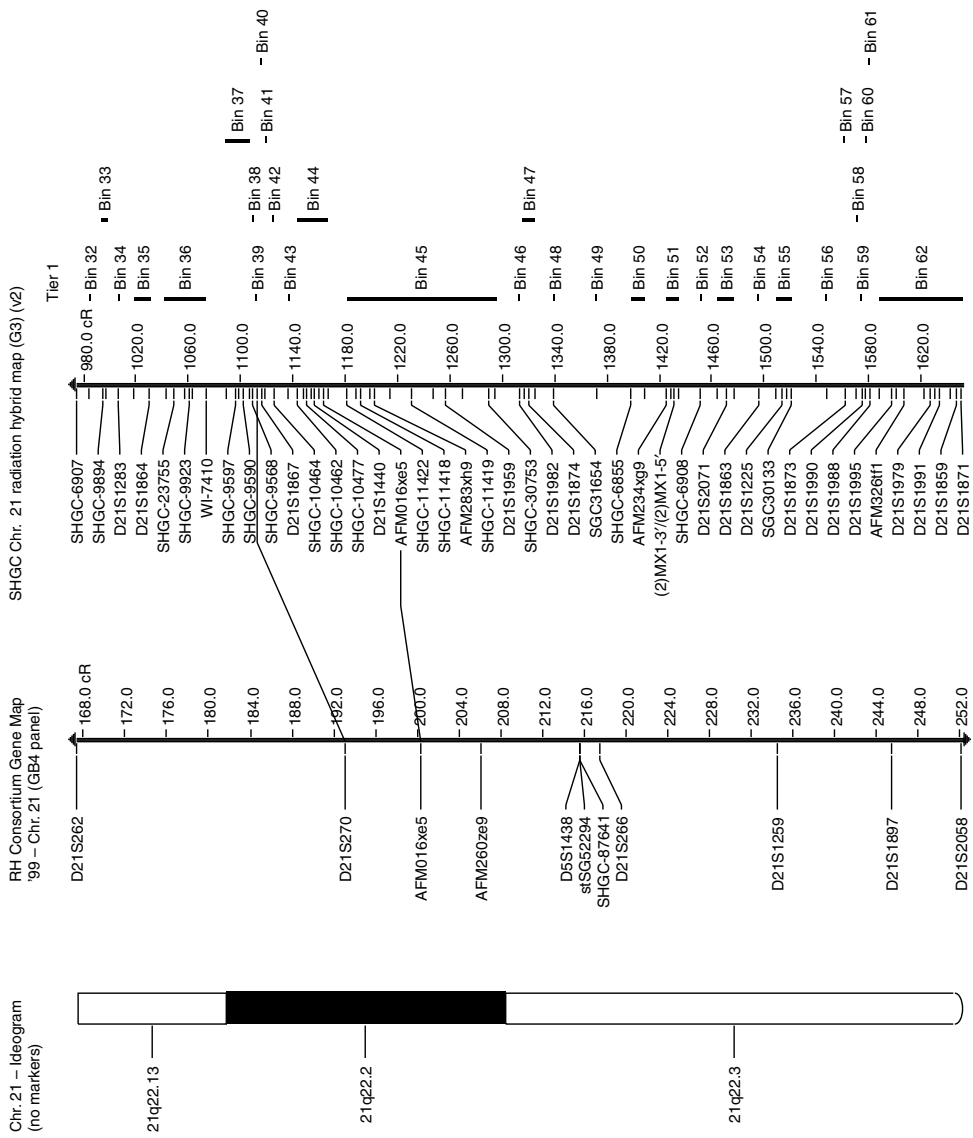
Figure 1.5 RH map of part of human chromosome 21. [Source: Mapview at `www.gdb.org/hugo/chr21/`.]

r for all fragments, the probabilities for the four possibilities are

$$p_{11} = 1 - 2(1 - r)^2 + [(1 - r)(1 - \theta r)]^2,$$

$$p_{10} = (1 - r)^2 - [(1 - r)(1 - \theta r)]^2,$$

$$p_{01} = (1 - r)^2 - [(1 - r)(1 - \theta r)]^2,$$

$$p_{00} = [(1 - r)(1 - \theta r)]^2.$$

For an arbitrary number of markers, hidden Markov models were used by Lange *et al.* (1995) to calculate the probability of any retention pattern.

When the retention probabilities are allowed to vary for different fragments, the number of parameters involved increases quadratically with the number of markers examined. Although a number of computational methods can be used for such problems (see Boehnke *et al.*, 1991; Leach and O'Connell, 1995), this raises a serious optimization problem. If the retention rate is assumed to be constant, the calculation can be simplified. Jones (1997) found that adopting simple models generally does not affect the ability to recover the true locus order, but affects the estimation of distances among the loci.

Bayesian methods have also been developed for RH mapping by Lange and Boehnke (1992) and Guerra *et al.* (1992). Tibshirani *et al.* (1999) proposed to maximize a pseudo-likelihood based on information from all marker pairs and then to use multidimensional scaling to provide starting positions for the markers. Lange (1997) has an excellent chapter on RH mapping, and also recommended is the review by Jones (2000). Other methods include Agarwala *et al.* (2000); Ben-Dor *et al.* (2000); and Ivansson and Lagergren (2004). Hitte *et al.* (2003) compared the performance of two methods for RH map constructions.

There are three human radiation hybrid panels available: Genebridge4 (93 hybrids), Stanford G3 (83 hybrids), and Stanford TNG4 (90 hybrids). These panels differ in radiation dose and retention probabilities. RH panels have also been made for mouse, rat, cow, pig, zebrafish, dog, cat, baboon, and horse. More detailed information can be found at `http://compgen.rutgers.edu/rhmap/`. An example of the RH maps are given in Figure 1.5.

1.5   OTHER PHYSICAL MAPPING APPROACHES

In *directed mapping* (see Palazzolo *et al.*, 1991; Mizukami *et al.*, 1993), random seed clones are selected first, and contigs are extended by anchors generated from contig ends. Then an unmapped clone is selected, and STSs from its ends are constructed and used to find the set of overlapping clones, usually by a PCR assay. The process continues until all clones have been either selected or identified as overlapping some selected clones. Nelson and Speed (1994b) found that a project using a directed strategy makes slower progress in the beginning, but closes the gaps much faster in later stages.

A *double end-sequencing strategy* combines sequencing and mapping by sequencing both ends of subclones and inferring clone overlaps from end-sequence comparisons; see Chen *et al.* (1993) and Roach (1995). A double end-sequencing strategy combined with directed finishing provides an efficient approach to sequencing a large piece of DNA (Yeh, 1999).

High-resolution mapping using FISH uses two or more fluorescent probes to hybridize to chromosomes at a particular stage in the cell cycle. The distance between the fluorescent dots in each cell is measured. The data from FISH experiments consist of a series of distance measurements between two or more probes. Such mapping data are now being linked to more traditional physical mapping data; see Kirsch *et al.* (2000).

1.6   GENE MAPS

Because of the possibilities of having two or more noncontiguous DNA fragments in a single clone, in 1994 the International Radiation Hybrid Mapping Consortium was formed to construct a human gene map in which cDNA-based STS markers from $3'$-untranslated regions of cDNAs were physically mapped and then integrated with the genetic map of polymorphic microsatellite markers. The consortium initially reported a map with about 16 000 genes by Schuler *et al.* (1996); a later map constructed by Deloukas *et al.* (1998) contains 30 181 gene-based markers. The resulting map density approached the target of one marker per 100 kb set as the objective for physical mapping at the outset of the human genome project. The GeneMap '98 or Human Transcript Map STSs are derived from transcribed sequences. Finally, we would like to mention the Integrated Molecular Analysis of Genomes and their Expression (IMAGE) Consortium, 'the world's largest public collection of genes'.

Acknowledgments

REFERENCES

Agarwala, R., Applegate, D.L., Maglott, D., Schuler, G.D. and Schäffer, A.A. (2000). A fast and scalable radiation hybrid map construction and integration strategy. *Genome Research* **10**, 350–364.

Alizadeh, F., Karp, R.M., Newberg, L.A. and Weisser, D.K. (1995). Physical mapping of chromosomes: a combinatorial problem in molecular biology. *Algorithmica* **13**, 52–76.

Anantharaman, T.S., Mishra, B. and Schwartz, D.C. (1997). Genomics via optical mapping. 2. Ordered restriction maps. *Journal of Computational Biology* **4**, 91–118.

Bailey, N.T.J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford University Press, London.

Baldmin, M. and Chovnick, A. (1967). Autosomal half-tetrad analysis in *Drosophila melanogaster*. *Genetics* **55**, 277–293.

Bateson, W., Saunders, E.R. and Punnett, R.C. (1905). Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society* **2**, 1–55, 88–99.

Beadle, G.W. and Emerson, S. (1935). Further studies of crossing over in attached-X chromosomes of *Drosophila melanogaster*. *Genetics* **20**, 192–206.

Bell, J. and Haldane, J.B.S. (1937). The linkage between the genes for colour blindness and haemophilia in man. *Proceedings of the Royal Society of London Series B* **123**, 119–150.

Ben-Dor, A., Chor, B. and Pelleg, D. (2000). RHO – radiation hybrid ordering. *Genome Research* **10**, 365–378.

Bernstein, F. (1931). Zur Grundlegung der Chromosomentheorie der Vererbung beim Menschen. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre* **57**, 113–138.

Bishop, D.T. and Thompson, E.A. (1988). Linkage information and bias in the presence of interference. *Genetic Epidemiology* **5**, 107–119.

Boehnke, M., Lange, K. and Cox, D.R. (1991). Statistical methods for multipoint radiation hybrid mapping. *American Journal of Human Genetics* **49**, 1174–1188.

Booth, K.S. and Lueker, G.S. (1976). Testing for the consecutive 1s property, interval graphs, and graph planarity using PQ-tree algorithm. *Journal of Computer and System Sciences* **13**, 335–379.

Branscomb, E., Slezak, T., Pae, R., Galas, D., Carrano, A.V. and Waterman, M. (1990). Optimizing restriction fragment fingerprinting methods for ordering large genomic libraries. *Genomics* **8**, 351–366.

Bridges, C.B. (1935). Salivary chromosome maps. *The Journal of Heredity* **26**, 60–64.

Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. and Weber, J.L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American Journal of Human Genetics* **63**, 861–869.

Broman, K.W. and Weber, J.L. (2000). Characterization of human crossover interference. *American Journal of Human Genetics* **66**, 1911–1926.

Browning, S. (2003). Pedigree data analysis with crossover interference. *Genetics* **164**, 1561–1566.

Carter, T.C. and Falconer, D.S. (1951). Stocks for detecting linkage in the mouse and the theory of their design. *Journal of Genetics* **50**, 307–323.

Chakravarti, A., Majumder, P.P., Slaugenhaupt, S.A., Deka, R., Warren, A.C., Surti, U., Ferrell, R.E. and Antonarakis, S.E. (1989). *Molecular and Cytogenetic Studies of Nondisjunction: Proceedings of the Fifth Annual National Down Syndrome Society Symposium*, T.J. Hassold and C.J. Epstein, eds. Alan R. Liss, New York, pp. 35–42.

Chakravarti, A. and Slaugenhaupt, S.A. (1987). Methods for studying recombination on chromosomes that undergo nondisjunction. *Genomics* **1**, 35–42.

Chen, E.Y., Schlessinger, D. and Kere, J. (1993). Ordered shotgun sequencing: a strategy for integrating mapping and sequencing of YAC clones. *Genomics* **17**, 651–656.

Christof, T., Jünger, M., Kececioglu, J., Mutzel, P. and Reinelt, G. (1997). A branch-and-cut approach to physical mapping of chromosome by unique end-probes. *Journal of Computational Biology* **4**, 433–447.

Cohen, D., Chumakov, I. and Weissenbach, J. (1993). A first-generation map of the human genome. *Nature* **366**, 698–701.

Copenhaver, G.P., Housworth, E.A. and Stahl, F.W. (2002). Crossover interference in Arabidopsis. *Genetics* **160**, 1631–1639.

Coulson, A., Sulston, J., Brenner, S. and Karn, J. (1986). Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 7821–7825.

Cox, D.R., Burmeister, M., Price, E.R., Kim, S. and Myers, R.M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250**, 245–250.

Cui, X., Gerwin, J., Navidi, W., Li, H., Kuehn, M. and Arnheim, N. (1992). Gene-centromere linkage mapping by PCR analysis of individual oocytes. *Genomics* **13**, 713–717.

Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.M. and White, R. (1990). Program description – Centre-d'Étude-du-Polymorphisme-Humain (CEPH) – collaborative genetic mapping of the human genome. *Genomics* **6**, 575–577.

Deloukas, P., Schuler, G.D., Gyapay, G., Beasley, E.M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T.C., McKusick, K.B., Beckmann, J.S., Bentolila, S., Bihoreau, M.-T., Birren, B.B.,

Browne, J., Butler, A., Castle, A.B., Chiannilkulchai, N., Clee, C., Day, P.J.R., Dehejia, A., Dibling, T., Drouot, N., Duprat, S., Fizames, C., Fox, S., Gelling, S., Green, L., Harrison, P., Hocking, R., Holloway, E., Hunt, S., Keil, S., Lijnzaad, P., Louis-Dit-Sully, C., Ma, J., Mendis, A., Miller, J., Morissette, J., Muselet, D., Nusbaum, H.C., Peck, A., Rozen, S., Simon, D., Slonim, D.K., Staples, R., Stein, L.D., Stewart, E.A., Suchard, M.A., Thangarajah, T., Vega-Czarny, N., Webber, C., Wu, X., Hudson, J., Auffray, C., Nomura, N., Sikela, J.M., Polymeropoulos, M.H., James, M.R., Lander, E.S., Hudson, T.J., Myers, R.M., Cox, D.R., Weissenbach, J., Boguski, M.S. and Bentley, D.R. (1998). A physical map of 30 000 genes. *Science* **282**, 744–746.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., Lander, E.S., Botstein, D., Akots, G., Rediker, K.S., Gravius, T., Brown, V.A., Rising, M.B., Parker, C., Powers, J.A., Watt, D.E., Kauffman, E.R. Bricker, A., Phipps, P., Muller-Kahle, H., Fulton, T.R., Ng, S., Schumm, J.W., Braman, J.C., Knowlton, R.G., Barker, D.F., Crooks, S.M., Lincoln, S.E., Daly, M.J. and Abrahamson, J. (1987). A genetic linkage map of the human genome. *Cell* **51**, 319–337.

Elston, R.C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523–542.

Eppig, J.T. and Eicher, E.M. (1983). Application of the ovarian teratoma mapping method in the mouse. *Genetics* **103**, 797–812.

Feingold, E., Brown, A.S. and Sherman, S.L. (2000). Multipoint estimation of genetic maps for human trisomies with one parent or other partial data. *American Journal of Human Genetics* **66**, 958–968.

Felsenstein, J. (1979). A mathematically tractable family of genetic mapping functions with different amount of interference. *Genetics* **91**, 769–775.

Fincham, J.R.S., Day, P.R. and Radford, A. (1979). *Fungal Genetics*. University of California Press, Berkeley, CA.

Fisher, R.A. (1922). The systematic location of genes by means of crossover relations. *American Naturalist* **56**, 406–411.

Fisher, R.A., Lyon, M.F. and Owen, A.R.G. (1947). The sex chromosome of the house mouse. *Heredity* **1**, 335–365.

Fujitani, Y., Mori, S. and Kobayashi, I. (2002). A reaction-diffusion model for interference in meiotic crossing over. *Genetics* **161**, 365–372.

Goldgar, D.E. and Fain, P.R. (1988). Models of multilocus recombination: non-randomness in chiasma number and crossover location. *American Journal of Human Genetics* **43**, 38–45.

Goldstein, D.R., Zhao, H. and Speed, T.P. (1995). Relative efficiencies of chi-square models of recombination for exclusion mapping and gene ordering. *Genomics* **27**, 265–273.

Goss, S.J. and Harris, H. (1975). New method for mapping genes in human chromosomes. *Nature* **255**, 680–684.

Goss, S.J. and Harris, H. (1977). Gene transfer by means of cell fusion II. The mapping of 8 loci on human chromosome 1 by statistical analysis of gene assortment in somatic cell hybrids. *Journal of Cell Science* **25**, 39–57.

Green, M.C. (1981). *The Mouse in Biomedical Research*, Vol. 1, H.L. Foster, J.D. Small and J.G. Fox, eds. Academic Press, New York, pp. 105–117.

Green, P. (1988). Rapid construction of multilocus genetic linkage maps. I. Maximum likelihood estimation. Draft manuscript.

Green, E. and Green, P. (1991). Sequence-tagged sites (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods and Applications* **1**, 77–90.

Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C. and Gelbart, W.M. (1996). *An Introduction to Genetic Analysis*, 6th edition. W.H. Freeman, New York.

Guerra, R., McPeek, M.S., Speed, T.P. and Stewart, P.M. (1992). A Bayesian analysis for mapping from radiation hybrid data. *Cytogenetics and Cell Genetics* **59**, 104–106.

Haldane, J.B.S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**, 299–309.

Haldane, J.B.S. (1931). The cytological basis of genetical interference. *Cytologia* **3**, 54–65.

Hitte, C., Lorentzen, T.D., Guyon, R., Kim, L., Cadieu, E., Parker, H.G., Quignon, P., Lowe, J.K., Gelfenbeyn, B., Andre, C., Ostrander, E.A. and Galibert, F. (2003). Comparison of MultiMap and TSP/CONCORDE for constructing radiation hybrid maps. *The Journal of Heredity* **94**, 9–13.

Housworth, E.A. and Stahl, F.W. (2003). Crossover interference in humans. *American Journal of Human Genetics* **73**, 188–197.

Irwin, M., Cox, N. and Kong, A. (1994). Sequential imputation for multilocus linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 11684–11688.

Ivansson, L. and Lagergren, J. (2004). Algorithms for RH mapping: new ideas and improved analysis. *SIAM Journal on Computing* **34**, 89–108.

Johnson, D.S. (1990). *Automata, Languages, and Programming*, M.S. Paterson, ed. Springer-Verlag, Berlin, pp. 446–461.

Jones, H.B. (1997). Estimating physical distance using radiation hybrid mapping data. *Genomics* **43**, 258–266.

Jones, H.B. (2000). A review of statistical methods for genome mapping. *International Statistical Review* **68**, 5–21.

Jones, G.H. and Franklin, F.C. (2006). Meiotic crossing-over: obligation and interference. *Cell* **126**, 246–248.

Karlin, S. and Liberman, U. (1978). Classification and comparisons of multilocus recombination distributions. *Proceedings of the National Academy of Sciences of the United States of America* **75**, 6332–6336.

Karlin, S. and Liberman, U. (1979). A natural class of multilocus recombination processes and related measure of crossover interference. *Advances in Applied Probability* **11**, 479–501.

King, J.S. and Mortimer, R.K. (1990). A polymerization model of chiasma interference and corresponding computer simulation. *Genetics* **126**, 1127–1138.

Kirsch, I.R., Green, E.D., Yonescu, R., Strausberg, R., Carter, N., Bentley, D., Leversha, M.A., Dunham, I., Braden, V.V., Hilgenfeld, E., Schuler, G., Lash, A.E., Shen, G.L., Martelli, M., Kuehl, W.M., Klausner, R.D. and Ried, T. (2000). A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome. *Nature Genetics* **24**, 339–340.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonnson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S.T., Frigge, M.L., Thorgeirsson, T.E., Gulcher, J.R. and Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nature Genetics* **31**, 241–247.

Kosambi, D.D. (1944). The estimation of the map distance from recombination values. *Annals of Eugenics* **12**, 172–175.

Kruglyak, L., Daly, M.J. and Lander, E.S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *American Journal of Human Genetics* **56**, 519–527.

Kruglyak, L. and Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics* **57**, 439–454.

Kruglyak, L., Reeve-Daly, M.J., Daly, M.P. and Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* **58**, 1347–1363.

Lai, Z.W., Jing, J., Aston, C., Clarke, V., Apodaca, J., Dimalanta, E.T., Carucci, D.J., Gardner, M.J., Mishra, B., Anantharaman, T.S., Paxia, S., Hoffman, S.L., Venter, J.C., Huff, E.J. and Schwartz, D.C. (1999). A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genetics* **23**, 309–313.

Lamb, N.E., Feingold, E. and Sherman, S.L. (1997). Estimating meiotic exchange patterns from recombination data: an application to humans. *Genetics* **146**, 1011–1017.

Lander, E.S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 2363–2367.

Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E. and Newburg, L. (1987). MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**, 174–181.

Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*. Springer-Verlag, New York.

Lange, K. and Boehnke, M. (1992). Bayesian methods and optimal experimental design for gene mapping by radiation hybrids. *Annals of Human Genetics* **56**, 119–144.

Lange, K., Boehnke, M., Cox, D.R. and Lunetta, K.L. (1995). Statistical analysis for polyploid radiation hybrid mapping. *Genome Research* **5**, 136–150.

Lange, K., Zhao, H. and Speed, T.P. (1997). The Poisson-skip model of crossing-over. *Annals of Applied Probability* **7**, 299–313.

Lathrop, G.M., Lalouel, J.-M., Julier, C. and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 3443–3446.

Lathrop, G.M., Lalouel, J.-M., Julier, C. and Ott, J. (1985). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *American Journal of Human Genetics* **37**, 482–498.

Lathrop, M., Nakamura, Y., Cartwright, P., O'Connell, P., Leppert, M., Jones, C., Tateishi, H., Bragg, T., Lalouel, J.M. and White, R. (1988). A primary genetic map of markers for human chromosome 10. *Genomics* **2**, 157–164.

Leach, R.J. and O'Connell, P. (1995). Mapping of mammalian genome with radiation (Goss and Harris) hybrids. *Advances in Genetics* **33**, 63–99.

Lee, J.K., Dancik, V. and Waterman, M.S. (1998). Estimation for restriction sites observed by optical mapping using reversible-jump Markov chain Monte Carlo. *Journal of Computational Biology* **5**, 505–515.

Li, J.M., Sherman, S.L., Lamb, N. and Zhao, H.Y. (2001). Multipoint genetic mapping with trisomy data. *American Journal of Human Genetics* **69**, 1255–1265.

Liberman, U. and Karlin, S. (1984). Theoretical models of genetic map functions. *Theoretical Population Biology* **25**, 331–346.

Lin, S. (1996). *Genetic Mapping and DNA Sequencing*, T.P. Speed and M.S. Waterman, eds. Springer-Verlag, New York, pp. 15–38.

Lin, S. and Speed, T.P. (1996). Incorporating crossover interference into pedigree analysis using the chi-square model. *Human Heredity* **46**, 315–322.

Lin, S. and Wijsman, E. (1994). Monte Carlo multipoint linkage analysis. *American Journal of Human Genetics* **55**, A40.

Ling, S. (2000). Constructing genetic maps for outbred experimental crosses. Ph.D. dissertation, The University of California, Berkeley, CA.

Ludwig, W. (1934). Über numerische Beziehungen der Crossover-Werte untereinander. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre* **67**, 58–95.

Luo, Z.W., Zhang, Z., Leach, L., Zhang, R.M., Bradshaw, J.E. and Kearsey, M.J. (2006). Constructing genetic linkage maps under a tetrasomic model. *Genetics* **172**, 2635–2645.

Marinov, M., Matise, T.C., Lathrop, G.M. and Weeks, D.E. (1999). A comparison of two algorithms, MultiMap and gene mapping system, for automated construction of genetic linkage maps. *Genetic Epidemiology* **17**, S649–S654.

Mather, K. (1933). The relationship between chiasmata and crossing-over in diploid and triploid *Drosophila melanogaster*. *Journal of Genetics* **27**, 243–259.

Mather, K. (1935). Reductional and equational separation of the chromosomes in bivalents and multivalents. *Journal of Genetics* **30**, 53–78.

Matise, T.C., Perlin, M. and Chakravarti, A. (1994). Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome map. *Nature Genetics* **6**, 384–390.

McPeek, M.S. and Speed, T.P. (1995). Modeling interference in genetic recombination. *Genetics* **139**, 1031–1044.

Mendel, G. (1866). Versuche über Pflanzen-Hybriden. *Verhandlungen des Naturforschenden Vereines in Brünn* **4**, 3–47.

Mester, D.I., Ronin, Y.I., Hu, Y., Peng, J., Nevo, E. and Korol, A.B. (2003). Efficient multipoint mapping: making use of dominant repulsion-phase markers. *Theoretical and Applied Genetics* **107**, 1102–1112.

Michiels, F., Craig, A.G., Zehetner, G., Smith, G.P. and Lehrach, H. (1987). Molecular approaches to genome analysis: a strategy for the construction of ordered overlapping clone libraries. *Computer Applications in the Biosciences* **3**, 203–210.

Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Pittard, W.S. and Devine, S.E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research* **16**, 1182–1190.

Mizukami, T., Chang, W.I., Garkavstev, I., Kaplan, N., Lombardi, D., Matsumoto, T., Niwa, O., Kounosu, A., Yanagida, M., Marr, T.G. and Beach, D. (1993). A 13 kb resolution cosmid map of the 14 Mb fission yeast genome by nonrandom sequence-tagged site mapping. *Cell* **73**, 121–132.

Mohr, J. (1954). *A Study of Linkage in Man*. Munksgaard, Copenhagen.

Morgan, T.H. (1911). An attempt to analyze the constitution of the chromosomes on the basis of sex limited inheritance in Drosophila. *Journal of Experimental Zoology* **11**, 365–414.

Morton, N.E., Collins, A., Lawrence, S. and Shields, D.C. (1992). Algorithms for a location database. *Annals of Human Genetics* **56**, 223–232.

Morton, N.E., Keats, B.J., Jacobs, P.A., Hassold, T., Pettay, D., Harvey, J. and Andrews, V. (1990). A centromere map of the X chromosome from trisomies of maternal origin. *Annals of Human Genetics* **54**, 39–47.

Mott, R., Grigoriev, A., Maier, E., Hoheisel, J. and Lehrach, H. (1993). Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of *Schizosaccharomyces pombe*. *Nucleic Acids Research* **21**, 1965–1974.

Muller, H.J. (1916). The mechanism of crossing over. *American Naturalist* **50**, 193–221; 284–305; 350–366; 421–434.

Nelson, D.O. and Speed, T.P. (1994a). Statistical issues in constructing high resolution physical maps. *Statistical Science* **9**, 334–354.

Nelson, D.O. and Speed, T.P. (1994b). Predicting progress in directed mapping projects. *Genomics* **24**, 41–52.

Nelson, D.O., Speed, T.P. and Yu, B. (1997). The limits of random fingerprinting. *Genomics* **40**, 1–12.

O'Connell, J.R. and Weeks, D.E. (1995). The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recording and fuzzy inheritance. *Nature Genetics* **11**, 402–408.

Olson, M.V., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R. and Frank, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 7826–7830.

Ott, J. (1999). *Analysis of Human Genetic Linkage*, 3rd edition. Johns Hopkins University Press, Baltimore, MD.

Palazzolo, M.J., Sawyer, S.A., Martin, C.H., Smoller, D.A. and Hartl, D.L. (1991). Optimized strategies for sequence-tagged-site selection in genome mapping. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 8034–8038.

Perkins, D.D. (1949). Biochemical mutants in the smut fungus *Ustilago maydis*. *Genetics* **34**, 607–626.

Rao, D.C., Morton, N.E., Lindsten, J., Hulten, M. and Yee, S. (1977). A mapping function for man. *Human Heredity* **27**, 99–104.

Reslewic, S., Zhou, S., Place, M., Zhang, Y., Briska, A., Goldstein, S., Churas, C., Runnheim, R., Forrest, D., Lim, A., Lapidus, A., Han, C.S., Roberts, G.P. and Schwartz, D.C. (2005).

Whole-genome shotgun optical mapping of Rhodospirillum rubrum. *Applied and Environmental Microbiology* **71**, 5511–5522.

Risch, N. (1991). A note on multiple testing procedures in linkage analysis. *American Journal of Human Genetics* **48**, 1058–1064.

Risch, N. and Lange, K. (1979). An alternative model of recombination and interference. *Annals of Human Genetics* **43**, 61–70.

Risch, N. and Lange, K. (1983). Statistical analysis of multilocus recombination. *Biometrics* **39**, 949–963.

Roach, J. (1995). Random subcloning. *Genome Research* **5**, 464–473.

Robinson, W.P., Bernascoli, F., Mutirangura, A., Ledbetter, D.H., Langlois, S., Malcolm, S., Morris, M.A. and Schinzel, A.A. (1993). Nondisjunction of chromosome 5: origin and recombination. *American Journal of Human Genetics* **53**, 740–751.

Saura, A.O., Saura, A.J. and Sorsa, V. (1997). Electron micrographs maps of *Drosophila melanogaster* polytene chromosomes. `http://www.helsinki.fi/~saura/EM/`.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tomé, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B.B., Butler, A., Castle, A.B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P.J.R., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., Edwards, C., Fan, J.-B., Fang, N., Fizames, C., Garrett, C., Green, L., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, E., Hui, L., Hussain, S., Louis-Dit-Sully, C., Ma, J., MacGilvery, A., Mader, C., Maratukulam, A., Matise, T.C., McKusick, K.B., Morissette, J., Mungall, A., Muselet, D., Nusbaum, H.C., Page, D.C., Peck, A., Perkins, S., Piercy, M., Qin, F., Quackenbush, J., Ranby, S., Reif, T., Rozen, S., Sanders, C., She, X., Silva, J., Slonim, D.K., Soderlund, C., Sun, W.-L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voyticky, S., Wilmer, T., Wu, X., Adams, M.D., Auffray, C., Walter, N.A.R., Brandon, R., Dehejia, A., Goodfellow, P.N., Houlgatte, R., Hudson, J.R., Hudson, J.r., Ide, S.E., Iorio, K.R., Lee, W.Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Phillips, C., Polymeropoulos, M.H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J.C., Sikela, J.M., Beckmann, J.S., Weissenbach, J., Myers, R.M., Cox, D.R., James, M.R., Bentley, D., Deloukas, P., Lander, E.S. and Hudson, T.J. (1996). A gene map of the human genome. *Science* **274**, 540–546.

Schwartz, D.C., Li, X., Hernandez, L.I., Ramnarain, S.P., Huff, E.J. and Wang, Y.-K. (1993). Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**, 110–114.

Shahar, S. and Morton, N.E. (1986). Origin of teratomas and twins. *Human Genetics* **74**, 215–218.

Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotype analysis, location scores, and marker sharing statistics. *American Journal of Human Genetics* **58**, 1323–1337.

Soderlund, C., Longden, I. and Mott, R. (1997). FPC: a system for building contigs from restriction fingerprinted clones. *Computer Applications in the Biosciences* **13**, 523–535.

Speed, T.P., McPeek, M.S. and Evans, S.N. (1992). Robustness of the no-interference model for ordering genetic markers. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 3103–3106.

Stahl, F.W. (1979). *Genetic Recombination: Thinking About it in Phage and Fungi*. Freeman, San Francisco, CA.

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. *The Plant Journal* **3**, 739–744.

Sturt, E. (1976). A mapping function for human chromosomes. *Annals of Human Genetics* **40**, 147–163.

Sturtevant, A.H. (1913). The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of Experimental Zoology* **14**, 43–59.

Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T. and Coulson, A. (1988). Software for genome mapping by fingerprinting techniques. *Computer Applications in the Biosciences* **4**, 125–132.

Tan, Y.D. and Fu, Y.X. (2006). *A New Strategy for Estimates of Recombination Fractions between Dominant markers from F2 Population*. Genetics Press.

Terwilliger, J.D. and Ott, J. (1994). *Handbook of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, MD.

Thompson, E.A. (1984). Information gain in joint linkage analysis. *IMA Journal of Mathematics Applied in Medicine and Biology* **1**, 31–49.

Thompson, E.A. (1994). Monte Carlo likelihood in genetic analysis. In *Probability, Statistics, Optimization: A Tribute to Peter Whittle*, F.P. Kelley, ed. John Wiley & Sons, New York, pp. 281–293.

Tibshirani, R., Lazzeroni, L., Hastie, T., Olshen, A. and Cox, D.R. (1999). The global pairwise approach to radiation hybrid mapping. Technical Report 201, Division of Biostatistics, Stanford University.

Trask, B.J. (1998). *Mapping Genomes*, *Genome Analysis: A Laboratory Manual Series*, Vol. 4, B. Birren, E.D. Green, P. Hieter, S. Klapholz, R.M. Myers, H. Riethman and J. Roskams, eds. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp. 303–413.

Valouev, A., Schwartz, D.C., Zhou, S. and Waterman, M.S. (2006). An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 15770–15775.

Vogel, F. and Motulsky, A.G. (1997). *Human Genetics: Problems and Approaches*, 3rd edition. Springer-Verlag, Berlin.

Walter, M.A., Spillett, D.J., Thomas, P. and Goodfellow, P.N. (1994). A method for constructing radiation hybrid maps of whole genomes. *Nature Genetics* **7**, 22–28.

Waterman, M.S. (1995). *Introduction to Computational Biology: Maps Sequences and Genomes*. Chapman & Hall, London.

Weeks, D.E. (1991). *Advanced Techniques in Chromosome Research*, K.W. Adolph, ed. Marcel Dekker, New York, pp. 297–330.

Weinstein, A. (1936). The theory of multiple-strand crossing over. *Genetics* **21**, 155–199.

Whitehouse, H.L.K. (1957). Mapping chromosome centromeres from tetratype frequencies. *Journal of Genetics* **55**, 348–360.

Wu, R. and Ma, C.X. (2005). A general framework for statistical linkage analysis in multivalent tetraploids. *Genetics* **170**, 899–907.

Wu, S.S., Wu, R.L., Ma, C.X., Zeng, Z.-B., Yang, M.C.K. and Casella, G. (2001). A multivalent pairing model of linkage analysis in autotetraploids. *Genetics* **159**, 1339–1350.

Yeh, R.-F. (1999). Statistical issues in genomic mapping and sequencing. Ph.D. dissertation, The University of California, Berkeley, CA.

York, T.L., Durrett, R.T., Tanksley, S. and Nielsen, R. (2005). Bayesian and maximum likelihood estimation of genetic maps. *Genetical Research* **85**, 159–168.

Zhao, H., Li, J. and Robinson, W.P. (2000). Multipoint genetic mapping with uniparental disomy data. *American Journal of Human Genetics* **67**, 851–861.

Zhao, H. and Speed, T.P. (1996). On genetic map functions. *Genetics* **142**, 1369–1377.

Zhao, H. and Speed, T.P. (1998a). Statistical analysis of half-tetrads. *Genetics* **150**, 473–485.

Zhao, H. and Speed, T.P. (1998b). Statistical analysis of ordered tetrads. *Genetics* **150**, 459–472.

Zhao, H., McPeek, M.S. and Speed, T.P. (1995a). Statistical analysis of chromatid interference. *Genetics* **139**, 1057–1065.

Zhao, H., Speed, T.P. and McPeek, M.S. (1995b). Statistical analysis of crossover interference using the chi-square model. *Genetics* **139**, 1045–1056.

Statistical Significance in Biological Sequence Comparison

W.R. Pearson and T.C. Wood

Department of Biochemistry, University of Virginia, Charlottesville, VA, USA

The chapter reviews the role of statistical significance estimates in biological sequence comparison, focusing on local similarity searches. A brief history of the concept of 'homology' is presented, and the relationship between 'statistical significance' and 'biological significance', is discussed, addressing the question: 'What biological inferences can be drawn from statistically significant sequence similarity?' Algorithms and scoring matrices used to quantify local sequence similarity are summarized, and the statistical basis for the use of the extreme-value distribution to describe local similarity scores, both without and with gaps, is presented in outline. Particular attention is given to λ and K, the two parameters of the extreme-value distribution used by Karlin and Altschul, and to the use of bit-scores, and other scale-independent measures of similarity. Strategies that have been used to estimate the significance of local sequence similarity scores are compared using several distant evolutionary relationships. The reliability of statistical estimates for local sequence similarity scores is discussed in detail; it is shown that, with the exception of highly biased protein sequences and sequences with low-complexity regions, real, unrelated protein sequences behave very similarly to sequences generated randomly, so that the assumptions underlying the statistical models are widely applicable and statistical significance estimates are generally reliable.

2.1 INTRODUCTION

The availability of comprehensive sequence databases, rapid sequence comparison methods, and accurate statistical estimates for sequence similarity has fundamentally changed the practice of biochemistry and molecular biology. With the possible exceptions of *Escherichia coli* and *Saccharomyces*, the vast majority of the genes in newly sequenced genomes are characterized by sequence similarity searching. BLAST, FASTA, and Smith–Waterman similarity searches provide the most informative and reliable method for inferring the biological function of an anonymous gene (or the protein that it encodes). Typically, 60–80 % of eubacterial (and yeast) genes share statistically significant

sequence similarity with sequences from another organism. Significant sequence similarity can be used to infer common ancestors and similar three-dimensional structures, and is routinely used to assign functions in metabolic pathways. Even for the first archebacterial genome sequenced (*Methanococcus jannaschii*; Bult *et al.*, 1996), similarity-based functional gene assignments could be made for about 50 % of the genes (Andrade *et al.*, 1997) and subsequent sequence analyses (Koonin, 1997) suggested functions for another 20 % of the genes.

Unfortunately, some investigators are uncomfortable inferring the relationship between two sequences from a probability or expectation value; they prefer to think in terms of percent identity (sometimes misstated as percent homology). When current versions of the BLAST and FASTA similarity searching programs are used, this concern is rarely justified. It is very unusual for a statistically significant sequence similarity not to reflect common ancestry, and thus common structure, for the two sequences.

This chapter will provide an overview of the role of statistical significance estimates in biological sequence comparison, focusing on local similarity searches. We will begin by discussing the relationship between 'statistical significance' and 'biological significance', addressing the question: 'What biological inferences can be drawn from *statistically significant* sequence similarity?' Next, we will survey strategies that have been used to estimate the significance of local sequence similarity scores. Finally, we will discuss the reliability of statistical estimates for local sequence similarity scores.

2.2   STATISTICAL SIGNIFICANCE AND BIOLOGICAL SIGNIFICANCE

BLAST, FASTA, and other sequence similarity searching programs are designed to identify distantly related – homologous – sequences based on sequence similarity. When we say that two sequences are homologous, we are stating our belief that the two sequences diverged from a common ancestor in the past. A remarkable result of microbial genome sequencing projects has been that a large fraction of proteins, typically 50–80 % of each newly sequenced genome, share statistically significant similarity with proteins in other organisms that diverged hundreds to thousands of millions of years ago. Thus, it is common to observe very strong sequence similarity between prokaryotic and eukaryotic proteins that diverged more than 2 billion years ago.

The inference of homology, at least as the term is commonly used in sequence analysis, implies that the homologous proteins have similar structures. Indeed, structural similarity is the gold standard for homology. Almost without exception, if two sequences share statistically significant sequence similarity, they will share significant structural similarity. However, the converse is not true; there are many examples of similar structures that do not share significant sequence similarity (though perhaps not as many examples as are presented in the literature).

The concept of homology was given wide exposure and common usage by Richard Owen, the first curator of the British Museum. Owen defined a homolog as simply 'the same organ in different animals' (Owen, 1843). He further divided homology into two types: special and serial. Special homology is essentially the definition of homology we use today, 'the same organ in different animals'. In contrast, serial homology specifically refers to similarity between structures in different body segments, such as the legs of a millipede. Darwin's theory of evolution by natural selection conferred upon homology

the specialized meaning of structures or organs that share a common ancestor. Although he regarded Darwin's theory as little more than speculation, Owen did admit that special homology was the result of common ancestry (Owen, 1866).

Implicit in all definitions of anatomical homology is some kind of recognizable similarity, e.g. similarity of form or ontogeny. The classic example of anatomical homology is the similarity of forelimbs in the higher vertebrates. Whether adapted for grasping, running, swimming, or flying, the same basic skeletal pattern can be readily observed. Although forelimbs are detectably similar in their adult forms, some homologous structures are only similar in embryonic stages.

Closely related concepts describe biological similarity that is not the result of a common evolutionary ancestry. Originally, Owen defined *analogy* as similarity of function, without regard to structure (Owen, 1843), and that definition was repeated by Neurath *et al.* (1967). The current definition of analogy adds the qualification that the similarity is not due to homology, that is, the similarity is primarily due to chance and is typically superficial (Kent, 1992). The horns of cows and rhinoceroses, and the limbs of insects and vertebrates are analogous.

Convergence is often invoked as a possible explanation of biological similarity, particularly when discussing protein sequence motifs. Properly understood, convergence refers to the process of evolution: two distantly related species developing a similar trait that was not present in their common ancestor. If convergence is observed over numerous stages of the evolution of two separate groups, it is termed *parallel evolution*. Examples of similarity from convergence include the body plans of sharks, dolphins, and ichthyosaurs. As each organism adapted to existence in the water, they developed a similar body plan by convergent evolution.

2.2.1 'Molecular' Homology

In the 1950s and 1960s, as protein sequences and three-dimensional structures were determined, researchers began to recognize surprising similarity between protein molecules. Though rigorous methods of understanding and detecting protein similarity were years away, the term 'homology' was quickly applied to similarities observed among the trypsin-like proteases and the globins, implying that members of protein families shared their remarkable similarity because of divergence from a common ancestor.

Just as the term *homology* has been misused and misunderstood among anatomists, among biochemists the usage of homology as a synonym for similarity has unfortunately remained common. One often reads of 'low homology' or even a quantified 'percent homology' in papers reporting new sequences. Since homology is qualitative (having a common ancestor), it cannot be quantified as similarity can. Any two sequences have some measurable similarity, but a statement of homology implies that the similarity has some special meaning, specifically common ancestry.

2.2.2 Examples of Similarity in Proteins

Modern biochemical studies have revealed numerous examples of homology and analogy. In general, it is widely accepted that the three-dimensional structures of homologous proteins are more highly conserved than their sequences. Practically speaking, this means that homologous proteins with very low sequence similarity can and do have very similar structures. It is also believed that *orthologous* proteins – sequences that differ as a

result of speciation events, in contrast to *paralogous* sequences, which result from gene duplication – share the same cellular function, and that new biological functions have arisen through the generation of paralogs by gene duplication.

Although mentioned frequently, convergent evolution as an explanation of protein similarity is not well defined. To avoid confusion, Doolittle (1994) proposed three categories of convergent evolution in proteins: mechanistic convergence, structural convergence, and functional convergence. The actual mechanisms that produce convergence are subjects of ongoing research (Sanderson and Hufford, 1996). Here we will qualify convergence as similarity that arises by some kind of common selection. We will reserve the term analogy for similarity by chance, when no common selection is apparent.

Mechanistic convergence refers to similar active sites and residues in otherwise unrelated proteins. The classic example given is the mechanistic similarity between the trypsins and subtilisins. Although these proteins are entirely structurally dissimilar and thus almost certainly unrelated, they have geometrically and chemically equivalent catalytic triads. In mechanistic convergence, one can conclude that the need to accomplish a particular biochemical reaction is the selection producing the convergence. From principles of chemistry, it is reasonable to conclude that there are a limited number of enzymatic mechanisms available to accomplish particular reactions; thus, the occurrence of proteins with similar catalytic sites but distinct evolutionary histories is not surprising.

Structural convergence refers to structural similarity that is not the result of common ancestry. The adaptive selection applied to the structure is not the protein's cellular or biochemical function as in mechanistic convergence, but rather the thermodynamic stability of the particular fold. Doolittle mentions the ubiquitous TIM barrels (named for their well-known example, triosephosphate isomerase) as examples of structural convergence. The structural similarity in convergent TIM barrels is typically both topological and geometric; that is, both the ordering of the secondary structural elements in the peptide chain and the atomic positions in space are similar. A second type of structural convergence is restricted to geometry only, proteins that have a similar three-dimensional arrangement of secondary structural elements but a different ordering of those elements in the peptide. Examples of geometric structural convergence include the pleckstrin homology domain (PHd) and verotoxin (Orengo *et al.*, 1995), and the N-terminal β-barrels of *E. coli* transcription termination factor rho and the F1 ATPase subunits. In each case, the arrangement of atoms in space is very similar, but the tracing of the peptide chain through those atoms is different. In the case of the rho/F1 similarity, the rho barrel is actually traced in reverse order with respect to the F1 barrel (Allison *et al.*, 1998).

A third category of protein convergence defined by Doolittle is *functional convergence*. Multiple examples of independent origins of the same or similar enzymatic activities are known. For example, Rawlings and Barrett (1993) used a sequence analysis and manual structural evaluation to assign peptidases to 64 different 'clans', each with an independent evolutionary origin. Although Doolittle calls this similarity 'functional convergence', no adaptive advantage or selection pressure is known or given for why so many different kinds of peptidases would exist. Analogy, or similarity by chance, seems a better description for this type of gross functional similarity.

2.2.3 Inferences from Protein Homology

The inference of protein homology from similarity is routinely used to assign biochemical and cellular functions of newly sequenced proteins when a protein of known function is

available for comparison. This is of critical importance for initial analysis of genomic sequences. For example, the vast majority of function assignments of the open reading frames (ORFs) of the *M. jannaschii* genome were made based on protein homologs detected by sequence similarity (Bult *et al.*, 1996). By properly using computational tools for sequence comparison, inferring homology from sequence similarity is the single most powerful tool we have today for understanding the function and origin of a protein without actually performing biochemical experiments.

Since protein structure is conserved in divergent evolution, identifying homologous proteins of known structure can give a general insight into the fold of the protein of interest as well as a detailed molecular model if the sequence similarity is high enough. Although a remarkable amount of information about the function of a protein or protein complex can be gained from traditional biochemical and genetic methods, nothing brings these data into such clear focus as an atomic-resolution protein structure. Unfortunately, solving a protein structure by nuclear magnetic resonance or crystallographic methods can be very time-consuming, much more so than determining the sequence. Deriving structural and mechanistic information from closely related proteins of known structure will remain an attractive means of understanding most proteins.

2.3   ESTIMATING STATISTICAL SIGNIFICANCE FOR LOCAL SIMILARITY SEARCHES

The inference of homology from statistically significant sequence similarity is an application of Occam's razor: given two competing hypotheses – first, that a particular sequence ordering arose twice independently by chance; and second, that the similarity reflects divergence from a common ancestor – it seems simpler to conclude that a particular structure arose only once in evolutionary history. Thus, in biological sequence analysis, we infer homology from statistically significant sequence similarity. The inference depends on two parts: our ability to measure sequence similarity; and accurate estimates for the statistical significance of the similarity measure to reduce the likelihood that the similarity could be expected by chance.

2.3.1   Measuring Sequence Similarity

2.3.1.1   *Sequence comparison algorithms*

Effective algorithms for comparing protein and DNA sequences have been available for more than 30 years, since the publication of a global sequence comparison algorithm by Needleman and Wunsch (1970). *Global* sequence comparison algorithms seek to align every residue in one sequence with every residue in a second, in contrast to the more commonly used *local* sequence alignment algorithms, which seek only the strongest region of similarity between two sequences. Global alignment algorithms are used for aligning families of sequences with similar lengths in preparation for phylogenetic analysis; global alignment scores can be transformed to the distance measures used for building evolutionary trees. Global similarity scores are rarely used to infer homology, however, because the distribution of global similarity scores is not well understood and thus it is difficult to assign a statistical significance to a global similarity score. Moreover, many proteins are made up of domains that are homologous only over a portion of the protein sequence.

The most widely used programs for searching protein and DNA sequence databases, including BLAST, FASTA, and implementations of the Smith–Waterman algorithm, measure *local* sequence similarity. First described by Smith and Waterman (1981), local sequence alignment algorithms seek to align the most similar regions of two sequences. Local alignment algorithms have two dramatic advantages over global alignment methods when searching sequence databases for statistically significant matches: first, the statistics of local similarity scores are well understood; and second, local alignments allow one to identify conserved domains in proteins, which may not extend over the entire sequence. BLAST and FASTA use heuristic methods that attempt to approximate the optimal local similarity shared by two sequences. BLAST is particularly efficient in identifying distantly related sequences because it spends very little time calculating similarity scores for sequences that are unlikely to share significant similarity. FASTA is considerably slower than BLAST, because it calculates an approximate similarity score for every sequence in the database. FASTA uses these approximate scores to estimate the parameters of the extreme-value distribution, λ and K, which describes the expected distribution of local similarity scores between random sequences.

2.3.1.2 Similarity scores for sequence comparison

All algorithms that calculate sequence similarity, global or local, optimal or heuristic, seek to maximize a measure of similarity. The earliest (and unfortunately most commonly cited even today) similarity measure was based on percent identical residues (Watson and Kendrew, 1961). Initially, the low percent identity of the myoglobin and hemoglobin sequences (typically less than 30 %) was a surprising feature of two proteins with such similar structures. Later, researchers began to develop means to describe the similarity of different amino acid residues; the first such efforts were based on the redundancy of the genetic code, e.g. the minimal number of nucleotide substitutions required to convert one amino acid in the protein sequence to another (Fitch, 1966). In the 1970s, Margaret Dayhoff developed the notion of an *accepted point mutation* or PAM (Dayhoff *et al.*, 1978). The PAM concept centered around the natural selection against certain amino acid substitutions (thus an *accepted* point mutation) rather than simply the probability of mutations in the underlying DNA sequence. More recently the BLOSUM series of matrices, which tabulate the frequency with which different substitutions occur in conserved blocks of protein sequences, has been shown to be very effective in identifying distant relationships (Henikoff and Henikoff, 1992).

Dayhoff's PAM matrices are based on a well-defined evolutionary model for protein sequences (Dayhoff *et al.*, 1978). Given an estimate for the probability that any amino acid will change to each of the other amino acids, or remain the same, after 1 % change (1 accepted mutation per 100 residues), one can estimate the probability that any amino acid will change into each of the others after 2 %, 10 %,..., 40 %,..., 200 % change by multiplying the transition probability matrix by itself 2, 10,..., 40, 200 times. After incorporating the probability p_i of seeing a particular residue, the resulting matrix gives the probability $q_{i,j}$ of residue i aligning with residue j after a specified amount of evolutionary change. These probabilities are converted to log-odds scores by normalizing the alignment probabilities by the probability of seeing two residues align by chance, $p_i p_j$, yielding a scoring matrix $s_{i,j} = \log(q_{i,j}/p_i p_j)$.

	A	R	N	D	E	I	L		A	R	N	D	E	I	L
A	8							A	2						
R	−9	12						R	−2	6					
N	−4	−7	11					N	0	0	2				
D	−4	−13	3	11				D	0	−1	2	4			
E	−3	−11	−2	4	11			E	0	−1	1	3	4		
I	−6	−7	−7	−10	−7	12		I	−1	−2	−2	−2	−2	5	
L	−8	−11	−9	−16	−12	−1	10	L	−2	−3	−3	−4	−3	2	6
(a)								(b)							

Figure 2.1 Similarity scoring matrices. (a) PAM40 and (b) PAM250 similarity scoring matrices for six amino acid residues. The substitution matrices are symmetric. Diagonal elements are the scores given to amino acid identities; off-diagonal elements are the scores used for amino acid substitutions. Both the PAM40 and PAM250 matrices are scaled to 0.33 bits per unit raw score. Thus, if $\log_2(q_{ij})/(p_i p_j) = 2$, the entry in the matrix would be 6.

Figure 2.1 shows parts of two PAM scoring matrices, PAM40, which incorporates transition probabilities between residues in sequences that have had 40 accepted mutations per 100 residues, and PAM250, which is 'targeted' for sequences that have had 250 accepted mutations per 100 residues.[1] The PAM40 and PAM250 matrices differ dramatically in the relative scores of identities and substitutions; replacements that are considered unlikely at PAM40, e.g. R to N with $s_{N,R} = -7$, are considered neutral, $s_{N,R} = 0$, at PAM250. Likewise, replacements that are expected less frequently than chance ($s_{I,L} = -1$) after 40 % change are more likely than chance substitutions ($s_{I,L} = 2$) after 250 % change. Although the Dayhoff PAM matrices are based on the relatively small number of transitions available in 1978, a modern equivalent is available (Jones *et al.*, 1992), which performs well when appropriate gap penalties are used (Pearson, 1995).

An alternative strategy for calculating scoring matrices was developed by Henikoff and Henikoff (1992). Rather than extrapolate transition probabilities for a very large amount of change from the frequencies obtained after a very small amount of change, they sought to measure transition probabilities directly, by building a very large set of conserved blocks of aligned amino acid residues and then tabulating the amino acid substitution frequencies by examining columns in the aligned blocks with different degrees of identity (Henikoff and Henikoff, 1992). These calculations were used to generate the BLOSUM series of scoring matrices; BLOSUM50, the default scoring matrix used by the FASTA family of sequence comparison programs, reports substitution frequencies for residues in conserved blocks of sequences that show 50 % identity or less; BLOSUM62, which is the default for the BLAST programs, is derived from blocks that are up to 62 % identical, and BLOSUM80 reflects a very high degree of sequence conservation by including sequences up to 80 % identical. The BLOSUM matrices are now more widely used than either the original or modern versions of the PAM matrices because they appear to perform better with many alignment algorithms (Henikoff and Henikoff, 1992) and over a broad range of gap penalties (Pearson, 1995).

---

[1] Because different amino acids have different mutation probabilities, and an amino acid can mutate to a different residue, which can then mutate again back to the original amino acid, sequences that have changed by 250 % are expected to remain about 20 % identical, on average (Dayhoff *et al.*, 1978).

Both the PAM and BLOSUM series of matrices provide similarity scores that are targeted for different levels of sequence identity (Altschul, 1991; Henikoff and Henikoff, 1992); PAM matrices range from low values PAM10–PAM40 for high identity to PAM200–PAM250 for low (25–20 %) identity. BLOSUM matrices range from high (BLOSUM80) values for high identity to low values BLOSUM50–45 for distant relationships. However, despite this apparent similarity, the meaning of a 'shallow' PAM20 matrix is quite different from that of very conservative BLOSUM80 substitution values. The PAM20 provides scores for sequences that have changed by only 20 %; the amount expected for a comparison of mouse and human proteins, for example. In contrast, BLOSUM80 is targeted towards the most highly conserved regions in proteins, blocks that remain up to 80 % identical within two sequences that may share less than 30 % identity overall. Thus, low PAM matrices, but not BLOSUM80, are appropriate for short divergence times.

Although the PAM and BLOSUM matrices were built to target specific models of evolution and conservation, Altschul has shown (Altschul, 1991; States *et al.*, 1991) that any scoring matrix can be written in the form $s_{i,j} = \log(q_{i,j}/p_i p_j)$ reflecting an implied target substitution frequency, which can be calculated using the formula $\lambda s_{i,j} = \log(q_{i,j}/p_i p_j)$. In particular, the BLASTN2.0 program for DNA substitutions, which uses $+1$ for a match and -3 for a mismatch, has $\lambda = 1.374$. Rearranging the equation above, the target frequency for any nucleotide match, assuming $p_{A,C,G,T} = 0.25$, is $q_{A,A} = q_{C,C} = q_{G,G} = q_{T,T} = p_A p_A e^{\lambda(+1)} = 0.2469$ and the overall target identity is $\sum_{b=A,C,G,T} p_{b,b} = 0.988$. Thus, BLASTN2.0 is optimally efficient at identifying homologous sequences that are 98.8 % identical, and considerably less efficient at identifying sequences that share 80 % identity or less. In contrast, the DNA match/mismatch values for BLAST1.4 and FASTA are $+5/-4$, which, with $\lambda = 0.1915$, are targeted for alignments averaging 65 % identity.

2.3.2 Statistical Significance of Local Similarity Scores

A major breakthrough in biological sequence comparison occurred when Karlin and Altschul (1990) published their statistical analysis of local sequence similarity scores without gaps, and the BLAST program incorporated those statistics (Altschul *et al.*, 1990). Although a method for evaluating the statistical significance of sequence similarity scores, the RDF program, was included with the FASTP program (Lipman and Pearson, 1985), along with the advice that sequence similarity scores that were 6 standard deviations above the mean of the distribution of shuffled sequence scores ($z > 6$) were 'probably' significant, there was no statistical basis for this observation. Work by Arratia *et al.* (1986) and by Karlin and Altschul (1990) demonstrated that local similarity scores, at least for alignments without gaps, were accurately described by the extreme-value distribution, which can be written as

$$P(S \geq x) = 1 - \exp(-Kmne^{-\lambda x}) \tag{2.1}$$

where λ and K can be calculated from the similarity scoring matrix $s_{i,j}$ and the amino acid compositions of the aligned sequences p_i, p_j, and m and n are the lengths of the two sequences.

Accurate similarity statistics allow us to discriminate reliably between statistically significant similarities, which reflect *homology*, and similarities that could have arisen by chance, *analogous* sequences. The availability of Karlin–Altschul statistics in the

BLAST program (Altschul *et al.*, 1990) separated 'first-generation' score-only programs from 'second-generation' methods. Without accurate statistics, it is impossible to do large-scale sequence interpretation.

2.3.2.1 Statistics of alignments without gaps

The first statistical models for local and global alignment scores applied to runs of similar amino acid or nucleotide residues, which are equivalent to alignments without gaps. Arratia *et al.* (1986; 1990); Karlin and Altschul (1990); and Karlin *et al.* (1991) demonstrated that local similarity scores are expected to follow the extreme-value distribution. Waterman (1995) presents an intuitive argument when, referring to Erdös and Rényi, he points out that the expected number of runs of heads of length l in n coin tosses is $E(l) \cong np^l$, where p is the probability of heads (Figure 2.2). This relationship follows from the logic that the expected number of heads is the product of the probability of heads at each toss, times the number of tosses. If the longest run R_l is expected once, $1 = np^{R_l}$ and thus $R_l = \log_{1/p}(n)$. The longest run of heads coin-toss example is equivalent to finding the highest scoring region (e.g. a hydrophobic patch) in a single protein sequence using a scoring matrix that assigns a positive value to some of the residues, and $-\infty$ to all of the others. The probability of a positive score, which corresponds to the probability of heads in the coin-toss example, is $\sum p_i$ for each of the residues p_i that obtain a positive score s_i.

The simple example of head runs, or scores with $-\infty$ mismatch penalties, shows that *local* similarity scores for single sequences are expected to increase with the logarithm of the sequence length n. In sequence comparison, we consider possible alignments of

(a)

(b)
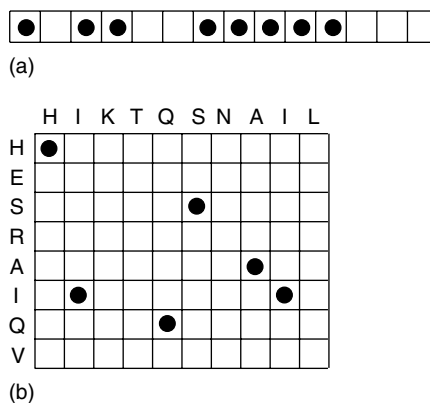
Figure 2.2 Sequence comparison as coin tosses. (a) Results from tossing a coin 14 times; black circles indicate heads. The probability of 5 heads in a row is $p(5) = 1/2^5 = 1/32$, but since there were 10 places that one could have obtained 5 heads in a row, the expected number of times that 5 heads occurs by chance is $E(5H) = 10 \times 1/32 = 0.31$. (b) Comparison of two protein sequences, with identities indicated as black circles. Assuming the residues were drawn from a population of 20, each with the same probability, the probability of an identical match is $p = 0.05$. In this example, there are $m = 10 \times n = 8$ boxes, so $E() = mnp = 80 \times 0.05 = 4$ matches are expected by chance. The probability of two successive matches is $p^2 = 1/20^2$ so a run of two matches is expected about $nmp^2 = 8 \times 10 \times 1/20^2 = 0.2$ times by chance.

two sequences, $a_{1..m}$ and $b_{1..n}$, but the probability calculation is quite similar. Rather than calculate the probability of obtaining the k heads, where $p_k = pp_{k-1}$, we consider the case of matching at m positions, or equivalently giving a *head* score if $a_i = b_j$. If the sequences are placed as in Figure 2.2(b), the head-run problem corresponds to the longest run of matches along any of the diagonals. If the letters (residues) in the two sequences have equal probabilities p, then the probability of a match of residue a_i with b_j is p and the probability of a match of length l from a_i, b_j to a_{i+l-1}, b_{j+l-1} is again p^l. In this case, however, there are $m - l + 1 \times n - l + 1$ places where that match could start, so $E(l) \cong mnp^l$. Thus, the expected length of the longest match between two random sequences of length m and n when the match score is positive and the mismatch score is $-\infty$ is $M_{mn} = \log_{1/p}(mn)$ or $2\log_{1/p}(n)$ when $m = n$ (Waterman, 1995). The shift from $\log_{1/p}(n)$ for one sequence to $\log_{1/p}(n^2)$ for two sequences of length n, reflects the larger number of positions where a run of length M_l with probability $P(M_l) = p^{M_l}$ could start. As in the single-sequence case, we can transform the problem from the probability of the longest match run to the probability of score $S_l \geq x$ by considering the probability $P(S \geq x)$ when a pair of residues $a_i b_j$ is matched with positive score $s_{i,j}$ and all negative scores are $-\infty$. For local pairwise alignment scores with a mismatch score of $-\infty$ and no gaps, the expected number of runs of score $S \geq x$ has the general form: $E(S \geq x) \propto mnp^x$, or equivalently $E(S \geq x) \propto mne^{x \ln p}$ or $mne^{-\lambda x}$ where $\lambda = -\ln p$.

Karlin and Altschul provided a natural extension of the problem of head runs, or match runs, or positive similarity scores bounded by $-\infty$ mismatch scores, to the more general case of local sequence patches or local similarity scores for nonintersecting alignments without gaps. To ensure the scores are local, the requirement that $E(s_{i,j}) = \sum_{i,j} p_i p_j s_{i,j} < 0$ must first be met. If so, the expected number of alignments with score S is

$$E(S \geq x) = Kmne^{-\lambda x}. \tag{2.2}$$

Karlin and Altschul (1990) derived analytical expressions for K and λ. $K < 1$ is a proportionality constant that corrects the mn 'space factor' for the fact that there are not really mn independent places that could have produced score $S \geq x$. Compared to λ, K has a modest effect on the statistical significance of a similarity score.

The λ parameter provides the scale factor by which a score must be multiplied to determine its probability. For ungapped alignments, λ is the unique positive solution to the equation

$$\sum_{i,j} p_i p_j e^{\lambda s_{i,j}} = 1. \tag{2.3}$$

λ thus depends both on the scoring matrix ($e^{s_{i,j}}$) and the residue compositions of the two sequences ($p_i p_j$). In some sense, λ can be interpreted as a factor that converts pairwise match scores to probabilities, so that $e^{-\lambda x}$ is similar to p^l in the coin-tossing example. Thus, just as in the coin-tossing case, the expectation of a run of heads (or an alignment run that produces score S) is the product of a space-factor term, Kmn, and a probability term $e^{-\lambda S}$.

The need for a scale factor to convert raw similarity scores into probabilities follows intuitively from the observation that multiplying or dividing every value in a similarity scoring matrix by a constant has no effect on the local alignments that would be produced by that matrix, or on the relative distribution of similarity scores in a library search – the highest scoring sequence will still be the highest, second highest second, etc. Thus it is

impossible, without some previous knowledge of the scoring matrix used and the particular scaling of the scoring matrix, to evaluate the statistical significance of a raw similarity score. However, by using a scaled similarity score λS_{raw}, one can readily compare alignments done with any scoring matrix. BLAST2.0 (Altschul *et al.*, 1997) and the current FASTA3 comparison program (Pearson, 1999) report the scaled score in terms of a *bit*-score that incorporates the space correction factor K: $S_{\text{bit}} = (\lambda S_{\text{raw}} - \ln K)/\ln 2$. Thus, substituting in equation (2.2),

$$E(S_{\text{bit}}) = mn2^{-S_{\text{bit}}} = \frac{mn}{2^{S_{\text{bit}}}} \qquad (2.4)$$

Equations (2.2) and (2.4) describe the number of times a score greater than or equal to S_{bit} would be expected by chance when two random sequences are compared.[2] This expectation can range from a very small value for very high scores (e.g. $S_{\text{bit}} = 1000$), to a value that approaches mn when $S = 0$. In a comparison of two average length protein sequences $n = m = 400$, $S_{\text{bit}} = 10$ would be expected $E(S_{\text{bit}} \geq 10) = mn2^{-S_{\text{bit}}} = 156$ times. To estimate the probability $P(S_{\text{bit}} \geq 10)$, which must range from 0 to 1, of obtaining at least one score $S \geq x$, we use the Poisson approximation.

The Poisson formula describes the probability of an event occuring a specified number of times, based on the average number of times μ it is expected to occur.[3] The Poisson probability of seeing n events when an event is expected μ times on average is $P(n) = e^{-\mu}\mu^n/n!$. In general, we are interested in the probability of seeing the event at least n times, and in the case of sequence comparisons, we ask for the probability of seeing a high score one or more times ($n \geq 1$). In this case, one can calculate the probability of not seeing the event zero times: $P(n \geq 1) = 1 - P(0)$, so $P(S \geq x) = 1 - P(n = 0) = 1 - e^{-\mu}\mu^0/0!$. Since $\mu = E(S \geq x) = Kmne^{-\lambda x}$ and $\mu^0 = 0! = 1$, the probability of seeing a raw similarity score $S \geq x$ is

$$P(S \geq x) = 1 - \exp(-\mu) = 1 - \exp(-Kmne^{-\lambda x}),$$

as seen earlier in equation (2.1).

Equation (2.1) describes the probability of obtaining a similarity score $S \geq x$ in a single pairwise comparison of a query sequence of length m against a library sequence of length n. This equation has the same form as the extreme-value distribution or Gumbel distribution, which is often presented as

$$P(S \geq x) = 1 - \exp(-e^{-(x-a)/b}), \qquad (2.5)$$

with a providing the 'location' of the mode, and b determining the scale, or width, of the distribution. For local similarity scores without gaps, $b = 1/\lambda$ and $a = \ln Kmn/\lambda$. The mean of the extreme-value distribution is $a - b\Gamma'(1)$, where $\Gamma'(1) = -0.577\,216$ is the first derivative of the gamma function $\Gamma(n = 1)$ with respect to n. The variance is $b^2\pi^2/6$ (Evans *et al.*, 1993). Thus, one can express the probability that an alignment

---

[2] More accurately, the statistical model assumes that the two sequences are made up of residues that are independent and identically distributed (i.i.d.). The identical distribution assumption can be violated by low-complexity regions in proteins and DNA or by strongly biased amino acid or nucleotide composition.

[3] λ is generally used to denote the characteristic parameter of a Poisson distribution, but we use μ here to avoid confusion with the λ scaling factor and to reinforce the fact that μ is the mean of the Poisson distribution.

obtains a score z standard deviations above the mean of the distribution of unrelated (or random) sequence scores as

$$P(Z \geq z) = 1 - \exp(-e^{-((\pi/\sqrt{6})z - \Gamma'(1))}). \tag{2.6}$$

These equations describe the probability that two sequences would obtain a similarity score by chance in a single comparison. However, in a sequence database search, the highest-scoring alignments are identified after a query sequence has been compared with each of the tens or hundreds of thousands of sequences in the database. Thus, in the context of a database search after $D = 100\,000 - 500\,000$ or more alignments have been scored, the number of times a score is expected to occur, the *expectation value*, is considerably higher:

$$E(S \geq x) = DP(S \geq x). \tag{2.7}$$

Thus, a similarity score 6 standard deviations above the mean ($z = 6$) has a probability $P(Z \geq 6) < 2.6 \times 10^{-4}$ in a single pairwise comparison. However, in 1985, with 3000 entries in the protein sequence database, $E(Z \geq 6) = 3000 \times 2.6 \times 10^{-4} = 0.77$. Therefore, a score 6 standard deviations above the mean should be seen by chance very frequently, and the advice provided with the description of the FASTP program overestimated statistical significance. Today, with protein sequence databases ranging in size from $100\,000$ to $500\,000$ sequences, a $z = 6$ score would be expected 25 times by chance when searching a $100\,000$ sequence database (equations 2.6 and 2.7), and $z \geq 12.1$ is required to achieve statistical significance of $E(100\,000) \leq 0.01$. (For $E(500\,000) \leq 0.01$, $z \geq 13.4$.)

2.3.2.2　Alignments with Gaps

The statistical analysis of local similarity scores summarized above was derived for alignments without gaps. Although searches that report only the best local alignment without gaps can perform very well, they do not perform as well as a Smith–Waterman search with modern scoring matrices and appropriate gap penalties (Pearson, 1995). Thus, there is considerable interest in the statistical parameters that describe the distribution of local similarity scores with gaps.

The first implementation of Smith and Waterman's (1981) algorithm that provided statistical estimates for similarity scores was developed by Collins *et al.* (1988). Although they did not use the extreme-value distribution, they recognized that the number of sequences S_x obtaining a score of $x \geq \bar{x}$, where $\bar{x}$ is the mean similarity score, decreases exponentially. A line fit to $\log(S_x)$, the declining number of scores excluding the top 3 %, can be used to extrapolate the expectation of obtaining a high score. This strategy works reasonably well because the number of sequences that obtain a score predicted from the probability density function (PDF) of the extreme-value distribution (Figure 2.3) has the form: $PDF(S = x) = \lambda Kmne^{-\lambda x}e^{-Kmne^{-\lambda x}}$. The second exponential term does not contribute significantly to the PDF when $x > \bar{x}$, so for high scores, the regression becomes $\log(PDF(S = x)) = \log(\lambda Kmn) - \lambda x$. Collins *et al.* (1988) recognized that the highest expected score by chance increased with the length of the query sequence, but they did not incorporate a length correction into their expectation calculation. The lack of a log n_l library sequence length correction significantly reduces the sensitivity of the search, as long unrelated sequences can have higher scores, by chance, than shorter related sequences (Pearson, 1995; 1998).
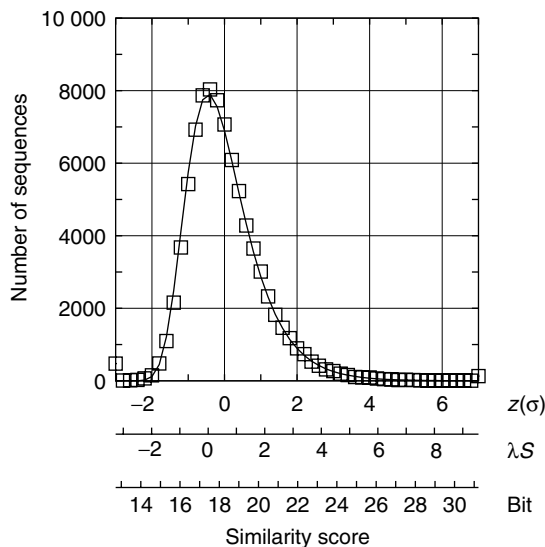
Figure 2.3 The extreme-value distribution. The observed distribution (squares) of similarity scores from a comparison of the human glucose transporter sequence `gtr1_human` against each of the ~84 000 sequences in SwissProt, and the expected (solid line) distribution of scores, based on the extreme-value distribution, are shown. Similarity scores were calculated with the Smith–Waterman algorithm, with the BLOSUM62 scoring matrix and a penalty of −12 for the first residue in a gap and −1 for each additional residue. The y-axis shows the number of SwissProt sequences obtaining the score shown on the x-axis. Three different scales for the similarity scores are shown: from top to bottom, these are the scores in terms of standard deviations (σ) above the mean (equation 2.6); the scale in terms of $\lambda S - \log(Kmn)$ (equation 2.1); and the *bit* score (equation 2.4).

Mott (1992) provided the first empirical evidence that the distribution of optimal local similarity scores with gaps could be well approximated by an extreme-value distribution. He considered an equation of the form $F(y, m, n, c) = \exp(-\mathrm{e}^{-(y-A)/B})$ where $A = a_0 + ca_1 + ca_2 \log(mn)$, $B = cb_1$ and $c = 1/\lambda_{\mathrm{ungapped}}$, defined as in equation (2.3). In this case, the $c = 1/\lambda$ parameter was calculated for sets of sequence pairs with identical compositions. In addition to correcting for the scaling of the $s_{i,j}$ scoring matrix, c reflects the amino acid composition of the two sequences being compared. Unfortunately, estimating λ or c using equation (2.3) is time-consuming. This approach may improve searches when query sequences have a biased amino acid composition, but it is not generally available in sequence comparison programs.

The most widely used estimates for λ and K for searches with gapped alignments are those provided for in the BLAST2 and PSI-BLAST comparison programs (Altschul *et al.*, 1997). These values are based on maximum likelihood estimates of λ, K, and H from simulations of random protein sequences of average composition (Altschul and Gish, 1996). H describes the relative entropy, or information content, of a scoring matrix and can be thought of as the average score per aligned residue (Altschul and Gish, 1996). In this case, the parameters of the extreme-value distribution are slightly different:

$$P(S \geq x) = 1 - \exp(-Km'n'\mathrm{e}^{-\lambda x}), \tag{2.8}$$

where for m, n, the query and library sequence lengths, $m' = m - \log(Kmn)/H$ and $n' = n - \log(Kmn)/H$. By correcting m and n for the expected length of an alignment between two random sequences $\log(Kmn)/H$, the search space term $Km'n'$ is estimated more accurately (Altschul and Gish, 1996).

The FASTA package of programs estimates the extreme-value parameters from the distribution of similarity scores calculated during the search (Pearson, 1996; 1998). This approach is efficient when scores are available for every sequence in the database, as is the case for a FASTA or Smith–Waterman search; no additional similarity scores must be calculated and the statistical parameters reflect the true distribution of similarity scores produced with the specific query sequence and the specific sequence database. However, a method that estimates statistical parameters from the actual distribution of similarity scores must avoid including scores from 'related' sequences in the estimation sample. This is straightforward in the typical case where a query sequence is compared to 50 000–500 000 sequences in a comprehensive database and fewer than 1000 sequences could be related in the worst case. However, these empirical statistical estimates cannot be used when a search is done against a special-purpose database that may contain only sequences from a single protein family. For this case, the FASTA programs provide an option to calculate a similarity score from a shuffled version of each sequence in the database; the distribution of these shuffled scores is then used for parameter estimation (Pearson, 1999).

By default, the FASTA programs estimate the location and scale parameters of the extreme-value distribution by fitting a line to the relationship between similarity score and $\log(n_l)$, the library sequence length, by calculating the mean and variance of similarity scores in bins of length $\log(n_l)$ of the library sequence (sequences in each bin differ in length by $\sim 10\%$: Figure 2.4). This line provides the location parameter, related to $\log(Kn_l)/\lambda$, and the residual variance $(\widehat{\sigma^2})$ of the $\log(n_l)$-normalized similarity scores, which can be used to calculate $\lambda = \pi/\sqrt{6\widehat{\sigma^2}}$. Binning similarity scores by $\log(n_l)$ provides a simple strategy for excluding related (high-scoring) sequences from the estimation process. $\log(n_l)$ length bins are initially weighted by the inverse variance of their similarity scores; length bins with very high scores have a high variance. After the initial $\log(n_l)$ regression is performed, bins that continue to have very high score variance are excluded and the number of bins and scores excluded is reported. Typically, this process excludes 0–2 of 50 length bins with about 5% of the library sequences. Once the $\log(n_l)$ regression line and $\widehat{\sigma^2}$, the average residual variance, have been determined, the probability of a single pairwise similarity score can be calculated using equation (2.6).

Alternatively, FASTA provides an option to estimate λ and K by maximum likelihood, using equation (2.1). This estimation is similar to that of Mott (1992), but omits the 'composition' data c and estimates the λ parameter directly. To avoid the scores from related sequences, the likelihood model implements a censored estimation strategy that excludes the lowest and highest 2.5% of the scores. This approach has the advantage that both K and λ are estimated directly and that there is no assumption that related sequences have well-defined lengths. (Families that are globally similar, e.g. globins and cytochrome 'c's, have characteristic lengths, but homologous domains, e.g. the EF-hand calcium binding domain, zinc-fingers, or protein kinase domains, may be in proteins with very different lengths.)

The major difference between the FASTA programs and the BLAST programs (aside from speed) is the strategy used for estimating statistical significance of similarity
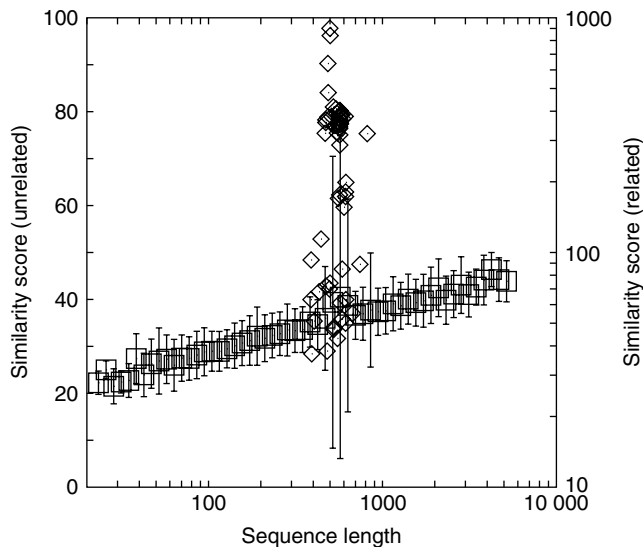
Figure 2.4 Empirical estimation of extreme-value parameters: sequence similarity scores plotted as a function of library sequence length. All the similarity scores calculated in the comparison of `gtr1_human` with an annotated subset of SwissProt (~24 000 sequences) are summarized. Scores from unrelated sequences are shown as averages (squares) with standard errors indicated; each score from a related sequence is plotted (diamonds). The unrelated sequence scores are plotted linearly against the left ordinate, the related scores are plotted on a logarithmic scale on the right ordinate.

scores. While BLAST pre-calculates λ and K from randomly shuffled sequences, FASTA calculates the extreme-value parameters from the actual distribution of similarity scores obtained in a search. Thus, FASTA must calculate at least an approximate similarity score for every sequence in a database. BLAST is fast because it calculates scores only for sequences that are likely to be homologous. While this strategy works well for protein sequences, it is more problematic for translated-DNA : protein comparisons, where the appropriate statistical model is more difficult to specify.

2.3.2.3 Pairwise Statistical Significance

The strategies outlined above can be used to estimate the statistical significance of a high similarity score obtained during a database search. If the BLAST2.0 (Altschul *et al.*, 1997) λ and K parameters are used as calculated in Altschul and Gish (1996), the statistical significance measurement reports the likelihood that a similarity score as good or better would be obtained by two random sequences with 'average' amino acid composition, and lengths similar to the lengths of the sequences that produced the score. However, if either of the two sequences have amino acid compositions significantly different from 'average', the statistical significance may be an over- or underestimate.

The empirical statistical estimates provided by programs in the FASTA package (Pearson, 1996; 1998) report a slightly different value: the expectation that a sequence with the length and composition of the query sequence would obtain a similarity score against an unrelated sequence drawn at random from the sequence database that was searched.

Again, if the query sequence has a slightly biased amino acid composition, e.g. because it is a membrane-spanning protein with several hydrophobic regions, then while the significance of the similarity with respect to 'average' composition proteins is accurate, the more biologically important question, the significance of the similarity when compared to unrelated membrane-spanning proteins, may be an overestimate. To address this problem one could use Mott's strategy to include the $c = 1/\lambda_{\text{ungapped}}$ composition/scaling parameter in the maximum likelihood fit, with c calculated using equation (2.3) for every pairwise comparison in the database. Although the composition calculation can be time-consuming, this option is available in the FASTA3 package.

The significance of a specific pairwise similarity score, in the context of the residue distributions in each of the two sequences, the query and library sequence, can also be estimated using a Monte Carlo approach. The two sequences are compared, and then one or both of the sequences will be shuffled hundreds of times to generate a sample of random sequences with the same length and residue composition. Similarity scores are calculated for alignments between the query sequence and each of the shuffled sequences. λ and K parameters can then be calculated from this distribution of scores using maximum likelihood, as is done by the PRSS program in the FASTA package (Pearson, 1996). The FASTA programs offer two shuffling options: (a) a *uniform* shuffle, in which each residue is randomly repositioned anywhere in the sequence; and (b) a *window* shuffle, in which the sequence is broken into n/w windows (n is the length of the sequence and w is the length of the window, typically 10–20 residues) and the sequence in each window is randomly shuffled. For 'average' composition query sequences, both uniform and window-shuffle estimates should be similar to those obtained from a database search. However, for scores of alignments between sequences of biased composition, significance estimates derived from the similarity scores of uniformly shuffled sequences should be more conservative than estimates based on the distribution of unrelated sequences from a comprehensive sequence database (Table 2.1). Window-shuffle estimates should be even more conservative, particularly if the similarity reflects a local patch of biased amino acid composition that would be homogenized by the uniform shuffling strategy.

Shuffling strategies rely on the assumption that the similarity scores of real unrelated protein sequences behave like the similarity scores of randomly generated sequences. While this is almost always true, some query sequences may have properties that are present in unrelated sequences but not in shuffled sequences. An alternative strategy for estimating λ and K from a comparison of two sequences has been proposed by Waterman and Vingron (1994),[4] based on a strategy they refer to as 'Poisson de-clumping'. They note that not only are the highest-scoring similarity scores from a sequence similarity search extreme-value distributed, but the highest $H_{(1)}$, second highest $H_{(2)}$, $H_{(3)}, \ldots, H_{(n)}$ alignment scores from a single pairwise comparison can be used to estimate λ and K, as long as the alignments do not overlap or intersect. An algorithm for calculating the n best nonintersecting local alignments between two sequences was described by Waterman and Eggert (1987); a space-efficient version is available as the SIM algorithm (Huang *et al.*, 1990). This approach has the advantage that it does not require the use of shuffled sequences, which may have different statistical properties than 'natural' protein sequences in some cases, and it calculates λ and K for the pair of sequences, with their specific lengths and residue compositions, rather than for an average distribution of library

---

[4] p and γ in Waterman and Vingron (1994) correspond to $e^{-\lambda}$ and K, respectively, in equation (2.1).

Table 2.1 Statistical significance estimates. Expectation values are shown for similarity scores between human glucose transporter type 1 (gtr1_human); three members of the glucose transporter family, quinate permease (qutd_emeni), maltose permease (cit1_ecoli), and α-ketoglutarate permease (kgtp_ecoli); and a probable nonmember, a hypothetical yeast protein (yb8g_yeast). The BLASTP2.0 search was done with the default scoring matrix (BLOSUM62) and gap penalties -12 for the first residue in a gap (-11 gap-open) and -1 for each additional residue (gap-extend). SSEARCH (Smith and Waterman, 1981; Pearson, 1996) searches used either the default matrix (BLOSUM50, BL50) and gap penalties ($-12/-2$) or the same scoring matrix and gap penalties as the BLASTP2.0 search (BL62). SSEARCH statistical estimates were calculated using the default linear regression method (BL50, BL62) or the maximum likelihood method (BL62*). Both BLASTP2.0 and SSEARCH searches examined alignments between sequences with low-complexity regions removed by the SEG program (Wootton, 1994). Expectation values are reported in the context of a search of the SwissProt (Bairoch and Apweiler, 1996) database ($\sim$84 000 entries). The λ scaling/composition-factor for each search is shown in the right column. Statistical significance was also estimated by a Monte Carlo approach (PRSS) in which the second sequence was shuffled 1000 times using either a uniform or 'window' (-w 20) shuffle. Expectations reported by PRSS have been multiplied by 84 to reflect the expectation from a search of the 84 000 entry SwissProt database.

gtr1_human:	qutd_emeni moderate	cit1_ecoli distant	kgtp_ecoli very weak	yb8g_yeast unrelated	λ
BLASTP2.0	2.0e-25	1e-05	0.077	2.0	0.2700
SSEARCH BL50	1.6e-28	6.1e-05	0.014	0.72	0.1544
raw-score	536	199	148	123	–
bit-score	127	48	40	35	–
% identity	27.1	22.1	24.1	22.1	–
SSEARCH BL62	4.7e-32	1.2e-4	1.3	3.1	0.2584
raw-score	356	120	75	72	–
bit-score	138	47	34	33	–
% identity	26.9	21.0	27.9	24.1	–
SSEARCH BL62*	2.8e-30	3.2e-4	2.3	5.2	0.2459
bit-score	356	46	33	32	–
PRSS BL50	7.2e-25	6.5e-03	0.0039	92	–
λ	0.1375	0.1237	0.1317	0.1263	–
window 20	3.9e-09	0.097	0.21	361	–
λ	0.0653	0.1064	0.1206	0.1110	–
BL62	6.6e-30	8.5e-4	0.36	49	–
λ	0.2405	0.2282	0.2343	0.2265	–
window 20	2.0e-25	9.8e-03	0.72	92	–
λ	0.2108	0.2011	0.2256	0.2172	–

sequences. However, the approach also assumes that, for some i, $H_{(i)}$ reflects the score of an alignment that occurs by chance, rather than because of homology. This is true for single-domain proteins that do not contain internal repeats, but it is not true for proteins containing internal duplications. For example, a comparison of calmodulin with troponin 'C' would produce $H_{(1)}, \ldots, H_{(4)}$ which reflect the homology of the four EF-hand calcium binding domains in each sequence, and $H_{(5)}, \ldots, H_{(n)}$, which could be used to estimate λ and K. A protein with a dozen copies of a duplicated domain would have more than 100 local alignments with scores that reflect homology.

2.3.2.4   Accuracy of λ and K

Reliable statistical estimates for similarity scores can dramatically improve the sensitivity of a similarity search, because they provide an accurate quantitative model for the behavior of scores from unrelated sequences. Thus, it is far more informative to state that a pair of distantly related sequences has a similarity score that is expected by chance only once in 10 000 database searches (E() $< 10^{-4}$) than it is to state that two sequences share 30 % identity. Unfortunately, percent identity remains the most commonly published measure of sequence similarity, despite the fact that identity measures are far less effective than similarity scores that reflect conservative replacements (Schwartz and Dayhoff, 1978; Pearson, 1995; Levitt and Gerstein, 1998). High levels of identity are frequently seen between unrelated sequences over short regions (Kabsch and Sander, 1984) and sequence alignments with less than 25 % identity may either be clearly statistically significant (Table 2.1), (gtr1_human versus cit1_ecoli, BL62, E() $< 10^{-9}$) or not significant (gtr1_human versus yb8g_yeast, BL62, E() < 0.25).

Before accurate statistical estimates for local similarity scores were available, it was routine to consider the tradeoffs between a search strategy's 'sensitivity', the ability to identify distantly related sequences (to avoid false negatives), and its 'selectivity', not assigning high scores to unrelated sequences (false positives). With an accurate model for the distribution of similarity scores from unrelated sequences, the threshold for statistical significance (typically 0.02–0.001) sets the selectivity or false positive rate; a threshold E() < 0.001 predicts a false positive every 1000 searches. Thus, a significance threshold of E() < 0.001 is expected to produce several false positives when characterizing all the proteins in *E. coli* or yeast (4000 and 6000 proteins), and 18 false positives are expected with E() < 0.001 when each of the 18 000 proteins in *Caenorhabditis elegans* is compared to the SwissProt database. However, the conservative strategy of reducing the significance threshold to 0.001/4000 for *E. coli*, or 0.001/18 000 for *C. elegans*, ensures that many homologous proteins will be missed (false negatives).

Of the λ and K parameters for the extreme-value distribution, the scale parameter λ has the largest effect on the statistical significance estimate. In searches using the BLOSUM62 scoring matrix and gap penalties of −12/−2 of a subset of the SwissProt with database 50 unrelated protein sequences with lengths ranging from 98 to 2252 (mean 432 ± 57), maximum likelihood estimates of λ ranged from 0.204 to 0.304 (mean 0.275) while K ranged from 0.0039 to 0.062 (mean 0.012). K and λ are strongly correlated; low values of K are found with low values of λ. Around the average values, however, reducing K by a factor of 2 reduces the E() value only twofold (1 bit), but a similar change in statistical significance would occur by reducing λ from 0.275 to 0.268, or about 2.5 %. Reducing λ by 20 %, which is well within the range of λs seen after shuffling with PRSS in Table 2.1, would reduce the statistical significance of a raw score of 100 by 250-fold, or by 8 bits.

Table 2.1 illustrates the importance of λ on significance estimates for three related and one unrelated sequence. The differences in expectation values reflect differences in estimates for λ and K; for a given scoring matrix (BLOSUM50 or BLOSUM62) the raw similarity scores for each pairwise comparison (e.g. gtr1_human:cit1_ecoli) do not change. The significant differences between the λ values for BLOSUM50 and BLOSUM62 reflect the different scaling of the two matrices. BLOSUM50 is scaled at 0.33 bits per unit raw score, so that a raw score of 148 produces a bit score of ∼148/3

(the actual value for these gap penalties is 0.27 bits/raw-score). BLOSUM62 is scaled at 0.5 bits/raw-score, with a raw score of 75 giving a bit score of 34.[5]

Two trends are apparent: λ estimates from PRSS shuffled comparisons tend to be smaller than λ estimates from database searches; and λ estimates for local (window) shuffles are somewhat lower, reducing the significance even further. These decreases in λ are expected because the query and library sequences used in this example have a somewhat biased amino acid composition; the proteins have multiple transmembrane domains with a bias towards hydrophobic amino acid residues (Kyte and Doolittle, 1982). Thus, the λs from SSEARCH are lower than the BLAST2.0 λ, because the simulations used to assign λ in BLAST2.0 assume an 'average' amino acid composition for both the query and library sequence; the empirical SSEARCH estimates correct for the composition bias of the query sequence, but still reflect the 'average' composition of the library sequences. λs determined by PRSS shuffling are lower still, because PRSS estimates account for the composition bias in both the query and library sequences. Window shuffling in PRSS reduces λ even further, presumably because the highest-scoring regions in each pairwise comparison are restricted to sequence patches with the most biased composition. However, despite these differences in λs, the SSEARCH and PRSS uniform-shuffle significance estimates for the intermediate and distantly related sequence pairs usually agree within a factor of 4. Window-shuffled estimates reduce statistical significance much more dramatically, about 2–4 orders of magnitude for moderately and weakly significant similarities.

The statistical estimates provided by the BLAST2.0 and FASTA sequence comparison programs are generally robust and reliable. To illustrate the factors affecting significance estimates, we have emphasized the modest differences in λ and E() in Table 2.1. However, Table 2.1 illustrates even for sequences with biased amino acid composition that share 20–25 % sequence identity, the significance estimates reported by either BLAST2.0 or programs in the FASTA package are very similar, and consistent with statistical estimates produced by uniform shuffling. Window-based shuffling produces a much more conservative statistical estimate.

2.3.3   Evaluating Statistical Estimates

The inference of homology (common ancestry) from statistically significant similarity rests on two assertions: that similarity scores, calculated with optimal (Smith–Waterman) or heuristic (BLAST or FASTA) algorithms using common scoring matrices (PAM250, BLOSUM62) and gap penalties, follow the extreme-value distribution; and that the behavior of similarity scores for random sequences holds as well for real, unrelated, protein sequences. This second assertion is critical – an accurate statistical theory for similarity scores of random sequences is of little value if real sequences have properties that distinguish their scores from those of random sequences. It seems reasonable that real protein sequences might have statistical properties that distinguish them from random sequences; of the $20^{400} = 2.6 \times 10^{520}$ potential sequences of length 400 that could be generated at random from 20 amino acids, fewer than 10^5–10^8 unrelated sequences are thought to exist in nature, and many structural biologists would argue that there are fewer than 10^3 distinct protein folds (Brenner *et al.*, 1997). Real protein sequences are constrained to fold into a compact three-dimensional structure with a physiological

---

[5] Because of this different scaling, a gap penalty of $-12/-2$ for BLOSUM50, the default with SSEARCH, is equivalent to a gap penalty of $-8/-1$ for BLOSUM62.

function; the fact that such a large fraction (typically $50-80\%$) of the sequences in most organisms can be found in other distantly related organisms suggests that the folding constraint substantially restricts the universe of protein sequences; it is far easier to produce a new protein sequence by duplicating an old one than by producing a sequence *de novo*. Thus, it would not be surprising to learn that the folding/function constraint produced real protein sequences whose similarity scores behave differently from those of random protein sequences.

The reliability of statistical estimates can be evaluated both by comparing the observed distribution of sequence similarity scores obtained in a search with the expected extreme-value distribution, and by examining the expectation value for the highest-scoring nonhomologous sequence. Figure 2.5 shows the distribution of sequence similarity scores for two query sequences, an 'average' protein sequence, `pyre_colgr`, orotate phosphoribosyltransferase, and a protein sequence with a biased amino acid composition, `prio_atepa`, major prion precursor. Two sets of similarity scores are shown for each sequence. One set shows the scores obtained when all the amino acid residues in the library are examined; the second shows the scores when low-complexity sequences, or regions of sequence with a reduced or biased amino acid composition, are removed (Wootton, 1994). With the 'average' protein, the distributions of the 'complete' database scores and 'high-complexity' scores are indistinguishable. With the prion protein (Figure 2.5b), there is some difference in the central portion of the distribution, but the greatest differences are seen for the highest-scoring sequences, where there are typically $2-3$ times as many 'raw' scores as expected between 4 and 5 standard deviations above the mean, and $5-10$ times as many scores as expected from 5 to 7 standard deviations above the mean. This effect of biased composition is largely removed by searching against a SEGed database that has had low-complexity regions removed.

The effect of biased composition is seen more dramatically by looking at the number of very high-scoring sequences and the expectation value of the highest-scoring unrelated sequence. When `prio_atepa` is used as a query, 198 library 'raw' sequence scores have $z \geq 7.0$;[6] this is reduced to 28, 26 of which are related to the query, when the SEGed database is used. Likewise, when the 'raw' sequences are examined, the highest-scoring unrelated sequence is a glycine-rich cell wall protein that obtains an expectation value of $E() < 10^{-8}$ and there are 90 unrelated sequences with $10^{-8} \leq E() \leq 0.01$. In contrast, with the SEGed database the highest-scoring unrelated sequence has an expectation value of $E() = 0.012$ and the second highest unrelated sequence has $E() = 0.99$.

Reliable statistical estimates – statistics that estimate $E() < 0.02$ about 2% of the time – allow much more sensitive searches. If an investigator can have confidence that an unrelated sequence will obtain a score of $E() < 0.001$ about once in 1000 searches, $E() < 0.001$ can be used to reliably infer homology. However, if unrelated sequences sometimes obtain $E() < 0.001$ by chance, a more conservative threshold may be adopted, e.g. $E() < 10^{-6}$ or even $E() < 10^{-10}$. While using a very stringent threshold for statistical significance ensures that one will rarely infer homology when the proteins are unrelated, it also ensures that moderately distant evolutionary relationships will be missed. Thus, both the FASTA and BLAST developers have given high priority to the accuracy of

---

[6] Similarity scores 7 standard deviations above the mean have an expectation value $E() < 1.7$ for this database of 23 981 sequences.
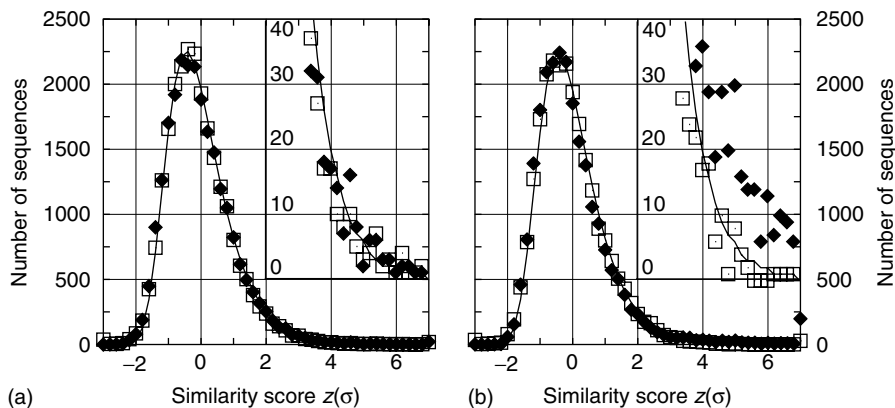
Figure 2.5 Distribution of library sequence similarity scores in searches with (a) an 'average' protein sequence, `pyre_colgr`, and (b) a sequence with biased amino acid composition, `prio_atepa`. Filled diamonds show the distribution of similarity scores that include all the residues in every sequence; open squares show the distribution of scores when low-complexity regions are removed with the PSEG program (Wootton, 1994). The solid line shows the expected distribution of scores predicted from the size of the database and the extreme-value distribution (2.6). The x-axis reports similarity scores scaled in standard deviations above the mean. Searches were done with the SSEARCH33 program (Smith–Waterman) using the BLOSUM62 scoring matrix with gap penalties of -12 for the first residue in a gap and -2 for each additional residue. λ and K were estimated by maximum likelihood ($-z\ 2$) option: (Pearson, 1999).

the statistical estimates, particularly for the highest-scoring unrelated sequences (Brenner *et al.*, 1998).

When evaluating the quality of statistical estimates for high-scoring unrelated sequences, it is important to examine real protein sequences, whose properties may differ from randomly generated sequences. Figure 2.6 summarizes the highest-scoring unrelated sequence similarity scores obtained when query sequences from 50 randomly selected Pfam protein families were used to search a database of sequences with carefully annotated evolutionary relationships. Searches were done with either random sequences generated from the 50 Pfam family queries, or with the queries themselves, against either a 'raw' protein sequence database, or one with low-complexity regions removed. Figure 2.6(a) shows that even when random sequences are used to search the database, similarity scores can be much higher than expected (and E() values much lower than expected) if low-complexity regions are present in the sequence database. Thus, when 50 random sequences were used, the lowest E() value was 0.006 from a match between a randomly shuffled human histone H1 (`h10_human`) and other histone H1 sequences. This may simply reflect the fact that it is difficult to randomly shuffle a sequence that is 30 % lysine. However, when low-complexity regions are removed, the observed and expected distributions of E() values agree extremely well.

When real sequences are used as the query, the statistical estimates are not as accurate, even when low-complexity regions are removed. Most of the time, however, the estimates are not far off. The log/log plot in Figure 2.6 emphasizes the searches that obtained the lowest E() values for unrelated sequences, but 80 % of the real query sequences
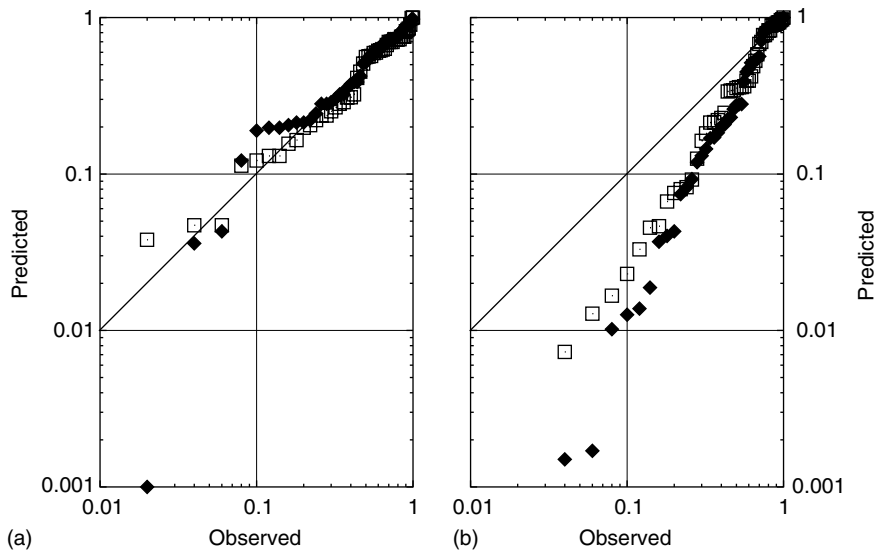
Figure 2.6 Quantile–quantile plot of expectation values for searches with (a) 50 random sequences and (b) 50 real protein sequences for which the highest scoring unrelated sequence is known. Searches were performed against a 'raw' annotated protein sequence database (filled diamonds) and the same database with low-complexity regions removed (open squares). For each search, the highest score (a) or highest scoring unrelated (b) sequence was recorded, and converted from an expectation (E()) to a probability of obtaining that E() using the Poisson formula $p(E) = 1 - e^{-E}$. Each set of 50 probabilities was sorted from lowest to highest and plotted. The 50 query sequences were chosen from 50 randomly selected PFAM families (Sonnhammer *et al.*, 1997) with 25 or more members. The random sequences were obtained by shuffling the 50 real PFAM derived sequences. Searches were done using the Smith–Waterman algorithm (SSEARCH33) using the default scoring matrix (BLOSUM50) and gap penalties $(-12/-2)$ with regression-scaled (binned) statistical estimates.

had expectation values E() > 0.1 (low by a factor of 2), and 90 % had E() > 0.02 (low by a factor of 5) when low-complexity sequences were removed from the database. In the search of the SEGed database, again the most 'significant' unrelated similarity score was involved alignments with h1o_human. In the search against the raw database, this alignment had an E() < 0.002 (low by factor of 10); against a SEGed database the score was even lower, E() < 0.0006. Histone H1 has an exceptionally biased amino acid composition, which cannot be completely corrected for by removing low-complexity regions from the database. However, for the vast majority of query sequences (80–90 %), unrelated sequences will have expectation values within a factor of 2–5 of their true frequency in database searches. Thus, thresholds of statistical significance in the range 0.001 < E() < 0.01 against SEGed sequence databases will be reliable with rare exceptions.

The observation that the statistical significance estimates (E() values) from similarity searches with real, unrelated sequences are 2–5 times less conservative than those obtained for genuinely random sequences suggests that to a large extent, real, unrelated protein sequences have many of the same statistical properties as random sequences. The major

difference between real protein sequences and random sequences seems to be the i.i.d. assumption for amino acid residue positions. In real, unrelated sequences, unusual amino acid compositions are distributed in low-complexity clumps. The SEG program, which masks out these regions, removing them from the similarity score calculation, can reduce the effect of clumps with biased composition, but not eliminate it. Fortunately, the deviation from being i.i.d. is modest in 80 % of protein sequences. Other than the biased composition effect, no other property of 'real' protein sequences has been identified that distinguishes them from sequences built from picking amino acids from a probability distribution at random.

2.4   SUMMARY: EXPLOITING STATISTICAL ESTIMATES

The inference of homology from statistically significant sequence similarity is one of the most reliable conclusions a scientist can draw. Indeed, the vast majority of bacterial, *C. elegans*, and *Drosophila* genes are annotated largely on the basis of statistically significant sequence similarity shared by other proteins with known structures or functions. This trend is certain to continue as sequence databases become more comprehensive.

While the inference of homology from significant sequence similarity is reliable – sequences that share much more similarity than expected by chance share a common ancestor – the inference tells us much more about structure than function. Without exception, sequences that share statistically significant similarity share significant structural similarity. However, homologous proteins need not perform the same or even similar functions. Functional inferences are most reliable when based on assignments of *orthology*. *Orthologous* sequences are sequences that differ because of species differences. This contrasts with *paralogous* sequences, which are produced by gene duplication events. While *homology* can be demonstrated by sequence similarity, an inference of *orthology* is best supported by phylogenetic analysis, which is considerably more challenging computationally. In addition, many proteins are built from evolutionarily independent domains with different structures and functions. The inference of homology is transitive – if protein A is homologous to B and B is homologous to C, even if A and C do not share significant similarity – but it is critical that such inferences be limited to the domain to which they apply. There is great concern that incorrect functional assignments are greatly reducing the value of sequence database annotations because functional assignments are inappropriately extended to new family members based on a correct, but functionally uninformative, inference of homology.

Statistical significance estimates, whether as expectation values or bit scores, are far more informative than the most commonly used measure of sequence similarity, percent identity. It has been known for more that 20 years (Dayhoff *et al.*, 1978) that percent identity is much less effective than measures of similarity that distinguish biochemically similar and dissimilar amino acids, and that recognize that some amino acids mutate far more rapidly than others. Moreover, high sequence identity is expected over very short regions by chance in unrelated sequences that share no structural similarity (Kabsch and Sander, 1984). Thus, the inference of homology should always be based on statistically significant sequence similarity using an appropriate scoring matrix (Altschul, 1991).

However, once homology has been established, measures of statistical significance are not good measures of evolutionary distance. Two sequences that have diverged by the same amount, and thus share the same average levels of sequence similarity, can have very different similarity scores, with very different levels of statistical significance, depending on their lengths. For example, two members of the orotate phosphoribosyltransferase family, `pyre_colgr` and `pyre_klula`, that share 48.5% identity over 223 amino acid residues, have similarity scores $S_{bit} = 161$ with $E() < 10^{-39}$, while two members of the twice as long glucose transporter family with slightly lower identity (47.4% over 502 amino acids) obtain a similarity score of 308 bits with $E() < 10^{-82}$. Thus, similarity scores and expectation values must be adjusted when comparing among different length protein sequences if they are used as surrogates for evolutionary divergence.

This review of sequence similarity statistics has focused on protein sequence comparison for two reasons. First, protein sequence comparison is far more sensitive than DNA sequence comparison – the evolutionary lookback time for protein sequences is typically 5–10 times greater than that for DNA sequences (Pearson, 1997). Moreover, protein databases are more compact, so that more rigorous algorithms can be used for similarity searching. Secondly, DNA sequences are well known to have higher-order sequence dependence due to codon bias and simple-sequence repeat regions. Because of the small nucleotide alphabet and the possible translation of normal-complexity DNA sequences into low-complexity protein sequences, it is much more difficult to detect and correct for deviations from i.i.d. in DNA sequences. Thus, in general, statistical estimates from protein sequence comparisons are more reliable than the similar comparisons with DNA.

Our understanding of the statistical properties of biological sequences has improved dramatically over the past decade, so that most sequence similarity searching methods now include reliable statistical estimates. However, there is still room for improvement, as more searches are done with more complex queries, e.g. profiles, position-specific scoring matrices (Altschul *et al.*, 1997), and three-dimensional sequence-structure alignments, whose statistical properties on real sequences are not well understood. Fortunately, there is no shortage of data that can be used to develop and validate new statistical approaches.

Acknowledgments

REFERENCES

Allison, T.J., Wood, T.C., Briercheck, D.M., Rastinejad, F., Richardson, J.P. and Rule, G.S. (1998). Crystal structure of the RNA-binding domain from transcription termination factor ρ. *Nature Structural Biology* **5**, 352–356.

Altschul, S.F. (1991). Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology* **219**, 555–565.

Altschul, S.F. and Gish, W. (1996). Local alignment statistics. *Methods in Enzymology* **266**, 460–480.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.

Andrade, M., Casari, G., de Daruvar, A., Sander, C., Schneider, R., Tamames, J., Valencia, A. and Ouzounis, C. (1997). Sequence analysis of the *Methanococcus jannaschii* genome and the prediction of protein function. *Computer Applications in the Biosciences* **13**, 481–483.

Arratia, R., Gordon, L. and Waterman, M.S. (1986). An extreme value theory for sequence matching. *Annals of Statistics* **14**, 971–993.

Arratia, R., Gordon, L. and Waterman, M.S. (1990). The Erdös–Rényi law in distribution, for coin tossing and sequence matching. *Annals of Statistics* **18**, 539–570.

Bairoch, A. and Apweiler, R. (1996). The Swiss-Prot protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Research* **24**, 21–25.

Brenner, S.E., Chothia, C. and Hubbard, T.J. (1997). Population statistics of protein structures: lessons from structural classifications. *Current Opinion in Structural Biology* **7**, 369–376.

Brenner, S.E., Chothia, C. and Hubbard, T.J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Sciences (USA)* **95**, 6073–6078.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sulton, G.G., Blake, J.A., Fitzgerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., Dougherty, B.A., Tomb, J.-F., Adams, M.D., Reisch, C.I., Overbeek, R., Kirkness, E.F., Weinstock, K.G., Merrick, J.M., Glodek, A., Scott, J.L., Geoghagen, N.S.M., Weidman, J.F., Fuhrmann, J.L., Nguyen, D., Utterback, T.R., Kelley, J.M., Peterson, J.D., Sadow, P.W., Hanna, M.C., Cotton, M.D., Roberts, K.M., Hurst, M.A., Kaine, B.P., Borodovsky, M., Klenk, H.-P., Fraser, C.M., Smith, H.O., Woese, C.R. and Venter, J.C. (1996). Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science* **273**, 1058–1073.

Collins, J.F., Coulson, A.F.W. and Lyall, A. (1988). The significance of protein sequence similarities. *Computer Applications in the Biosciences* **4**, 67–71.

Dayhoff, M., Schwartz, R.M. and Orcutt, B.C. (1978). In *Atlas of Protein Sequence and Structure*, Vol. 5, supplement 3, M. Dayhoff, ed. National Biomedical Research Foundation, Silver Spring, MD, pp. 345–352.

Doolittle, R.F. (1994). Convergent evolution: the need to be explicit. *Trends in Biochemical Sciences* **19**, 15–18.

Evans, M., Hastings, N. and Peacock, B. (1993). *Statistical Distributions*, 2nd edition. Wiley, New York.

Fitch, W.M. (1966). An improved method of testing for evolutionary homology. *Journal of Molecular Biology* **16**, 9–16.

Henikoff, S. and Henikoff, J.G. (1992). Amino acid substitutions matrices from protein blocks. *Proceedings of the National Academy of Sciences (USA)* **89**, 10 915–10 919.

Huang, X., Hardison, R.C. and Miller, W. (1990). A space-efficient algorithm for local similarities. *Computer Applications in the Biosciences* **6**, 373–381.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**, 275–282.

Kabsch, W. and Sander, C. (1984). On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proceedings of the National Academy of Sciences (USA)* **81**, 1075–1078.

Karlin, S. and Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences (USA)* **87**, 2264–2268.

Karlin, S., Bucher, P., Brendel, V. and Altschul, S.F. (1991). Statistical methods and insights for protein and DNA sequences. *Annual Review of Biophysics and Biophysical Chemistry* **20**, 175–203.

Kent, G.C. (1992). *Comparative Anatomy of the Vertebrates*. Mosby, St. Louis, MO.

Koonin, E.V. (1997). Big time for small genomes. *Genome Research* **7**, 418–421.

Kyte, J. and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105–132.

Levitt, M. and Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciences (USA)* **95**, 5913–5920.

Lipman, D.J. and Pearson, W.R. (1985). Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441.

Mott, R. (1992). Maximum likelihood estimation of the statistical distribution of Smith–Waterman local sequence similarity scores. *Bulletin of Mathematical Biology* **54**, 59–75.

Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology* **48**, 444–453.

Neurath, H., Walsh, K.A. and Winter, W.P. (1967). Evolution of structure and function of proteases. *Science* **158**, 1638–1644.

Orengo, C.A., Swindells, M.B., Michie, A.D., Zvelebil, M.J., Driscoll, P.C., Waterfield, M.D. and Thornton, J.M. (1995). Structural similarity between the pleckstrin homology domain and verotoxin: the problem of measuring and evaluating structural similarity. *Protein Science* **4**, 1977–1983.

Owen, R. (1843). *Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals*. Longman, Brown, Green and Co., London.

Owen, R. (1866). *On the Anatomy of Vertebrates*. Longmans, Green and Co., London.

Pearson, W.R. (1995). Comparison of methods for searching protein sequence databases. *Protein Science* **4**, 1145–1160.

Pearson, W.R. (1996). Effective protein sequence comparison. *Methods in Enzymology* **266**, 227–258.

Pearson, W.R. (1997). Identifying distantly related protein sequences. *Computer Applications in the Biosciences* **13**, 325–332.

Pearson, W.R. (1998). Empirical statistical estimates for sequence similarity searches. *Journal of Molecular Biology* **276**, 71–84.

Pearson, W.R. (1999). In *Bioinformatics Methods and Protocols*, S. Misener and S.A. Krawetz, eds. Humana Press, Totowa, NJ, pp. 185–219.

Rawlings, N.D. and Barrett, A. (1993). Evolutionary families of peptidases. *Biochemical Journal* **290**, 205–218.

Sanderson, M. and Hufford, L. (eds) (1996). *Homoplasy: The Recurrence of Similarity in Evolution*. Academic Press, New York.

Schwartz, R.M. and Dayhoff, M. (1978). In *Atlas of Protein Sequence and Structure*, Vol. 5, supplement 3, M. Dayhoff, ed. National Biomedical Research Foundation, Silver Spring, MD, pp. 353–358.

Smith, T.F. and Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197.

Sonnhammer, E.L., Eddy, S.R. and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405–420.

States, D.J., Gish, W. and Altschul, S.F. (1991). Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *METHODS: A Companion to Methods in Enzymology* **3**, 66–70.

Waterman, M.S. (1995). *Introduction to Computational Biology*. Chapman and Hall, London.

Waterman, M.S. and Eggert, M. (1987). A new algorithm for best subsequences alignment with application to tRNA–rRNA comparisons. *Journal of Molecular Biology* **197**, 723–728.

Waterman, M.S. and Vingron, M. (1994). Rapid and accurate estimates of the statistical significance
    for sequence database searches. *Proceedings of the National Academy of Sciences (USA)* **91**,
    4625–4628.
Watson, H.C. and Kendrew, J. (1961). Comparison between the amino acid sequences of sperm
    whale myoglobin and of human haemoglobin. *Nature* **190**, 670–672.
Wootton, J.C. (1994). Non-globular domains in protein sequences: automated segmentation using
    complexity measures. *Computers and Chemistry* **18**, 269–285.

Bayesian Methods in Biological Sequence Analysis

Jun S. Liu and T. Logvinenko

Department of Statistics, Harvard University, Cambridge, MA, USA

Hidden Markov models, the expectation-maximization algorithm, and the Gibbs sampler were introduced for biological sequence analysis in early 1990s. Since then the use of formal statistical models and inference procedures has revolutionized the field of computational biology. This chapter reviews the hidden Markov and related models, as well as their Bayesian inference procedures and algorithms, for sequence alignments and gene regulatory binding motif discoveries. We emphasize that the combination of Markov chain Monte Carlo and dynamic-programming techniques often results in effective algorithms for nondeterministic polynomial (NP)-hard problems in sequence analysis. We will also discuss some recent approaches to infer regulatory modules and to combine expression data with sequence data.

3.1 INTRODUCTION

In the past two decades, we have witnessed the development of the likelihood approach to pairwise sequence alignments (Bishop and Thompson, 1986; Thorne *et al.*, 1991); probabilistic models for RNA secondary structure predictions (Zuker, 1989; Lowe and Eddy, 1997; Ding and Lawrence, 2001; Pedersen *et al.*, 2004); the expectation-maximization (EM) algorithm for finding regulatory binding motifs (Lawrence and Reilly, 1990; Cardon and Stormo, 1992), the Gibbs sampling strategies for detecting subtle sequence similarities (Lawrence *et al.*, 1993; Liu, 1994; Neuwald *et al.*, 1997); the hidden Markov models (HMMs) for DNA composition analysis, multiple sequence alignments, gene prediction, and protein secondary structure prediction (Churchill, 1989; Krogh *et al.*, 1994a; Baldi *et al.*, 1994; Burge and Karlin, 1997; Schmidler *et al.*, 2000; **Chapters 4** and **5**); regression and Bayesian network approaches to gene regulation networks (Bussemaker *et al.*, 2001; Segal *et al.*, 2003; Conlon *et al.*, 2003; Beer and Tavazoie, 2004; Zhong *et al.*, 2005); and many statistical-model based approaches to gene expression microarray analyses (Li and Wong, 2001; Lu *et al.*, 2004; Speed, 2003). All these developments show

that algorithms resulting from statistical modeling efforts constitute a significant portion of today's bioinformatics toolbox. This chapter aims at introducing the readers to these modeling techniques and related Bayesian methodologies.

Section 3.2 gives an overview of the Bayesian inference procedure, including model building, prior specification, model selection, and Bayesian computation. Section 3.3 introduces the general HMM framework with an example on DNA compositional heterogeneity. Section 3.4 reviews Bayesian pairwise alignment methods. Section 3.5 demonstrates how HMMs are used in multiple sequence alignment (MSA) and how Bayesian inferences can be made on model parameters. Section 3.6 outlines some Bayesian methods for finding subtle repetitive motifs, which often correspond to transcription factor (TF) binding sites and binding modules, in DNA sequences. Section 3.7 provides a brief overview of recent activities in combining gene expression microarray information with promoter sequence analysis. Section 3.8 concludes the chapter with a brief summary. In this chapter, we emphasize the usefulness of dynamic-programming-like recursive algorithms in Bayesian and likelihood-based inferences and the importance of combining these efficient computational techniques with more flexible Markov chain Monte Carlo MCMC tools for bioinformatics problems.

3.2   OVERVIEW OF THE BAYESIAN METHODOLOGY

In Bayesian analysis, a joint probability distribution $f(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\tau})$ is employed to describe relationships among all variables under consideration: those that we observe (data and knowledge, $\mathbf{y}$), those about which we wish to learn (scientific hypotheses, $\boldsymbol{\theta}$), and those that are needed in order to construct the model (missing data or nuisance parameters, $\boldsymbol{\tau}$). The basic probability theory then leads us to an efficient use of the available information and to a precise quantification of uncertainties in estimation and prediction (Gelman *et al.*, 1995). The Bayesian approach has following advantages: (1) its explicit use of probability models to formulate scientific problems (i.e. a quantitative story-telling); (2) its coherent way of incorporating all sources of information and treating nuisance parameters and missing data; and (3) its ability to quantify uncertainties in all estimates. Some other aspects of the Bayesian method are discussed in **Chapters 8**, **15**, **19**, **20**, and **26**.

3.2.1   The Procedure

Bayesian analysis treats parameters $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ as realized values of random variables that follow a *prior distribution*, $f_0(\boldsymbol{\theta}, \boldsymbol{\tau})$, typically regarded as known to the researcher independently of the data under analysis. The joint probability distribution can then be represented as *Joint = likelihood × prior*, that is,

$$p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\tau}) = f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\tau}) f_0(\boldsymbol{\theta}, \boldsymbol{\tau}).$$

The theorem that combines the prior and the data to form the conditional distribution $p(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \mathbf{y})$, also called the *posterior distribution* of $\boldsymbol{\theta}$, is a simple mathematical result first given by Thomas Bayes in his famous article "Essay Towards Solving a Problem in the Doctrine of Chances" (1763), published posthumously in the *Philosophical Transactions of the Royal Society of London*. As per Bayes theorem,

$$p(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\tau})}{p(\mathbf{y})} = \frac{f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\tau}) f_0(\boldsymbol{\theta}, \boldsymbol{\tau})}{\int f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\tau}) f_0(\boldsymbol{\theta}, \boldsymbol{\tau}) \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\boldsymbol{\tau}}. \tag{3.1}$$

The denominator $p(\mathbf{y})$, which is a normalizing constant for the function, is sometimes called the *marginal likelihood* and can be used for model selection. If we are interested in only $\boldsymbol{\theta}$, we can obtain its marginal posterior distribution as

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \int p(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \mathbf{y})\mathrm{d}\boldsymbol{\tau}. \tag{3.2}$$

Formula (3.2) can give us not only a point estimate of $\boldsymbol{\theta}$ (e.g. posterior mean), but also an explicit measure of uncertainty (e.g. a 95 % probability interval). In contrast, frequentist approaches often face conceptual difficulties in dealing with nuisance parameters and in quantifying uncertainties.

Statistical procedures based on the systematic use of this theorem to manipulate subjective probabilities are often termed *Bayesian*, although they were first developed by Laplace in the early 1800s, after Bayes' death. Despite the deceptively simple-looking form of (3.1) and (3.2), the challenging aspects of Bayesian statistics are (i) the development of a model, $f(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\tau})f_0(\boldsymbol{\theta}, \boldsymbol{\tau})$, that must effectively capture the key features of the underlying scientific problem; and (ii) the necessary computation for deriving the posterior distribution.

3.2.2 Model Building and Prior

It is often a useful model building strategy to distinguish two kinds of unknowns: population parameters and missing data. Although there is no absolute distinction between the two types, missing data are usually directly related to individual observations. They can be 'imputed' either conceptually or computationally so as to ease the statistical analysis. On the other hand, the parameters usually characterize the entire population under study and are fixed in number. For example, in a multiple alignment problem, alignment variables that must be specified for each observed sequence can be viewed as missing data. Residue frequencies for the aligned positions or the choice of scoring matrices, which apply to all the sequences, are population parameters.

The main controversial aspect of the Bayesian method is the need for *prior* distributions for unknown parameters. Since the choice of priors injects subjective judgments into the analyses, Bayesian methods have long been regarded as less 'objective' than their frequentist counterpart and disfavored. However, the emotive words 'subjective' and 'objective' should not be taken too seriously since there are considerable subjective elements and personal judgments in all phases of scientific investigations. These subjective elements, if made explicit and treated with care, should not undermine the results of the investigation. More importantly, it should be regarded as a good scientific practice for investigators to make their subjective inputs explicit. A truly objective evaluation of any procedure is based on how well it attains the goals of the original scientific problem.

Although it is worthwhile to think of prescribing 'objective' priors (usually with the adjective *'noninformative'*), such choices are usually unattainable in practice. We advocate the use of sensitivity analysis, that is, an analysis of how the inferential statements vary for a reasonable range of prior distributions, to validate the conclusion of a Bayesian analysis.

3.2.3 Model Selection and Bayes Evidence

Classical hypothesis testing can be seen as a model selection procedure in which one chooses between the null and the alternative hypotheses based on the degree of 'surprise' of the observed data. In contrast, Bayesian model selection can be achieved in a coherent

probabilistic framework. First, all the candidate models are embedded into an aggregate model. Second, the posterior probability of each candidate model is computed and used to discriminate or combine the models (Kass and Raftery, 1995).

Consider the situation where there are K competing models. Let $M = m$ indicate the mth model, where $m = 1, \ldots K$. We first write down the full joint distribution with all the models: $P(\mathbf{y}, \boldsymbol{\theta}, M) = P(\mathbf{y} \mid \boldsymbol{\theta}, M) P(\boldsymbol{\theta}, M)$. Since each model may have its own set of parameters, we rewrite the earlier equation as

$$P(\mathbf{y}, \boldsymbol{\theta}, M) = P(\mathbf{y} \mid \boldsymbol{\theta}_m) P(\boldsymbol{\theta}_m \mid M = m) P(M = m),$$

where $P(\boldsymbol{\theta}_m \mid M = m)$ is the prior distribution for the parameters in model m, and $P(M = m)$ is the prior probability of model m. Note that the dimensionality of $\boldsymbol{\theta}_m$ can be different for different m. The posterior probability of model m is simply

$$P(M = m \mid \mathbf{y}) \propto P(\mathbf{y} \mid M = m) P(M = m)$$
$$= \left\{ \int P(\mathbf{y} \mid \boldsymbol{\theta}_m) P(\boldsymbol{\theta}_m \mid M = m) \, \mathrm{d}\boldsymbol{\theta}_m \right\} P(M = m).$$

Sometimes one may not want to select and use a single model. Then the foregoing Bayesian formulation can be used to conduct 'model averaging' (Kass and Raftery, 1995). The model prior $P(M = m)$ is determined independently of the data in study. A frequent choice is $P(M = m) = 1/K$, implying that all models are equally likely *a priori*. In many cases, however, we may want to set smaller prior probabilities for models with higher complexities.

The classic hypothesis testing problem (i.e. null versus alternative models) can be easily fitted into the foregoing framework by letting M take on two possible values. The only caveat is that, in accordance with classical conventions, a small prior probability (e.g. 0.05) for the alternative model is preferable.

3.2.4 Bayesian Computation

In many applications, computation is the main obstacle in applying either the Bayesian or other sophisticated statistical methods. In fact, until recently these computations have often been so difficult that sophisticated statistical modeling and Bayesian methods were largely used by theoreticians and philosophers. The introduction of the bootstrap method (Efron, 1979), the EM algorithm (Dempster *et al.*, 1977), and the (MCMC) methods (Gilks *et al.*, 1998; Liu, 2001) has brought many powerful models into the mainstream of statistical analysis. As illustrated in later sections, by appealing to the rich history of computation in bioinformatics, many required optimizations and integrations can be done exactly, which gives rise to either the exact maximum likelihood estimation (MLE) or the exact posterior distribution, or a better approximation to both.

MCMC refers to a class of algorithms for simulating random variables from a distribution, $\pi(\mathbf{x})$, known up to a normalizing constant. These algorithms are well suited for Bayesian analyses since Bayesian inference can be trivially constructed if we can draw random samples from the posterior distribution (3.1) without computing its denominator. The basic idea behind all MCMC algorithms is the simulation of a Markov chain whose equilibrium distribution is the target distribution $\pi(\mathbf{x})$. Two main strategies for constructing such chains are the Metropolis–Hastings (M-H) algorithm and

the Gibbs sampler (Liu, 2001), both being widely used in diverse fields and reviewed in the Appendix. More discussions on MCMC and its fascinating applications can be found in **Chapters 15** and **26** by Huelsenbeck & Bollback, and Stephens respectively.

3.3 HIDDEN MARKOV MODEL: A GENERAL INTRODUCTION

A sequence of random variables $h_1, h_2, \ldots,$ is said to follow a lth order *Markov chain* if

$$P(h_i \mid h_1, \ldots, h_{i-1}) = P(h_i \mid h_{i-1}, \ldots h_{i-l}).$$

For example, one may assume that an observed sequence of nucleotide bases forms a first-order Markov chain with transition probabilities $P(h_{i+1} \mid h_i) = \theta_{h_i, h_{i+1}}$. With an observed realization of the chain (e.g. a segment of DNA sequence), we can obtain the MLEs of the θ by counting the frequencies of dimer occurrences.

It has long been known that the simple independent model is insufficient to describe genomic sequences. In coding regions, every nonoverlapping triplet of nucleotides codes for one of the 20 amino acid residues or a stop signal. Thus, a second-order Markov chain is perhaps desirable. However, correlation between the neighboring bases is still highly significant in noncoding regions. Some recent studies (Liu *et al.*, 2001; 2002; Huang *et al.*, 2004) show that using a second-order or a third-order Markov chain to model the promoter regions (hundreds to thousands bases upstream of the starting codon of the gene) can significantly improve the accuracy of gene regulatory binding motif discovery. It is clear that the genome is much more complex than a third-order Markov chain. Although it is desirable to model the genome sequences by even higher-order Markov chains, the number of unknown parameters increases so fast that a large amount of sequence data is required. Additionally, even high-order Markov models cannot capture certain local structures, regularities, and inhomogeneities of DNA sequences. Researchers found that it is often more suitable to use HMMs to capture various sequence features.

The HMM was initially introduced in the late 1960s and has been widely used in signal processing, speech recognition, and time series analysis (Rabiner, 1989). The method was first applied to model DNA sequences by Churchill (1989) and has become very popular in early 1990s owing to the pioneering work of Krogh *et al.* (1994a) and Baldi *et al.* (1994) on MSAs. The basic form of an HMM can be written as

$$r_i \sim f_i(r \mid h_i, \theta); \qquad h_i \sim g_i(h \mid h_{i-1}, \tau),$$

where f_i and g_i are probability distributions, θ and τ are parameters, and the r_i are observations. The h_i form a Markov chain and are often unobservable (i.e. hidden). What is of interest is the inference of θ, τ, and perhaps the h_i.

To be specific, let us examine an HMM that can accommodate compositional heterogeneity in DNA sequences. In particular, consider a sequence R consisting of two types of segments, each represented by a nucleotide frequency vector. We can only observe R and are interested in making inference on the locations of the segment change points and the composition parameters for each type of segment. A simple HMM first proposed by Churchill (1989) is shown in Figure 3.1.

In this model, we assume that the hidden layer $\mathbf{h} = (h_0, h_1, \ldots, h_n)$ is a Markov chain. Each h_i takes on only two possible values: $h_i = 0$ implies that residue $r_i \sim \text{Multinom}(\boldsymbol{\theta}_0)$;
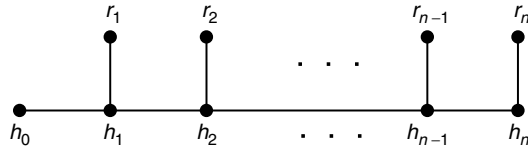
Figure 3.1   A graphical illustration of the hidden Markov model.

and $h_i = 1$ indicates that $r_i \sim \text{Multinom}(\boldsymbol{\theta}_1)$. Here $\boldsymbol{\theta}_k = (\theta_{ka}, \theta_{kc}, \theta_{kg}, \theta_{kt})$. A 2×2 transition matrix, $\boldsymbol{\tau} = (\tau_{kl})$, where $\tau_{kl} = P(h_i = k \to h_{i+1} = l)$, dictates the generation of $\mathbf{h}$. A similar model has been developed by Krogh *et al.* (1994b) to predict protein coding regions in the *E. Coli* genome.

Let $\Theta = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\tau})$. The likelihood function of Θ can be written as

$$L(\Theta \mid R) = \sum_{\mathbf{h}} P(R \mid \mathbf{h}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) P(\mathbf{h} \mid \boldsymbol{\tau}), = \sum_{\mathbf{h}} p_0(h_0) \prod_{i=1}^{n} (\theta_{h_i r_i} \tau_{h_{i-1} h_i}),$$

where h_0 is assumed to follow a known distribution $p_0(h_0)$. This function can be evaluated using a recursive summation method shown later in (3.3). With a prior distribution $f_0(\Theta)$, which may be a product of three independent Dirichlet distributions, we can write down the joint posterior distribution of all unknowns:

$$P(\Theta, \mathbf{h} \mid R) \propto P(R \mid \mathbf{h}, \Theta) P(\mathbf{h} \mid \Theta) f_0(\Theta).$$

In order to sample from this distribution, we can implement a data augmentation method (Tanner and Wong, 1987, a special Gibbs sampler), which iterates the following steps:

- *Path imputation*. Draw $\mathbf{h}^{(t+1)} \sim P(\mathbf{h} \mid R, \Theta^{(t)})$;

- *Posterior sampling*. Draw $\Theta^{(t+1)} \sim P(\Theta \mid R, \mathbf{h}^{(t+1)})$.

The path imputation step samples a hidden chain, or path, $\mathbf{h}$, from its *posterior* distribution with given parameter value, and this task can be achieved by a recursive method. More precisely, we note that

$$P(\mathbf{h} \mid R, \Theta) = c \; P(\mathbf{h}, R \mid \Theta) = c \; P(R \mid \mathbf{h}, \Theta) p(\mathbf{h} \mid \Theta)$$

$$= c \; p_0(h_0) \prod_{i=1}^{n} \left\{ P(r_i|h_i) P(h_i|h_{i-1}) \right\} = c \; p_0(h_0) \prod_{i=1}^{n} \left(\theta_{h_i r_i} \tau_{h_{i-1} h_i} \right),$$

where $c^{-1} = \sum_{\mathbf{h}} \left\{ p_0(h_0) \prod_{i=1}^{n} \left(\theta_{h_i r_i} \tau_{h_{i-1} h_i} \right) \right\}$. Define $F_0(h) = p_0(h)$ and then, recursively, let

$$F_{k+1}(h) = \sum_{h_i=0}^{1} \{ F_k(h_i) \tau_{h_i h} \theta_{h r_{i+1}} \}, \quad \text{for } i = 1, \ldots, n. \tag{3.3}$$

At the end of the recursion, we have $c^{-1} = F_n(0) + F_n(1)$ and

$$P(h_n \mid R, \Theta) = \frac{F_n(h_n)}{F_n(0) + F_n(1)}. \tag{3.4}$$

In order to sample $\mathbf{h}$ from $P(\mathbf{h} \mid R, \Theta)$, we first draw h_n from distribution (3.4) and then draw h_i recursively backward from distribution

$$P(h_i \mid h_{i+1}, R, \Theta) = \frac{F_k(h_i)\tau_{h_i h_{i+1}}}{F_k(0)\tau_{0 h_{k+1}} + F_k(1)\tau_{1 h_{k+1}}}. \qquad (3.5)$$

The posterior sampling step in data augmentation involves only the sampling from appropriate Dirichlet distributions. For example, $\boldsymbol{\theta}_0$ should be drawn from Dirichlet $(n_{0a} + \alpha_a, \ldots, n_{0t} + \alpha_t)$, where, n_{0a}, say, is the counts of the r_i whose type is A and whose hidden state h_i is zero. It is straightforward to extend this model to a k-state HMM so as to analyze a sequence with regions of k different compositional types.

3.4 PAIRWISE ALIGNMENT OF BIOLOGICAL SEQUENCES

Ever since the creation of GenBank, which grew from 680 338 base pairs in 1982 to 22 billion base pairs in 2002 (Benson *et al.*, 2002), and other related databases, sequence comparisons and sequence database search have played a pivotal role in contemporary biological research. By observing sequence similarities between a new target gene (or a segment of it) and some well-studied ones, biologists can gain important insights into its function and structure (see **Chapter 2** for more biological importance). The two most well-known 'rigorous' pairwise alignment methods are the method of Needleman and Wunsch (1970) for global alignment and that of Smith and Waterman (1981) for local alignment. They are based on dynamic programming and are guaranteed to find the global optimum of certain alignment scoring functions. Two most popular 'heuristic' pairwise alignment algorithms are BLAST (Altschul *et al.*, 1990) and FASTA (Pearson and Lipman, 1988), and both are an order of magnitude faster than the rigorous ones. **Chapter 2** provides some detailed discussions on these algorithms and issues related to the statistical significance of the resulting scores. This section discusses a Bayesian version of the Needleman–Wunsch algorithm and a motif-based Bayesian pairwise alignment method. Throughout the section, the observed data consist of two DNA or protein sequences, $R^{(1)} = (r_1^{(1)}, r_2^{(1)}, \ldots, r_{n_1}^{(1)})$ and $R^{(2)} = (r_1^{(2)}, r_2^{(2)}, \ldots, r_{n_2}^{(2)})$.

3.4.1 Bayesian Pairwise Alignment

Since traditional pairwise alignment methods are not based on any statistical models, it is difficult to judge whether they have incorporated all the relevant information, and, more importantly, whether the parameters have been tuned optimally. The use of statistical models and Bayesian methodology can help in these aspects. A Bayesian pairwise alignment method starts with a model, $P(R^{(1)}, R^{(2)} \mid \Theta, \boldsymbol{\tau})$ that describes the relationship between two sequences. To be concrete, we let Θ be a 20×20 joint symmetric probability matrix analogous to a scoring matrix such as the PAM and BLOSUM, describing the joint distribution of a pair of aligned amino acids. For example, according to BLOSUM62, the joint probability $\theta_{r_1 r_2}$ of $r_1 = \text{I}$ (isoleucine) occurring in sequence 1 and $r_2 = \text{L}$ (leucine) in the same position of sequence 2 is about 4 times the product of their respective marginal frequencies, $\theta_{r_1} \times \theta_{r_2}$, where θ_{r_1} is the sum of the entries in the row for r_1 (or the column for r_1, due to symmetry) of Θ. Generally, the (i, j)th entry of a BLOSUMx matrix stores

$2 \log_2 \frac{\theta_{i,j}}{\theta_i \theta_j}$ for amino acid pair (i, j). The number 'x' reflects that these joint frequencies are estimated from the set of protein sequences among which no pair has more than x % alignment positions with identical residues (for more details, see **Chapter 2**).

Conceptually, the alignment of $R^{(1)}$ and $R^{(2)}$ is characterized by an alignment matrix A, where element $a_{i,j}$ are set to 1 if residue i of sequence 1 'aligns' with residue j of sequence 2 and 0 otherwise. A restriction is that the aligned residues have to be 'collinear', that is, if $r_{i_1}^{(1)}$ is aligned with $r_{j_1}^{(2)}$ and $r_{i_2}^{(1)}$ with $r_{j_2}^{(2)}$, then $(i_1 - i_2)(j_1 - j_2) > 0$.

3.4.1.1 *Gap-based Global Alignment*

Let $\Lambda = (\lambda_o, \lambda_e)$ be probabilities of gap opening and gap extension, respectively, which govern the formation of the alignment matrix A. Here we show how the global alignment problem may be treated in a Bayesian way by using the statistical models pioneered by Thorne *et al.* (1991). First, the joint distribution is defined as

$$P(R^{(1)}, R^{(2)}, A, \Theta, \Lambda) = P(R^{(1)}, R^{(2)} \mid A, \Theta)P(\Theta) \ P(A \mid \Lambda)P(\Lambda).$$

Traditional alignment procedures can be seen as optimizing an objective function that is analogous to a log-likelihood (Holmes and Durbin, 1998). More precisely, for a set of fixed values $\Theta = \Theta^0$ and $\Lambda = \Lambda^0$, one finds A^* so that

$$\log P(R^{(1)}, R^{(2)}, A^* \mid \Theta^0, \Lambda^0) = \max_A \{\log P(R^{(1)}, R^{(2)} \mid A, \Theta^0) + \log P(A \mid \Lambda^0)\}.$$
$$(3.6)$$

The need for setting parameter values Θ^0 and Λ^0 has been the subject of much discussion in the field of bioinformatics. A distinctive advantage of the Bayesian procedure is its added modeling flexibility in the specification of parameters.

Let A be the alignment indicator matrix, which can also be seen as a *'path'* in a dynamic-programming setting. With given $\Lambda = (\lambda_o, \lambda_e)$, the probability of any allowable path, prior to seeing the content of the two sequences to be aligned, but conditional on their lengths n_1 and n_2, is

$$P(A \mid \lambda_o, \lambda_e) = \frac{\lambda_o^{k_g(A)} \lambda_e^{l_g(A) - k_g(A)}}{\sum_{A'} \lambda_o^{k_g(A')} \lambda_e^{l_g(A') - k_g(A')}} \tag{3.7}$$

where $k_g(A)$ and $l_g(A)$ are the total number and the total length of the gaps in A, respectively. The summation in the denominator is over all possible alignments of the two sequences. In the derivation given in subsequent text, we condition on the length information, n_1 and n_2.

If we let Θ take values in a series of BLOSUMx matrices (after log-odds transformations), then we can compute the marginal likelihood of Λ as

$$P(R^{(1)}, R^{(2)} \mid \Lambda) = \sum_{\Theta} \sum_A P(R^{(1)}, R^{(2)} \mid A, \Theta)P(A \mid \Lambda)P(\Theta),$$

$$= \frac{\sum_{\Theta} \sum_A P(R^{(1)}, R^{(2)} \mid A, \Theta)P(\Theta)\lambda_o^{k_g(A)} \lambda_e^{l_g(A) - k_g(A)}}{\sum_{A'} \lambda_o^{k_g(A')} \lambda_e^{l_g(A') - k_g(A')}} \tag{3.8}$$

In the numerator, Θ is marginalized out by summing over all the scoring matrices in a given set, each with a prior 'weight' $P(\Theta)$. Both the numerator and the denominator of (3.7) can be computed via a recursive algorithm similar to the Needleman–Wunsch algorithm (Liu and Lawrence, 1999).

As with traditional dynamic-programming alignment algorithms (see Figure 3.3(b), for example), we can describe a path as consecutive moves of three types: $\rightarrow$ (deletion), $\downarrow$ (insertion), and $\searrow$ (match). To ensure uniqueness, one often adds the restriction that an insertion cannot follow a deletion. For example, to obtain the numerator of (3.7), we start with $p(0, 0) = p_\mathrm{m}(0, 0) = 1$, $p_\mathrm{i}(0, 0) = p_\mathrm{d}(0, 0) = 0$, and compute recursively:

$$p_\mathrm{m}(k, l) = p(k - 1, l - 1)\theta_{r_k^{(1)} r_l^{(2)}},$$

$$p_\mathrm{i}(k, l) = \{\lambda_\mathrm{e} p_\mathrm{i}(k - 1, l) + \lambda_\mathrm{o} p_\mathrm{m}(k - 1, l)\}\theta_{r_k^{(1)}},$$

$$p_\mathrm{d}(k, l) = \{\lambda_\mathrm{e} p_\mathrm{d}(k, l - 1) + \lambda_\mathrm{o} p_\mathrm{m}(k, l - 1) + \lambda_\mathrm{o} p_\mathrm{i}(k, l - 1)\}\theta_{r_l^{(2)}},$$

$$p(k, l) = p_\mathrm{m}(k, l) + p_\mathrm{i}(k, l) + p_\mathrm{d}(k, l).$$

In the equations, $p_\mathrm{m}(k, l)$ is the score of entry (k, l) when the last move is a 'match'; $p_\mathrm{i}(k, l)$ is the score when the last move is an 'insertion' for sequence 1; and $p_\mathrm{d}(k, l)$ is the score when the last move is a 'deletion' for sequence 1. Thus, $P(R^{(1)}, R^{(2)} \mid \Lambda, \Theta) = p(n_1, n_2)$, and the posterior distribution of Λ is

$$P(\Lambda \mid R^{(1)}, R^{(2)}) \propto P(R^{(1)}, R^{(2)} \mid \Lambda) f_0(\Lambda).$$

This Bayesian approach provides us with not only the maximum *a posteriori* alignment of the two sequences, but also a *distribution* of the pairwise alignments, which can be represented by a random sample from the posterior alignment distribution and used to measure uncertainty in the alignment result. A Bayesian version of the Smith–Waterman algorithm is given in Webb *et al.* (2002). They showed that the new method outperformed the optimally tuned Smith–Waterman algorithm in detecting distantly related proteins.

3.4.1.2 Motif-based Local Alignment

While the gap-based approaches have dominated alignment methods for many years, Bayesian statistics opens up new directions in dealing with insertions and deletions in alignments. Zhu *et al.* (1998) attacked the Bayesian alignment problem by directly specifying a prior alignment distribution: all alignments with k gaps are equally likely, and all k in the given range are equally likely. This prior penalizes an alignment with many gaps by a factor inversely proportional to the number of that type of alignment. The summation over all alignments is carried out by the dynamic-programming method of Sankoff (1972). Input requirements for the scoring matrices are more flexible in the Bayesian setting than in traditional methods. For example, Zhu *et al.* (1998) examine the use of a series of either the PAM or the BLOSUM matrices as prior input in which all the matrices are assigned equal probability *a priori*. They report that the posterior distribution of the scoring matrices is often flat and sometimes multimodal, indicating that no one matrix is clearly more preferable to others when aligning two sequences.

Consider the expression for the posterior distribution of the scoring matrix:

$$P(\Theta \mid R^{(1)}, R^{(2)}) = \frac{1}{P(R)} \sum_k \sum_A P(R^{(1)}, R^{(2)} \mid A, \Theta) P(\Theta) P(A \mid \Lambda = k) P(\Lambda = k),$$

where Θ indicates one of the series of scoring matrices (PAM or BLOSUM), reflecting to a certain extent the distance between the two sequences, and Λ denotes the number of gaps allowed in the alignment. Since this posterior is obtained by averaging over all alignments, a 'good' alignment is not required for assessing the distance between the two sequences. This feature may be of value in distance-based methods employed in molecular evolutionary studies because the requirement that a pair of sequences must be sufficiently close to permit a good alignment is removed.

3.5   MULTIPLE SEQUENCE ALIGNMENT

Although pairwise alignment methods have been tremendously successful in modern biological researches, these methods treat all the positions of the query sequence as equally important, whereas in biological reality many 'unimportant' regions of a protein can tolerate severe distortions and are not well conserved even for those within a specialized protein family. A more attractive approach for detecting remotely related proteins is to use features common to a set of related proteins to perform the search. These common features can be most effectively represented by a position-specific consensus 'profile' extracted by a comparative study of all the protein sequences in consideration, that is, by the MSA.

Currently, the most widely used MSA method is ClustalW (Thompson *et al.*, 1994a). At first, the algorithm compares all pairs of sequences in the set using a dynamic-programming algorithm (similar to Smith–Waterman). The resulting pairwise comparison scores are transformed into evolutionary distances (Kimura, 1985). Then, a phylogenetic tree is constructed using the neighbor-joining method (Saitou and Nei, 1987). Following the branching order of the built phylogenetic tree, ClustalW progressively performs pairwise alignments of sequences or the consensus profile matrix at each node. To improve the quality of the alignments, ClustalW uses a sequence weighting strategy to avoid overrepresentation of closely related homologs (Thompson *et al.*, 1994b). At the alignment stage, a variety of substitution matrices (chosen depending on the closeness of the relationship between sequences/profiles to be aligned) is used and position-specific gap penalties are incorporated.

Another widely used method for constructing MSA is PSI-BLAST (Altschul *et al.*, 1997). It first uses BLAST (Altschul *et al.*, 1990) to collect all the sequences in the searched database that are significantly related to the query sequence. Then an MSA is created by 'anchoring' selected pairwise alignments to the query sequence (to avoid overrepresentation of closely related proteins, sequences with more than 98 % identity are thrown out). On the basis of the weighted counts of the residues in columns of the MSA (the weighting procedure of Henikoff and Henikoff (1992) is used), a position-specific profile is constructed. Using a modification of BLAST, another iteration of database search is performed using the constructed profile and the next order relatives are collected. They are, in turn, anchored to the current MSA, and used to update the profile. The procedure is repeated until no more relatives can be found from the database.

3.5.1   The Rationale of Using HMM for Sequence Alignment

The heuristics used in both PSI-BLAST and ClustalW incorporate a significant amount of biological knowledge. However, both methods lack principled ways to synthesize more delicate biological information and to tune parameters. In addition, the MSAs resulting from these methods tend to be heavily influenced by either the query sequence or the sequence order. The introduction of the profile HMM for the sequence alignment problem in early 1990s (Krogh *et al.*, 1994a; Baldi *et al.*, 1994) revolutionized the field and provided the scientist fresh ways of looking at many biological problems including MSA.

In the evolution of protein sequences, segment transpositions are rare, so we can safely assume in most cases that the discrepancy between two homologous sequences is caused by insertions/deletions (indels) and point mutations. Thus, although the sequences are misaligned via indels, conserved residues remain in order. By capturing this characteristic, the HMM not only captures an important feature of protein evolution, but also results in an effective algorithm.

As shown in Figure 3.2, in the HMM framework, one treats the sequences to be aligned as iid observations from a probabilistic mechanism (i.e. 'insert', 'delete', and 'match') that perturbs a hypothetical common 'ancestral' model sequence (called *model*), denoted as $M = (\mathbf{m}_1, \ldots, \mathbf{m}_L)$. The 'match' state implies that the current residue at this position has evolved from (i.e. corresponds to) the 'ancestral' residue via mutations, not indels. Note that it does not mean that the residue matches with the ancestral ones. Two major distinctions between this HMM framework and the standard evolutionary model are noteworthy: (a) since the observed sequences are iid given the ancestral model, this HMM does not capture the important tree structure of a realistic evolutionary process, which is often essential to reflect correlations among the observed sequences; (b) the ancestral model is not meant to be an ancestral sequence. In other words, each model position $\mathbf{m}_j$ is not meant to recover what the ancestral residue most likely is, but is used to model this position's residue preference, which results from both the evolutionary force and functional constraints (selection). With principled statistical inference methods, one can estimate optimally all the parameters in this model, and consequently in the ancestral profile model, based on the observed sequences.

The diagram in Figure 3.3(a) describes the generative procedure for the underlying (hidden) Markov chain. In this model, each $\mathbf{m}_l$ is regarded as an abstract residue and is associated with an 'emission' probability vector $\boldsymbol{\theta}_l$ of length p ($p = 4$ for DNA sequences, and $p = 20$ for proteins). For simplicity, we assume that all 'insert' states are associated with a common 'emission' probability $\boldsymbol{\theta}_0$. When generating biological sequences, the
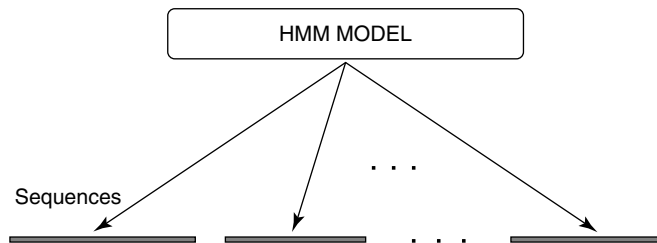
Figure 3.2   Independent generation of protein sequences from an HMM.

types of perturbations allowed are point mutations, insertions, and deletions. A residue in an observed sequence is generated either by a 'match' state, $\mathbf{m}_j$ say, or by an 'insert' state, $\mathbf{i}_k$, say. Another set of transition probabilities, one associated with each arrow in Figure 3.3(a), for example, $\tau_{\mathbf{m}_j\mathbf{m}_{j+1}}$, $\tau_{\mathbf{d}_j\mathbf{m}_{j+1}}$, $\tau_{\mathbf{i}_j\mathbf{i}_j}$, and so on, are needed to describe the underlying Markov chain. Note that probabilities coming out of a node have to sum up to 1, and they can be either given in advance or estimated ('trained') from many unaligned sequences.

Starting from the 'Begin' state, a protein sequence is 'produced' from the profile HMM as follows. According to the transition probabilities, a series of hidden states is generated until the 'End' state is reached. In turn, each of the 'match' and 'insert' states generates an amino acid based on its state specific 'emission' probability distribution. Figure 3.4 shows how the sequence $R = (r_1, r_2, \ldots, r_6) \equiv \texttt{VLSFFD}$ is generated by a profile HMM with three match states.
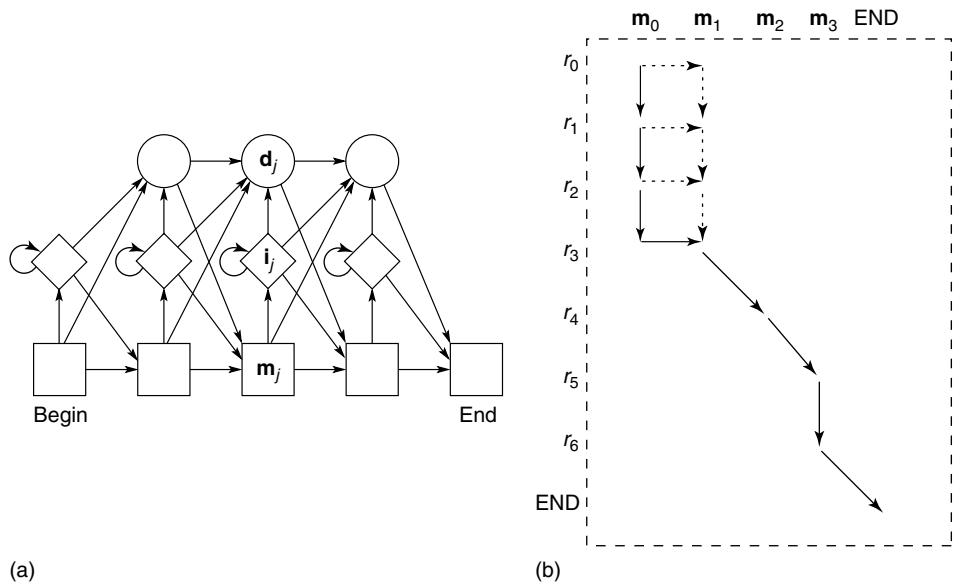
Figure 3.3 (a) A profile HMM model for sequence alignment. 'Match', 'insert', and 'delete' states are represented by squares, diamonds, and circles, respectively. (b) A 'Path' representing the alignment of a sequence to a profile HMM.
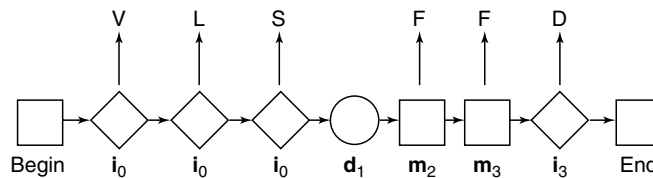
Figure 3.4 A toy sequence of length 6 generated by a profile HMM with 3 match states. Note that $\mathbf{m}_1$ is replaced by state d_1, implying that first match state is deleted in generating the observed sequence.

One can also think of the process of generating sequence R from the profile HMM as that of choosing a path through an $(n + 1) \times (L + 1)$ table starting from the upper left corner and ending with the lower right corner (Figure 3.3b), very much similar to the diagram used to illustrate the Smith–Waterman and Needleman–Wunsch algorithms. The columns for this table are denoted by $\mathbf{m}_0, \ldots, \mathbf{m}_L$, which correspond to a void start position and L model positions. The rows, $r_0, \ldots, r_n$, correspond to a void starting residue and the n observed sequence residues. At any position (k, j) of this table, the next step allowed is (a) position $(k, j + 1)$, implying that a deletion of model position $\mathbf{m}_{j+1}$ has occurred ($\rightarrow$); (b) position $(k + 1, j)$, implying that r_{k+1} has occurred ($\downarrow$); or, lastly, (c) position $(k + 1, j + 1)$, implying that r_{k+1} is generated by model position $\mathbf{m}_{j+1}$ ($\searrow$). Thus the path depicted by the solid arrows in Figure 3.3(b) corresponds to the following probabilistic transitions:

$$\mathbf{m}_0 \rightarrow \mathbf{i}_0 \rightarrow \mathbf{i}_0 \rightarrow \mathbf{i}_0 \rightarrow \mathbf{d}_1 \rightarrow \mathbf{m}_2 \rightarrow \mathbf{m}_3 \rightarrow \mathbf{i}_3 \rightarrow \text{END},$$

which is exactly the same as in Figure 3.4.

It is worthwhile to note that the hidden states for this HMM are NOT the $\mathbf{m}_l$'s, but the allowable 'paths' that traverse the $(n + 1) \times (L + 1)$ table. Liu *et al.* (1999) provided a slightly different formulation of this model so as to make it conform with the conventional HMM shown in Figure 3.1. In other words, they constructed a hidden chain, $h_0 \rightarrow h_1 \rightarrow \cdots \rightarrow h_n$, that generates the observed sequence R, where each h_i records the number of deletions that have occurred till residue r_i and whether r_i is generated by an insert or a match state.

3.5.2 Bayesian Estimation of HMM Parameters

Let $\mathbf{R} = (R_1, \ldots, R_m)$ be the set of sequences to be aligned, let Θ denote the collection of all parameters including the transition probabilities $\boldsymbol{\tau}$ and the emission probabilities $\boldsymbol{\theta}_j$, and let $\mathcal{A} = (A_1, \ldots, A_m)$ denote the unobserved alignment variable (i.e. for each of the sequences an alignment path as shown in Figure 3.3(b)). For simplicity, we assume that the emission probabilities from the 'insert' state are the same for all positions and denoted as $\boldsymbol{\theta}_0$. It is observed that, once the sequences are aligned (i.e. given $\mathcal{A}$), the transition probabilities, $\tau_{k,l} = P(s_k \rightarrow s_l)$ (from any state s_k to any state s_l), model emission probabilities, $\boldsymbol{\theta}_j(r) = P(\text{residue } r | s_k = \mathbf{m}_j)$, and insertion emission probabilities $\boldsymbol{\theta}_0$ are easy to estimate:

$$\tau_{k,l} = \frac{T_{k,l}}{\sum_{\text{states } l'} T_{k,l'}},$$

$$\boldsymbol{\theta}_j(r) = \frac{E_j(r)}{\sum_{\text{residues } r'} E_j(r')},$$

where $T_{k,l}$ is the number of transitions from state s_k to s_l, and $E_j(r)$ is the number of residues of type r emitted from match state $\mathbf{m}_j$ in the sequence alignment. With Dirichlet priors (or Dirichlet Mixture priors, Sjolander *et al.*, 1996), a Bayesian estimate of these parameters are also easy. On the other hand, once the parameters of the HMM are given, it is also feasible to find either the optimal or the average alignment of each sequence to the model by using a dynamic-programming technique shown earlier (i.e. finding the optimal path in Figure 3.3(b)). On the basis of these insights, Krogh *et al.* (1994a) modified a

Baum–Welch algorithm (an earlier version of the EM algorithm, Baum, 1972) for training the profile HMM. Baldi *et al.* (1994) developed a gradient descent method (Baldi *et al.*, 1994) for finding the MLE of Θ.

The foregoing heuristics that it is 'easy' to estimate Θ given $\mathcal{A}$ and to handle $\mathcal{A}$ given Θ also facilitate a Bayesian MCMC approach (Churchill and Lazareva, 1999). The general procedure is outlined in subsequent text.

- Initialization. Choose model length L (the number of 'match' states) and set initial parameter values. In the absence of prior knowledge, one may let L be the average length of all the sequences.

- Data Augmentation. Starting with $\Theta^{(0)}$, we proceed for $t = 1, \ldots, N$:

  - Alignment imputation. Sample $\mathcal{A}^{(t+1)}$ from $P(\mathcal{A} \mid \mathbf{R}, \Theta^{(t)})$.

  - Parameter simulation. Draw $\Theta^{(t+1)}$ from $P(\Theta \mid \mathbf{R}, \mathcal{A}^{(t+1)})$.

  The number of iterations N may be determined dynamically on the basis of certain convergence criteria (Liu, 2001).

- The final alignment model can be estimated in one of the following ways.

  - Use the average of the $\Theta^{(t)}$ in the last K iterations ($K = 500$, say).

  - Estimate Θ by its posterior mode.

  - One can also find an optimal alignment $\mathcal{A}^*$, and then estimate Θ as $E(\Theta \mid \mathbf{R}, \mathcal{A}^*)$.

Figure 3.5 shows a simple alignment of three sequences, from which the HMM parameters can be estimated.

In most applications, the number of match states L is not known and it is desirable to let the data speak for itself. In some available packages for HMM-based MSA, L is updated iteratively using heuristic rules. For example, once an alignment such as the one shown in Figure 3.5 is produced, one can either remove certain 'match' positions if the corresponding column contains more than $x\,\%$ deletions or add a new 'match' position if more than $y\,\%$ of the sequences have insertions of similar types at the same place. It is also possible to treat L as a new parameter and infer the size of HMM in a coherent Bayesian framework. Operationally, one can insert between the alignment imputation and
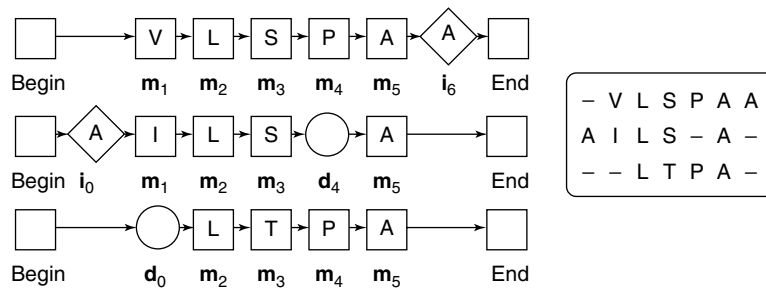
Figure 3.5   The alignment of three sequences to the profile HMM.

parameter simulation steps of data augmentation a 'match' state updating step to either add or delete a match column using the M-H rule (see the Appendix).

A key step in implementing the data augmentation procedure is the computation of the likelihood:

$$P(\mathbf{R} \mid \mathbf{\Theta}) = \sum_{\mathcal{A}} P(\mathbf{R} \mid \mathbf{\Theta}, \mathcal{A}) P(\mathcal{A} \mid \mathbf{\Theta}),$$

because this is needed in imputing $\mathcal{A}$ from $P(\mathcal{A} \mid \mathbf{R}, \mathbf{\Theta}) = P(\mathcal{A}, \mathbf{R} \mid \mathbf{\Theta})/P(\mathbf{R} \mid \mathbf{\Theta})$. Dynamic-programming recursions can be applied to complete the task with complexity $O(nmL)$, where m is the number of sequences, n is their average length, and L is the number of match states. Since all the sequences in consideration are mutually independent conditional on $\mathbf{\Theta}$, we may focus only on one sequence, that is, the computation of $P(R \mid \mathbf{\Theta})$. The imputation of A for this sequence can be achieved by the following two-stage algorithm, which is analogous to the forward–backward algorithm in Section 3.3.

Forward summation

- Initialization with $f_m(0, 0) = 1$ and $f_i(0, 0) = f_d(0, 0) = 0$.

- Recursion:

$$f_{\mathrm{m}}(j + 1, k + 1) = \boldsymbol{\theta}_{\mathrm{k}}(j + 1)[f_{\mathrm{m}}(j, k)\tau_{\mathbf{m}_k \mathbf{m}_{k+1}} + f_{\mathrm{i}}(j, k)\tau_{\mathbf{i}_k \mathbf{m}_{k+1}} + f_{\mathrm{d}}(j, k)\tau_{\mathbf{d}_k \mathbf{m}_{k+1}}];$$

$$f_{\mathrm{i}}(j + 1, k + 1) = \boldsymbol{\theta}_0(j + 1)[f_{\mathrm{m}}(j, k + 1)\tau_{\mathbf{m}_{k+1} \mathbf{i}_{k+1}} + f_{\mathrm{i}}(j, k + 1)\tau\mathbf{i}_{k+1} \mathbf{i}_{k+1}];$$

$$f_{\mathrm{d}}(j + 1, k + 1) = f_{\mathrm{m}}(j + 1, k)\tau_{\mathbf{m}_k \mathbf{d}_{k+1}} + f_{\mathrm{i}}(j + 1, k)\tau_{\mathbf{i}_k \mathbf{d}_{k+1}} + f_{\mathrm{d}}(i + 1, k)\tau_{\mathbf{d}_k \mathbf{d}_{k+1}}.$$

- The algorithm terminates when the (L, n)th entry is reached.

At the end, we have $P(R \mid \mathbf{\Theta}) = f_{\mathrm{m}}(n, L) + f_{\mathrm{i}}(n, L) + f_{\mathrm{d}}(n, L)$. After the forward summation, one can carry out the backward-sampling step to impute the alignment path A. For a simple description, we let $v()$ be a backward-tracing function of the three types of states: $v(\mathbf{m}) = (-1, -1)$, $v(\mathbf{i}) = (-1, 0)$, and $v(\mathbf{d}) = (0, -1)$.

Backward sampling

- Start from position (n, L) and proceed as follows.

- Suppose we are at a position (x, y). Sample B from one of the three symbols '$\mathbf{m}$', '$\mathbf{i}$' and '$\mathbf{d}$' with probabilities proportional to $f_{\mathrm{m}}(x, y)$, $f_{\mathrm{i}}(x, y)$, and $f_{\mathrm{d}}(x, y)$ respectively; set an arrow from (x, y) to $(x, y) + v(B)$.

- Terminate when the $(0, 0)$th entry is reached. The path indicated by the collection of arrows is a sample of A from distribution $P(A \mid R, \mathbf{\Theta})$.

Several HMM-based software packages for producing multiple protein sequence alignments are available, although none of them are Bayesian. HMMER was developed by Eddy (1998) and is available at `http://hmmer.wustl.edu/`. SAM was developed by the UC Santa Cruz group and is available at `http://www.soe.ucsc.edu/projects/compbio/HMM-apps/HMM-applications.html`. It uses Dirichlet Mixture prior on amino acid emission probabilities.

3.5.3   PROBE and Beyond: Motif-based MSA Methods

Aligning distantly related sequences presents unique algorithmic and statistical challenges because often such proteins only share a minimal structural core with sizable insertions occurring between, and even within, core elements. Classical dynamic programming-based multiple alignment procedures typically have considerable difficulty spanning these insert regions because the log-odds scores associated with weakly conserved core elements are often too low to offset the substantial gap penalties that such insert regions incur. This problem is further exacerbated when core elements contain short insertions or deletions within them.

Although the standard HMM is flexible enough to model multiple related sequences, a large number (over 100) of unaligned sequences is usually required in order to train an HMM with flexible gap penalties due to the large number of parameters present in the model. In order to detect and align more distantly related proteins, one has to further constrain the profile HMM model in a biologically meaningful way. The block-motif propagation model utilized in PROBE (Neuwald *et al.*, 1997; Liu *et al.*, 1999) achieves this type of parameter reduction by focusing on the alignment composed of block motifs.

A block motif is a special HMM that allows no insertions or deletions among the match states. The propagation model consists of a number of block motifs, as shown in Figure 3.6. The gaps between blocks, corresponding to gaps in an HMM, can be modeled by a probability distribution other than the geometric one implied by the standard HMM. PROBE also uses a Bayesian procedure for selecting the sizes of the blocks and the number of blocks.

The propagation model assumes that there are L conserved 'blocks' for each sequence to be aligned, and the lth block of residues is of length w_l. We can imagine that L motif elements propagate along a sequence. Insertions are reflected by gaps between adjacent blocks. No deletions are allowed, but it is possible to relax this constraint by treating each block as a mini-HMM. Again, let Θ be the collection of all parameters needed in the propagation model (e.g. the motif profile matrices, background frequencies, and parameters characterizing the distribution of gaps between motif blocks) and let the alignment variable $\mathcal{A} = (A_1, \ldots, A_L) = (a_{k,l})_{K \times L}$ be a matrix with $a_{k,l}$ indicating the starting position of the lth motif element in sequence k (Figure 3.6). Note that both Θ and $\mathcal{A}$ in this model are of much smaller dimensions compared to their counterparts in the HMM because of the block-motif constraint.

The alignment variable $\mathcal{A}$ as represented by the locations of the motif elements is not observable. But once it is known, we can write down the likelihood $P(\mathbf{R} \mid \mathcal{A}, \Theta)$ (details can be found in Liu *et al.* 1999), from which we can make Bayesian inference on Θ easily. On the other hand, once Θ were given, we could impute $\mathcal{A}$ by a draw from $P(\mathcal{A} \mid \mathbf{R}, \Theta)$, which can be achieved using a forward–backward method similar to that in Section 3.3. Thus, we can again implement a data augmentation strategy to iterate between sampling of $\mathcal{A}$ and Θ. In PROBE, a 'collapsing' technique (Liu, 1994) is employed to further improve the efficiency of the computation.

Figure 3.6   An illustration of the propagation model in Liu *et al.* (1999).

The number of motifs L and the total number of motif columns $W = w_1 + \cdots + w_L$ are treated as random variables and selected according to an approximate Bayesian criterion. PROBE is also one of the earliest algorithms implementing the iterative database search process. Similar to PSI-BLAST, PROBE first uses BLAST to get the first-order relatives of the query sequence and aligns them to produce a block-motif profile. Then, it uses the constructed profile to find more relatives and further refine the alignment. It stops when no new relatives are found. It has been shown that PROBE is a very sensitive method for detecting distantly related proteins and capturing very weak similarities (Hudak and McClure, 1999; Neuwald *et al.*, 1997). The program is available at `ftp://ftp.ncbi.nih.gov/pub/neuwald/probe1.0/`.

A serious limitation of PROBE is that it does not allow gaps within conserved blocks. Neuwald and Liu (2004) recently extended the method to allow indels in conserved blocks–previous ungapped blocks are replaced by local HMMs. They have also described a few MCMC strategies for optimizing the alignment and for adjusting position-specific amino acid frequencies and gap penalties, and implemented them in the software package 'GISMO'. The new alignment method was applied to a statistically based approach called *contrast hierarchical alignment and interaction network* (CHAIN) analysis, which infers the strengths of various categories of selective constraints from co-conserved patterns in a multiple alignment (Neuwald *et al.*, 2003). The power of this approach strongly depends on the quality of the multiple alignments.

The programs MUSCLE (Edgar, 2004a; 2004b) and MAFFT (Katoh *et al.*, 2002) are also designed to avoid alignment of nonhomologous regions, and in other respects they are generally superior to more widely used multiple alignment programs, such as ClustalW and T-coffee (Notredame *et al.*, 2000). Because MUSCLE and MAFFT can handle large data sets, Neuwald explored the use of these programs for CHAIN analysis (personal communication). Somewhat surprisingly, these failed to achieve the degree of accuracy needed to detect subtle, co-conserved patterns, such as those recently identified and structurally confirmed within P loop GTPases (Neuwald *et al.*, 2003). We found that, although these programs align regions globally conserved in the sequences well, for several large test sets they failed to accurately align regions conserved only within more closely related subsets. This is, of course, a major drawback to their general application for CHAIN analysis. By contrast, PSI-BLAST (Altschul *et al.*, 1997), which seems less likely to produce high-quality global alignments given its simple alignment procedure, nevertheless, in many cases, does a better job of aligning database sequences relative to the query.

MUSCLE and MAFFT perform well on small sets of relatively diverse representative sequences, such as the BALIBASE benchmark sets (Bahr *et al.*, 2001), because they incorporate heuristics that unfortunately can also compromise statistical rigor and, as a result, confuse random noise with biologically valid homology. Statistically, the best alignment for random sequences is the 'null alignment', that is, the procedure should leave such sequences unaligned.

3.5.4 Bayesian Progressive Alignment

Owing to its use of MCMC in profile estimation, PROBE (and its later extension) is very slow. Another limitation of PROBE is its inflexibility in introducing gaps within block motifs. As mentioned earlier, the HMM-based model can allow for flexible gaps, but it needs many sequences in order to estimate the excessive number of parameters well. In

contrast, PSI-BLAST produces multiple alignment progressively without having to iterate and can accommodate gaps in motif in a biologically meaningful fashion because of its use of BLAST (a motif-based approach). To combine the attractive features of PSI-BLAST and PROBE, Logvinenko (2002) proposed a Bayesian progressive alignment procedure based on a sequential Monte Carlo principle (Liu, 2001).

The algorithm starts by aligning the query sequence $R^{(1)}$ to each of the other sequences in a set, $R^{(2)}, \cdots, R^{(n)}$, by a gap-based Bayesian local alignment method (similar to the one described in Section 3.4.1.1). The set of sequences can be either given in advance or obtained by an iterative BLAST search as in PSI-BLAST. The sequences are then sorted according to their distances from $R^{(1)}$ derived on the basis of the pairwise alignments. A position-specific profile matrix $\boldsymbol{\Theta} = (\hat{\theta}_{dl})_{20, l_1}$ is constructed on the basis of the alignment of $R^{(1)}$ to its closest relative, $R^{(2)}$, and then to $R^{(3)}$, and so on. Each time a new sequence is introduced, the profile is updated on the basis of the new alignment. As in PSI-BLAST, sequence-weighing strategy of Henikoff and Henikoff (1994) is used in profile computation to avoid overrepresentation of closely related sequences.

In producing the progressive alignment, a Dirichlet Mixture prior distribution (Sjolander *et al.*, 1996) is used for amino acid emission probabilities:

$$P(\boldsymbol{\theta}) \sim \sum_{k=1}^{K} p_k \frac{\Gamma(\alpha_1^{(k)} + \cdots + \alpha_{20}^{(k)})}{\Gamma(\alpha_1^{(k)}) \cdots \Gamma(\alpha_{20}^{(k)})} \theta_1^{\alpha_1^{(k)}-1} \cdots \theta_{20}^{\alpha_{20}^{(k)}-1}$$

For every column l of the alignment, residue counts $\mathbf{m}^{(l)} = (m_1^{(l)}, \cdots, m_{20}^{(l)})$ follow Multinom $(\boldsymbol{\theta}_l)$ distribution. Profile entries are set to be predictive probabilities of the residues:

$$\hat{\theta}_{d,l} = \int_{\boldsymbol{\theta}_l} \theta_{d,l} P(\boldsymbol{\theta}_l \mid \mathbf{m}_l) d\boldsymbol{\theta}_l \propto \int_{\boldsymbol{\theta}_l} \theta_{d,l} P(\mathbf{m} \mid \boldsymbol{\theta}_l) P(\boldsymbol{\theta}_l) d\boldsymbol{\theta}_l.$$

As in Section 3.4.1.1 a dynamic-programming type recursion is used to compute $P(R, \boldsymbol{\Theta} \mid \Lambda) = p(n, L)$:

$$p_{\mathrm{m}}(k, l) = p(k-1, l-1)\hat{\theta}_{r_k l},$$
$$p_{\mathrm{i}}(k, l) = \{\lambda_{\mathrm{e}} p_{\mathrm{i}}(k-1, l) + \lambda_{\mathrm{o}} p_{\mathrm{m}}(k-1, l)\},$$
$$p_{\mathrm{d}}(k, l) = \{\lambda_{\mathrm{e}} p_{\mathrm{d}}(k, l-1) + \lambda_{\mathrm{o}} p_{\mathrm{m}}(k, l-1) + \lambda_{\mathrm{o}} p_{\mathrm{i}}(k, l-1)\},$$
$$p(k, l) = p_{\mathrm{m}}(k, l) + p_{\mathrm{i}}(k, l) + p_{\mathrm{d}}(k, l);$$

where $\Lambda = (\lambda_{\mathrm{o}}, \lambda_{\mathrm{e}})$ are the gap opening and extension probabilities; L is the number of positions in the profile; and $\hat{\theta}_{r_k l}$ is a profile entry corresponding to residue r_k in lth column of the alignment. A backward-tracing procedure analogous to that in **Chapter 2** is used to obtain an alignment.

After all the sequences are incorporated into an alignment, a Gibbs Sampling step is used to improve on its quality. Each sequence $R^{(i)}$ is removed in turn and realigned to the others. The profile is updated accordingly. Although this additional step makes the procedure relatively slow, the resulting alignment is less prone to be only locally optimal.

3.6 FINDING RECURRING PATTERNS IN BIOLOGICAL SEQUENCES

Our focus here is on the discovery of repetitive patterns (often termed as *motif elements*) in a given set of biopolymer sequences. The main motivation for this task is that these motif elements often correspond to the functionally or structurally important part of these molecules. For instance, repetitive patterns in the upstream regions of coregulated genes may correspond to a 'regulatory motif' to which certain regulatory protein binds so as to control gene expressions (Stormo and Hartzell, 1989; Lawrence and Reilly, 1990). Without loss of generality, we concatenate all the sequences in the dataset to form one 'supersequence' R. The multiple occurrences of a motif element in R is thus analogous to the multiple occurrences of a *word* in a long sentence. It is of interest to find out what this motif is and where it has occurred. What makes this problem challenging is that the motif occurrences are not necessarily identical to each other. In other words, there are often some 'typos' in each occurrence of the word. It is therefore rather natural for us to employ probabilistic models to handle this problem.

3.6.1 Block-motif Model with iid Background

A simple model conveying the basic idea of a motif that repeats itself with random variations is the block-motif model as shown in Figure 3.7. It was first developed by Liu *et al.* (1995) and has been employed to find subtle repetitive patterns, such as helix-turn-helix structural motifs (Neuwald *et al.*, 1995) in protein sequences and gene regulation motifs (Roth *et al.*, 1998; Liu *et al.*, 2001; 2002) in DNA sequences.

The model assumes that at unknown locations $A = (a_1, \ldots, a_K)$ there are realized instances of a motif so that the sequence segments at these locations look similar to each other. In other parts of the sequence, the residues (or base pairs) follow a 'background model' represented as iid observations from a multinomial distribution. Since the motif region is a very small fraction of the whole sequence, it is not unreasonable to assume that the background frequency $\boldsymbol{\theta}_0 = (\theta_{0,a}, \ldots, \theta_{0,t})$ is known in advance. For the motif of width w, we let $\Theta = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_w]$, where each $\boldsymbol{\theta}_j$ describes the base frequency at position j of the motif. The matrix Θ is simply the profile matrix for the motif.

To facilitate analysis, we introduce an indicator vector $\mathbf{I} = (I_1, \ldots, I_n)$, where n is the length of R. We let $\mathbf{I}_{[-i]}$ be the vector of I's without I_i. Here, $I_i = 1$ means that position i is the start of a motif pattern, and $I_i = 0$ means otherwise. We assume *a priori* each I_i has a small probability p_0 to be equal to 1. With this setup, we can write down the joint posterior distribution:

$$P(\Theta, \mathbf{I} \mid R) \propto P(R \mid \mathbf{I}, \Theta) P(\mathbf{I} \mid \Theta), f_0(\boldsymbol{\theta}), \qquad (3.9)$$

where $P(\mathbf{I} \mid \Theta) \propto \prod_{i=1}^{n} p_0^{I_i} (1 - p_0)^{1-I_i}$. If we do not allow overlapping motifs, we need to restrict that in $\mathbf{I}$ there is no pair $I_i = 1$ and $I_j = 1$ with $|i - j| < w$.

Figure 3.7 A graphical illustration of the repetitive motif model.

With a prior Dirichlet ($\boldsymbol{\alpha}$) for each $\boldsymbol{\theta}_j$, we can easily obtain the posterior distribution of the $\boldsymbol{\theta}_j$ *if* we know the positions of the motif (Liu *et al.*, 1995). Thus, a simple Gibbs sampling algorithm can be designed to draw Θ and $\mathbf{I}$ from (3.9):

- For a current realization of Θ, we update each I_i, $i = 1, \ldots, n$, by a random draw from its conditional distribution, $P(I_i \mid \mathbf{I}_{[-i]}, R, \Theta)$, where

$$\frac{P(I_i = 1 \mid \mathbf{I}_{[-i]}, R, \Theta)}{P(I_i = 0 \mid \mathbf{I}_{[-i]}, R, \Theta)} = \frac{p_0}{1 - p_0} \prod_{k=1}^{w} \left(\frac{\theta_{k, r_{i+k-1}}}{\theta_{0, r_{i+k-1}}} \right). \tag{3.10}$$

  Intuitively, this odds ratio is simply the 'signal-to-noise' ratio.

- On the basis of the current value of $\mathbf{I}$, we update the profile matrix Θ column by column. In other words, each $\boldsymbol{\theta}_j$, $j = 1, \ldots, w$, is drawn from an appropriate posterior Dirichlet distribution determined by $\mathbf{I}$ and R.

After a burn-in period (until the sampler stabilizes), we continue to run the sampler for m iterations and use the average of the sampled Θ's to estimate the profile matrix. The estimated Θ can then be used to scan the sequence to find the locations of the motif.

3.6.2 Block-motif Model with a Markovian Background

It is well known that the iid multinomial distribution cannot model a DNA or protein sequence (background) well. In particular, Liu *et al.* (2001; 2002) showed that use of a second- or third-order Markov model for the background can significantly improve the motif-finding capability of the method. For simplicity, here we describe only the first-order Markov background model, where a 4×4 transition matrix, $B_0 = (\beta_{jj'})$, needs to be estimated. Since the total number of base pairs occupied by motif sites is a very small fraction of all the base pairs in R, we may estimate B_0 from the raw data directly, assuming that the whole sequence of R is homogeneous. With this approximation, the transition probabilities can be estimated as $\hat{\beta}_{j_1 j_2} = n_{j_1 j_2} / n_{j_1 \cdot}$, similar to that in Section 3.3. We may then treat B_0 as a known parameter.

The joint posterior distribution of $(\Theta, \mathbf{I})$ in this case differs from (3.9) only in the description of the residues in the background. The Gibbs sampler can also be similarly implemented with a slight modification in $P(I_i \mid \mathbf{I}_{[-i]}, R, \Theta)$. In other words, conditional on R, we slide Θ through the whole sequence position by position to update I_i according to a random draw from $P(I_i \mid \mathbf{I}_{[-i]}, R, \Theta)$, which satisfies

$$\frac{P(I_i = 1 \mid \mathbf{I}_{[-i]}, R, \Theta)}{P(I_i = 0 \mid \mathbf{I}_{[-i]}, R, \Theta)} = \frac{p_0}{1 - p_0} \prod_{k=1}^{w} \left(\frac{\theta_{k, r_{i+k-1}}}{\hat{\beta}_{r_{i+k-2} r_{i+k-1}}} \right).$$

For given $\mathbf{I}$, we update the profile matrix Θ in the same way as in Section 3.6.1.

3.6.3 Block-motif Model with Inhomogeneous Background

It has long been noticed that DNA sequences contain regions of distinctive compositions. As discussed in Section 3.3, an HMM can be employed to delineate a sequence with k types of regions. Suppose we decide to use an HMM to model sequence inhomogeneity.

As mentioned earlier, we may estimate the background model parameters using all the sequences by the methods in Section 3.3, assuming that R does not contain any motifs. Then, we can treat these parameters as known at the estimated values and use one of the following two strategies to modify the odds ratio formula (3.10).

In the first strategy, we treat each position in the sequence as a 'probabilistic base pair' (i.e. having probabilities to be one of the four letters) and derive its frequency. In other words, we need to find $\theta_{ij}^* = p(r_i^* = j \mid R)$ for a future r_i^* and then treat residue r_i in the background as an independent observation from Multinom $(\boldsymbol{\theta}_i^*)$, with $\boldsymbol{\theta}_i^* = (\theta_{ia}^*, \ldots, \theta_{it}^*)$. But this computation is nontrivial because

$$\theta_{ij}^* = P(r_i^* = j \mid R) = \theta_{0j} P(h_i = 0 \mid R) + \theta_{1j} P(h_i = 1 \mid R), \qquad (3.11)$$

where $P(h_i)$ can be computed via a recursive procedure similar to (3.3). More precisely, in addition to the series of forward functions F_i, we can define the backward functions B_i. Let $B_n(h) = \sum_{h_n} \tau_{hh_n} \theta_{h_n r_n}$, and let

$$B_k(h) = \sum_{h_i} \left\{ \tau_{hh_i} \theta_{h_i r_i} B_{k+1}(h_i) \right\}, \quad \text{for } k = n - 1, \ldots, 1. \qquad (3.12)$$

Then, we have

$$P(h_i = 1 \mid R) = \frac{F_i(1) B_{i+1}(1)}{F_i(1) B_{i+1}(1) + F_i(0) B_{i+1}(0)}. \qquad (3.13)$$

This is the *marginal* posterior distribution of h_i and can be used to predict whether position i is in state 1 or 0. Thus, in the Gibbs sampling algorithm we only need to modify the denominator of the right-hand side of (3.10) as $\prod_{k=i}^{i+w-1} \theta_{k,r_i}^*$.

In the second strategy, we seek to obtain the joint probability of a whole segment, $R_{[i:i+w-1]} \equiv (r_i, \ldots, r_{i+w-1})$, conditional on the remaining part of the sequence. Then we modify (3.10) accordingly. Clearly, compared with the first strategy, the second one is more faithful to the compositional HMM assumption. The required probability evaluation can be achieved as follows:

$$\begin{aligned} P(R_{[i:i+w-1]} \mid R_{[1:i-1]}, R_{[i+w:n]}) &= \frac{P(R)}{P(R_{[1:i-1]}, R_{[i+w:n]})} = \frac{P(R)}{\sum_{\mathbf{h}} P(R_{[1:i-1]}, R_{[i+w:n]}, \mathbf{h})} \\ &= \frac{F_n(0) + F_n(1)}{\sum_{h_1, \ldots, h_w} F_i(h_1) \tau_{h_1 h_2} \cdots \tau_{h_{w-1} h_w} B_{i+w}(h_w)}, \qquad (3.14) \end{aligned}$$

where the denominator can also be obtained via recursions.

3.6.4 Extension to Multiple Motifs

Earlier, we have assumed that there is only one kind of motif in the sequence and the prior probability for each $I_i = 1$ is known as p_0. Both of these assumptions can be relaxed. Suppose we want to detect and align m different types of motifs of lengths $w_1, \ldots, w_m$, respectively, with each occurring unknown number of times in R. We can similarly introduce the indicator vector $\mathbf{I}$, where $I_i = j$ indicates that an element from motif j starts at position i, and $I_i = 0$ means that no element starts from position i. For simplicity, we consider only the independent background model.

Let $P(I_i = j) = \varepsilon_j$, where $\varepsilon_0 + \cdots + \varepsilon_m = 1$, is an unknown probability vector. Given what is known about the biology of the sequences being analyzed, a crude guess k_j for the number of elements for motif j is usually possible. Let $k_0 = n - k_1 - \cdots - k_m$. We can represent this prior opinion about the number of occurrences of each type of elements by a Dirichlet distribution on $\boldsymbol{\varepsilon} = (\varepsilon_0, \ldots, \varepsilon_m)$, which has the form Dirichlet $(b_0, \ldots, b_m)$ with $b_j = J_0 \frac{k_j}{n}$, where J_0 represents the 'weight' (or 'pseudo-counts') to be put on this prior belief. Then the same predictive updating approach as illustrated in Section 3.6.1 can be applied. Precisely, the update formula (3.10) for $\mathbf{I}$ is changed to

$$\frac{P(I_i = j \mid \mathbf{I}_{[-i]}, R)}{P(I_i = 0 \mid \mathbf{I}_{[-i]}, R)} = \frac{\varepsilon_j}{\varepsilon_0} \prod_{k=1}^{w_j} \left(\frac{\theta_{k,r_{i+k-1}}^{(j)}}{\theta_{0,r_{i+k-1}}} \right),$$

where $\Theta^{(j)} = [\boldsymbol{\theta}_1^{(j)}, \ldots, \boldsymbol{\theta}_{w_j}^{(j)}]$ is the profile matrix for the jth motif. Conditional on $\mathbf{I}$, we can then update $\boldsymbol{\varepsilon}$ by a random sample from Dirichlet$(b_0 + n_0, \ldots, b_m + n_m)$, where n_j $(j > 0)$ is the number of motif type j found in the sequence, that is, the total number of is such that $I_i = j$, and $n_0 = n - \sum n_j$.

3.6.5 HMM for *cis* Regulatory Module Discovery

Motif predictions in high eukaryotes such as human and mouse are more challenging than that for simpler organisms, partly owing to the weak motif signals, larger sizes of promoter regions, and difficulties in identifying transcription start sites. In these higher organisms, regulatory proteins often work in combination to regulate target genes, and their binding sites have often been observed to occur in spatial clusters, or *cis-regulatory modules* (CRMs). One approach to locating CRMs is by predicting novel motifs and looking for co-occurrences (Sinha and Tompa, 2002). However, since individual motifs in the cluster may not be well conserved, such an approach often leads to a large number of false negatives. To cope with these difficulties, one can use HMMs to capture both the spatial and contextual dependencies of the motifs in a CRM and use MCMC sampling to infer the CRM models and locations (Thompson *et al.*, 2004; Zhou and Wong, 2004). Gupta and Liu (2005) introduced a competing strategy, which first uses existing *de novo* motif-finding algorithms and/or TF databases to compose a list of putative binding motifs, $\mathcal{D} = \{\Theta_1, \ldots, \Theta_D\}$, where D is in the range of 50 to 100, and then simultaneously updates these motifs and estimates the posterior probability for each of them for inclusion in the CRM.

Let $\mathbf{S}$ denote the set of n sequences with lengths $L_1, L_2, \ldots, L_n$, respectively, corresponding to the upstream regions of n coregulated genes. We assume that the CRM consists of K ($<D$) different kinds of TFs with distinctive position-specific weight matrices (PWMs, which are just a special sequence HMM that do not allow for any gaps). Both the PWMs and K are unknown and need to be inferred from the data. Let $\mathbf{a} = \{a_{ij}; i = 1, \ldots, n; \ j = 1, \ldots, L_i\}$, where $a_{i,j}$ denotes the location of the jth motif site (irrespective of motif type) in the ith sequence. Associated with each site is its *type* indicator $T_{i,j}$, with $T_{i,j}$ taking one of the K values (Let $\mathbf{T} = (T_{ij})$). We model the dependence between $T_{i,j}$ and $T_{i,j+1}$ by a $K \times K$ transition matrix $\boldsymbol{\tau}$. The distance between neighboring TF binding sites in a CRM, $d_{ij} = a_{i,j+1} - a_{i,j}$, is assumed to follow the distribution $Q(d; \lambda, w) = (1 - \lambda)^{d-w} \lambda$ $(d = w, w + 1, \ldots)$. The *background* sequence follows a multinomial distribution with unknown parameter $\boldsymbol{\rho} = (\rho_A, \ldots, \rho_T)$. Finally,

we let **u** be a binary vector indicating which motifs are included in the module, that is, $\mathbf{u} = (u_1, \ldots u_D)^T$, where $u_j = 1$ if the jth motif type is present in the module, and 0 otherwise. By construction, $|\mathbf{u}| = K$. The set of PWMs for the CRM is then $\boldsymbol{\theta} = \{\Theta_j : u_j = 1\}$.

Since we now restrict our inference of CRM to a subset of $\mathcal{D}$, the probability model for the observed sequence data can be written out explicitly as in Gupta and Liu (2005). To implement the Bayesian analysis, we prescribe a joint prior distribution on the unknown parameters, $\Omega = (\mathcal{D}, \boldsymbol{\tau}, \lambda, \boldsymbol{\rho})$, and a prior probability of π for each $u_i = 1$. A Gibbs sampling approach was developed in Thompson *et al.* (2004) to sample both Ω and **u** from their joint posterior distribution. But given the flexibility of the model and the size of the parameter space for an unknown **u**, it is unlikely that a standard MCMC approach can converge to a good solution in a reasonable amount of time. If we ignore the ordering of sites **T** and assume components of **a** to be independent, this model is reduced to the original motif model, which can be updated through the previous Gibbs or data augmentation (DA) procedure.

The following strategy was developed in Gupta and Liu (2005). With a starting set of putative binding motifs $\mathcal{D}$, we iterate the following Monte Carlo sampling steps: (1) Given the current collection $\mathcal{D}$ of motif PWMs (or sites), sample motifs into the CRM; (2) Given the CRM configuration and the PWMs, update the motif site locations through DA; and (3) Given motif site locations, update the corresponding PWMs and other parameters. Since the construction of a CRM in our formulation is done by using an indicator variable $\mathcal{D}$, it is natural to use a genetic-type algorithm to speed up computation. So we implemented an evolutionary Monte Carlo (Liang and Wong, 2000) strategy for the module inference, and obtained very good results for a range of examples.

3.7 JOINT ANALYSIS OF SEQUENCE MOTIFS AND EXPRESSION MICROARRAYS

A highly successful tactic for TF motif discoveries is to cluster genes based on their expression profiles, and search for motifs in the sequences upstream of tightly clustered genes (Spellman *et al.*, 1998). When noise is introduced into the cluster through spurious correlations, however, such an approach may result in many false positives. A filtering method based on the specificity of motif occurrences has been shown to effectively eliminate false positives (Hughes *et al.*, 2000). An iterative procedure for simultaneous gene clustering and motif finding has been suggested (Holmes and Bruno, 2000), but no effective algorithms were implemented to demonstrate its advantage.

Two methods for discovering TF motifs via the association of gene expression values with k-mer abundance have been proposed by Bussemaker *et al.* (2001) and Keles *et al.* (2002). These approaches operate under the explicit assumption that, in response to a given biological condition, the effect of a TF motif is strongest among genes with the most dramatic increase or decrease in mRNA expression. In Bussemaker *et al.* (2001), all the k-mers (k ranging from 5 to 7, say) are first enumerated. Then, for any k-mer, the number n_g of its occurrence in the promoter region (defined as the 800 base pair segment upstream of the transcription start site for the baker's yeast) of each gene g is counted. A regression model is then fit between the expression level y_g of the gene and n_g, and those k-mers whose occurrences are significantly correlated with the gene expression values are regarded as potential TF motifs.

As an alternative to the *k*-mer approach, Conlon *et al.* (2003) provide a motif regression approach to further utilize gene expression information or experimental data from chromatin immunoprecipitation combined with microarrays (often called *ChIP-chip*) to help motif discovery. They first use a fast and sensitive motif-finding method, such as BioProspector, MEME, or MDscan (Liu *et al.*, 2002) to generate a large set of putative motifs that are enriched in the DNA sequence upstream of genes with the highest fold changes in mRNA level relative to a control condition. Then, they conduct a stepwise linear regression to select candidates that correlate with the microarray expression data. Tadesse *et al.* (2004) later presented a Bayesian version of a similar approach. To alleviate the dependence of the linearity and Gaussian assumptions, Das *et al.* (2004) suggested an approach using MARS, and Zhong *et al.* (2005) designed a modified slice inverse regression approach, which also effectively reduces the dimensionality without assuming linearity.

An interesting and bold approach for combining multiple expression microarrays and TF motif analysis is given by Beer and Tavazoie (2004), where they aspired to 'predict gene expression' using only sequence information. To state briefly, they first collected 255 cDNA microarray datasets and used a tight clustering procedure to produce 49 clusters for ~2600 yeast genes (the remaining 3000+ genes were excluded from this analysis because they cannot be clustered 'tightly' enough). For each cluster, they used AlignACE (Roth *et al.*, 1998) to conduct *de novo* motif search among the promoter regions of the genes in the cluster, retaining 5–15 top motifs. The 49 clusters gave rise to a total of 615 motifs. They further conducted an optimization procedure to make each predicted motif more specific to their corresponding cluster, and added 51 experimentally derived motifs reported by Harbison *et al.* (2004). With the 666 motifs in hand, they then trained 49 Bayes network models to predict each gene's cluster membership based only on its motif occurrence scores, putative motif site positions, and orientations. They showed that their procedure yielded an impressive 71 % accuracy in a fivefold cross-validation experiment (compared to ~20 % under random gene order). However, it is interesting to note that their cross-validation procedure is flawed as they have used the cluster information in their motif-finding and motif optimization step, which could lead to an overly optimistic result. Furthermore, their Bayes network model seems to be too heavy for the data in hand. Indeed, our recent preliminary study showed that perhaps they overestimated their procedure's accuracy by about 15 % and that a simpler Naive Bayes procedure can be as accurate or more accurate than their Bayes network model.

3.8   SUMMARY

This article has reviewed a few statistical models used in biological sequence analysis, the corresponding Bayesian formulations, and related computational strategies. As in classical statistics, optimization has been the primary tool in computational biology, where point estimates of very high-dimensional objects obtained by dynamic programming or other clever computational methods are used. Characterizations of uncertainty in these estimates are mostly limited to simple significance test or completely ignored. The marginalization of nuisance parameters is also problematic, and is most frequently done using the *profile likelihood* method in which the nuisance parameters are fixed at their best estimates. In comparison, the Bayesian method has no difficulties in these important aspects:

the uncertainty in estimation is addressed by posterior calculations and the nuisance parameters are removed by summation and integration. In exchange for these advantages, however, one needs to set prior distributions and overcome computational hurdles, neither of which are trivial in practice. Recursion-based Bayesian algorithms generally have time and space requirements of the same order as their dynamic-programming counterparts. For those problems where there is no polynomial time solution, MCMC methods (and other Monte Carlo methods) provide an effective means to implement Bayesian analysis.

Acknowledgments

APPENDIX A: MARKOV CHAIN MONTE CARLO METHODS

Metropolis−Hastings Algorithm. Let $\pi(\mathbf{x}) = c \exp\{-h(\mathbf{x})\}$ be the target distribution with unknown constant c. Metropolis *et al.* (1953) introduced the fundamental idea of Markov chain sampling and prescribed the first general construction of such a chain. Hastings (1970) later provided an important generalization. Starting with any configuration $\mathbf{x}^{(0)}$, the M-H algorithm evolves from the current state $\mathbf{x}^{(t)} = \mathbf{x}$ to the next state $\mathbf{x}^{(t+1)}$ as follows:

- Propose a new state $\mathbf{x}'$ that can be viewed as a small and random 'perturbation' of the current state. More precisely, $\mathbf{x}'$ is generated from a *proposal* function $T(\mathbf{x}^{(t)} \to \mathbf{x}')$ (i.e. it is required that $T \geq 0$ and $\sum_{\text{all } \mathbf{y}} T(\mathbf{x} \to \mathbf{y}) = 1$ for all $\mathbf{x}$) determined by the user.

- Compute the Metropolis ratio

$$r(\mathbf{x}, \mathbf{x}') = \frac{\pi(\mathbf{x}')T(\mathbf{x}' \to \mathbf{x})}{\pi(\mathbf{x})T(\mathbf{x} \to \mathbf{x}')}. \tag{3A.1}$$

- Generate a random number $u \sim \text{Unif } [0,1]$.

  - Let $\mathbf{x}^{(t+1)} = \mathbf{x}'$ if $u \leq r(\mathbf{x}, \mathbf{x}')$.
  - Let $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$ otherwise.

A more well-known form of the Metropolis algorithm is obtained by iterating the following steps: (a) a small random perturbation of the current configuration is made; (b) the 'gain' (or loss) in an objective function (i.e. $-h(\mathbf{x})$) resulting from this perturbation is computed; (c) a random number U is generated independently; and (d) the new configuration is accepted if $\log(U)$ is smaller than or equal to the 'gain', and is rejected otherwise. The well-known *simulated annealing* algorithm (Kirkpatrick *et al.*, 1983) is built upon this basic Metropolis iteration with an additional twist of including an adjustable exponential scaling parameter to the objective function (i.e. $\pi(\mathbf{x})$ is scaled to $\pi^{\alpha}(\mathbf{x})$ and

$\alpha \to 0$). Metropolis *et al.* (1953) restricted their choices of the 'perturbation' function to be the symmetric ones, that is, $T(\mathbf{x} \to \mathbf{x}') = T(\mathbf{x}' \to \mathbf{x})$. Hastings (1970) generalized the choice of T to all those that satisfy the property: $T(\mathbf{x} \to \mathbf{x}') > 0$ if and only if $T(\mathbf{x}' \to \mathbf{x}) > 0$.

Heuristically, π can be seen as a 'fixed point' under the M-H operation in the space of all distributions. It follows from the standard Markov chain theory that if the chain is *irreducible* (i.e. it is possible to go from anywhere to anywhere else in a finite number of steps), *aperiodic* (i.e. there is no parity problem), and not drifting away, then in the long run the chain will settle in its invariant distribution (Liu, 2001). The random samples so obtained eventually are like those drawn directly from π.

Gibbs Sampler. Suppose $\mathbf{x} = (x_1, \ldots, x_d)$. In the Gibbs sampler, one randomly or systematically chooses a coordinate, say x_1, and then updates its value with a new sample x_1' drawn from the conditional distribution $\pi(\cdot \mid \mathbf{x}_{[-1]})$, where $\mathbf{x}_{[-A]}$ refers to $\{x_j, \ j \in A^c\}$. Algorithmically, the Gibbs sampler can be implemented as follows:

Random Scan Gibbs sampler. Suppose currently $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots x_d^{(t)})$.

- Randomly select i from $\{1, \ldots, d\}$ according to a given probability vector $(\alpha_1, \ldots, \alpha_d)$.

- Let $x_i^{(t+1)}$ be drawn from the conditional distribution $\pi(\cdot \mid \mathbf{x}_{[-i]}^{(t)})$, and let $x_{[-i]}^{(t+1)} = \mathbf{x}_{[-i]}^{(t)}$.

Systematic Scan Gibbs sampler. Let the current state be $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots x_d^{(t)})$.

- For $i = 1, \ldots, d$, we draw $x_i^{(t+1)}$ from the conditional distribution

$$\pi(x_i \mid x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_d^{(t)}).$$

The Gibbs sampler's popularity in statistics community stems from its extensive use of *conditional distributions* in each iteration. Tanner and Wong (1987)'s data augmentation first linked the Gibbs sampling structure with missing data problems and the EM algorithm. Gelfand and Smith (1990) further popularized the method by pointing out that the conditionals needed in Gibbs iterations are commonly available in many Bayesian and likelihood computations.

Other Techniques. A main problem with all MCMC algorithms is that they may, for some problems, move very slowly in the configuration space or may be trapped in the region of a local mode. This phenomena is generally called *slow-mixing* of the chain. When chain is slow-mixing, estimation based on the resulting Monte Carlo samples can be very inaccurate. Some recent techniques suitable for designing more efficient MCMC samplers in bioinformatics applications include simulated tempering, parallel tempering, multicanonical sampling, multiple-try method, and evolutionary Monte Carlo. These and some other techniques are summarized in Liu (2001). Some more discussions and applications of MCMC can also be found in **Chapters 8**, **15**, and **26**.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.

Bahr, A., Thompson, J.D., Thierry, J.C. and Poch, O. (2001). BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Research* **29**, 323–326.

Baldi, P., Chauvin, Y., McClure, M. and Hunkapiller, T. (1994). Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 1059–1063.

Baum, L.E. (1972). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41**, 164–171.

Beer, M.A. and Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell* **117**, 185–198.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. (2002). GenBank. *Nucleic Acids Research* **30**, 17–20.

Bishop, M.J. and Thompson, E.A. (1986). Maximum likelihood alignment of DNA sequences. *Journal of Molecular Biology* **190**, 159–165.

Bussemaker, H.J., Li, H. and Siggia, E.D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics* **27**, 167–171.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78–94.

Cardon, L.R. and Stormo, G.D. (1992). Expectation maximization algorithm for identifying-binding sites with variable lengths from unaligned DNA fragments. *Journal of Molecular Biology* **223**, 159–170.

Churchill, G.A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51**, 79–94.

Churchill, G.A. and Lazareva, B. (1999). Bayesian restoration of a Hidden Markov Chain with applications to DNA sequencing. *Journal of Computational Biology* **6**, 261–277.

Conlon, E.M., Liu, X.S., Lieb, J.D. and Liu, J.S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences of the United States of America* **100**(6), 3339–3344.

Das, D., Banerjee, N. and Zhang, M.Q. (2004). Interacting models of cooperative gene regulation. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 16234–16239.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series A* **39**, 1–38.

Ding, Y. and Lawrence, C.E. (2001). Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucleic Acids Research* **29**(5), 1034–1104.

Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics* **14**, 755–763.

Edgar, R.C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113.

Edgar, R.C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797.

Efron, B. (1979). Bootstrap method: another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approach to calculating marginal densities. *Journal American Statistical Association* **85**, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, New York.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1998). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, Boca Raton, FL.

Gupta, M. and Liu, J.S. (2005). De-novo cis-regulatory module elicitation for eukaryotic genomes. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 7079–7084.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N., Macisaac, K.D., Danford, T.D., Hannett, N.M., Tagne, J.-B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E. and Young, R.A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Henikoff, S. and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10915–10919.

Henikoff, S. and Henikoff, J.G. (1994). Position-based sequence weights. *Journal of Molecular Biology* **243**, 574–578.

Holmes, I. and Bruno, W. (2000). Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proceedings International Conference on Intelligent Systems for Molecular Biology* **8**, 202–210.

Holmes, I. and Durbin, R. (1998). Dynamic programming alignment accuracy. *Proceedings of the 2nd Annual International Conference on Computational Molecular Biology* **2**, 102–108.

Huang, H., Kao, M.J., Zhou, X., Liu, J.S. and Wong, W.H. (2004). Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification. *Journal of Computational Biology* **11**, 1–14.

Hudak, J. and McClure, M.A. (1999). A comparative analysis of computational motif-detection methods. *Pacific Symposium on Biocomputing* **4**, 138–149.

Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology* **296**, 1205–1214.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal American Statistical Association* **90**, 773–795.

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059–3306.

Keles, S., van der Laan, M. and Eisen, M.B. (2002). Identification of regulatory elements using a feature selection method. *Bioinformatics* **18**, 1167–1175.

Kimura, M. (1985). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Krogh, A., Brown, M., Mian, S., Sjolander, K. and Haussler, D. (1994a). Protein modeling using hidden Markov models. *Journal of Molecular Biology* **235**, 1501–1531.

Krogh, A., Mian, I.S. and Haussler, D. (1994b). A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Research* **22**, 4768–4778.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214.

Lawrence, C.E. and Reilly, A.A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **7**, 41–51.

Li, C. and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America* **98**(1), 31–36.

Liang, F. and Wong, W.H. (2000). Evolutionary Monte Carlo: applications to C_p model sampling and change point problem. *Statistica Sinica* **10**, 317–342.

Liu, J.S. (1994). The collapsed Gibbs sampler with applications to a gene regulation problem. *Journal American Statistical Association* **89**, 958–966.

Liu, J.S. (2001). *Monte Carlo Methods for Scientific Computing*. Springer-Verlag, New York.

Liu, X.S., Brutlag, D.L. and Liu, J.S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Proceedings of the Pacific Symposium on Bioinformatics* **6**, 127–138.

Liu, X.S., Brutlag, D.L. and Liu, J.S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nature Biotechnology* **20**, 835–839.

Liu, J.S. and Lawrence, C.E. (1999). Bayesian inference on biopolymer models. *Bioinformatics* **15**, 38–52.

Liu, J.S., Neuwald, A.F. and Lawrence, C.E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal American Statistical Association* **90**, 1156–1170.

Liu, J.S., Neuwald, A.F. and Lawrence, C.E. (1999). Markovian structures in biological sequence alignments. *Journal American Statistical Association* **94**, 1–15.

Logvinenko, T. (2002). Sequential Monte Carlo and dirichlet mixtures for extracting protein alignment models. Ph.D. Thesis, Stanford University.

Lowe, T.M. and Eddy, S.R. (1997). tRNA-scan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **5**, 955–964.

Lu, X., Zhang, W., Qin, Z.S., Kwast, K.E. and Liu J.S. (2004). Statistical resynchronization and Bayesian detection of periodically expressed genes. *Nucleic Acids Research* **32**(2), 447–455.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1091.

Needleman, S.B. and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453.

Neuwald, A.F., Kannan, N., Poleksic, A., Hata, N. and Liu, J.S. (2003). Ran's C-terminal, basic patch and nucleotide exchange mechanisms in light of a canonical structure for Rab, Rho, Ras and Ran GTPases. *Genome Research* **13**, 673–692.

Neuwald, A.F. and Liu, J.S. (2004). Gapped alignment of protein sequence motifs through Monte Carlo optimization of a hidden Markov model. *BMC Bioinformatics* **5**, 157.

Neuwald, A.F., Liu, J.S. and Lawrence, C.E. (1995). Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Science* **4**, 1618–1632.

Neuwald, A.F., Liu, J.S., Lipman, D.J. and Lawrence, C.E. (1997). Extracting protein alignment models from the sequence database. *Nucleic Acids Research* **25**, 1665–1677.

Notredame, C., Higgins, D.G. and Heringa, J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**, 205–217.

Pearson, W.R. and Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 2444–2448.

Pedersen, J.S., Meyer, I.M., Forsberg, R., Simmonds, P. and Hein, J. (2004). A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Research* **32**(16), 4925–4936.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257–286.

Roth, F.P., Hughes, J.D., Estep, P.W. and Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* **16**, 939–945.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406–425.

Sankoff, D. (1972). Matching sequences under deletion/insertion constraints. *Proceedings of the National Academy of Sciences of the United States of America* **69**, 4–6.

Schmidler, S.C., Liu, J.S. and Brutlag, D.L. (2000). Bayesian segmentation of protein secondary structure. *Journal of Computational Biology* **7**, 233–248.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* **34**(2), 166–176.

Sinha, S. and Tompa, M. (2002). Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research* **30**, 5549–5560.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S. and Haussler, D. (1996). Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. *Computer Applications in the Biosciences* **12**, 327–345.

Smith, T.F. and Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197.

Speed, T. (ed) (2003). *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC, London.

Spellman, P.T., Sherlock G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown P.O., Botstein D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9**(12), 3273–3297.

Stormo, G.D. and Hartzell, G.W. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 1183–1187.

Tadesse, M.G., Vannucci, M. and Lio, P. (2004). Identification of DNA regulatory motifs using Bayesian variable selection. *Bioinformatics* **20**, 2553–2561.

Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal American Statistical Association* **82**, 528–550.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994a). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994b). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Computer Applications in the Biosciences* **10**, 19–29.

Thompson, W., Palumbo, M.J., Wasserman, W.W., Liu, J.S. and Lawrence, C.E. (2004). Decoding human regulatory circuits. *Genome Research* **10**, 1967–1974.

Thorne, J.L., Kishino, H. and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* **33**, 114–124.

Webb, B.M., Liu, J.S. and Lawrence, C.E. (2002). BALSA: Bayesian algorithm for local sequence alignment. *Nucleic Acids Research* **30**, 1268–1277.

Zhong, W., Zeng, P., Ma, P., Liu, J.S. and Zhu, Y. (2005). RSIR: regularized sliced inverse regression for motif discovery. *Bioinformatics* **21**, 4169–4175.

Zhou, Q. and Wong, W.H. (2004). CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12114–12119.

Zhu, J., Liu, J.S. and Lawrence, C.E. (1998). Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14**, 25–31.

Zuker, M. (1989). Computer prediction of RNA structure. *Methods in Enzymology* **180**, 262–288.

<div align="right">

4

</div>

Statistical Approaches in Eukaryotic Gene Prediction

V. Solovyev

Department of Computer Science, University of London, Surrey, UK

Finding genes in genomic DNA is a foremost problem of molecular biology. With the ongoing genome sequencing projects producing large quantities of sequence data, computational gene prediction is the major instrument for the identification of new genes. Usually, gene-finding programs accurately predict most coding exons in analyzed sequences, while producing a complete set of exact gene structures in any genome is still unsolved and difficult task, complicated by the large amount of gene variants generated by alternative splicing, alternation promoters and alternative polyadenylation sites. Nevertheless using gene prediction, the scientific community is now able to start experimental work with the majority of genes in dozens of sequenced genomes. Therefore, computational methods of gene identification have attracted significant attention of the genomics and bioinformatics communities. This chapter presents a comprehensive description of advanced probabilistic and discriminative gene-prediction approaches such as Hidden-Markov Models and pattern-based algorithms. We have described the structure of functional signals and significant gene features incorporated into the programs to recognize protein-coding genes. We have presented comparative performance data for a variety of gene structure identification programs and discussed some experiences in annotation of sequences from genome sequencing projects. A complex approach for finding promoters and pseudogenes have been considered as well as evaluation of their accuracy in annotation of human genome sequences. Finally, we described structural features and expression of miRNA genes and some computational methods for miRNA gene identification in genomic sequences as well as computational methods of finding miRNA targets.

4.1 STRUCTURAL ORGANIZATION AND EXPRESSION OF EUKARYOTIC GENES

The gene is the unit of inheritance encoded by a segment of nucleic sequence that carries the information representing a particular polypeptide or RNA molecule. A two-stage process comprising transcription and translation makes use of this information.

Transcription (or pre-mRNA synthesis on a DNA template) involves initiation, elongation and termination steps. RNA polymerase catalyzing RNA synthesis binds a special region (promoter) at the start of the gene and moves along the template, synthesizing RNA, until it reaches a terminator sequence. Posttranscriptional processing of mRNA precursors includes capping, 3′-polyadenylation and splicing. The processing events of mRNA capping and polyA addition take place before pre-mRNA splicing finally produces the mature mRNA. The mature mRNA includes sequences that correspond exactly to the protein product according to the rules of the genetic codes, called *exons*. The genomic gene sequence often includes noncoding regions called *introns* that are removed from the primary transcript during RNA splicing. Eukaryotic pre-mRNA is processed in the nucleus and then transported to the cytoplasm for translation (protein synthesis). The sequence of mRNA contains a series of triplet codons that interact with the anticodons of aminoacyl-tRNAs (carrying the amino acids) so that the corresponding series of amino acids is incorporated into a polypeptide chain. The small subunit of the ribosome binds to the 5′-end of mRNA and then migrates to the special sequence on mRNA (prior to the start codon) called the *ribosome binding site*, where it is joined by a large ribosome subunit forming a complete ribosome. The ribosome initiates protein synthesis at the start codon (AUG in eukaryotes) and moves along the mRNA, synthesizing the polypeptide chain, until it reaches a stop codon sequence (TAA, TGA or TAG), where release of polypeptide and dissociation of the ribosome from the mRNA take place. Many proteins undergo posttranslational processing (i.e. covalent modifications such as proteolytic cleavage, attachment of carbohydrates and phosphates) before they become functional. The expression stages and the structural organization of a typical eukaryotic protein-coding gene including associated regulatory regions is shown in Figure 4.1. Figure 4.2 illustrates how one DNA sequence may code for multiple proteins due to alternative promoters or terminators and alternative splicing. These processes significantly complicate *ab initio* computational gene finding.
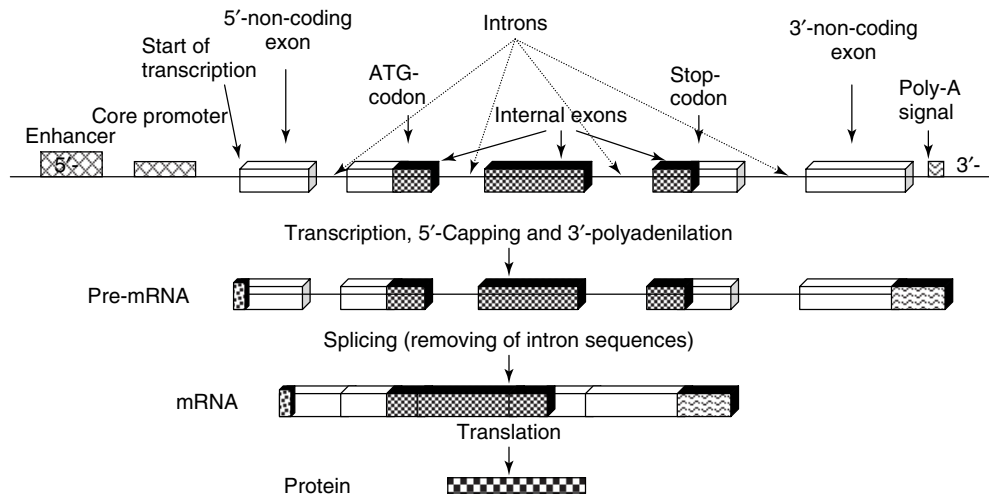
Figure 4.1 Expression stages and structural organization of typical eukaryotic protein-coding gene including its regulatory regions.
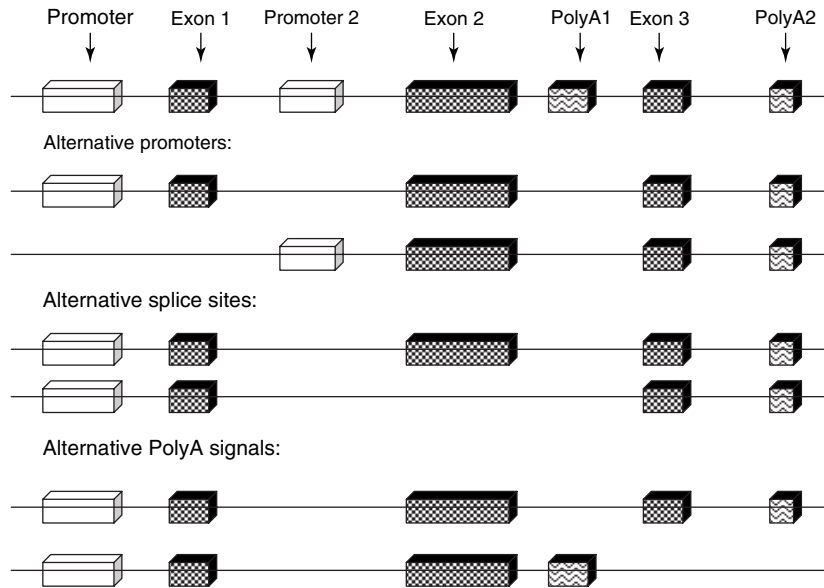
Figure 4.2   Alternative gene products coded by the same DNA region.

Information describing structural gene characteristics is accumulated in the GenBank (Benson *et al.*, 1999) and the EMBL (Kanz *et al.*, 2005) nucleotide sequence databases. However, these databases mostly contain annotations of genome fragments. Therefore, one gene can be described in dozens of entries containing partially sequenced gene regions or alternative splicing forms of its mRNA. The value of public availability of predicted genes was recognized during human genome sequencing by creating the InfoGene database (Solovyev and Salamov, 1999), which contained descriptions of known and predicted genes and their basic functional signals. Later, a few big research groups developed powerful specialized WEB accessible recourses, where various annotations of different genomes are stored and can be interactively analyzed: University of California Santa Cruz (USSC) Genome Browser (Kent *et al.*, 2002) and Ensembl genome database (Hubbard *et al.*, 2002). Some big sequencing centers such as The Institute for Genomic Research (TIGR) and Joint Genomic Institute (JGI) produce and present annotation of specific genomes at their Web servers.

Table 4.1 shows the major structural characteristics of human genes deposited to GenBank (Release 116).

41 % of sequenced human DNA consists of different kinds of repeats. Only ∼3 % of the genome sequence contains protein-coding exon sequences. Characteristics of genes in major model organisms such as mouse, *Drosophila melanogaster, Caenorhabditis elegans, S. cerevisiae* and *Arabidopsis thaliana* are presented in Table 4.2.

In general, there is no big difference in the size of protein-coding mRNAs in different types of organisms, but the gene sizes are often higher in vertebrates and especially in primates. Human coding exons are significantly shorter than the sizes of genes. The average size of the exons is about 190 bp that is close to the DNA length associated with the nucleosome particle. There are many exons as short as a few bases. For example, the human pleiotrophin gene (*HUMPLEIOT*) includes a 1-bp exon and one of the alternative

Table 4.1 Structural characteristics of human genes.[*]

Gene features	Numbers from Infogen
CDS/partially sequenced CDS	48 088/26 584
CDS length (minimal, maximal, average)	15, 80 781, 1482
Exons/partially sequenced exons	72 488/19 392
Genes/partially sequenced genes	18 429/14 385
Alternative splicing	12 %
Pseudogenes	8.5 %
Genes without introns	8 %
Number of exons (maximal, average)	117, 5.4
Exon length (range, average)	1–10 088, 195
Intron length (maximal, average)	185 838, 2010
Gene length (maximal, average)	401 910, 7865
Repeats in genome	41 % of total DNA
DNA occupied by coding exons	3 %

[*]The numbers reflect genes described in GenBank, which might deviate from the average parameters for the organism. Gene numbers were calculated for DNA loci only. Many long genes have partially sequenced introns, therefore average sizes of genes and introns are actually bigger. The average numbers of exons and gene lengths were calculated for completely sequenced genes only.
CDS: protein CoDing Sequences.

forms of the human folate receptor (*HSU20391*) gene contains a 3-bp exon. Coding exons can also be very short.

The human myosin-binding protein gene (*HSMYBPC3*) includes 2 exons that are 3-bp long. At the same time we observe a coding region about 90 000 bp for the titin gene (NM_003319) and an exon 8000 bp in human gene encoding microtubule-associated protein 1a (HSU38291). Very often protein-coding exons occupy a small percent of the gene size. The human fragile X mental retardation gene (*HUMFMR1S*) presents a typical example: 17 exons (40–60-bp long) occupy just 3 % of 67 000 bp gene sequence.

The structural characteristics of eukaryotic genes, discussed above, create difficult problems in computational gene identification. Low density of coding regions (3 % in human DNA) will generate a lot of false-positive predictions in fragments of noncoding DNA. The number of these false positives might be even comparable with the true number of exons. Recognition of small exons (1–20 bp) cannot be achieved using any composition-based method that is relatively successful for identification of prokaryote coding regions. It is necessary to develop gene-prediction approaches that rely significantly on the recognition of functional signals encoded in the DNA sequence.

4.2 METHODS OF FUNCTIONAL SIGNAL RECOGNITION

In this paragraph, we will describe several approaches for gene functional signal recognition and some features of these signals used in gene identification. The simplest way to find functional sequences is based on a consensus sequence or weight matrix reflecting conservative bases of the signal. Using a consensus sequence or weight matrix we can scan a given sequence and select high scoring regions as potential functional signals.

Table 4.2 Structural characteristics of genes in eukaryotic model organisms.[*]

	Mus musculus	D. melanogaster	C. elegans	S. cerevisiae	A.thaliana
CDS/partial	20 340/11 192	5095/1057(13 601)	18 146/377	12 629/1040	14 590/1076
Exons/partial	14 940/7812	8661/3694 (56 673)	108 934/33 821	13 444/13 028	62 151/19 505
Genes/partial	5571/4077	2802/948 (13601)	17 336/1003	12 495/1024	12 221/616
Alternative splicing	11 %	15 % (7 %)	5.6 %	5.8 %	1.4 %
No introns genes	10 %	20 %	4 %	90 %	20 %
Number of exons	64, 4.59	26, 3.1 (4.2)	48, 6.1	3, 1.03	70, 5.1
Exon length	1–6642, 199.1	6–9462, 454 (425)	1–14 975, 22 125	1–7471, 1500.0	2–5130, 190.8
Intron length	29 382, 678.4	73650, 457 (488)	19 397, 244.0	7317, 300	118 637, 186.39
Gene size	8020, 3388	74 691, 2150	45 315, 2496	14 733, 1462	170 191, 2040

[*]Description of features is given in Table 4.1. For Drosophila genes, the numbers in () are taken from computer and manual annotation of Drosophila genome (Adams et al., 2000).

To select only significant matches, a statistical method for estimating the significance of similarity between the consensus of functional signal and a sequence fragment was developed (Shahmuradov *et al.*, 1986; Solovyev and Kolchanov, 1994). Computation of this statistic was implemented in the program NSITE (`http://www.softberry.com/berry.phtml?topic=nsite&group=programs&subgroup=promoter`) (Shahmuradov and Solovyev, 1999) that identifies nonrandom similarity between fragments of a given sequence and consensuses of regulatory motifs from various databases such as Transcription Factor Database (TFD) (Ghosh, 2000), Transfac (Wingender *et al.*, 1996), RegsiteAnimal and RegsitePlant (Solovyev *et al.*, 2003). Here we briefly describe application of weight matrices, which usually contain more information about the structure of functional signal than consensus sequences. Procedures using weight matrices are implemented in many modern gene-prediction approaches that score potential functional signals.

4.2.1 Position-specific Measures

Weight matrices are typically used for functional signal description (Staden, 1984a; Zhang and Marr, 1993; Burge, 1997). We can consider weight matrix as a simple model based on a set of position-specific probability distributions $\{p_s^i\}$, that provide probabilities of observing a particular type of nucleotide in a particular position of functional signal sequence (S). The probability of generating a sequence $X(x_1, \ldots, x_k)$ under this model is

$$P(X/S) = \prod_{i=1}^{k} p_{x_i}^i, \tag{4.1}$$

where each position of the signal is considered to be independent. A corresponding model can also be constructed for a sequence having no functional signal (N): $\{\pi_s^i\}$. An appropriate discriminative score based on these models is the log likelihood ratio:

$$\text{LLR}(X) = \log \frac{P(X/S)}{P(X/N)}. \tag{4.2}$$

To evaluate a given sequence fragment, a score can be computed as an average sum of weights of observed nucleotides using the corresponding weight matrix $w_{(i,s)} = \{log(p_s^i/\pi_s^i)\}$:

$$\text{Score} = \text{LLR}(X) = \frac{1}{k} \sum_{i=1}^{k} w(i, x_i). \tag{4.3}$$

Different weight functions have been used to score the sequence, for example, weights can be obtained by some optimization procedures such as a perceptron or neural network (Stormo *et al.*, 1982). Different position-specific probability distributions $\{p_s^i\}$ can also be considered.

A generalization of the weight matrix uses position-specific probability distributions $\{p_s^i\}$ of oligonucleotides (instead of single nucleotides). Another approach is to exploit Markov chain models, where the probability of generating a particular nucleotide x_i of the signal sequence depends on the $k_0 - 1$ previous bases (i.e. it depends on an oligonucleotide ($k_0 - 1$ base long) ending at the position $i - 1$). Then the probability of generating the

signal sequence X is:

$$P(X/S) = p_0 \prod_{i=k_0}^{k} p_{s_{i-1}, x_i}^{i-1, i} : \qquad (4.4)$$

where $p_{s_{i-1}, x_i}^{i-1, i}$ is the conditional probability of generating nucleotide x_i in position i given oligonucleotide s_{i-1} ended at position $i-1$, p_0 is the probability of generating oligonucleotide $x_{1...}x_{k0-1}$. For example, a simple weight matrix represents independent mononucleotide model (or 0-order Markov chain), where $k_0 = 1$, $p_0 = 1$ and $p_{x_{i-1}, x_i}^{i-1, i} = p_{x_i}^i$. When we use dinucleotides (1st order Markov chain) $k_0 = 2$, $p_0 = p_{x_1}^1$, and $p_{x_{i-1}, x_i}^{i-1, i}$ is the conditional probability of generating nucleotide x_i in position i given nucleotide x_{i-1} at position $i-1$. The conditional probability can be estimated from the ratio of observed frequency of oligonucleotide k_0 bases long ($k_0 > 1$) ending at position i to the frequency of the oligonucleotide $k_0 - 1$ bases long ending at position $i-1$ in a set of aligned sequences of some functional signal.

$$p_{s_{i-1}, x_i}^{i-1, i} = \frac{f(s_{i-1}, x_i)}{f(s_{i-1})}.$$

Using the same procedure we can construct a model for nonsite sequences for computing $P(X/N)$, where often 0-order Markov chain with genomic base frequencies (or even equal frequencies (0.25)) is used.

A log likelihood ratio (3) with Markov chains was applied to select CpG island regions (Durbin *et al.*, 1998). The same approach was used in a description of promoters, splice sites and start and stop of translation in gene-finding programs such as Genscan (Burge and Karlin, 1997), Fgenesh (Find GENES Hmm) (Salamov and Solovyev, 2000) and GeneFinder (Green and Hillier, 1998).

A useful discriminative measure taking into account *a priori* knowledge is based on the computation of Bayesian probabilities as components of position-specific distributions $\{p_s^i\}$:

$$P(S/o_s^i) = \frac{P(o_s^i/S)P(S)}{\left(P(o_s^i/S)P(S) + P(o_s^i/N)P(N)\right)}, \qquad (4.5)$$

where $P(o_s^i/S)$ and $P(o_s^i/N)$ can be estimated as position-specific frequencies of oligonucleotides o_s^i in the set of aligned sites and nonsites; $P(s)$ and $P(N)$ are the *a priori* probabilities of site and nonsite sequences, respectively. S is a type of the oligonucleotide starting (or ending) in ith position (Solovyev and Lawrence, 1993a). The probability that a sequence X belongs to a signal, if one assumes independence of oligonucleotides in different positions, is:

$$P(S/X) = \prod_{i=1}^{k} P(S/o_m^i).$$

Another empirical discriminator called *preference* uses the average positional probability of belonging to a signal:

$$Pr(S/X) = \frac{1}{k} \sum_{i=1}^{k} P(S/o_m^i).$$
(4.6)

This measure was used in constructing discriminant functions for the Fgenes gene-finding program (Solovyev *et al.*, 1995). It can be more stable than the previous measure on short sequences and has simple interpretation: if the $Pr > 0.5$, then our sequence is more likely to belong to a signal than to a nonsignal sequence.

4.2.2 Content-specific Measures

Some functional signal sequences have a distinctive general oligonucleotide composition. For example, many eukaryotic promoters are found in GC-rich chromosome fragments. We can characterize these regions by applying similar methods to the above scoring functions, but using probability distributions and their estimates by oligonucleotide frequencies computed on the whole sequence of the functional signal. For example, the Markov-chain-based probability of generating the signal sequence X will be:

$$P(X/S) = p_0 \prod_{i=k_0}^{k} p_{s_{i-1}}, x_i.$$
(4.7)

4.2.3 Frame-specific Measures

The coding sequence is a sequence of triplets (codons) read continuously from a fixed starting point. Three different reading frames with different codons are possible for any nucleotide sequence (or 6 if the complementary chain is also considered). The nucleotides are distributed unevenly relative to the positions within codons. Therefore the probability of observing a specific oligonucleotide in coding sequence depends on its position relative to the coding frame (three possible variants) as well as on neighboring nucleotides (Shepherd, 1981; Borodovskii *et al.*, 1986; Borodovsky and McIninch, 1993). Asymmetry in base composition between codon positions arises because of uneven usage of amino acids and synonymous codons, as well as the specific nature of the genetic code (Guigo, 1999). Fickett and Tung (1992) did a comprehensive assessment of the various protein-coding measures. They estimated the quality of more than 20 measures and showed that the most powerful is 'in phase hexanucleotide composition'. In Markov chain approaches, the frame-dependent probabilities $p_{s_{i-1},x_i}^f (f = \{1,2,3\})$ are used to model coding regions. The probability of generating a protein-coding sequence X is

$$P(X/C) = p_0 \prod_{i=k_0}^{k} p_{s_{i-1},x_i}^f,$$
(4.8)

where f is equal to 1, 2 or 3 for oligonucleotides ending at codon position 1, 2 or 3, respectively.

4.2.4 Performance Measures

Several measures to estimate the accuracy of a recognition function were introduced in genomic research (Fickett and Tung, 1992; Snyder and Stormo, 1993; Dong and Searls, 1994). Consider that we have S sites (positive examples) and N nonsites (negative

examples). By applying the recognition function, we correctly identify Tp sites (true positives) and T_n nonsites (true negatives). At the same time F_p (false positives) sites are wrongly classified as nonsites and F_n (false negative) nonsites are wrongly classified as sites. $T_p + F_n = S$ and $T_n + F_n = N$. Sensitivity (S_n) measures the fraction of the true positive examples that are correctly predicted: $S_n = T_P/(T_P + F_n)$. Specificity (S_p) measures the fraction of the predicted sites that are correct amongst those predicted: $S_p = T_P/(T_P + F_P)$. Note that the definition of S_p used in gene-prediction research is different from the usual $S_p = T_n/(T_n + F_p)$. Only the simultaneous consideration of both S_n and S_p values makes sense when we provide some accuracy information. Using only one value of accuracy estimation means that the average accuracy of prediction of true sites and nonsites is $AC = 0.5(T_P/S + T_n/N)$. However, this measure does not take into account the possible difference in sizes of site and nonsites sets. A more correct single measure (correlation coefficient) takes the relation between correctly predictive positives and negatives as well as false positives and negatives into account (Matthews, 1975):

$$CC = \frac{(T_p T_n - F_p F_n)}{\sqrt{(T_p + F_p)(T_n + F_n)(T_p + F_n)(T_n + F_p)}}.$$

4.3 LINEAR DISCRIMINANT ANALYSIS

Different features of a functional signal may have different significance for recognition and may not be independent. Classical linear discriminant analysis provides a method to combine such features in a discriminant function. A discriminant function, when applied to a pattern, yields an output that is an estimate of the class membership of this pattern. The discriminative technique provides minimization of the error rate of classification (Afifi and Azen, 1979). Let us assume that each given sequence can be described by vector X of p characteristics $(x_1, x_2, \ldots, x_p)$, that can be measured. The linear discriminant analysis procedure finds a linear combination of the measures (called the *linear discriminant function* or *LDF*), that provides maximum discrimination between site sequences (class 1) and nonsite examples (class 2). The LDF classifies (X) into class 1 if $Z > c$ and into class 2 if $Z > c$. The vector of coefficients $(\alpha_1, \alpha_2, \ldots, \alpha_p)$ and threshold constant c are derived from the training set by maximizing the ratio of the between-class variation of z to within-class variation and are equal to (Afifi and Aizen, 1979):

$$\vec{a} = s^{-1}(\vec{m_1} - \vec{m_2}),$$

and

$$\vec{c} = \vec{a}(\vec{m_1} - \vec{m_2})/2,$$

where $\vec{m_i}$ are the sample mean vectors of characteristics for class 1 and class 2, respectively; s is the pooled covariance matrix of characteristics

$$s = \frac{1}{n_1 + n_2 - 2}(s_1 + s_2)$$

s_i is the covariation matrix, and n_i is the sample size of class i. On the basis of these equations, we can calculate the coefficients of LDF and threshold constant c using the values of characteristics of site and nonsite sequences from the training sets and then test the accuracy of LDF on the test set data. Significance of a given characteristic or a set of characteristics can be estimated by the generalized distance between two classes (called the *Mahalonobis distance* or D^2):

$$\vec{D^2} = (\vec{m_1} - \vec{m_2})s^{-1}(\vec{m_1} - \vec{m_2}),$$

that is computed on the basis of values of the characteristics in the training sequences of classes 1 and 2. To find sequence features a lot of possible characteristics as score of weigh matrices, distances, oligonucleotide preferences at different subregions are generated. Selection of the subset of significant characteristics (among those tested) is performed by a stepwise discriminant procedure including only those characteristics that significantly increase the Mahalonobis distance (Afifi and Aizen, 1979).

4.4   PREDICTION OF DONOR AND ACCEPTOR SPLICE JUNCTIONS

4.4.1   Splice-sites Characteristics

Splice-site patterns are mainly defined by nucleotides at the ends of introns, because deletions of large parts of intron do not affect their selection (Breathnach and Chambon, 1981; Wieringa *et al.*, 1984). A sequence of eight nucleotides is highly conserved at the boundary between an exon and an intron (donor or 5′-splice site). This is AG|GTRAGT and a sequence of 4 nucleotides, preceded by a pyrimidine rich region, is also highly conserved between an exon and an intron (acceptor or 3′-splice site): YYTTYYYYYYNC|AGG (Senapathy *et al.*, 1990). The third less-conserved sequence of about 5−8 nucleotides, and containing an adenosine residue, lies within the intron, usually between 10 and 50 nucleotides upstream of the acceptor splice site (branch site). These sequences provide specific molecular signals by which the RNA splicing machinery can select the splice sites with precision.

Two very conservative dinucleotides are observed in practically all introns. The donor site has GT just after the point where the spliceosomes cut the 5′-end of intron sequences and the acceptor site has AG just before the point where the spliceo-somes cut the 3′-end of intron sequences (Breathnach *et al.*, 1978; Breathnach and Chambon, 1981).

Additionally, a rare type of splice pair AT−AC has been discovered. It is processed by related but different splicing machinery (Jackson, 1991; Hall and Padget, 1994). Introns flanked by the standard GT−AG pairs excised from pre-mRNA by the spliceosome including U1, U2, U4/U6 and U5 snRNPs (Nilsen, 1994). A novel type of spliceosome composed of snRNPs U11, U12, U4atac/U6atac and U5 (Hall and Padgett, 1996, Tarn and Steitz, 1996a; 1996b; 1997) excises AT−AC introns. For AT−AC group a different conserved positions have been noticed: |ATATCCTTT for donor site and YAC| for acceptor site (Dietrich *et al.*, 1997; Sharp and Burge, 1997; Wu and Krainer, 1997).

Burset *et al.* (2000) have done a comprehensive investigation of canonical and noncanonical splice sites. They have extracted 43 427 pairs of exon−intron bound-aries and their sequences from the InfoGene (Solovyev and Salamov, 1999) database

including all the annotated genes in mammalian genomic sequences. Annotation errors present a real problem in getting accurate information about eukaryotic gene functional signals from nucleotide sequence databases, such as GenBank or EMBL (Benson *et al.*, 1999; Kanz *et al.*, 2005). The authors generated a spliced construct for every splice pair combining 40 nt. of the left exon and 40 nt. of the right exon producing the same sequence as the splicing machinery generated by removing intron region (Figure 4.3). To verify the extracted splice sites, the alignments of splice constructs with known mammalian expressed sequence tags (ESTs (Boguski *et al.*, 1993) were used. For 43 427 pairs of donor and acceptor splice sites (splice pairs), 1215 were annotated as nonstandard donor sites (2.80 %) and 1027 were annotated as nonstandard acceptor sites (2.36 %). 41 767 splice pairs (96.18 %) contained the standard splice-site pair GT−AG. As a result of the analysis, from 1660 noncanonical pairs, 441 were supported by ESTs (27.35 %) and just 292 (18 %) were supported by ESTs after removing potential annotation errors and cases with ambiguities in the position of the splice junction (Table 4.3).

It is interesting to note that the EST-supported rate is clearly higher for canonical splice pairs. There were 22 374 out of 42 212 canonical pairs supported by ESTs (53.63 %) and just 27.3 % of noncanonical pairs. About a half (43.15 %) of all noncanonical splice pairs belongs to the GC−AG group (126). The next biggest in size of the noncanonical group GG−AG contains significantly less cases (12). There were many other groups in the same size range, including those processed by the special splicing machinery, the AT−AC group. Weight matrices for GT−AG and GC−AG pairs are presented in Table 4.4 and a consensus sequence for AT−AC pair in Figure 4.4.

Most other noncanonical splice pairs have a canonical conserved dinucleotide shifted by one base from the annotated splice junction. For example, for 12 EST-supported GG−AG pairs, 10 have a shifted canonical donor splice site with the GT dinucleotide, 1 major noncanonical site has the GC dinucleotide, and one GA−AG case (Figure 4.5).

One is prompted to explain the observations with the shifted canonical dinucleotides by an annotation error of inserting/deleting one nucleotide that is actually absent/present in real genomic sequence. This hypothesis was tested by comparing human gene sequences deposited to GenBank earlier with the sequences of the same region obtained in high throughput genome sequencing projects. Several examples of clear annotation and sequencing errors identified by the comparison are presented in Figure 4.6. We found 88
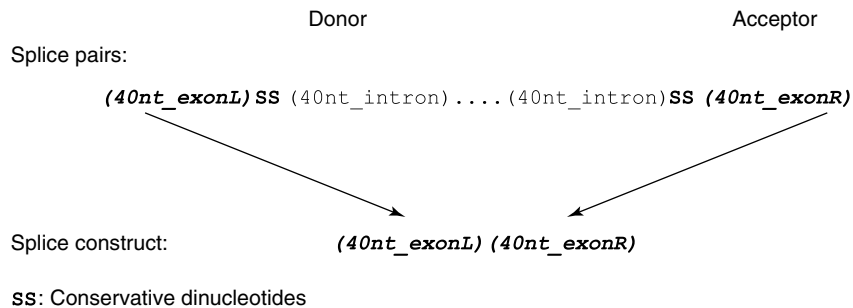
Figure 4.3 Structure of spliced constructs. Two sequence regions of a splice pair (marked as Donor and Acceptor) with the corresponding splice-site dinucleotides surrounded by 40 nt. of gene sequence at each side. Joining exon part of donor (40nt_ exonL) and exon part of acceptor (40nt_exonR) produce a sequence of splice construct to be verified by ESTs.

Table 4.3   Canonical and noncanonical splice sites in mammalian genomes.
(a) Annotated splice pairs.

Splice sites	Donors	Acceptors	Pairs
Canonical	42160 (97.28 %)	42344 (97.71 %)	41722 (96.27 %)
Noncanonical	1177 (2.72 %)	993 (2.29 %)	1615 (3.73 %)
EST - supported canonical	22437 (98.34 %)	22568 (98.92 %)	22374 (98.07 %)
EST - supported noncanonical	378 (1.66 %)	247 (1.08 %)	441 (1.93 %)
EST - supported canonical after correction	22306 (**98.94 %**)	22441 (**99.54 %**)	22253 (**98.70 %**)
EST - supported noncanonical after correction	239 (**1.06 %**)	104 (**0.46 %**)	292 (**1.30 %**)

(b) Generalization of analysis of human noncanonical splice pairs.

GT−AG	22310	99.20 %
GC−AG	140	0.62 %
AT−AC	18	0.08 %
Other Noncanonical	7	0.03 %
Errors	14	0.06 %
TOTAL	**22489**	**100 %**

AT-AC group:

```
AC002397   TGCCAAGATG| atatccttgtgt    ctgtctgctcac |CTTGGAGAAG
AC004976   GAAAGAACCC| atatcctttctg    actacttcatac |AAAACAGTCA
AF136179   TATGGTAGAG| atatcctttact    actgtttcggac |ATTGACCAAA
AL021578   ACGCTGAACC| atatcctttggg    ttaaccgctcac |TGGCCCAGCT
L10295     ATTGGTGAAG| atatccttttag    aatcattactac |ATGTGAATCC
U39892     AGATTAGAGA| atatcctttctt    aactgccagcac |ATTTTGTCAG
U47924     TGCCAAGATG| atatccttctgc    aaccctcctcac |CTTGGAGAAG
U53004     GGAAGTGGTC| atatccttcctg    aactctgcacac |GAAGCTCACG
```

Consensus of donor site:

G_{50}|$A_{100} T_{100} A_{100} T_{100} C_{100} C_{100} T_{100} T_{100} T_{62}$

Consensus of acceptor site:

$C_{62} T_{75} T_{37} C_{75} T_{37} C_{62} A_{100} C_{100}$|$A_{50} T_{62}$

Figure 4.4   Consensus sequences for the AT−AC pair of the alternative splicing machinery.

examples of independent gene sequencing with sequences overlapping splice junctions. All human EST-supported GC−AG cases having HTS matches were supported by them (39 cases). 31 errors damaging the standard splice pairs were found. 7 cases had one or both intronic GenBank sequences completely unsupported by HTS, 8 cases had intronic GenBank sequences supported, but there was a gap between exonic and intronic parts and finally 16 cases had small errors as some insertions, deletions or substitutions. 5 AT−AC pairs (3 pairs were correctly annotated in original noncanonical set and 2 were recovered from errors) were identified. In additition, 2 cases were annotated as introns, but in HTS the exonic parts were continuous (accession numbers: U70997 and M13300). 7 cases of HTS were themselves GenBank sequences and for this reason they were excluded from the analysis.

Table 4.4 Characteristics of major splice-pair groups.

GT−AG group. Number of supported cases: 22 268

Donor frequency matrix

A	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0
C	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5
G	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9
U	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2

Acceptor frequency matrix

A	9.0	8.4	7.5	6.8	7.6	8.0	9.7	9.2	7.6	7.8	23.7	4.2	100	0.0	23.9
C	31.0	31.0	30.7	29.3	32.6	33.0	37.3	38.5	41.0	35.2	30.9	70.8	0.0	0.0	13.8
G	12.5	11.5	10.6	10.4	11.0	11.3	11.3	8.5	6.6	6.4	21.2	0.3	0.0	100	52.0
U	42.3	44.0	47.0	49.4	47.1	46.3	40.8	42.9	44.5	50.4	24.0	24.6	0.0	0.0	10.4

GC−AG group. Number of supported cases: 126

Donor frequency matrix

A	40.5	88.9	1.6	0.0	0.0	87.3	84.1	1.6	7.9
C	42.1	0.8	0.8	0.0	100	0.0	3.2	0.8	11.9
G	15.9	1.6	97.6	100	0.0	12.7	6.3	96.8	9.5
U	1.6	8.7	0.0	0.0	0.0	0.0	6.3	0.8	70.6

Acceptor frequency matrix

A	11.1	12.7	3.2	4.8	12.7	8.7	16.7	16.7	12.7	9.5	26.2	6.3	100	0.0	21.4
C	36.5	30.9	19.1	23.0	34.9	39.7	34.9	40.5	40.5	36.5	33.3	68.2	0.0	0.0	7.9
G	9.5	10.3	15.1	12.7	8.7	9.5	16.7	4.8	2.4	6.3	13.5	0.0	0.0	100	62.7
U	38.9	41.3	58.7	55.6	42.1	40.5	30.9	37.3	44.4	47.6	27.0	25.4	0.0	0.0	7.9

	Donor	Acceptor	Donor + 1 shift
U07083	**AAGGG|** ggtaagg	ctttaag **|GGTGT**	GG ⇒ GT
X98208	**CGGCA|** ggtcaga	aatgcag **|GTGTA**	GG ⇒ GT
L43831	**CAAAG|** ggtactg	tctgcag **|CTTTG**	GG ⇒ GT
U37431	**AAACA|** ggtcagt	gccccag **|GGGAA**	GG ⇒ GT
U02978	**AGGCC|** ggtgagt	gggccag **|GGGTC**	GG ⇒ GT
AJ000060	**AGTAT|** ggtaagg	tttccag **|GGAGA**	GG ⇒ GT
U12599	**GCTGG|** ggtaagt	tccccag **|TCATA**	GG ⇒ GT
U01247	**TCACA|** ggtatgc	attctag **|GAGAA**	GG ⇒ GT
U28721	**CGCAG|** ggcaagg	ctaacag **|GTCTA**	GG ⇒ **GC**
M20214	**AACAG|** ggaaggc	acgctag **|GGAAA**	GG ⇒ **GA**
M62601	**TGCAG|** ggtatac	cctttag **|ACAAT**	GG ⇒ GT
U66878	**TAGTG|** ggtgagt	ccttcag **|GAGTG**	GG ⇒ GT

Figure 4.5 Shifted splice sites. Example for GG−AG verified splice sites (12 cases). In donor, exactly after the cut point was always found a GG pair. To obtain which splicing pair are characteristic to this donor we should produce a shift of 1 nucleotide downstream. After this we reclassify sites as 10 GT−AG canonical splice sites, 1 GC−AG site and 1 apparently strange GA−AG site.

By generalizing these results we conclude that the overwhelming majority of splice sites contain the conserved dinucleotides GT−AG (99.2 %). The other major group includes GC−AG pairs (0.62 %), the alternative splicing mechanism group AC−AT (about

```
Sequences of homeodomain protein, HOXA9EC (AF010258)
                        Donor                        Acceptor
Genbank:         CGATCCCAAT|  aa-tgtctcct       cccgcagaat  |AACCCAGCAG
High throughput: CGATCCCA|  gtaagtgtctcct       cccgcag  |AT-AACCCAGCAG


Sequences of  poly(A) binding protein II, PABP2 (AF026029)
                        Donor                        Acceptor
Genbank:         TCCAGGCAAT|  gctgagtaac       tttcctgata  |GCTGGCCCGG
High throughput: TCCAGGCAATG|  gtgagtaac       tttcctgatag  |CTGGCCCGG
```

Figure 4.6 Errors found by comparing GenBank and the human high-throughput sequences for several annotated noncanonical splice sites.

0.08 %) and a very small number of other noncanonical splice sites (about 0.03 %) (Table 4.3.d). Therefore, gene-finding approaches using only standard GT−AG splice sites can potentially predict 97 % genes correctly (if we assume 4 exons per gene, on average). Including the GC−AG splice pair will increase this level to 99 %. 22 253 verified examples of canonical splice pairs were presented in a database (SpliceDB), which is available for public use through the www (`http://www.softberry.com/berry.phtml?topic=splicedb&group=data&subgroup=spldb`) (Burset *et al.*, 2000). It also includes 1615 annotated and 292 EST-supported and shift-verified noncanonical pairs. This set can be used to investigate the reality of these sites as well as to further understand the splicing machinery.

Analysis of splice-site sequences demonstrated that their consensuses are somewhat specific for different classes of organisms (Senapathy *et al.*, 1990; Mount, 1993) and some important information is encoded by the sequences outside the short conserved regions. Scoring schemes based on consensus sequences or weight matrices which take into account the information about open reading frames, free energy of base pairing of RNA with snRNA and other peculiarities, give an accuracy of about 80 % for the prediction splice-site positions (Nakata *et al.*, 1985; Gelfand, 1989). More accurate prediction is produced by neural network algorithms (Lapedes *et al.*, 1988; Brunak *et al.*, 1991; Farber *et al.*, 1992). The integral view on the difference of triplet composition in splice and pseudosplice sequences is shown in Figure 4.7. This figure demonstrates the various functional parts of splice sites. We can see that the only short regions around splice junctions have a great difference in triplet composition. Their consensus sequences are usually used as determinants of donor or acceptor splice-site positions. However, dissimilarity in many other regions can also be seen. For the donor site – coding region, a G-rich intron region may be distinguished. For acceptor sites – a G-rich intron region, a branch point region, a polyT/C tract and coding sequence. Splice-site prediction methods using a linear function that combines several of such features is described below (Solovyev and Lawrence, 1993a; Solovyev *et al.*, 1994).

4.4.2   Donor Splice-site Characteristics

Seven characteristics were selected for donor splice-site identification. Their values were calculated for 1375 authentic donor site and for 60 532 pseudosite sequences from the learning set. The Mahalonobis distances showing the significance of each characteristic are given in Table 4.5. The strongest characteristic for donor sites is a triplet composition in consensus region ($D^2 = 9.3$) followed by the adjacent intron region ($D^2 = 2.6$) and the
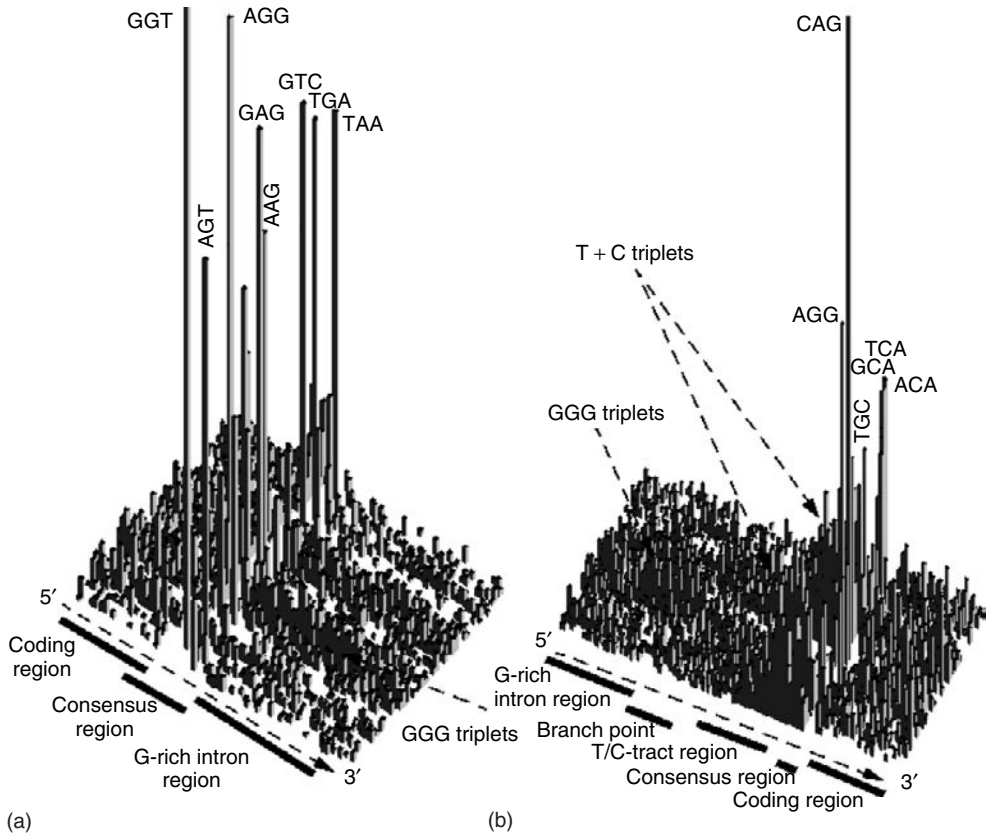
Figure 4.7 Difference of the triplet composition around donor and GT-containing non-donor sequences (a); around acceptor and AG-containing non acceptor sequences (b) in 692 human genes. Each column presents the difference of specific triplet numbers between sites and pseudosites in a specific position. For comparison the numbers were calculated for equal quantities of sites and pseudosites.

Table 4.5   Significance of various characteristics of donor splice sites.

Characteristics	1	2	3	4	5	6	7
Individual D^2	9.3	2.6	2.5	0.01	1.5	0.01	0.4
Combined D^2	9.3	11.8	13.6	14.9	15.5	16.6	16.8

1, 2, 3 are the triplet preferences (13) of consensus (-4 to $+6$), intron G-rich ($+7$ to $+50$) and coding regions (-30 to -5), respectively. 4 is the number of significant triplets in the consensus region. 5 and 6 are the octanucleotide preferences for being coding 54 bp region on the left and for being intron 54 bp region on the right of donor splice-site junction. 7 is the number of G bases, GG doublets and GGG triplets in $+6$ to $+50$ intron G-rich region.

coding region ($D^2 = 2.5$). Other significant characteristics are the number of significant triplets in conserved consensus region; the number of G bases, GG doublets and GGG triplets; and the quality of the coding and intron regions.

Rigorous testing of several splice-site prediction programs on the same sets of new data demonstrated that the linear discriminant function (implemented in SPL program: `http://www.softberry.com/berry.phtml?topic=spl&group=programs&subgroup=gfind`) provides the most accurate local donor site recognizer (Table 4.6) (Milanesi and Rogozin, 1998).

Although a simple weight matrix provides less accurate recognition than more sophisticated approaches, it can be easily recomputed for new organisms and is very convenient to use in probabilistic HMM-gene-prediction methods. An interesting extension of this approach was suggested on the basis of analysis of dependencies between splice-site positions (Burge and Karlin, 1997). Using a maximal dependence decomposition procedure (Burge, 1998), 5 weight matrices corresponding to different subsets of splice-site sequences were generated. The subclassification of donor signals and the matrices constructed based on 22 306 EST-supported splice sites are presented in Figure 4.8. Performance of these matrices compared with other methods was evaluated on the Burset and Guigo (1996) data set (Figure 4.9). We can observe that several weight matrices definitely provide better splice-site discrimination than just one. However, their discriminatory power is similar to that of the matrix of triplets and lower than that of the linear discriminant function described above.

4.4.3 Acceptor Splice-site Recognition

Seven characteristics were selected for acceptor splice-sites recognition. Their values were calculated for 1386 authentic acceptor site and 89 791 pseudosite sequences from the learning set. The Mahalonobis distances showing the individual significance for each characteristic are given in Table 4.7. The strongest characteristics for acceptor sites are the triplet composition in the polyT/C tract region ($D^2 = 5.1$); consensus region ($D^2 = 2.7$); adjacent coding region ($D^2 = 2.3$); and branch point region ($D^2 = 1.0$). Some significance is found using the number of T and C in the adjacent intron region ($D^2 = 2.4$) and the quality of the coding region ($D^2 = 2.6$).

Table 4.8 illustrates the performance of different methods for acceptor site recognition (Milanesi and Rogozin, 1998). The linear discriminant function described above provides the best accuracy. Also, we can observe that acceptor site recognition accuracy is lower than that for donor sites.

It was shown that the first-order Markov chain model (11) based on dinucleotide frequencies of $[-20, +3]$ acceptor site region gives slightly better discrimination than the simple weight matrix model (Burge, 1998). Such a model was incorporated in

Table 4.6 Comparing the accuracy of local donor splice-site recognizers. The accuracy is averaged for 3 tested sets.

Method	False positives (%)	False negatives (%)	CC	Reference
Weight matrix	2.3	53	0.13	Guigo *et al.* (1992)
Consensus MAG/GURAGU	6.0	18	0.27	Mount (1982)
Five consensuses	4.2	15	0.31	Milanesi and Rogozin (1998)
Neural network	25.0	2.7	0.51	Brunak *et al.* (1991)
Discriminant analysis	10.0	3.0	0.56	Solovyev *et al.* (1994)

	−4	−3	−2	−1	1	2	3	4	5	6	7	8	9	10
A	27.6	34	60.5	9.2	0	0	52.5	71.3	7.1	16.1	27.4	20.8	20.1	19.9
T	28.7	36.2	12.8	3.3	0	0	2.8	7.6	5.5	16.6	21.1	27.1	28.9	25.5
G	24	18.3	12.4	80.3	100	0	42.2	11.8	81.5	21	32.7	25.5	26.7	29.3
C	19.7	11.4	14.3	7.2	0	100	2.5	9.3	5.9	46.3	18.7	26.5	24.2	25.3

G_5H_{-1}
4029

	A	T	G	C
−3	32.2	29.1	21.6	17.1
−2	42.9	30.9	17.2	9.1
−1	46.6	16.6	0	36.8
3	54.5	0.4	44.7	0.4
4	90.4	2.9	3	3.6
6	5.6	7.9	6.3	80.2

H_5
3930

	A	T	G	C
−3	39.4	40.9	14.2	5.5
−2	83.6	4.8	4.9	6.7
−1	1.9	0.7	95.9	1.5
3	79.7	1.3	16.8	2.2
4	53.2	24.6	9.6	12.7
5	38.1	30	0	31.9
6	21	18.3	30.8	29.8

$G_5G_{-1}B_{-2}$
5474

	A	T	G	C
−3	31.4	28.9	16.9	22.8
−2	0	23.6	32.2	44.3
3	45.3	2	51.8	0.9
4	81.5	3.1	8	7.4
6	14.2	16.5	18.1	51.2

$G_5G_{-1}A_{-2}V_6$
5183

	A	T	G	C
−3	34.4	43.4	19.3	2.9
3	47	4.4	46	2.5
4	67	4.7	18.5	9.8
6	30.9	30.5	38.5	0

$G_5G_{-1}A_{-2}T_6$
2636

	A	T	G	C
−3	34.3	39.8	21	5
3	36.2	7.2	47	9.7
4	56	4.6	23.1	16.2

Figure 4.8 Classification of donor splice sites by several weight matrices reflecting different splice-site groups (Burge and Karlin, 1997).

Genscan gene-prediction method (Burge and Karlin, 1997). Thanaraj (2000) performed a comprehensive analysis of computational splice-site identification. The HSPL program remains the best local recognizer. Of course, most complex gene-prediction systems use a lot of other information about optimal exon (or splice site) combinations that provides a better level of accuracy. However it cannot be applied to study the possible spectrum of all alternative splice sites for a particular gene. Local recognizers seem useful for such tasks.

4.5　IDENTIFICATION OF PROMOTER REGIONS IN HUMAN DNA

Computational recognition of eukaryotic polymerase II (PolII) promoter sequences in genomic DNA is an extremely difficult problem. Promoter 5′-flanking regions may contain dozens of short motifs (5–10 bases) that serve as recognition sites for proteins providing initiation of transcription as well as specific regulation of gene expression. Each promoter has its own composition and arrangement of these elements providing a unique regime
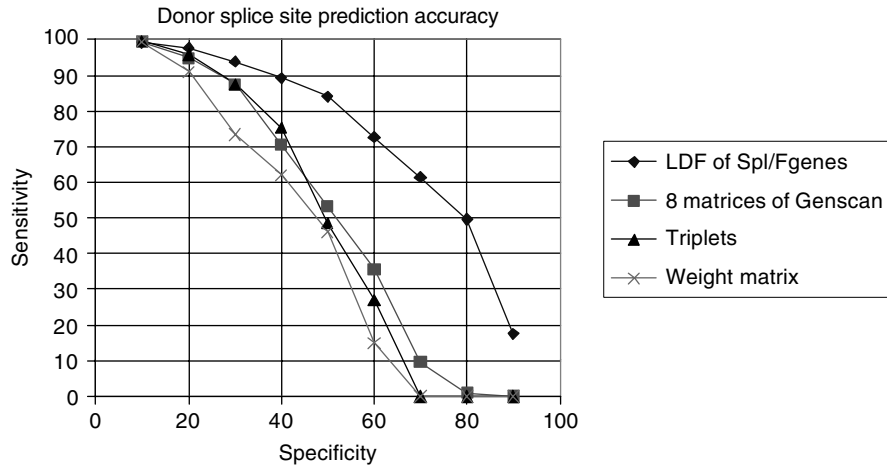
Figure 4.9 Comparison of the accuracy of donor splice-site recognizers: single weight matrix, five matrices suggested by Burge and Karlin (1997), matrix of triplets, linear discriminant function.

Table 4.7 Significance of various characteristics of acceptor splice sites.

Characteristics	1	2	3	4	5	6	7
Individual D^2	5.1	2.6	2.7	2.3	0.01	1.05	2.4
Combined D^2	5.1	8.1	10.0	11.3	12.5	12.8	13.6

1, 3, 4, 6 are the triplet preferences (13) of (-33 to -7) polyT/C tract, consensus (-6 to $+5$), coding ($+6$ to $+30$) and branch point (-48 to -34) regions, respectively. 7 is the number of T and C in intron polyT/C tract region. 2 and 5 are the octanucleotide preferences for being coding 54 bp region on the left and 54 bp region for being intron on the right side of donor splice-site junction.

Table 4.8 Comparing the accuracy of local acceptor splice-site recognizers. The accuracy is averaged for 3 tested sets.

Method	False positives (%)	False negatives (%)	CC	Reference
Weight matrix	5.0	20	0.22	Guigo *et al.* (1992)
Neural network	16.3	6.7	0.35	Brunak *et al.* (1991)
Discriminant analysis	22.0	2.3	0.51	Solovyev *et al.* (1994)

of gene expression. Here we will consider some general features of PolII promoters that can be exploited in promoter prediction programs.

The minimal promoter region that is capable of initiating basal transcription is referred to as the core promoter. It contains a transcription start site (TSS), often located in initiator region (Inr) and typically spans from -60 to $+40$ bp relative to TSS. About 30–50 % of all known promoters also contain a TATA box at a position about 30 bp upstream from TSS. The TATA box is apparently the most conserved functional signal in eukaryotic promoters. In some cases it can direct accurate transcription initiation by PolII, even in the absence of other control elements. Many highly expressed genes contain a

strong TATA box in their core promoter. However for large groups of genes, like most housekeeping genes, some oncogenes and growth factor genes, a TATA box is absent and the corresponding promoters are referred as *TATA-less promoters*. In these promoters, Inr may control the exact position of the transcription start point or the recently found downstream promoter element (DPE), usually located 30 bp downstream of TSS. Many human genes are transcribed from multiple promoters often producing alternative first exons. Moreover, transcription initiation appears to be less precise than initially assumed. In the human genome, it is not uncommon that the 50 ends of mRNAs transcribed from the same promoter region are spread over dozens or hundreds bp (Suzuki *et al.*, 2001; Cooper *et al.*, 2006; Schmid *et al.*, 2006).

The region 200–300 bp immediately upstream of the core promoter constitutes the proximal promoter. The proximal promoter usually contains multiple transcription factor binding sites that are responsible for transcription regulation. Further upstream is the distal part of promoter that may also contain transcription factor binding sites as well as enhancer elements. The typical organization of PolII promoters is shown in Figure 4.10. Because the distal part is usually the most variable region of promoters and generally poorly described, most current promoter recognition tools use the characteristics of only the core and/or proximal regions. Comprehensive reviews of eukaryotic promoters, specifically written from the prediction point of view, have appeared in the literature (Werner, 1999; Pedersen *et al.*, 1999; Cooper *et al.*, 2006).

A collection of experimentally mapped TSSs and surrounding sequences called *Eukaryotic Promoter Database* (EPD) was created by Bucher and Trifonov (1986). In 2000, this database comprised about 800 independent promoter sequences including about 150 human promoters. There is no information about specific regulatory features in this database (Perier *et al.*, 2000). Up to release 72 (October 2002) EPD was a manually compiled database, relying exclusively on experimental evidence published in scientific journals. With release 73, they started to exploit 5′-ESTs from full-length cDNA clones
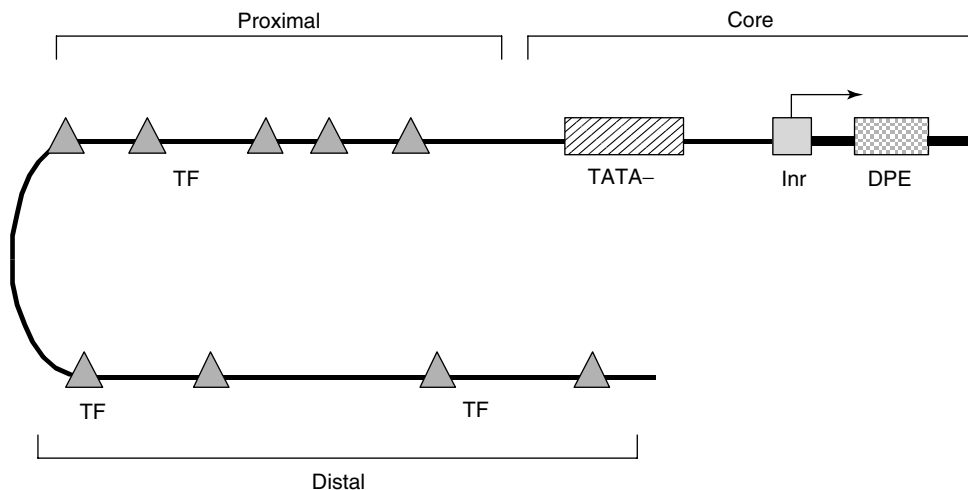
Figure 4.10 Schematic organization of polymerase II promoter. Inr – initiator region, usually containing TSS. DPE – downstream promoter element, often appearing in TATA-less promoters, TF binding sites – transcription factor binding sites.

as a new resource for defining promoters. These data are automatically processed by computer programs and already a year after the introduction of this new method, more than half of the EPD entries (1634) are based on 5′-EST sequences (Schmid *et al.*, 2004). EPD is not the only database providing information about experimentally mapped TSSs. DBTSS (Suzuki *et al.*, 2004) and PromoSer (Halees *et al.*, 2003) are large collections of mammalian promoters based on clustering of EST and full-length cDNA sequences. These databases define the TSS as the furthest 50 position in the genome which can be aligned with the 50 end of a cDNA from the corresponding gene. In contrast, EPD considers the most frequent cDNA 50 end as the TSS and further applies a specialized algorithm to infer multiple promoters for a given gene (Schmid *et al.*, 2006). There is a plant-specific PlantProm (Shahmuradov *et al.*, 2003) database of promoters based on published TSS mapping data.

Regulatory promoter elements are relatively short sequence motifs (typically 5−15 bp in length) (Wingender, 1988; Tjian, 1995; Fickett and Hatzigeorgiou, 1997). A relational TFD including collection of regulatory factors and their binding sites was created by Ghosh (1990; 2000). Over 7900 sequences of transcriptional elements have been described in TRANSFAC database (Wingender *et al.*, 1996; Matys *et al.*, 2006). It gives information about localization and sequence of individual regulatory elements within gene and transcription factors, which bind to them. RegsiteDB (Plant) contains about 1300 various regulatory motifs of plant genes and detail descriptions of their functional properties (http://www.softberry.com/berry.phtml?topic=regsite). Practically it is difficult to use most of these motifs to annotate long genomic sequences, because of their short length and degenerate nature. For example, even using well described the TATA box weight matrix there exists one false positive every 120−130 bp (Prestridge and Burks, 1993). Nevertheless, such resources are invaluable for detail analysis of gene regulation and interpretation of experimental data.

To generate regulatory diversity of gene expression, combinations of simple motifs can be used. Transcription factors and regulatory sequences are composed of modular components to achieve the high level of specificity by a relatively small number of different transcription factors (Tjian and Maniatis, 1994). Therefore, to understand gene function we should concentrate our attention on patterns of regulatory sequences rather than on single elements. Searching for such patterns should be much more effective in annotation of new sequences compared to the poor recognition of single motifs. The simplest examples of regulatory patterns are observed in composite regulatory elements (CE) of vertebrate promoters. Composite elements are modular arrangements of contiguous or overlapping binding sites for various distinct factors, raising the possibility that the bound regulatory factors may interact directly, producing novel patterns of regulation. For example, the composite element of proliferin promoter comprises glucocorticoid receptor (GR) and AP-1 factor binding sites. Both GR and AP-1 are expressed in most cell types, but the composite element demonstrated remarkable cell specificity: the hormone−receptor complex repressed the reporter gene expression in CV-1 cells, but enhanced its expression in HeLa cells and had no effect in the F9 cell (Diamond *et al.*, 1990). The database of composite elements (COMPEL) was set up as a common effort by the groups of Wingender (Germany) and Kolchanov (Russia) (Kel *et al.*, 1995). Currently the compilation contains information about several hundred experimentally identified composite elements, where each element consists of two functionally linked sites.

Development of the Transcription Regulatory Regions Database (TRRD), which describes observed regulatory elements in gene regulatory regions, was started in the Kolchanov group (Russia) in 1994 (Kolchanov *et al.*, 2000). TRRD was created by scanning the literature and covers just a fraction of genes taking into account the rather limited resources and very complex nature of the problem of comprehensive and accurate annotation. We are faced with exponentially growing data on transcriptional control owing to the advancement of experimental technologies. This requires us to unite efforts and expertise in creating knowledge databases in this field.

In one of the first attempts to predict eukaryotic promoters, Prestridge (1995) used the density of specific transcription factor binding sites in combination with the TATA box weight matrix. The program PROMOTERSCAN uses the promoter preferences for each binding site listed in TFD (Ghosh, 1990) previously calculated on the set of promoter and nonpromoter sequences. The other general-purpose promoter recognition tools take into account the oligonucleotide content of promoter sequences (Hutchinson, 1996; Audic and Claverie, 1997; Knudsen, 1999; Ohler *et al.*, 1999). In an earlier version of the linear discriminant recognizer, the signal-specific (TATA box weight matrix, binding site preferences) and content-specific characteristics (hexamer preferences) were combined for recognition of TSS (Solovyev and Salamov, 1997).

Fickett and Hatzigeorgiou (1997) presented a performance review of many general-purpose promoter prediction programs. Among these were oligonucleotide content-based (Hutchinson, 1996; Audic and Claverie, 1997), neural network (Guigo *et al.*, 1992; Reese *et al.*, 1996) and the linear discriminant approaches (Solovyev and Salamov, 1997). Although several problems were identified through the relatively small test set (18 sequences) (Ohler *et al.*, 1999), the results demonstrated that the programs can recognize just 50 % of promoters with false-positive rate about 1 per 700–1000 bp. If the average size of a human gene is more than 7000 bases and many genes occupy hundreds of kilobases, then we will expect significantly more false-positive predictions than the number of real promoters. However, these programs can be used to find promoter position (start of transcription and TATA box) in a given 5′-region or to help selecting the correct 5′-exons in gene-prediction approaches.

We will describe a current version of the promoter recognition program TSSW (Transcription Start Site, W stands for using functional motifs from the Wingender *et al.* (1996) database) (Solovyev and Salamov, 1997) to show sequence features that can be used to identify eukaryotic promoter regions. In this version, it was suggested that TATA+ and TATA− promoters have very different sequence features and these groups were analyzed separately. Potential TATA+ promoter sequences were selected by the value of score computed using the TATA box weight matrix (Bucher, 1990) with the threshold closed to the minimal score value for the TATA+ promoters in the learning set. Such a threshold divides the learning sets of known promoters into approximately equal parts. Significant characteristics of both groups found by discriminant analysis are presented in Table 4.9. This analysis demonstrated that TATA− promoters have much weaker general features compared with TATA+ promoters. Probably TATA− promoters possess more gene-specific structure and they will be extremely difficult to predict by any general-purpose method.

The TSSW program classifies each position of a given sequence as TSS or non-TSS based on two linear discriminant functions (for TATA+ and TATA− promoters) with characteristics calculated in the (−200, +50) region around a given position. If the TATA

Table 4.9 Significance of characteristics of promoter sequences used by TSSW programs for identification of TATA+ and TATA− promoters.

Characteristics	D^2 for TATA+ promoters	D^2 for TATA− promoters
Hexaplets −200 to −45	2.6	1.4 (−100 to −1)
TATA box score	3.4	0.9
Triplets around TSS	4.1	0.7
Hexaplets +1 to +40		0.9
Sp1-motif content		0.9
TATA fixed location	0.7	
CpG content	1.4	0.7
Similarity −200 to −100	0.3	0.7
Motif Density(MD) −200 to +1	4.5	3.2
Direct/Inverted MD −100 to +1	4.0	3.3 (−100 to −1)
Total Mahalonobis distance	11.2	4.3
Number promoters/nonpromoters	203/4000	193/74 000

box weight matrix gives a score higher than some threshold, then the position is classified based on LDF for TATA+ promoters, otherwise the LDF for TATA-less promoters is used. Only one prediction with the highest LDF score is retained within any 300 bp region. If we observe a lower scoring promoter predicted by the TATA-less LDF near a higher scoring promoter predicted by TATA+ LDF, then the first prediction is also retained as a potential enhancer region.

The recognition quality of the program was tested on 200 promoters, which were not included in the learning set. We provide the accuracy values for different levels of true predicted promoters in Table 4.10. The data demonstrate a poor quality of TATA− promoter recognition on long sequences and show that their recognition function can provide relatively unambiguous predictions within regions less than 500 bp. Contrarily, 90 % of TATA+ promoters can be identified within the range 0–2000 bp that makes their incorporation into gene-finding programs valid.

Ohler *et al.* (1999) used interpolated Markov chains in their approach and slightly improved the previous results. They identified 50 % in Fickett and Hatzigeorgiou (1997) promoter set, while having one false-positive prediction every 849 bp. Knudsen (1999), applying a combination of neural networks and genetic algorithms, designed another

Table 4.10 Performance of promoter identification by TSSW program.

Type of promoter	Number of test sites	True predicted (%)	1 false positive per bp
TATA+	101	98	1000
		90	2200
		75	3400
		52	6100
TATA−	96	52	500
		40	1000

program (Promoter2.0). Promoter 2.0 was tested on a complete Adenovirus genome 35 937 bases long. The program predicted all 5 known promoter sites on the plus strand and 30 false-positive promoters. The average distance between a real and the closest predicted promoter is about 115 bp. The TSSW program with the threshold to predict all 5 promoters produced 35 false positives. It gives an average distance between predicted TSS and real promoter of just 4 bp (2 predicted exactly, 1 with 1 bp shift, 1 with 5 bp shift and the weakest promoter was predicted with 15 bp shift).

Figure 4.11 shows an example of the results of the TSSW program for the sequence of human laminin beta 2 chain (GenBank accession number Z68155). The structure of this gene including its promoter region has been extensively studied. The length of gene is 11 986 bp, the first 1724 bp of which constitute a promoter region. TSSW predicts one enhancer at position 931 and one potential TSS at position 1197 with corresponding TATA box at the position 1167. Although both the predicted sites fall inside the designated promoter region, the second prediction is probably a false positive, because the

```
>H.sapiens LAMB2 gene for    laminin beta 2 chain
 Length of sequence    -     11986bp
 Thresholds for TATA+ promoters     -  0.45, for TATA-/enhancers -  3.70

 2 promoter/enhancer(s) are predicted

 Enhancer Pos :    931 LDF score - 3.78
 Promoter Pos :   1197 LDF score - 1.13   TATA box at   1167   Score -
18.96

 Transcription factor binding sites:
 for enhancer  at position -     931
   874 (+) CHICK$ACRA    CCGCCC
   778 (-) Y$ADH2_01     TCTCC
   631 (-) Y$ADH2_01     TCTCC
   831 (+) RAT$ANTEN_    ccacagttgggatttCCCAACctgaccag
   842 (+) RAT$ANTEN_    ccacagttgggatttCCCAACctgaccag
   879 (-) HS$APOE_08    GGGCGG
   876 (-) Y$CYC1_09     ctcatttggcgagcGTTGGt
   865 (-) Y$CYC1_09     ctcatttggcgagcGTTGGt
   842 (-) Y$CYC1_09     ctcatttggcgagcGTTGGt
   657 (+) AD$E2L_04     TGACgcA
   833 (+) AD$E2L_04     TGACgcA
   835 (-) HS$EGFR_15    TCAAT
   840 (-) RAT$EAI_09    GTCAG
   649 (+) Y$GAL1_10     AGCCT
   929 (-) MOUSE$AAG_    gcaacTGATAaggat
   928 (-) MOUSE$AAG_    cctgTGATAagga
   841 (+) HS$BG_01      ccaCACCCg
   852 (+) HS$BG_01      ccaCACCCg
   863 (+) HS$BG_01      ccaCACCCg
   907 (-) HS$BG_01      ccaCACCCg
   773 (-) HS$BG_01      ccaCACCCg
   661 (-) HS$BG_01      ccaCACCCg
```

Figure 4.11 An example of output of the TSSW program for the sequence of human laminin beta 2 chain (GenBank accession number Z68155).

predicted TATA box is located far upstream (500 bp) from the experimentally determined beginning of the 5′-UTR. TSSW also optionally lists all potential TF binding sites around the predicted promoters or enhancers (Figure 4.11). It outputs the position, the strand (+/−), the TRANSFAC identifier and the consensus sequences of sites found. The information about these sites may be of interest for researchers studying the transcription of a particular gene.

There is a high false-positive rate of promoter prediction in long genomic sequences. It is more useful to remove some false-positive predictions using knowledge of the positions of the coding regions. TSSW was additionally tested on the several GenBank entries that have information about experimentally verified TSS and were not included in the learning set (Table 4.11). The lengths of the sequences varied from 950 to 28 438 bp with a median length of 2938 bp. According to the criteria defined by Fickett and Hatzigeorgiou (1997), all true TSS in these sequences can be considered as correctly predicted, with an average 1.5 false positives per sequence or 1 false positive per 3340 bp. The distances between true TSS and those correctly predicted varied from exact matching to 196 bp, with the median deviation of 9 bp. This can be considered to be quite a good prediction taking into account that experimental mapping of TSS has an estimated precision of +/− 5 bp (Perier *et al.*, 2000).

Accurate prediction of promoters is fundamental to understanding gene expression patterns, where confidence estimation of the produced predictions is one of the main requirements for many applications. Using recently developed transductive confidence machine (TCM) techniques, we developed a new program TSSP-TCM (Shahmuradov *et al.*, 2005) for the prediction of plant promoters that also provides confidence of the prediction. The method presented in the paper identifies ∼85 % of tested promoters with one false positive per ∼5000 bp. It allows us not only to make predictions, but more importantly, it also gives valid measures of confidence in the predictions for each individual example in the test set. Validity in our method means that if we set up a confidence level, say, 95 %, then we can guarantee that we are not going to have more than 5 errors out of 100 examples.

Recently there was an attempt to make a critical assessment of the promoter prediction accuracy in its current state relative to the manual Havana gene annotation (Bajic *et al.*, 2006). There were only 4 programs in this EGASP project: 2 variants of McPromoter

Table 4.11 Results of TSSW predictions on some GenBank entries with experimentally verified TSS.

Gene	GenBank accession number	Length (bp)	True TSS	Predicted TSS	Number of false positives
CXCR4	AJ224869	8747	2632	2631	4
HOX3D	X61755	4968	2280	2278	2
DAF	M64356	2003	733	744	1
GJB1	L47127	950	404	428	0
SFRS7	L41887	8213	< 415	417	4
ID4	AF030295	1473	1066	1081	1
C inhibitor	M68516	15 571	2200	2004	4
MBD1	AJ132338	2951	1964	1876	1
Id-3	X73428	2481	665	663	0

program (Ohler *et al.*, 1999; 2002), N-scan (Arumugam *et al.*, 2006) and Fprom (Solovyev *et al.*, 2006), which is a modification of TSSW program described above that use TFD transcriptional motif database (Ghosh, 2000). McPromoter and Fprom derived its predictions from a sequence of the genome under analysis; N-scan used corresponding sequences of several genomes (such as human, mouse and chicken). When the maximum allowed mismatch of the prediction from the reference TSS for counting true positive predictions on test sequences was 1000 bp, the N-scan produced ∼3 % higher accuracy than the next most accurate predictor Fprom, but for the distance criterion 250 bp Fprom shows the best performance on most prediction accuracy measures (Bajic *et al.*, 2006). We should note that the sensitivity of computational promoter predictions was only 30–50 % (relatively 5′-gene ends of Havana annotation), but we should take into account that TSS annotation from two experimentally derived databases also produced a sensitivity of only 48–58 %. This leaves the open issue to create a reliable TSS reference dataset. The lesson from this EGASP experiment relative to promoter predictions is that it is beneficial to combine the TSS/promoter predictions with gene-finding programs as was done in generating N-scan or Fprom predictions.

Despite recent improvements in promoter prediction programs, their current accuracy is still not enough for their successful implementation as independent submodules in gene-recognition software tools. The rather small amount of experimentally verified promoters in databases such as EPD and GenBank hindered computational promoter identification progress, while now there is an order more promoter data became available (Seki *et al.*, 2002) generated by CapTrapper technique (Carninci *et al.*, 1996) that provided a relatively reliable method for promoter identification. Many of the above-mentioned promoter prediction algorithms use the propensities of each TF binding site independently and does not take into account their mutual orientation and positioning. It is well known that the transcriptional regulation is a highly cooperative process, involving simultaneous binding of several transcription factors to their corresponding sites. Specific groups of promoters may have specific patterns of regulatory sequences, where mutual orientation and location of individual regulatory elements are necessary requirements for successful transcription initiation or regulation.

4.6   RECOGNITION OF POLYA SITES

Another functionally important signal of eukaryotic transcripts is the 3′-untranslated region (3′UTR), which has a diversity of cytoplasmic functions affecting the localization, stability and translation of mRNAs (Decker and Parker, 1995). Almost all eukaryotic mRNAs undergo 3′-end processing which involves endonucleotide cleavage followed by the polyadenylation of the upstream cleavage product (Wahle, 1995; Manley, 1995). The essential sequences are involved in the formation of several large RNA-protein complexes (Wilusz *et al.*, 1990). RNA sequences directing binding of specific proteins are frequently poorly conserved and often recognized in a cooperative fashion (Wahle, 1995). Therefore we have been forced to use statistical characteristics of the polyA regions that may involve some unknown significant sequence elements.

Numerous experiments have revealed three types of RNA sequences defining a 3′-processing site (Wahle, 1995; Proudfoot, 1991) (Figure 4.12). The most conserved
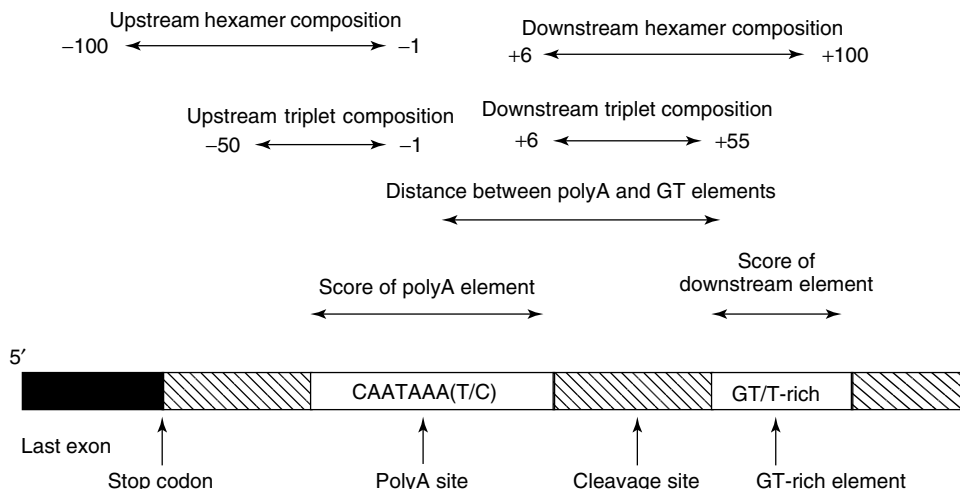
Figure 4.12   Characteristics of polyA signal sequences.

is the hexamer signal AAUAAA (polyA signal), situated 10–30 nucleotides upstream of the 3′-cleavage site. About 90 % of sequenced mRNAs have a perfect copy of this signal. Two other types, the upstream and the downstream elements, are degenerate and have not been properly characterized. Downstream elements are frequently located within approximately 50 nucleotides 3′ of the cleavage site (Wahle and Keller, 1992). These elements are often GU- or U-rich, although they may have various base compositions and locations. On the basis of sequence comparisons McLauchlan *et al.* (1985) have suggested that one of the possible consensuses of the downstream element is YGUGUUYY. The efficiency of polyadenylation in a number of genes can be also increased by sequences upstream of AAUAAA, which are generally U-rich (Wahle, 1995). All these RNA sequences serve as nucleation sites for a multicomponent protein complex catalyzing the polyadenylation reaction.

There have been a few attempts to predict 3′-processing sites by computational methods. Yada *et al.* (1994) conducted a statistical analysis of human DNA sequences in the vicinity of polyA signal in order to distinguish them from AATAAA sequences that are not active in polyadenylation (pseudo polyA signals). They found that a base C frequently appears on the upstream side of the AATAAA signal and a base T or C often appears on the downstream side, implying that CAATAAA(T/C) can be regarded as a consensus of the polyA signal. Kondrakhin *et al.* (1994) constructed a generalized consensus matrix using 63 sequences of cleavage/polyadenylation sites in vertebrate pre-mRNA. The elements of the matrix were absolute frequencies of triplets at each site position. Using this matrix, they have provided a multiplicative measure for recognition of polyadenylation regions. However this method has a very high false-positive rate.

Salamov and Solovyev (1997) developed LDF recognition function for polyA signal. The data sets for 3′-processing sites and 'pseudo' polyA signals were extracted from GenBank (Version 82). 3′-processing sites were taken from the human DNA entries, containing a description of the polyA signal in the feature table. Pseudosites were taken out of human genes as the sequences comprising (−100, +200) around the patterns revealed by polyA weight matrix (see below), but not assigned to polyA sites in the feature table.

4.7 CHARACTERISTICS FOR RECOGNITION OF 3′-PROCESSING SITES

As the hexamer AATAAA is the most conservative element of 3′-processing sites, it was considered as the main block in our complex recognition function. Although the hexamer is highly conserved, other variants of this signal were observed. For example, in the training set, 43 out of 248 polyA sites had hexamer variants of AATAAA with one mismatch. To consider such variants the position weight matrix for recognizing this signal has been used. The other characteristics such as content statistics of hexanucleotides and positional triplets in the upstream and downstream regions were defined relative to the position of the conservative hexamer sequence (Figure 4.12).

1. Position weight matrix for scoring of polyA signal $[-1, +7]$.

2. Position weight matrix [8] for scoring of downstream GT/T-rich element.

3. Distance between polyA signal and predicted downstream GT/T-rich element.

4. Hexanucleotide composition of downstream $(+6, +100)$ region.

5. Hexanucleotide composition of upstream $(-100, -1)$ region.

6. Positional triplet composition of downstream $(+6, +55)$ region.

7. Positional triplet composition of upstream $(-50, -1)$ region.

8. Positional triplet composition of the GT/T-rich downstream element

In Table 4.12, the Mahalonobis distances for each characteristic calculated on the training set are given. The most significant characteristic is the score of AATAAA pattern (estimated by the position weight matrix) that indicates the importance of occurrences almost perfect polyA signal (AATAAA). The second valuable characteristic is the hexanucleotide preferences of the downstream $(+6, +100)$ region. Although the discriminating ability of GT-rich downstream element itself (characteristic 2) is very weak, combining it with the other characteristics significantly increases the total Mahalonobis distance.

Kondrakhin *et al.* (1994) reported the error rates of their method at different thresholds for polyA signal selection. If the threshold is set to predict 8 of 9 real sites, their function also predicts 968 additional false sites. The algorithm-based LDF for 3′-processing site identification is implemented in the POLYAH program (http://genomic.sanger.ac.uk). First, it searches for the pattern similar to AATAAA using the weight matrix and, if the pattern is found, it computes the value of the linear discriminant function defined by the characteristics around this position. A polyA site is predicted if the value of this function is greater than an empirically selected threshold. The method demonstrates $S_n = 0.86$ and $S_p = 0.63$ when applied to a set of 131 positive and 1466 negative examples

Table 4.12 Significance of various characteristics of polyA signal.

Characteristics	1	4	2	5	3	6	8	7
Individual D^2	7.61	3.46	0.01	2.27	0.44	1.61	0.16	0.17
Combined D^2	7.61	10.78	11.67	12.36	12.68	12.97	13.09	13.1

that were not used in the training. The POLYAH program has been tested also on the sequence of the Ad2 genome, where for 8 correctly identified sites it predicts only 4 false sites.

4.8 IDENTIFICATION OF MULTIPLE GENES IN GENOMIC SEQUENCES

Computational gene finding started a long time ago with looking for open reading frames with an organism-specific codon usage (Staden and McLachlan, 1982). The approach worked well for bacterial genes (Staden, 1984b; Borodovskii *et al.*, 1986), but short eukaryotic exons and spliced genes require algorithms combining information about functional signals and the regularities of coding and intron regions. Several internal exon-predicting algorithms have been developed. The program SORFIND (Hutchinson and Hayden, 1992) was designed to predict internal exons based on codon usage plus Berg and von Hippel (1987) suggested discrimination energy for intron–exon boundary recognition. The accuracy of exact internal exon prediction (at both 5′- and 3-′ splice junctions and in the correct reading frame) by the SORFIND program reaches 59 % with a specificity of 20 %. Snyder and Stormo (1993) applied a dynamic programming approach (alternative to the rule-based approach) to internal exon prediction in GeneParser algorithm. It recognized 76 % of internal exons, but the structure of only 46 % exons was exactly predicted when tested on the entire GenBank sequence entries. HEXON (Human EXON) program (Solovyev *et al.*, 1994) based on linear discriminant analysis was the most accurate in exact internal exon prediction at that time.

Later a number of single gene-prediction programs has been developed to assemble potential eukaryotic coding regions into translatable mRNA sequence selecting optimal combinations of compatible exons (Fields and Soderlund, 1990; Gelfand, 1990; Guigo *et al.*, 1992; Dong and Searls, 1994). Dynamic programming was suggested as a fast method of finding an optimal combination of preselected exons (Gelfand and Roytberg, 1993; Solovyev and Lawrence, 1993b; Xu *et al.*, 1994). This is different from the approach suggested by Snyder and Stormo (1993) in the GeneParser algorithm to recursively search for exon–intron boundary positions. FGENEH (Find GENE in Human) algorithm incorporated 5′-, internal and 3′-exon identification linear discriminant functions and a dynamic programming approach (Solovyev *et al.*, 1994; 1995). Burset and Guigo (1996) have made a comprehensive test of gene-finding algorithms. The FGENEH program was one of the best in the tested group having the exact exon prediction accuracy 10 % higher than the other programs and the best level of accuracy at the protein level. A novel step in gene-prediction approaches was application of generalized Hidden Markov Models implemented in Genie algorithm. It was similar in design to GeneParser, but was based on a rigorous probabilistic framework (Kulp *et al.*, 1996). The algorithm demonstrated similar performance to FGENEH.

4.9 DISCRIMINATIVE AND PROBABILISTIC APPROACHES FOR MULTIPLE GENE PREDICTION

Genome sequencing projects require gene-finding approaches able to identify many genes encoded in the transcribed sequences. The value of sequence information for the

biomedical community is strongly dependent on the availability of candidate genes that are computationally predicted. The best multiple gene-prediction programs involve HMM-based probabilistic approaches as implemented in Genscan (Burge and Karlin, 1997) and Fgenesh (Salamov and Solovyev, 2000), Fgenes (discriminative approach) (Solovyev *et al.*, 1995) and Genie (Generalized HMM with neural network splice-site detectors) (Reese *et al.*, 2000). Initially we will describe a general scheme for HMM-based gene prediction (Stormo and Haussler, 1994) (first implemented by the Haussler group (Krogh *et al.*, 1994; Kulp *et al.*, 1996)) as the most general description of the gene model. A pattern-based approach can be considered as a particular case of this approach, where transition probabilities are not taken into account.

4.9.1 HMM-based Multiple Gene Prediction

Different components (states) of gene structure such as exons, introns and 5′-untranslated regions occupy k subsequences of a sequence X: $X = \bigcup_{i=1,k} x_i$. There are 35 states that describe eukaryotic gene model considering direct and reverse strands as possible gene locations (Figure 4.13). However, in the current gene-prediction approaches, noncoding 5′- and 3′-exons (and introns) are not considered because the absence of protein-coding characteristics makes their prediction less accurate. In addition, the major practical goal of gene prediction is to identify the protein-coding sequences. The remaining 27 states include 6 exon states (first, last single and 3 types of internal exons in 3 possible reading frame) and 7 noncoding states (3 intron, noncoding 5′- and 3′-, promoter and polyA) in each strand plus the noncoding intergenic region.

The predicted gene structure can be considered as the ordered set of states/sequence pairs, $\phi = \{(q_1, x_1), (q_2, x_2), \ldots, (q_k, x_k)\}$, called the *parse*, such that the probability $P(X, \phi)$ of generating X according to ϕ is maximal over all possible parses (or a score is optimal in some meaningful sense that best explains the observations (Rabiner, 1989)):

$$P(X,\phi) = P(q_1)\left(\prod_{i=1}^{k-1} P(x_i|l(x_i), q_i)P(l(x_i)|q_i)(P(q_{i+1}, q_i)\right)P(x_i|l(x_k), q_k)P(l(x_k)|q_k),$$

where $P(q_1)$ denotes the initial state probabilities; $P(x_i|l(x_i), q_i)P(l(x_i)|q_i)$ and $P(q_{i+1}, q_i)$ are the independent joint probabilities of generating the subsequence x_i of length l in the state q_i and transitioning to $i+1$ state.

Successive states of this HMM model are generated according to the Markov process with the inclusion of explicit state duration density. A simple technique based on the dynamic programming method for finding the optimal parse (or the single best state sequence) is called the *Viterbi algorithm* (Forney, 1973). The algorithm requires on the order of N^2D^2L calculations, where N is the number of states, D is the longest duration and L is the sequence length (Rabiner and Juang, 1993). A useful technique was introduced by Burge (1997) to reduce the number of states and simplify computations by modeling noncoding state length by a geometrical distribution. The algorithm for gene finding using this technique was initially implemented in the Genscan program (Burge and Karlin, 1997) and used later in Fgenesh program (Salamov and Solovyev, 2000). Since any valid parse will consist only of an alternating series of Noncoding and Coding states: NCNCNC, . . , NCN, we need only 11 variables, corresponding to the different types of N states. At each step corresponding to some sequence position, we select the maximum joint probability to continue the current state or to move to another noncoding
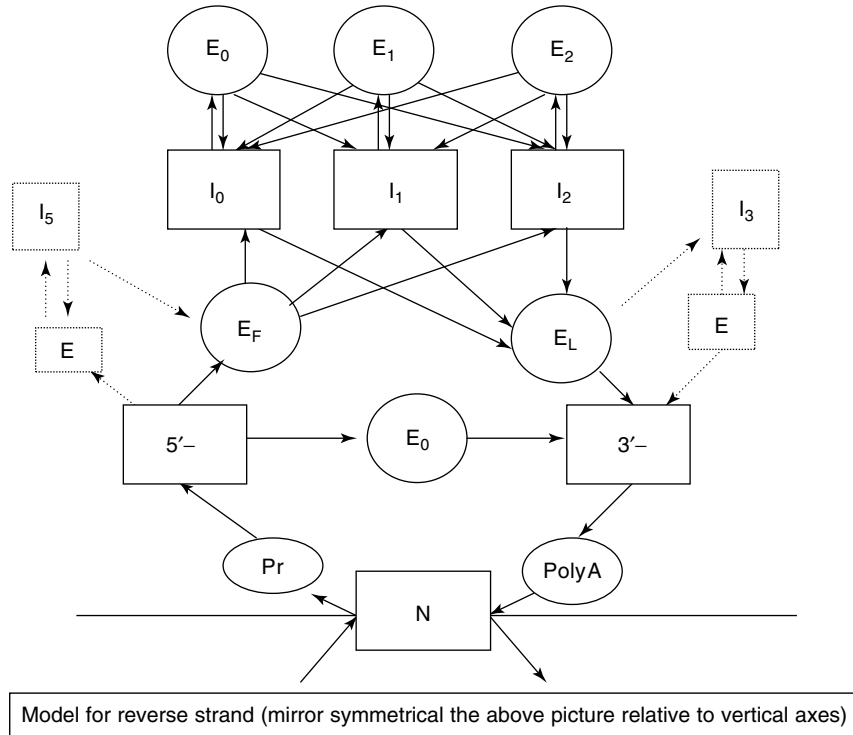
Figure 4.13 Different states and transitions in eukaryotic HMM genes model. E_i and I_i are different exon and intron states, respectively ($i = 0,1,2$ reflect 3 possible different ORF). E marks noncoding exons and I5/I3 are 5′- and 3′-introns adjacent to noncoding exons.

state defined by a coding state (from a precomputed list of possible coding states) that ends in analyzed sequence position.

Define the best score (highest joint probability) $\gamma_i(j)$ of optimal parse of the subsequence $s_{1,j}$, which ends in state q_i at position j. We have a set A_j of coding states $\{c_k\}$ of lengths $\{d_k\}$, starting at positions $\{m_k\}$ and ending at position j, which have the previous states $\{b_k\}$. The length distribution of state c_k is denoted by $f_{c_k}(d)$. The searching procedure can be stated as follows:

Initialization:

$$\gamma_i(1) = \pi_i P_i(S_1) p_i, i = 1, \ldots 11.$$

Recursion:

$$\gamma_i(j+1) = \mathbf{max}\{\gamma_i(j) p_i P_i(S_{j+1}),$$
$$\max_{c_k \in A_j}\{\gamma_i(m_k - 1)(1 - p_{b_k}) t_{b_k,c_k} f_{c_k}(d_k) P(S_{m_k,j}) t_{c_k,i} p_i P_i(S_{j+1})\}\}$$
$$i = 1, \ldots 11, j = 1, \ldots, L - 1.$$

Termination:

$$\gamma_i(L+1) = \max\{\gamma_i(L),$$
$$\max_{c_k \in A_j}\{\gamma_i(m_k-1)(1-p_{b_k})t_{b_k,c_k}f_{c_k}(d_k)P(S_{m_k,j})t_{c_k,i}\}\}$$
$$i = 1, \ldots 11.$$

At each step, we record the location and type of transition maximizing the functional to restore the optimal set of states (gene structure) by a backtracking procedure. Most parameters of these equations can be calculated on the learning set of known gene structures. Instead of scores of coding states $P(S_{m_k,j})$ it is better to use log likelihood ratios, which do not produce scores below the limits of computer precision.

Genscan (Burge and Karlin, 1997) was the first published algorithm to predict multiple eukaryotic genes. Several HMM-based gene-prediction programs were developed later: Veil (Henderson *et al.*, 1997), HMMgene (Krogh, 1997), Fgenesh (Salamov and Solovyev, 2000), a variant of Genie (Kulp *et al.*, 1996) and GeneMark (Lukashin and Borodovsky, 1998). Fgenesh is currently one of the most accurate programs. It is different from Genscan because, in the model of gene structure, a signal term (such as splice site or start site score) has some advantage over a content term (such as coding potential). In log likelihood terms, the splice sites and other exon functional signals have an additional score, depending on the environments of the sites. Also, in computing the coding scores of potential exons, *a priori* probabilities of exons are taken into account according to Bayes theorem. As a result, the coding scores of potential exons are generally lower than in Genscan. Fgenesh works with separately trained parameters for each model organism such as human, drosophila, chicken, nematode, dicot and monocot plants, and dozen yeast/fungi (currently using known genes or predicted protein-supported genes, Fgenesh gene-finding parameters has been computed for about 40 various organisms: `http://sun1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind`). Coding potentials were calculated separately for 4 isochores (human) and for 2 isochores (other species). The run time of Fgenesh is practically linear and the current version has no practical limit on the length of analyzed sequence. Prediction of about 800 genes in 34 MB of Chromosome 22 sequence takes about 1.5 minutes of Dec-alpha processor EV6 for the latest Fgenesh version.

4.9.2 Pattern-based Multiple Gene-prediction Approach

FGENES (Solovyev, 1997) is the multiple gene-prediction program based on dynamic programming. It uses discriminant classifiers to generate a set of exon candidates. Similar discriminant functions were developed initially in Fexh (Find Exon), Fgeneh (Find GENE) program (h stands for version to analyze human genes) and described in details earlier (Solovyev and Lawrence, 1993a; Solovyev *et al.*, 1995, Solovyev and Salamov, 1997).

The following major steps describe the analysis of genomic sequences by the Fgenes algorithm:

1. Create a list of potential exons, selecting all ORFs: ATG.. GT, AG–GT, AG.. Stop with exons scores higher than the specific thresholds depending on GC content (4 groups);

2. Find the set of compatible exons with maximal total score. Guigo (1998) described an effective algorithm for finding such set. Fgenes uses a simpler variant of a similar algorithm: Order all exon candidates according to their 3′-end positions; Going from the first to the last exon select for each exon the maximal score path (compatible exons combination) terminated by this exon using the dynamic programming approach. Include in the optimal gene structure either this exon or the exon with the same 3′-splicing pattern ending at the same position or earlier (which has the higher maximal score path).

3. Take into account promoter or polyA scores (if predicted) in the terminal exon scores.

The run time of the algorithm grows approximately linearly with the sequence length. Fgenes is based on the linear discriminant functions developed earlier for the identification of splice sites, exons, promoter and polyA sites (Solovyev *et al.*, 1994; Salamov and Solovyev, 1997). We consider these functions in the following sections to see what sequence features are important in exon prediction.

4.10   INTERNAL EXON RECOGNITION

For internal intron prediction, we consider all open reading frames in a given sequence that are flanked by AG (on the left) and by GT (on the right) as potential internal exons. The structure of such exons is presented in Figure 4.14. The values of 5 exon characteristics were calculated for 952 authentic exons and for 690 714 pseudoexon training sequences from the set. The Mahalonobis distances showing significance of each characteristic are given in Table 4.13. We can see that the strongest characteristics for exons are the values of the recognition functions for the flanking donor and acceptor splice sites ($D^2 = 15.04$ and $D^2 = 12.06$, respectively). The preference of an ORF being a coding region has $D^2 = 1.47$ and adjacent left intron region has $D^2 = 0.41$ and right intron region has $D^2 = 0.18$.
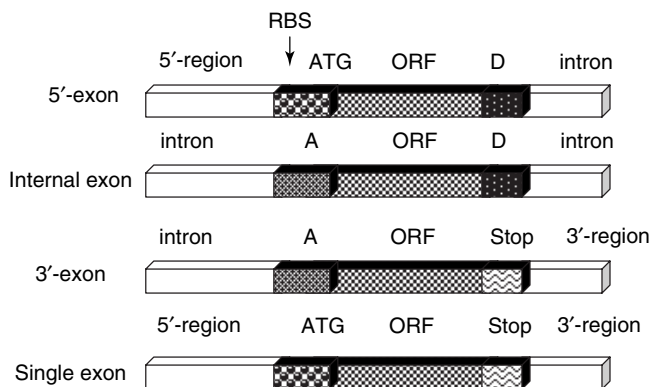
Figure 4.14   Different functional regions of the first, internal, last and single exons corresponding to components of recognition functions.

Table 4.13 Significance of internal exon characteristics.

	Characteristics	1	2	3	4	5
a	Individual D^2	15.0	12.1	0.4	0.2	1.5
b	Combined D^2	15.0	25.3	25.8	25.8	25.9

Characteristics 1 and 2 are the values of the donor and acceptor site recognition functions. Characteristic 3 gives the octanucleotide preferences for being coding for each potential exon. Characteristic 4 gives the octanucleotide preferences for being an intron 70-bp region on the left and a 70-bp region on the right of the potential exon region.

The performance of the discriminant function based on these characteristics was estimated using 451 exon and 246 693 pseudoexon sequences from the test set. The general accuracy of the exact internal exon prediction is 77 % with specificity 79 %. At the level of individual nucleotides, the sensitivity of exon prediction is 89 % with specificity 89 %; and the sensitivity of prediction of intron positions is 98 % with specificity 98 %. This accuracy is better than in the most accurate dynamic programming and neural network-based methods (Snyder and Stormo, 1993), which have 75 % accuracy of the exact internal exon prediction with specificity 67 %. The method has 12 % less false exon assignments with a better level of true exon prediction.

4.11   RECOGNITION OF FLANKING EXONS

Figure 4.15 shows the 3-dimensional histograms reflecting the oligonucleotide composition of the gene flanking regions based on a graphical fractal representation of nucleotide sequences (Jeffrey, 1990; Solovyev *et al.*, 1991; Solovyev, 1993). The clear differences in compositions were exploited to develop of recognizers of these regions.

4.11.1   5′-terminal Exon-coding Region Recognition

For 5′-exon prediction, all the open reading frames in a given sequence starting with the ATG codon and ending with the GT dinucleotide were considered as potential first exons. The structure of such exons is presented in Figure 4.14. The exon characteristics and their Mahalonobis distances are given in Table 4.14. The accuracy of the discriminant function based on these characteristics was computed using the recognition of 312 first exons and 246 693 pseudoexon sequences. The gene sequences were scanned and the 5′ exon with the maximal weight was selected for each of them. The accuracy of the prediction of the true first coding exon is 59 %. Competition with the internal exons was not considered in this test.

4.11.2   3′-exon-coding Region Recognition

All ORF regions that are flanked by GT (on the left) and finish with a stop codon were considered as potential last exons. The structure of such exons is presented in Figure 4.14. The characteristics of the discriminant functions and their Mahalonobis distances are presented in Table 4.15. The accuracy of the discriminant function was tested on the
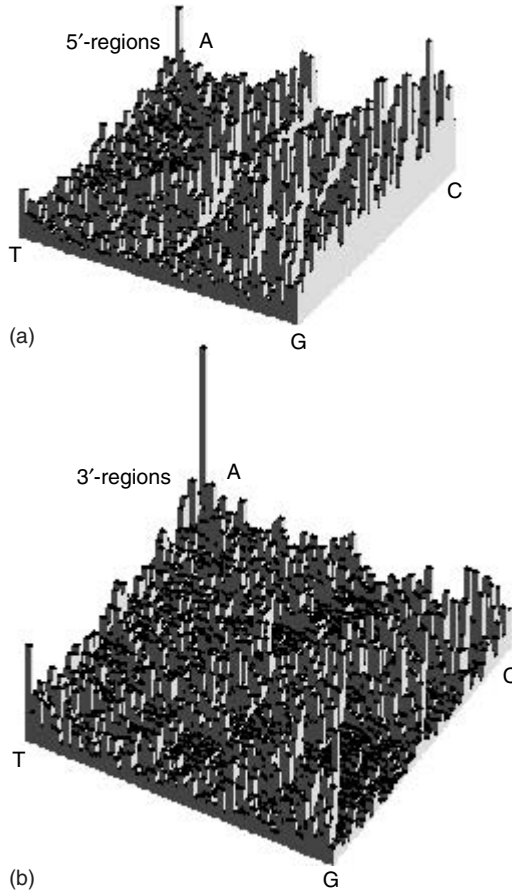
Figure 4.15 Graphical representation of the number of different oligonucleotides 6 bp long in 5'
(a) and 3' (b) gene regions. Each colon is the number of a particular oligonucleotide in the set of
sequences.

Table 4.14   Significance of 5'- exon characteristics.

	Characteristics	1	2	3	4	5	6	7
a	Individual D^2	5.1	2.6	2.7	2.3	0.01	1.05	2.4
b	Combined D^2	5.1	8.1	10.0	11.3	12.5	12.8	13.6

Characteristic 1 is the value of donor site recognition function. 2 is the average value of positional triplet
preferences in -15 to $+10$ region around ATG codon. 4 gives the octanucleotide preferences for being intron
in 70 bp region on the right of potential exon. 3, 5 and 7 are the hexanucleotide preferences in -150 to -101
bp, -100 to -51 bp and -50 to -1 bp regions on the left of potential exon, respectively; 6 is the octanucleotide
preferences for being coding in exon region.

recognition of 322 last exons and 2 47 644 pseudoexon sequences. The gene sequences
were scanned and the 3' exon with the maximal weight was selected for each of them.
The function can identify 60 % of annotated last exons.

Table 4.15   Significance of 3′-exon characteristics.

	Characteristics	1	2	3	4	5	6	7
a	Individual D^2	10.0	3.2	0.8	2.2	1.2	0.2	1.6
b	Combined D^2	10.0	11.4	12.0	13.8	14.3	14.5	14.6

Characteristic 1 is the value of acceptor site recognition. Characteristic 2 is the octanucleotide preferences for being coding of ORF region. 3, 5 and 7 are the hexanucleotide preferences in +100 to 150 bp, +50 to +100 bp and +1 to +50 bp regions on the left of coding region, respectively. 4 is the average value of positional triplet preferences in −10 to +30 region around the stop codon. 6 is the octanucleotide preferences for being intron in 70 bp region on the left of exon sequence.

The recognition function for single exons combines the corresponding characteristics of 5′- and 3′-exons.

4.12   PERFORMANCE OF GENE IDENTIFICATION PROGRAMS

Most gene-recognition programs were tested on a specially selected set of 570 single gene sequences (Burset and Guigo, 1996) of mammalian genes (Table 4.16). The best programs predict accurately on average 93 % of the exon nucleotides ($S_n = 0.93$) with just 7 % of false-positive predictions. Because the most difficult task is to predict small exons and exactly identify exon 5′ and 3′ ends, the accuracy at the exon level is usually lower than at the nucleotide level.

The table demonstrates that the modern multiple gene-prediction programs as Fgenesh, Fgenes and Genscan significantly outperform the older approaches. The exon identification rate is actually even higher than the data presented as the overlapped exons were not

Table 4.16   Accuracy of the best gene-prediction programs for single gene sequences from (Burset and Guigo, 1996) data set. S_n (sensitivity) = number of exactly predicted exons/number of true exons (or nucleotide); S_p (specificity) = number of exactly predicted exons/number of all predicted exons. Accuracy data for programs developed before 1996 were estimated by Burset and Guigo (1996). The other data were received by authors of programs.

Algorithm	S_n (exons)	S_p (exons)	S_n nucleotides	S_p nucleotides	Authors/year
Fgenesh	0.84	0.86	0.94	0.95	Solovyev and Salamov (1999)
Fgenes	0.83	0.82	0.93	0.93	Solovyev (1997)
Genscan	0.78	0.81	0.93	0.93	Burge and Karlin (1997)
Fgeneh	0.61	0.64	0.77	0.88	Solovyev *et al.* (1995)
Morgan	0.58	0.51	0.83	0.79	Salsberg *et al.* (1998)
Veil	0.53	0.49	0.83	0.79	Henderson *et al.* (1997)
Genie	0.55	0.48	0.76	0.77	Kulp *et al.* (1996)
GenLang	0.51	0.52	0.72	0.79	Dong and Searls (1994)
Sorfind	0.42	0.47	0.71	0.85	Hutchinson and Hyden (1992)
GeneID	0.44	0.46	0.63	0.81	Guigo *et al.* (1992)
Grail2	0.36	0.43	0.72	0.87	Xu *et al.* (1994)
GeneParser2	0.35	0.40	0.66	0.79	Snyder and Stormo (1995)
Xpound	0.15	0.18	0.61	0.87	Thomas and Skolnick (1994)

counted in exact exon predictions. However, there is a lot of room for future improvement. The accuracy at the level of exact gene prediction is only 59 % for Fgenesh, 56 % for Fgenes and 45 % for the Genscan program even on this relatively simple test set.

The real challenge for *ab initio* gene identification is to find multiple genes in long genomic sequences containing genes on both DNA strands. Often, there is no complete information about the real genes in such sequences. One example studied experimentally at the Sanger Centre (UK) is the human BRACA2 region (1.4 MB) that contains eight genes and 169 experimentally verified exons. This region is one of the worse cases for genome annotation because it has genes with many exons and almost all genes show no similarity of their products with known proteins. Moreover, it contains four pseudogenes and at least two of the genes have alternative splicing variants. The results of gene prediction initially provided by T. Hubbard and R. Bruskiewich (Sanger Centre Genome Annotation Group) are shown in Table 4.17.

Fgenesh predicts 20 % less false-positive exons in this region than the Genscan approach, with the same level of true predicted exons. Even for such difficult region about 80 % of exons were identified exactly by *ab initio* approaches.

The accuracy of the gene-finding programs depends not only on underlying algorithm, but also strongly affected by parameter file computed on the learning set of known genes. While Fgenesh and Genscan demonstrate similar performance for human gene prediction, Fgenesh has shown significantly better accuracy than many other tested gene-finders (including Genscan) in predicting rice genes (Yu *et al.*, 2002).

4.13   USING PROTEIN SIMILARITY INFORMATION TO IMPROVE GENE PREDICTION

The lessons from manual annotations show that it is often advantageous to take into account all the available information to improve gene identification. Automatic gene-prediction approaches can take into account the information about exon similarity with

Table 4.17   Accuracy of gene-prediction programs for BRACA2 1.4 MB human genomic sequence. Masked is prediction when repeats have been defined by RepeatMasker (Smit and Green, 1997) program in the analyzed sequence and excluded from potential exon location during prediction. The region consisted of 20 sequences with 8 verified genes, 4 pseudogenes and 169 exons. Later one sequence was constructed and three additional exons were identified. The results of prediction on this sequence marked bold.

	CC	S_{nb}	S_{pb}	P_e	$C_{e,ov}$	S_{ne}	$S_{n,ov}$	$S_{pe}/S_{pe,ov}$
Genscan	0.68	90	53	271/**271**	109/**131**	65	80/**76**	40/**49**
Fgenesh	0.80	89	73	188/**195**	115/**131**	69	80/**76**	61/**67**
Fgenes	0.69	79	62	298/**281**	110/**136**	66	86/**78**	37/**48**
Genscan masked	0.76	90	66	217	109	65	80	50
Fgenesh masked	0.84	89	82	172/**168**	114/**131**	68	79/**73**	66/**76**
Fgenes masked	0.73	80	68	257/**228**	107/**133**	64	85/**75**	42/**58**

CC is the correlation coefficient reflecting the accuracy of prediction at the nucleotide level. S_{nb}, S_{pb} – sensitivity and specificity at the base level (in %), P_e – number of predicted exons, C_e –number of correctly predicted exons, S_{ne}, S_{pe} – sensitivity and specificity at the exon level, S_{nep} – exon sensitivity, including partially correct predicted exons (in %). Ov is including overlapped exons.

known proteins or ESTs (Gelfand *et al.*, 1996; Krogh, 2000). Fgenesh+ (Salamov and Solovyev, 2000) is a version of Fgenesh which uses additional information from the available protein homologs. When exons initially predicted by Fgenesh show high similarity to a protein from the database, it is often advantageous to use this information to improve the accuracy of prediction. Fgenesh+ requires an additional file with protein homolog, and aligns all predicted potential exons with the protein homolog using own alignment algorithm. To decrease the computational time, all overlapped exons in the same reading frame are combined into one sequence and align only once.

The main additions to the algorithm, relative to Fgenesh, include:

1. Augmentation of the scores of exons with detected similarity by an additional term proportional to the alignment score.

2. An additional penalty included for the adjacent exons in the dynamic programming (Viterbi algorithm), if the corresponding aligned protein segments are not close in the corresponding protein.

Fgenesh+ was tested on the selected set of 61 GenBank human sequences, for which Fgenesh predictions were not accurate (correlation coefficient $0.0 <= CC < 0.90$) and which have protein homologs from another organism. The percentage identity between the encoded proteins and their homologs varied from 99 % to 40 %. The prediction accuracy using this set is presented in Table 4.18. The results show that if the alignment covers the whole length of both proteins, then Fgenesh+ usually increases the accuracy relative to Fgenesh and does not depend significantly on the level of identity (for $ID > 0$ %). This result makes knowledge of proteins from distant organisms valuable for improving the accuracy of gene identification. A similar approach exploiting known EST/cDNA information was implemented in Fgenesh_c program (Salamov and Solovyev, 2000).

4.13.1 Components of Fgenesh++ Gene-prediction Pipeline

Ab initio gene-prediction program such as Fgenesh predicts $\sim$93 % of all coding exon bases and exactly predicts $\sim$80 % of human exons when applied to single gene sequences (Table 4.16). Analysis of multigene, long genomic sequences is a more complicated task. A program can erroneously join neighboring genes or split a gene into two or more. To improve automatic annotation accuracy, we developed a pipeline Fgenesh++, which can take into account available supporting data such as mRNA or homologous protein sequences. Fgenesh++ is a pipeline for automatic, without human modification of results, prediction of genes in eukaryotic genomes. It uses thefollowing sequence analysis software.

Table 4.18 Comparison of accuracy of Fgenesh and Fgenesh+ on the set of 'difficult' human genes with known protein homologs from another organism.

	CG	S_{ne}	S_{pe}	S_{nb}	S_{pb}	*CC*
Fgenesh	0	63	68	86	83	0.74
Fgenesh+	46	82	85	96	98	0.95

The set contains 61 genes and 370 exons. *CG* – percent of correctly predicted genes; S_{ne}, S_{pe} –sensitivity and specificity at the exon level (in %); S_{nb}, S_{pb} – sensitivity and specificity at the base level (in %); *CC* – correlation coefficient.

Fgenesh++ script to execute the pipeline;

Fgenesh: HMM-based *ab initio* gene-prediction program;

Fgenesh+: gene-prediction program that uses homologous protein sequence to improve performance;

Est_map: a program for mapping known mRNAs/ESTs to a genome, producing genome alignment with splice sites identification;

Prot_map: a program for mapping a protein database to genomic sequence.

Est_map can map a set of mRNAs/ESTs to a chromosome sequence. For example, 11 000 full-length mRNA sequences from NCBI reference set were mapped to 52 MB unmasked Y chromosome fragment in ∼20 minutes **Est_map** takes into account statistical features of splice sites for more accurate mapping. **Prot_map** uses a genomic sequence and a set of protein sequences as its input data, and reconstructs gene structure based on protein identity or homology, in contrast to a set of unordered alignment fragments generated by Blast (Altschul *et al.*, 1997). The program is very fast and produces gene structures with similar accuracy to those of relatively slow GeneWise program (Birney and Durbin, 2000), but does not require knowledge of protein genomic location. The accuracy of gene reconstruction can be significantly improved further using Fgenesh+ program on output of Prot_map, that is, using a fragment of genomic sequence (where prot_map found a gene) and the cooreponding protein sequence mapped to it.

Comparison of accuracy of gene prediction by *ab initio* Fgenesh and gene prediction with protein support by Fgenesh+ or GeneWise (Birney and Durbin, 2000) and Prot_ map was performed on a large set of human genes with homologous proteins from mouse or drosophila. We can see that Fgenesh+ shows the best performance with mouse proteins (Table 4.19). With Drosophila proteins, *ab initio* gene prediction by Fgenesh works better than GeneWise for all ranges of similarity and Fgenesh+ is the best predictor if similarity is higher than 60 % (Table 4.20).

Table 4.19   Accuracy of human gene prediction using similar mouse proteins.

(a) Similarity of mouse protein > 90 % in 921 sequences *

	Sn_{ex}	Sp_{ex}	Sn_{nuc}	Sp_{nuc}	CC	$\%CG$
Fgenesh	86.2	88.6	93.9	93.4	0.9334	34
Genwise	93.9	95.9	99.0	99.6	0.9926	66
Fgenesh+	97.3	98.0	99.1	99.6	0.9936	81
Prot_map	95.9	96.9	99.1	99.5	0.9924	73

(b) 80 %¡ similarity of mouse protein < 90 % in 1441 sequences

	Sn_{ex}	Sp_{ex}	Sn_{nuc}	Sp_{nuc}	CC	$\%CG$
Fgenesh	85.8	87.7	94.0	93.4	0.9334	30
Genewise	92.6	94.1	98.9	99.5	0.9912	58
Fgenesh+	96.8	97.2	99.1	99.5	0.9929	77
Prot_map	93.9	94.1	98.9	99.3	0.9898	60

* Sn_{ex}, Sensitivity on exon level (exact exon predictions); Sno_{ex}, sensitivity with exon overlap; Sp_{ex}, specificity, exon level; Sn_{nuc}, sensitivity, nucleotides; Sp_{nuc}, specificity, nucleotides; CC, correlation coefficient; $\%CG$, percent of genes predicted completely correctly (no missing and no extra exons, and all exon boundaries are predicted exactly correctly).

Table 4.20 Accuracy of gene prediction using similar Drosphilla pfroteins.

(a) Similarity of Drosophila protein $> 80\%$ – 66 sequences.

	Sn_{ex}	Sp_{ex}	Sn_{nuc}	Sp_{nuc}	CC	CG%
Fgenesh	90.5	95.1	97.9	96.9	0.950	55
Genewise	79.3	86.8	97.3	99.5	0.985	23
Fgenesh+	95.1	97.0	98.9	99.5	0.9914	70
Prot_map	86.4	88.1	97.6	99.0	0.982	41

(b) $60\% <$ similarity of Drosophila protein $< 80\%$ – 290 sequences.

	Sn_{ex}	Sp_{ex}	Sn_{nuc}	Sp_{nuc}	CC	CG%
Fgenesh	88.6	90.8	94.9	93.8	0.941	34
Genewise	76.3	82.9	92.8	99.4	0.959	7
Fgenesh+	89.2	92.7	95.5	98.5	0.968	44
Prot_map	75.1	74.9	91.4	97.5	0.941	10

In addition to the programs listed above, Fgenesh++ package also includes files with gene-finding parameters for specific genome, configuration files for programs and a number of Perl scripts. In addition, Fgenesh++ pipeline uses the following public software and data: BLAST executables blastall and bl2seq (Altschul *et al.*, 1997), NCBI NR database (nonredundant protein database) formatted for BLAST, and NCBI RefSeq database (Pruitt *et al.*, 2005).

Fgenesh++ analyzes genome sequences and, optionally, same sequences with repeats masked by N. Sequences can be either complete chromosomes or their fragments such as scaffolds, contigs, etc. When preparing repeats-masked sequences, it is recommended not to mask low complexity regions and simple repeats, as they can be parts of coding sequences.

There are three main steps in running the pipeline:

1. mapping known mRNAs/cDNAs (e.g. from RefSeq) to genomic sequences;

2. prediction of genes based on homology to known proteins (e.g. from NR);

3. *ab initio* gene prediction in regions having neither mapped mRNAs nor genes predicted on the basis of protein homology.

The output of the pipeline consists of predicted gene structures and corresponding proteins. It also indicates whether particular gene structure was assigned on the basis of mRNA mapping, protein homology, or *ab initio* gene prediction.

4.14 GENOME ANNOTATION ASSESSMENT PROJECT (EGASP)

NHGRI (The National Human Genome Research Institute) has initiated the ENCODE project to discover all human genome functional elements (The ENCODE Project Consortium, 2004). Its pilot phase is focused on performance evaluation of different techniques of genome annotation, including computational analysis, on a specified 30 MB of human genome sequence. The community experiment (EGASP05) was organized (Guigo and Reese, 2005) to evaluate how well automatic annotation methods are able

to reproduce manual annotations. The best performance in most categories has been demonstrated by predictors that used the most sources of available information. Some of them included conservation of corresponding coding regions in several available genomes: Augustus (Stanke *et al.*, 2006), Jigsaw (Allen *et al.*, 2006) and Paragon (Arumugam *et al.*, 2006). The sensitivity of Fgenesh++ pipeline (which uses one genome information) is similar with them, but the above multigenome programs demonstrated better specificity (Guigo *et al.*, 2006). Performance of Fgenesh++ pipeline for mRNA or protein-supported predictions and *ab initio* predictions (in sequence regions where similar mRNA/protein were not found) is presented in Table 4.21 (Solovyev *et al.*, 2006). All the above-mentioned pipelines demonstrate ∼90–95 % sensitivity on the nucleotide level and 75–80 % on the exact exon prediction level. They can exactly predict ∼70 % genes (when we count at least one transcript per locus predicted exactly) and ∼50 % of all annotated transcripts. No annotation strategy produces perfect gene predictions even using a lot of supportive information that is available for human genome. It is worthwhile to note that the human genes are the most difficult to predict (due to regular occurrences ofvery short exons and very long intron sequences), while the accuracy on simpler organisms is usually much better. While human genes present the most difficult case, the other sequenced genomes have much less available experimental information.

4.15   ANNOTATION OF SEQUENCES FROM GENOME SEQUENCING PROJECTS

Knowledge of gene sequences has opened a new way of performing biological studies called *functional genomics* and the major challenge is to find out what all the newly discovered genes do, how they interact and how they are regulated (Wadman, 1998). Comparisons between genes from different genomes can provide additional insights into the details of gene structure and function.

   The successful completion of the Human Genome Project has demonstrated that large-scale sequencing projects can generate high-quality data at a reasonable cost. In addition to human genome, researchers have already sequenced the genomes of a number of important model organisms that are commonly used as test beds in studying human biology. These are the chimpanzee, the mouse, the rat, two puffer fish, two fruit flies, two sea squirts, two roundworms and baker's yeast. Recently, sequencing centers completed working drafts of the chicken, the dog, the honey bee, the sea urchin and a set of four

Table 4.21 Performance data for annotating 44 ENCODE sequences by either mRNA and protein-supported predictions or by *ab initio* predictions.

	mRNA + protein- supported, Sn and Sp (%)	*Ab initio*, Sn and Sp (%)
nucleotide level	91.14	88.44
	89.54	74.46
CDS EXACT	77.19	67.54
	86.48	64.22
CDS OVERLAP	90.60	85.00
	91.4	71.71

fungi. A variety of other genomes are currently in the sequencing pipelines. Many new genomes lack such rich experimental information as the human genome, and therefore their initial computational annotation is even more important as a starting point for further research to uncover their biology. The more comprehensive and accurate such computational analysis is performed, the less time consuming and costly experimental work will have to be done to determine all functional elements in new genomes. Using computational predictions, the scientific community can get at least partial knowledge of majority of real genes because usually gene-finding programs correctly predict most exons of each gene. Fgenesh++ gene-prediction software has been used in annotation of dozen new genomes such as human, rice, Medicago, silkworm, many yeast genomes, bee and sea urchin (see for example, Sodergren *et al.*, 2006; The Honeybee Genome Sequencing Consortium, 2006). Annotation of many genomes is quite a complex procedure. For example, five gene lists were combined to produce bee genome master gene set. Three of them present gene predictors from NCBI, Fgenesh++ and ENSEMBL. Two others comprised evolutionary conserved gene set and Drosophila orthologs. These gene sets merged by special procedure (GLEAN), that construct consensus prediction based on combination of evidences provided by the five gene lists. The Glean set of 10 157 genes is considered as based on experimental evidence, the official *ab initio* gene set comprised 15 500 gene models that did not overlap models of the Glean set (The Honeybee Genome Sequencing Consortium, 2006).

4.15.1 Finding Pseudogenes

Pseudogenes prediction can use two types of initial information (Solovyev *et al.*, 2006). One type contains exon–intron structures of annotated genes and their protein sequences for a genome under analysis. To get such information, we can execute a gene-finding pipeline such as Fgenesh++. In this case, we run Prot_ map program with a set of protein sequences to find possible significant genome–protein alignments that do not correspond to a location of a gene for mapped protein. Other type of initial data can be a set of known proteins for a given organism. Having such data, we can restore gene structure of a given protein using Prot_ map program. For each mapped protein, we can select the best scoring mapping and the computed exon/intron structure as the 'parent' gene structure of this protein. If the alignment of a protein with its own parent has obvious internal stop codons or frameshifts, this locus could be included in the list of potential pseudogenes, but we need to keep in mind more trivial explanations like sequencing errors. Such loci cannot be analyzed on their K_a/K_s or checked for intron losses. In any case, for each of two approaches we have a set of protein sequences, their parent gene structures, and protein–genome alignments for further analysis to identify pseudogenes. Most other pseudogene-finding methods do not include gene-finding and rely on the available protein databases (Harrison *et al.*, 2003) or search only for processed pseudogenes (Baren and Brent, 2006). Example of two types of pseudogenes, processed and nonprocessed, and their characteristics are presented in Figures 4.16 and 4.17.

4.15.2 Selecting Potential Pseudogenes

Using genome–protein alignments generated by Prot_ map program, PSF program produces a list of alignments possessing the following properties for each protein.

Alignment vs. protein encoded by the parent gene.

Identity: 83.7 %

Coverage of protein sequence: 93.9 %

Number of internal stop codons: 2

Number of frameshifts: 1

K_a/K_s: 0.484

[DD] Sequence: 11931(1), S: 21.993, L:99 C14000887 chr14 2 exon (s) 75425067 - 75425530 ORF: 1 - 297
98 aa, chain + ## BY PROTMAP: gi|18597373|ref|XP_090893.1| similar to 60S acidic ribosomal protein

```
1 58970658   58970665   58970695   58970725   58970755   58970785   58970815   58970835
nnnnnnn(..)ccgcgcc?[MASVSELACIY*ALILHDDEVTVTEDKINALIKAAGVNIEPF*PGLFAKAtggtcNVNIGSLICSVEAGG
.......(..).......  |||7||||||||0||||||5||||||0||2||||||||||7||0|||||||....|||0|||||5|0|||
-------(..)-------  MASISELACIYSALILHDNEVTVTEYKIKALIKAAGVNVEPFRPGLFAKAp---aNVNIRSLICNVGAGG
1                   1          11         21         31         41         51         58

58970865   58970889   58970919   58970947   58970956   63811645
AAP--AEEKKVEAKKEESEDGDDDMRFGLtttcactga]acctctt(..)nnnnnnnn
0||..|||||5||||||0||2|||||0|||...   .......(..).......
PAPaaAEEKKMEAKKEEFEDSDDDMGFGLsd*----- -------(..)-------
68         78         88         98         100        100
```

Figure 4.16 An example of processed pseudogene.

Alignment vs. protein encoded by the parent gene.

Identity: 86.4 %

Coverage of protein sequence: 97.6 %

Number of internal stop codons: 3

Number of frameshifts: 4

K_a/K_s: 0.594

[RD] Sequence: 35522(1), S: 50.463, L:423 C7000711 chr7 3 exon (s) 51197888 – 51195897 ORF: 1 – 1269
422 aa, chain – ## BY PROTMAP: gi|27481026|ref|XP_209794.1| similar to hypothetical protein DKFZp43

```
  1   63659329 63659336 63659366 63659385 63659392 63659422 63659452 63659472
nnnnnn(..)tacagtc?[PTSASQQILHAQcatctac(..)gtggaccPQAKLPTFQQLLHTQLPPASGLFRPAtggggcSFLTTAFP
......(..).......|2|||||50|||||......(..)......|5|0|2|50|022||0|||||||......||||||||
------(..)-----mg PASASQRTLHAQlala---(..)---slrpPQSKAPAFRPLRQAQLLPASGLFRP------sSFLTTAFP
  1    1          3         13       19       23        33       43       48

63659498 63659528 63659558 63659588 63659618 63659648 63659678 63659708 63659738
GPVFPFRRPLRAQNLLKSASPDPLAPSGRSLRAQLFFLVGSPGPIPASQQPLWTQCLPISWRPWSAHSFLKPSSPGPGQASRWPLQDELL
|7|||5|||5|||0|||0|||5|||2022||||0|||||||||0||||||||||0|||||||||||||||||||||6||
GPIFPFQRPLQAQNLLKLASPGPLAPSGRPLQAQLFLPAASPGPTPASQQPLWTQCLPISWRPWSAHSFLKPSSPGPGQASRWPLQDQLL
57    67    77    87    97    107    117    127    137

63659768 63659798 63659828 63659858 63659888 63659907 63659952 63659971 63660001
PSDGlSRPQMVSGRWAPPRQGWASRRLPQAQVVLKSGSPGPASQQ]gtaagca(..)tttgtag[APNFLQPSSEGPPASWMPVQF*HW
|||7||||||||||02|||||||00|||||||2||||||0|||  ||||||||2|||||||||0|||000|
PSDGVSRPQMVSGRWAPPRAWASRRPLQAQVVLKSASPGPASQQ  ------(..)------    APNFLQPSSGPPPASRWPVQAQLW
147   157   167   177   187   192       192   197   207

63660031 63660061 63660089 63660119 63660147 63662724 63662731 63662748 63662766
LENSLCRPRPCLPGGPLQAQLLPPRRPPGAKSLPASQQPgc]gtgcggc(..)gtgcggc[gPDSGccgactccagVPTTSLDSAPAQLP
|||||0|||||0|||||0|5|||||||||5|||......(..)...... .|||......5|00||||||||||
LENSLCRPRSCLP-GPLQAQLSPPQRPPGAKSLPASRQP--  ------(..)------    aPDSG-------LPIRSLDSAPAQLP
217   227   236   246   255       255   255   260   264

63662796 63662826 63662856 63662884 63662914 63662944 63662974 63663004 63663034
AALVGPQLP*AKLPRPSGLAVASPGSAGPGALR*HLQAPNGLRSVGSSRPSLGLPAASAGPNRPEVSLSRLSSSLPAASAGPSRPQVGLE
|||||||0||||||||2|||||0||||||.||0|||||||||||||||||||||||2|||0|2|||||||0||||||||
AALVGPQLPEAKLPRPSGLTVASPGSAP-ALRRHLQAPNGLRSVGSSRPSLGLPAASAGPNRPEVGLSRPSSGLPAASAGLSRPQVGLE
274   284   294   303   313   323   333   343   353

63663064 63663094 63663124 63663154 63663184 63663214 63663244 63811645
VGLEEQQVGLPGPSSVLSTASPGAKLPRVSLSRPSSSCLPVASFSPAQLMALGGLRRPCF*]ctttgg(..)nnnnnnn
||||0|||||||||||||2|||||||||||||||||2|0|||0| ......(..).......------
VGLEELQVGLPGPSSVLSAASPGAKLPRVSLSRPSSSCLPVASFGPAQLMALGSLPRPRF*  ------(..)-------
363   373   383   393   403   413   423   424
```

Figure 4.17 An example of not processed pseudogene.

1. Identity in blocks of alignment exceeds certain value

2. Substantial portion of protein sequence is included in the alignment

3. Genomic location of alignment differs from that of parent gene

4. At least one of four events is observed:

    i. *Damage to ORF.* There is one or more frameshifts or internal stop codons;

    ii. *Single exon with close PolyA site.* PolyA site is too close to a 3′-end of an alignment, while C-terminus of protein sequence is aligned to the last amino acid, and a single exon covers 95 % of protein sequence.

    iii. *Loss of introns.* Protein coverage by alignment is at least 95 %, and a number of exons is fewer than in parent gene by a certain number.

    iv. *Protein sequence is not preserved.* The ratio of nonsynonymous to synonymous replacements exceeds certain threshold ($K_a/K_s > 0.5$). K_a/K_s is calculated relative to a parent gene by method presented by Nei and Gojobori (1986).

4.15.3 Selecting a Reliable Part of Alignment

The procedures described above apply to a so-called reliable part of alignment. Necessity of introducing this concept is caused by imperfections in aligning a protein against a chromosome sequence. There are complex cases where accurate alignment cannot be produced, such as very short (1–3 bp) exons separated by a large intron, or some errors in protein or genome draft sequence that prevent perfect alignment. For instance, if a protein as a whole is well aligned to a chromosome, but ∼20 amino acids on its 5′-end cannot be aligned in one continuous block, Prot_ map will most likely try to align these 20 amino acids by scattering them along several short blocks. Most likely, these blocks will not have any relation to a gene or a pseudogene. Therefore, in search for pseudogenes, we remove short insignificant trailer blocks. The rest of alignment is considered as its reliable part. To find a reliable part of alignment, we evaluate the quality of alignment blocks (exons). For each exon found by Prot_map, we calculate the number of aligned amino acid (M), number of nonaligned amino acids (AI) and nucleotides (NI) within an exon, number of aligned amino acids (AO) and nucleotides (NO) located outside of exon region to the left and to the right side of an exon. We also compute the 'correctness' of splice sites conserved dinucleotides (SSC) that flank an exon. If an exon is N- or C-terminal one, we also compute 'correctness' of corresponding start or stop codons. The length of an intron (IL) that separates an exon from nearest exon in the direction of the longest mapped exon is also computed. The empirical 'quality' measure is defined by the following formula:

$$Q = M - P_{AI}(AI) - P_{NI}(NI) - P_{AO}(AO) - P_{NO}(NO) + B_{SSC}(SSC) - P_{IL}(IL).$$

Where P_{AI}, P_{NI}, P_{AO}, P_{NO} are the penalties for the internal and external unaligned amino acids and nucleotides, BSSC is a bonus for correctness of splice sites or start/stop codons, and P_{IL} is a penalty for high intron length. The reliable part of alignment consists of neighboring exons alignments that each have $Q > 5$. After Prot_map mapping, many loci on a chromosome include alignments to more than one protein. In such cases, we choose only one most reliable alignment, based on a sum of included exon's qualities.

The PSF (pseudogene finding) approach described above has been applied to identify pseudogenes in 44 ENCODE sequences (Solovyev *et al.*, 2006). As a result, it was found 181 potential pseudogenes, 118 of which had a significant overlap with annotated 145 HAVANA pseudogenes. 68 (58 %) of these 118 pseudogenes had only one exon and could be classified as processed pseudogenes: 58 had the parent gene with more than one exon and seven others had polyA tail. 106 (90 %) of 118 pseudogenes had one or more defects in their ORFs. Among the remaining 12, there are four pseudogenes with a single exon (while their parents have four or more exons), four contain both polyA signal and polyA tract, four have only a polyA tract, and two have only high K_a/K_s ratios (0.59 and 1.04). The PSF has not found 27 HAVANA annotated pseudogenes. Three of them were not reported because they are located in introns of larger pseudogenes (AC006326.4-001, AC006326.2-001 and AL162151.3-001). The other ten represent fragments of some human proteins and are missing stop codons or frameshifts. We did not include pseudogenes corresponding to fragments of proteins in our pseudogene set. The remaining 14 HAVANA pseudogenes were not found probably due to some limitation of our program and the used datasets of predicted genes and known proteins. Missed pseudogenes might have parent genes that were absent from our initial protein set compiled by Fgenesh++ gene-prediction pipeline. Some of 63 pseudogenes that have been predicted by PSF but were absent from HAVANA set might have appeared because of imperfect predictions by the pipeline, which produced frameshifts when a pseudogene candidate and its parent gene were aligned. However, some of these 'over-predicted' pseudogenes might be actual pseudogenes missed by HAVANA annotators (for example, see Figure 4.18).

To summarize, PSF pseudogene prediction program has found 81 % of annotated pseudogenes. Its quality can further be improved by improving the quality of parent gene/protein sets.

4.16   CHARACTERISTICS AND COMPUTATIONAL IDENTIFICATION OF miRNA GENES

MicroRNA (miRNA) are a class of small (~22 nt), noncoding RNAs that can regulate gene expression by directing mRNA degradation or inhibiting productive translation (Mallory and Vaucheret., 2004). They are sequence-specific regulators of posttranscriptional gene expression in many eukaryotes. Some components of miRNA machinery have been found even in archaea and eubacteria, revealing their very ancient origination. They are believed to control the expression of thousands of target mRNAs, with each mRNA possible targeted by multiple miRNAs (Pillai, 2005). miRNA discovery by molecular cloning has been supplemented by computational approaches that identify evolutionary conserved miRNA genes by searching for patterns of sequence and secondary structure conservation These approaches indicate that miRNA constitute nearly 1–3 % of all identified genes in nematodes, flies and mammalians (Jones-Rhoades and Bartel, 2004). Only in humans the latest miRNA count exceeded 800 genes (Pillai, 2005). The first two miRNA genes (lin-4 and let-7) were discovered in *C. elegans*, where their mutations cause defects in the temporal regulation of larval stage-specific programs of cell divisions. These miRNAs affect by base pairing to partially complementary sites in the 3′ untranslated region (UTR) of their target mRNAs and repressing their translation (Lee *et al.*, 1993; Reinhart *et al.*, 2000).

```
[DD] Sequence: 622 (1), S: 27.323, L:153 C6000781 chr6 6 exon (s) 840966 - 845318 ORF: 1 - 459  152 aa,
chain + ## gi|6755368|ref|NP_035426.1| ribosomal protein S18 [Mus musculus] gi|11968182|ref ## 152

 1    151509    151516    151546    151576    151606    151636    151664    151694    151724
caaannn(..)tcctgct?[MSLVIPEKFQRILRILNSNINGQQKIGFAITAIKDVG*QYTHaVLRKADVDLTKWAGELTEDEMERVMTIM
.......(..)........ |||||||||2|||7||5||5|55||2|||||0||05|2|.|||||7|||0|||||||5|||5|||
-------(..)------- MSLVIPEKFQHILRVLNTNIDGRRKIAFAITAIKGVGRRYAHvVLRKADIDLTKRAGELTEDEVERVITIM
 1     1          11        21        31        41        51        61        71

    151754    151784    151814    151844    151874    151904    151934    151964
QNPCQYKIPDWFLNRRKDVKDGKYSQVLASGLDKKLRADVERLKKIQAHRGPHHFWGLRVRGQHTKTTGHHGCTMGGSKKK*]gtctgca(..)aaaataa
|||0|||||||||5|||||||||5||||||2|||0|5|||5|||||02||||||||||||||22|0|5|0|||| .......(..)........
QNPRQYKIPDWFLNRQKDVKDGKYSQVLANGLDNKLREDLERLKKIRAHRGLRHFWGLRVRGQHTKTTGRRGRTVGVSKKK* -------(..)-------
    81        91        101       111       121       131       141       151
```

Figure 4.18 A pseudogene in ENm004 sequence that is absent in the manual HAVANA annotation. The alignment has a stop codon close to position 151636.

The majority of miRNA genes are located in intergenic regions or in antisence orientation to annotated genes, indication that they for independent transcriptional units. Most of the other miRNA genes are found in intronic regions, which may be transcribed as part of the annotated gene. Independent miRNA genes are initially transcribed by RNA PolII (Lee *et al.*, 2004) as part of a long primary transcript, which contain the mature miRNA as part of a predicted RNA hairpin. This transcript is cropped into the hairpin-shaped pre-miRNAs by nuclear RNaseIII Drosha (Lee *et al.*, 2003). The hairpin RNAs of approximately 70 nt bearing the 2-nucleotide 3′ overhang are exported to the cytoplasm by a Ran dependent nuclear transport receptor family. Once in the cytoplasm, pre-miRNAs are subsequently cleaved by cytoplasmic RNase III Dicer into ~22 nt miRNA duplex, one strand of which is degraded by a nuclease, while the other strand remains as a mature miRNA (Lee *et al.*, 2004; Denli *et al.*, 2004). A typical structure of miRNA gene and its processing is presented in Figure 4.19.

Despite the plenty of miRNAs that have identified from cloning, such technique is likely to be far from saturated, as it is biased to abundant miRNA. Therefore, computational approaches have been developed that predict miRNAs encoded in animal and plant genomes (Grad *et al.*, 2003; Jones-Rhoades and Bartel, 2004; Ohler *et al.*, 2004). There are several variations of these methods: one is based on analysis of sequence and secondary structure properties of typical pre-miRNA. However, the short length and high degree of sequence and structure variation limit the accuracy of computational predictions based on such characteristics along. To decrease the number of false-positive predictions, the candidate miRNAs are selected to be conserved across species (the presence in two or more genomes of very similar sequences embedded in the same stems of predicted hairpins). A flowchart of computational selection miRNA candidates for plant miRNA predictions is presented in Figure 4.20 (Jones-Rhoades and Bartel, 2004). Another algorithm is based on the search for possible homologs (including hairpin selection and Smith–Waterman sequence alignment) of a few hundreds of known miRNAs cloned from *C. elegans, D. melanogaster, M. musculus and H. sapiens* (for identification of miRNAs in animal genomes) (Grad *et al.*, 2003). Recently, using similar approaches the FindmiRNA and the TargetmiRNA programs were developed
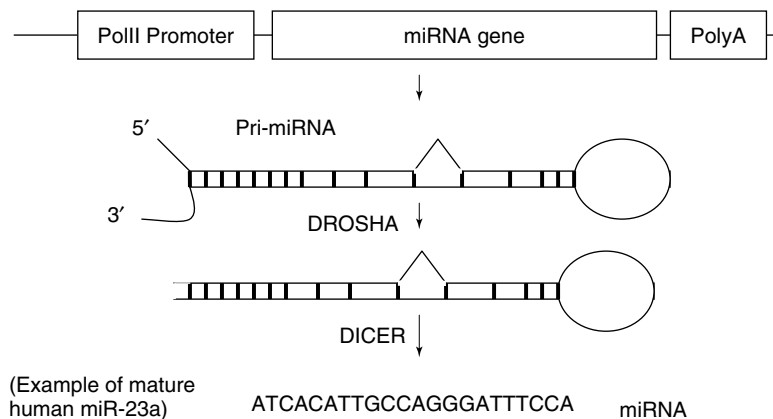

Figure 4.19 A model of expression of miRNA gene and processing of miRNA.
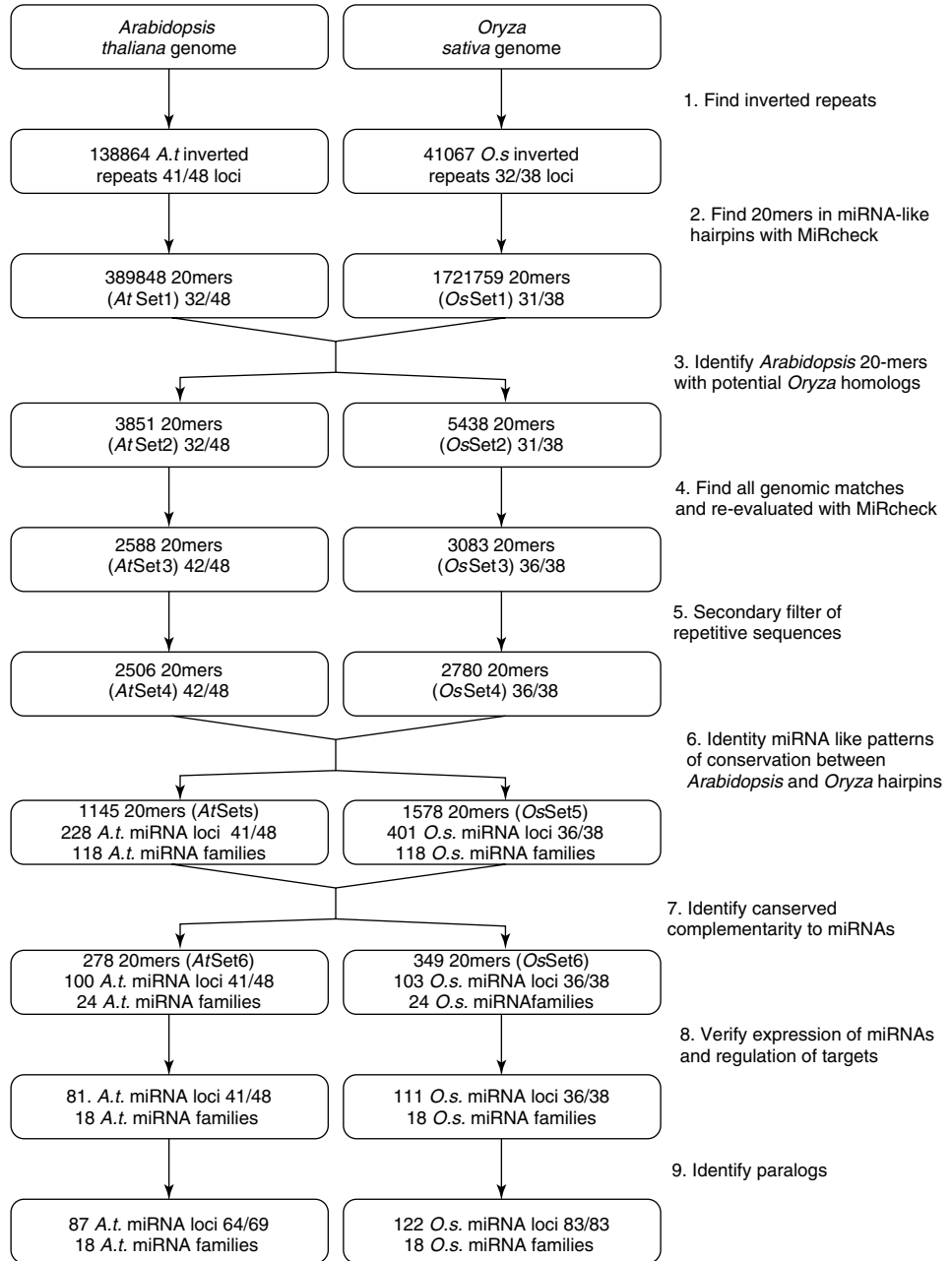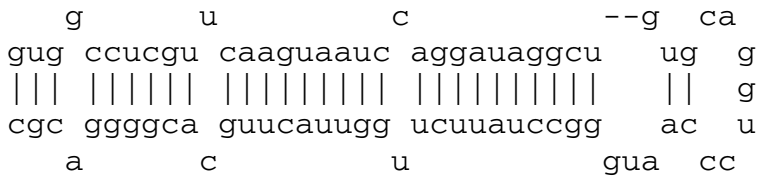
Figure 4.20   Flowchart of the miRNA prediction approach using two plant genomes. (Reprinted from Jones-Rhoades, M.W. and Bartel, D.P. (2004) Computational identification of plant microRNAs and their targets, including a stressinduced miRNA. *Mol. Cell* **14**: 787-799, with permission form Elsevier.)

to search for miRNA and their targets in sequences of a range of model eukaryotic organisms (http://www.softberry.com/berry.phtml?topic=index&group =programs&subgroup=rnastruct).

4.17 PREDICTION OF microRNA TARGETS

While hundreds of miRNAs have been deposited in the databases, their regulatory targets have not been established or predicted for many of them. Finding regulatory targets for plant miRNA is simply performed by looking for near-perfect complementarity to the mRNAs. For example, in a search for the targets of 13 *Arabodopsis* miRNA families, 49 unique targets were found with just a few false predictions (Rhoades *et al.*, 2002). However, animal miRNA targets have complementarity to the miRNAs only in the 'seed' sequence (usually 2–8 nucleotides numbered from the 5′ end) and often have multiple regions of complementarity, therefore more sophisticated search methods considering these features have recently been published (Stark *et al.*, 2003; Enright *et al.*, 2003; Lewis *et al.*, 2003; Rehmsmeier *et al.*, 2004). In general, miRNA, target genes are selected on the basis of three properties: sequence complementarity using a position-weighted local alignment algorithm, free energies of RNA–RNA duplexes, and conservation of target sites in related genomes. Lewis *et al.* (2003) in their TargetScan software took into account multiple miRNA–mRNA UTR complementary regions summing Z-scores (exp(-G/T)) produced by each such region in evaluating a potential target mRNA, where G

Homo sapiens miR-26a-1 stem-loop structure:

```
    g       u          c              --g  ca
 gug ccucgu caaguaauc aggauaggcu    ug  g
 ||| |||||| ||||||||| ||||||||||    ||  g
 cgc ggggca guucauugg ucuuauccgg    ac  u
    a       c          u              gua  cc
```

Two predicted target sites:

```
5'    UGCCU---CUGGAAAACUAUUGAGCCUUGCAUGUACUUGAAG
       |||     |||||                   |||||||||
      UCGGAUAGGACCUA------------------ AUGAACUU 5'
```
                                    −21.8 kcal/mol

```
5'    GAGCCUU-----GAUAAUACUUGAC
      |||||        |||  ||||||||
      UCGGAUAGGACCUA--AUGAACUU 5'
```
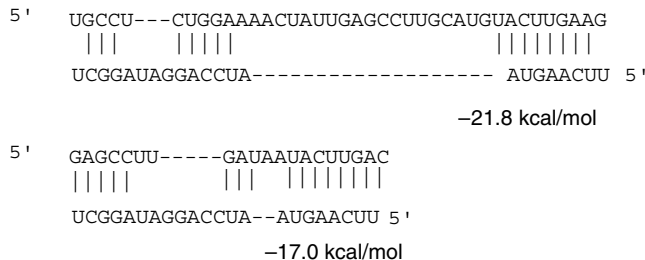   −17.0 kcal/mol

Figure 4.21 An example of stem-loop structure and predicted target sites for miR26a in human SMAD1 gene.

Table 4.22   Web server software for eukaryotic gene and functional signals prediction.

| Program/task	WWW address
Fgenesh/HMM-based gene prediction (Human, Drosophila, Dicots, Monocots, *C.elegans*, *S. pombe* and etc.)	http://sun1.softberry.com/berry.phtml?topic= fgenesh &group=programs&subgroup=gfind
Genscan/HMM-based gene prediction (Human, Arabidipsis, Maize)	`http://genes.mit.edu/GENSCAN.html`
HMM-gene/HMM-based gene prediction (Human, C.elegans)	`http://www.cbs.dtu.dk/ services/HMMgene/`
Fgenes/Disciminative gene prediction (Human)	`http://sun1.softberry.com/ berry.phtml?topic=fgenes&group= programs&subgroup=gfind`
Fgenesh-M/Prediction of alternative gene structures (Human)	`http://sun1.softberry.com/ berry.phtml?topic=fgenesh- m&group=programs&subgroup=gfind`
Fgenesh+/Fgenesh_c/ gene prediction with the help of similar protein/EST	`http://sun1.softberry.com/berry. phtml?topic=index&group= programs&subgroup=gfind`
Fgenesh-2/gene prediction using 2 sequences of close species	`http://sun1.softberry.com/ berry.phtml?topic=fgenes_c&group= programs&subgroup=gfs`
BESTORF/Finding best CDS/ORF in EST (Human, Plants, Drosophila)	`http://sun1.softberry.com/ berry.phtml?topic=bestorf&group= programs&subgroup=gfind`
FgenesB/gene, operon, promoter and terminator prediction in bacterial sequences	`http://sun1.softberry.com/ berry.phtml?topic=index&group= programs&subgroup=gfindb`
Mzef/internal exon prediction (Human, Mouse, Arabidopsis, Yeast	`http://rulai.cshl.org/tools/ genefinder/`
FPROM/TSSP/ promoter prediction	`http://sun1.softberry.com/ berry.phtml?topic=index&group= programs&subgroup=promoter`
NSITE/search for functional motifs	
Promoter 2.0/promoter prediction	`http://www.cbs.dtu.dk/services/ Promoter/`
CorePromoter/promoter prediction	`http://rulai.cshl.org/ tools/genefinder/CPROMOTER/ index.htm`
SPL/SPLM/splice-site prediction (Human, Drosophila, Plants nd etc.)	`http://www.softberry.com/ berry.phtml?topic=spl&group= programs&subgroup=gfind`
NetGene2/NetPGene/splice-site prediction (Human, C.elegans, Plants)	`http://www.cbs.dtu.dk/services/ NetPGene/`
Scan2 searching for similarity in genomic sequences and its visualization	`http://sun1.softberry.com/ berry.phtml?topic=scan2&group= programs&subgroup=scanh`
RNAhybrid prediction of microRNA target duplexes	`http://bibiserv.techfak.uni- bielefeld.de/rnahybrid/`
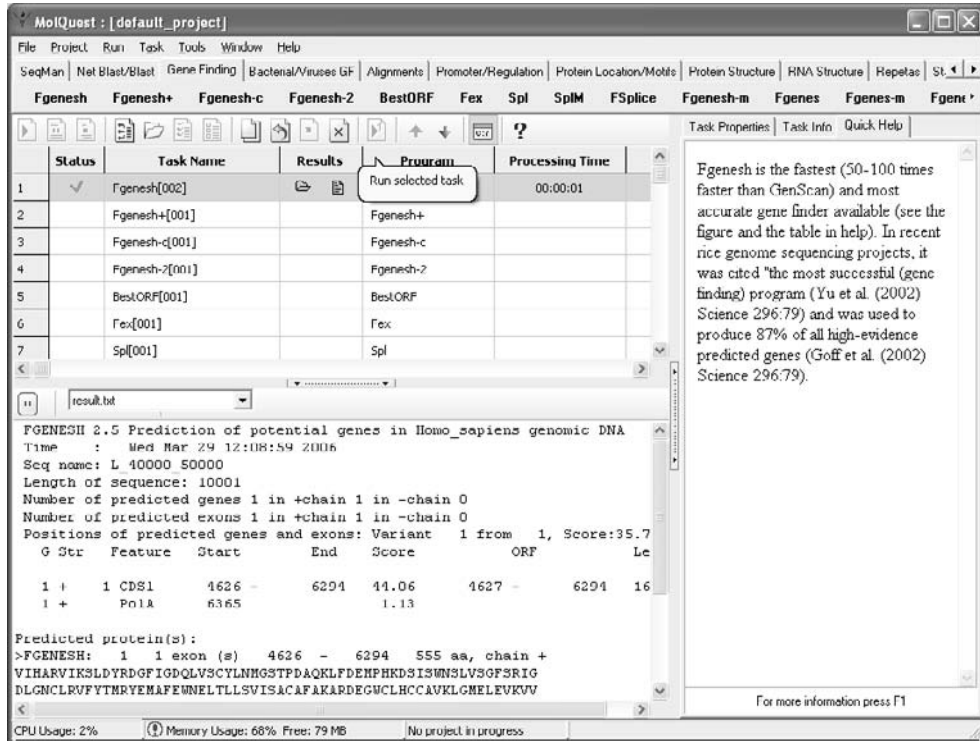
Figure 4.22 A user interface of **MolQuest** comprehensive desktop package for gene finding, sequence analysis and molecular biology data management.

is the free energy of miRNA:target site interaction. An example of stem-loop structure and predicted target sites for miR26a in human *SMAD1* gene is presented in Figure 4.21. Using TargetScan ~400 regulatory target genes have been predicted for the conserved vertebrate miRNAs. Eleven predicted targets (out of 15 tested) were supported experimentally (Lee *et al.*, 2003).

4.18   INTERNET RESOURCES FOR GENE FINDING AND FUNCTIONAL SITE PREDICTION

Prediction of genes, ORF, promoter, splice sites finding by the methods described in the preceding text is mostly available via World Wide Web. Table 4.22 presents WEB addresses of some of them. Many of these programs can be used within window-based **Molquest** computer package (www.molquest.com). It is the most comprehensive, easy-to-use desktop application for desktop sequence analysis (see Figure 4.22). The package includes gene-finders family (fgenesh/fgenesh+) programs for many organisms as well as pipelines (fgenesh++ and fgeneshb_annotator) that often used for fully automatic annotation eukaryotic and bacterial genomes (or genome
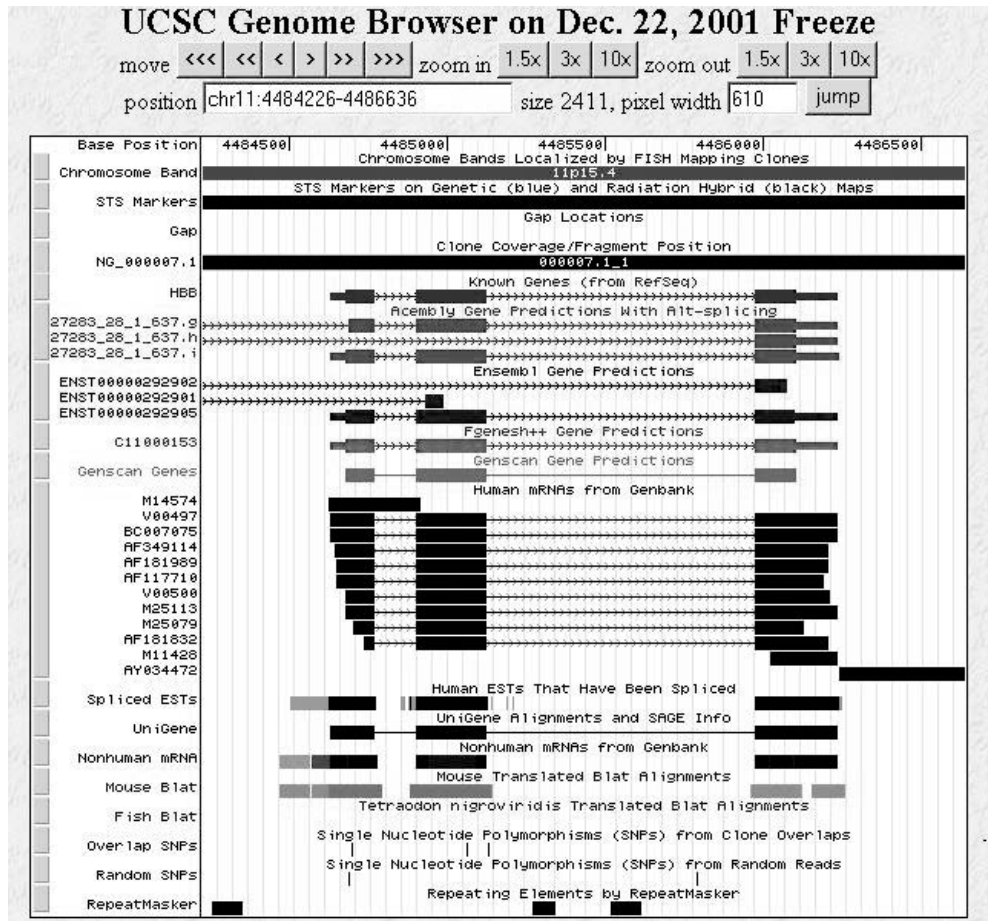
Figure 4.23 A screenshot of UCSC Genome Browser displaying gene predictions computed by various approaches.

communities) (The Honeybee Genome Sequencing Consortium, 2006; Tyson *et al.*, 2004). The package provides a user-friendly interface for sequence editing, primer design, internet database searches, gene prediction, promoter identification, regulatory elements mapping, patterns discovery protein analysis, multiple sequence alignment, phylogenetic reconstruction, and a wide variety of other functions. A lot of information generated during new genomes annotations (including gene predictions) is available through various genome browsers. A screenshot of popular UCSC Genome Browser (http://genome.ucsc.edu/) is presented in Figure 4.23. Another such interactive tool Genome Explorer (http://sun1.softberry.com/berry.phtml?topic= human&group=genomexp) can show a graph of expression data for selected genes (Figure 4.24). Its version for annotations of bacterial genomes is demonstrated in Figure 4.25. These web browsers provide search of numerous genome elements, visualization and retrieval of gene and protein sequences and fast comparison with auser-provided
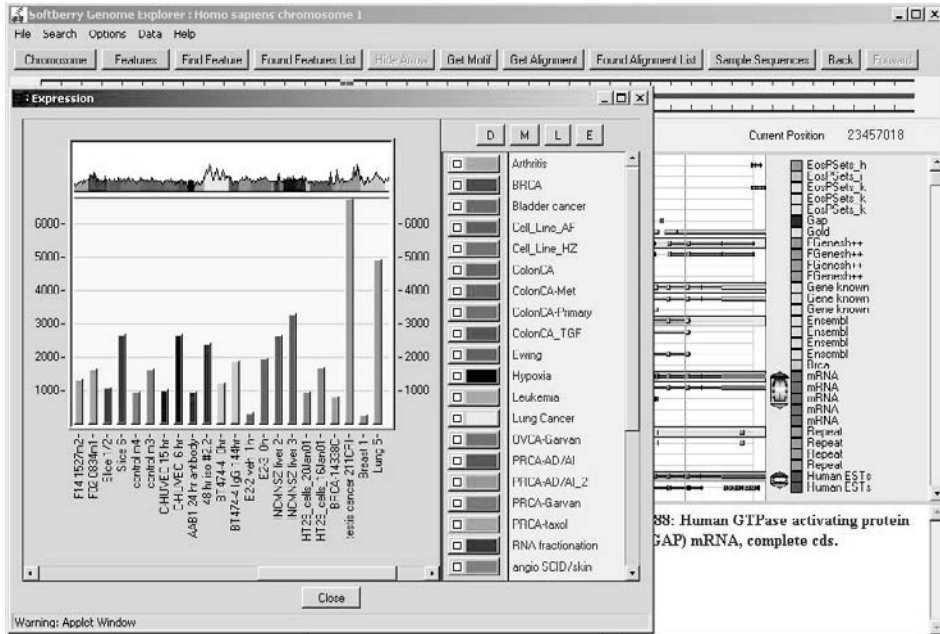
Figure 4.24 The Genome Explorer annotation browser showing graph of expression data for selected genes.
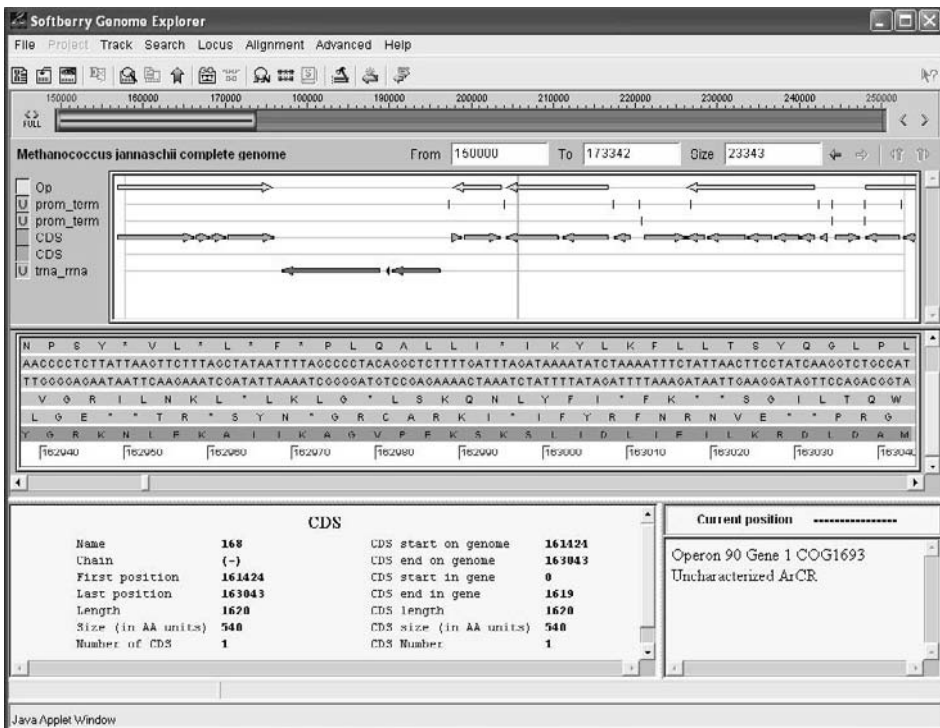

Figure 4.25 The Bacterial genome explorer displaying predicted operons, genes, promoters and terminators.

sequences. They are actively used not only by academic research community but also by many drug discovery and biotechnology companies for identification of drug candidates.

Acknowledgments

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Abril, J.F., Agbayani, A., An, H.J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D., Scheeler, F., Shen, H., Shue, B.C., Siden-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., WoodageT, Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye, J., Yeh, R.F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M., Venter and J.C. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.

Afifi, A.A. and Azen, S.P. (1979). *Statistical Analysis. A Computer Oriented Approach*. Academic Press, New York.

Allen, J.E., Majoros, W.H., Pertea, M. and Salzberg, S.L. (2006). JIGSAW, GeneZilla and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. *Genome Biology* **7**(Suppl. 1), S9.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**(17), 3389–3402.

Arumugam, M., Wei, C., Brown, R.H. and Brent, M.R. (2006). Pairagon+N-SCAN_EST: a model-based gene annotation pipeline. *Genome Biology* **7**(Suppl. 1), S5.1–S5.10.

Audic, S. and Claverie, J. (1997). Detection of eukaryotic promoters using Markov transition matrices. *Computers and Chemistry* **21**, 223–227.

Bajic, V., Brent, M., Brown, R., Frankish, A., Harrow, J., Ohler, U., Solovyev, V. and Tan, S. (2006). Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biology* **7**(Suppl. 1), S3.1–S3.13.

Baren, M. and Brent, M. (2006). Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Research* **16**, 678–685.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., Rapp, B.A. and Wheeler, D.L. (1999). GenBank. *Nucleic Acids Research* **27**(1), 12–17.

Berg, O.G. and von Hippel, P.H. (1987). Selection of DNA binding sites by regulatory proteins. *Journal of Molecular Biology* **193**, 723–750.

Birney, E. and Durbin, R. (2000). Using GeneWise in the *Drosophila* annotation experiment. *Genome Research* **10**, 547–548.

Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993). dbEST–database for "expressed sequence tags". *Nature Genetics* **4**(4), 332–333.

Borodovsky, M. and McIninch, J. (1993). GENMARK: parallel gene recognition for both DNA strands. *Computers and Chemistry* **17**, 123–133.

Borodovskii, M., Sprizhitskii, Yu., Golovanov, E. and Alexandrov, N. (1986). Statistical patterns in the primary structures of functional regions of the genome in *Escherichia coli. II*. nonuniform Markov models. *Molekulyarnaya Biologia* **20**, 1114–1123.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978). Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proceedings of the National Academy of Sciences* **75**(10), 4853–4857.

Breathnach, R. and Chambon, P. (1981). Organization and expression of eukaryotic split genes coding for proteins. *Annual Review of Biochemistry* **50**, 349–393.

Brunak, S., Engelbrecht, J. and Knudsen, S. (1991). Prediction of Human mRNA donor and acceptor sites from the DNA sequence. *Journal of Molecular Biology* **220**, 49–65.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology* **212**, 563–578.

Bucher, P. and Trifonov, E. (1986). Compilation and analysis of eukaryotic PolII promoter sequences. *Nucleic Acids Research* **14**, 10009–10026.

Burge, C. (1997). Identification of genes in human genomic DNA. Ph.D. Thesis, Stanford pp 152.

Burge, C. (1998). Modelling dependencies in pre-mRNA splicing signals. In *Computational Methods in Molecular Biology*, S. Salzberg, D. Searls, and S. Kasif, eds. Elsevier, 129–164.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78–94.

Burset, M. and Guigo, R. (1996). Evaluation of gene structure prediction programs. *Genomics* **34**(3), 353–367.

Burset, M., Seledtsov, I. and Solovyev, V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research* **28**(21), 4364–4375.

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N. and Izawa, M., (1996). High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37**(3), 327–336.

Cooper, S., Trinklein, N., Anton, E., Nguyen, L. and Myers, R. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1 % of the human genome. *Genome Research* **16**, 1–10.

Decker, C.J. and Parker, R. (1995). Diversity of cytopasmatic functions for the 3′-untranslated region of eukaryotic transcripts. *Current Opinions in Cell Biology* **7**, 386–392.

Diamond, M., Miner, J., Yoshinaga, S. and Yamamoto, K. (1990). Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. *Science* **249**, 1266–1272.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F. and Hannon, G.J. (2004). Processing of primary microRNAs by the microprocessor complex. *Nature* **432**, 231–235.

Dietrich, R., Incorvaia, R. and Padgett, R. (1997). Terminal intron dinucleotides sequences do not distinguish between U2- and U12-dependent introns. *Molecular Cell* **1**, 151–160.

Dong, S. and Searls, D. (1994). Gene structure prediction by linguistic methods. *Genomics* **23**, 540–551.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, pp 344.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D.S. (2003). MicroRNA targets in Drosophila. *Genome Biology* **5**, R1.

Farber, R., Lapedes, A. and Sirotkin, K. (1992). Determination of eukaryotic protein coding regions using neural networks and information theory. *Journal of Molecular Biology* **226**, 471–479.

Fickett, J. and Hatzigeorgiou, A. (1997). Eukaryotic promoter recognition. *Genome Research* **7**, 861–878.

Fickett, J.W. and Tung, C.S. (1992). Assesment of protein coding measures. *Nucleic Acids Research* **20**, 6441–6450.

Fields, C. and Soderlund, C. (1990). GM: a practical tool for automating DNA sequence analysis. *CABIOS* **6**, 263–270.

Forney, G.D. (1973). The Viterbi algorithm. *Proceedings of the IEEE* **61**, 268–278.

Gelfand, M. (1989). Statistical analysis of mammalian pre-mRNA splicing sites. *Nucleic Acids Research* **17**, 6369–6382.

Gelfand, M. (1990). Global methods for the computer prediction of protein-coding regions in nucleotide sequences. *Biotechnology Software* **7**, 3–11.

Gelfand, M. and Roytberg, M. (1993). Prediction of the exon-intron structure by a dynamic programming approach. *BioSystems* **30**(1–3), 173–182.

Gelfand, M., Mironov, A. and Pevzner, P. (1996). Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 9061–9066.

Ghosh, D. (1990). A relational database of transcription factors. *Nucleic Acids Research* **18**, 1749–1756.

Ghosh, D. (2000). Object-oriented transcriptional factors database (ooTFD). *Nucleic Acids Research* **28**, 308–310.

Grad, Y., Aach, J., Hayes, G.D., Reinhart, B.J., Church, G.M., Ruvkun, G. and Kim, J. (2003). Computational and experimental identification of C.elegans microRNAs. *Molecular Cell* **11**, 1253–1263.

Green, P., Hillier, L. (1998). *Genefinder, unpublished software*. It is still unpublished.

Guigo, R. (1998). Assembling genes from predicted exons in linear time with dynamic programming. *Journal of Computational Biology* **5**, 681–702.

Guigo, R. (1999). DNA composition, codon usage and exon prediction. In *Genetics Databases*, Academic Press, pp. 54–80.

Guigo, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T.R., Harrow, J., Hubbard, T., Lewis, S.E. and Reese, M.G. (2006). EGASP: the human ENCODE genome annotation assessment project. *Genome Biology* **7**(Suppl. 1), S2-1–S2-31.

Guigo, R., Knudsen, S., Drake, N. and Smith, T. (1992). Prediction of gene structure. *Journal of Molecular Biology* **226**, 141–157.

Guigo, R. and Reese, M.G. (2005). EGASP collaboration through competition to find human genes. *Nature Methods* **2**(8), 577.

Halees, A.S., Leyfer, D. and Weng, Z. (2003). PromoSer: a large-scale mammalian promoter and transcription start site identification service. *Nucleic Acids Research* **31**, 3554–3559.

Hall, S.L. and Padgett, R.A. (1994). Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *Journal of Molecular Biology* **239**(3), 357–365.

Hall, S.L. and Padgett, R.A. (1996). Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. *Science* **271**, 1716–1718.

Harrison, P., Milburn, D., Zhang, Z., Bertone, P. and Gerstein, M. (2003). Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Research* **31**(3), 1033–1037.

Henderson, J., Salzberg, S. and Fasman, K. (1997). Finding genes in DNA with a hidden Markov model. *Journal of Computational Biology* **4**, 127–141.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. and Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Research* **30**(1), 38–41.

Hutchinson, G. (1996). The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Computer Applications in the Biosciences* **12**, 391–398.

Hutchinson, G.B. and Hayden, M.R. (1992). The prediction of exons through an analysis of splicible open reading frames. *Nucleic Acids Research* **20**, 3453–3462.

Jackson, I.J. (1991). A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Research* **19**(14), 3795–3798.

Jeffrey, H.J. (1990). Chaos game representation of gene structure. *Nucleic Acids Research* **18**, 2163–2170.

Jones-Rhoades, M.W. and Bartel, D.P. (2004). Computational identification of plant microRNAs and their targets, including a stressinduced miRNA. *Molecular Cell* **14**, 787–799.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F.G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R. (2005). The EMBL nucleotide sequence database. *Nucleic Acids Research* **33**, D29–D33.

Kel, O., Romaschenko, A., Kel, A., Wingender, E. and Kolchanov, N. (1995). A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Research* **23**, 4097–4103.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002). The human genome browser at UCSC. *Genome Research* **12**(6), 996–1006.

Knudsen, S. (1999). Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics* **15**, 356–361.

Kolchanov, N.A., Podkolodnaya, O.A., Ananko, E.A., Ignatieva, E.V., Stepanenko, I.L., Kel-Margoulis, O.V., Kel, A.E., Merkulova, T.I. and Goryachkovskaya, T.N. (2000). Transcription regulatory regions regions database (TRRD): its status in 2000. *Nucleic Acids Research* **28**, 298–301.

Kondrakhin, Y.V., Shamin, V.V. and Kolchanov, N.A. (1994). Construction of a generalized consensus matrix for recognition of vertebrate pre-mRNA 3′-terminal processing sites. *Computer Applications in the Biosciences* **10**, 597–603.

Krogh, A. (1997). Two methods for improving performance of an HMM and their application for gene finding. *Intelligent Systems in Molecular Biology* **5**, 179–186.

Krogh, A. (2000). Using database matches with HMMgene for automated gene detection in *Drosophila*. *Genome Research* **4**, 523–528.

Krogh, A., Mian, I.S. and Haussler, D. (1994). A hidden Markov Model that finds genes in *Escherichia coli* DNA. *Nucleic Acids Research* **22**, 4768–4778.

Kulp, D., Haussler, D., Rees, M. and Eeckman, F. (1996). A generalized hidden Markov model or the recognition of human genes in DNA. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, D. States, P. Agarwal, T. Gaasterland, L. Hunter and R. Smith, eds. AAAI Press, St. Louis, MO, pp. 134–142.

Lapedes, A., Barnes, C., Burks, C., Farber, R. and Sirotkin, K. (1988). Application of neural network and other machine learning algorithms to DNA sequence analysis. *In Proceedings Santa Fe Institute* **7**, 157–182.

Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S. and Kim, V.N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**(6956), 415–419.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H. and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *EMBO Journal* **23**, 4051–4060.

Lewis, B.P., Shih, I., Jones-Rhoades, M.W., Bartel, D.P. and Burge, C.B. (2003). Prediction of mammalian microRNA targets. *Cell* **115**, 787–798.

Lukashin, A.V. and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Research* **26**, 1107–1115.

Mallory, A.C. and Vaucheret, H. (2004). MicroRNAs: something important between the genes. *Current Opinion in Plant Biology* **7**, 120–125.

Manley, J.L. (1995). A complex protein assembly catalyzes polyadenylation of mRNA precursors. *Current Opinion in Genetics and Development* **5**, 222–228.

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta* **405**, 442–451.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E. and Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* **34**, D108–D110.

McLauchlan, J., Gaffney, D., Whitton, J.L. and Clements, J.B. (1985). The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3′ termini. *Nucleic Acids Research* **13**, 1347–1367.

Milanesi, L. and Rogozin, I.B. (1998). Prediction of human gene structure. In *Guide to Human Genome Computing*, 2nd edition. M.J. Bishop, ed. Academic Press, London, pp. 215–259.

Mount, S. (1982). A catalogue of splice junction sequences. *Nucleic Acids Research* **10**, 459–472.

Mount, S.M. (1993). Messenger RNA splicing signal in Drosophila genes. In *An Atlas of Drosophila Genes*, G. Maroni. Oxford University Press, Oxford.

Nakata, K., Kanehisa, M. and DeLisi, C. (1985). Prediction of splice junctions in mRNA sequences. *Nucleic Acids Research* **13**, 5327–5340.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**, 418–426.

Nilsen, T.W. (1994). RNA-RNA interactions in the spliceosome: unraveling the ties that bind. *Cell* **78**, 1–4.

Ohler, U., Harbeck, S., Niemann, H., Noth, E. and Reese, M. (1999). Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics* **15**, 362–369.

Ohler, U., Liao, G.C., Niemann, H. and Rubin, G.M. (2002). Computational analysis of core promoters in the *Drosophila* genome. *Genome Biology*, **3**(12), research0087.1–research0087.12.

Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P. and Burge, C.B. (2004). Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* **10**, 1309–1322.

Pedersen, A.G., Baldi, P., Chauvin, Y. and Brunak, S. (1999). The biology of eukaryotic promoter prediction – a review. *Computers and Chemistry* **23**, 191–207.

Perier, C.R., Praz, V., Junier, T., Bonnard, C. and Bucher, P. (2000). The eukaryotic promoter database (EPD). *Nucleic Acids Research* **28**, 302–303.

Pillai, R. (2005). MicroRNA function: multiple mechanisms for a tiny RNA? *RNA* **11**, 1753–1761.

Prestridge, D. (1995). Predicting Pol II promoter sequences using transcription factor binding sites. *Journal of Molecular Biology* **249**, 923–932.

Prestridge, D. and Burks, C. (1993). The density of transcriptional elements in promoter and non-promoter sequences. *Human Molecular Genetics* **2**, 1449–1453.

Proudfoot, N.J. (1991). Poly(A) signals. *Cell* **64**, 617–674.

Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **33**(1), D501–D504.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**(2), 257–285.

Rabiner, L., Juang, B. (1993). *Fundamentals of speech recognition*. Prentice Hall, New Jersey, p. 507.

Reese, M.G., Harris, N.L. and Eeckman, F.H. (1996). *Large Scale Sequencing Specific Neural Networks for Promoter and Splice Site Recognition. Biocomputing: Proceedings of the 1996 Pacific Symposium*, L. Hunter and T. Klein, eds. World Scientific Publishing Company, Singapore.

Reese, M., Kulp, D., Tammana, H. and Haussler, D. (2000). Genie – Gene finding in *Drosophila melanogaster*. *Genome Research* **10**, 529–538.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R. and Ruvkun, G. (2000). The 21- nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**, 901–906.

Rehmsmeier, M., Steffen, P., Hochsmann, M. and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507–1517.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B. and Bartel, D.P. (2002). Prediction of plant microRNA targets. *Cell* **110**, 513–520.

Salamov, A.A. and Solovyev, V.V. (1997). Recognition of 3′-end cleavage and polyadenilation region of human mRNA precursors. *CABIOS* **13**(1), 23–28.

Salamov, A. and Solovyev, V. (2000). Ab initio gene finding in Drosophila genomic DNA. *Genome Research* **10**, 516–522.

Salsberg, S., Delcher, A., Fasman, K. and Henderson, J. (1998). A decision tree system for finding genes in DNA. *Journal of Computational Biology* **5**, 667–680.

Schmid, C.D., Perier, R., Praz, V. and Bucher, P. (2006). EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Research* **34**, D82–D85.

Schmid, C.D., Praz, V., Delorenzi, M., Perier, R. and Bucher, P. (2004). The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Research* **32**, D82–D85.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A. and Shinozaki, K. (2002). Functional annotation of a full-length Arabidopsis cDNA collection. *Science* **296**, 141–145.

Shahmuradov, I.A., Gammerman, A.J., Hancock, J.M., Bramley, P.M. and Solovyev, V.V. (2003). PlantProm: a database of plant promoter sequences. *Nucleic Acids Research* **31**, 114–117.

Shahmuradov, I., Solovyev, V. and Gammerman, A. (2005). Plant promoter prediction with confidence estimation. *Nucleic Acids Research* **33**(3), 1069–1076.

Shahmuradov, I.A., Kolchanov, N.A., Solovyev, V.V. and Ratner, V.A. (1986). Enhancer-like structures in middle repetitive sequences of the eukaryotic genomes. *Genetics (Russ)* **22**, 357–368.

Shahmuradov, I.A., Solovyev, V.V. (1999). NSITE program for identification of functional motifs with estimation of their statistical significance `http://sun1.softberry.com/berry.phtml?topic=nsite&group=programs&subgroup=promoter`.

Sharp, P.A. and Burge, C.B. (1997). Classification of introns: U2-type or U12-type. *Cell* **91**, 875–879.

Senapathy, P., Sahpiro, M. and Harris, N. (1990). Splice junctions, brunch point sites, and exons: sequence statistics, identification, and application to genome project. *Methods in Enzymology* **183**, 252–278.

Shepherd, J.C.W. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and ist possible evolutionary justification. *Proceedings of the National Academy of Sciences of the United States of America* **78**, 1596–1600.

Smit, A. and Green, (1997). RepeatMasker Web server: `http:// repeatmasker.genome. washington.edu/cgi-bin/RepeatMasker`.

Snyder, E.E. and Stormo, G.D. (1993). Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Research* **21**, 607–613.

Snyder, E. and Stormo, G. (1995). Identification of protein coding regions in genomic DNA. *Journal of Molecular Biology* **21**, 1–18.

Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., Gibbs, R.A., Angerer, R.C., Angerer, L.M., Arnone, M.I., Burgess, D.R., Burke, R.D., Coffman, J.A., Dean, M., Elphick, M.R., Ettensohn, C.A., Foltz, K.R., Hamdoun, A., Hynes, R.O., Klein, W.H., Marzluff, W., McClay, D.R., Morris, R.L., Mushegian, A., Rast, J.P., Smith, L.C., Thorndyke, M.C., Vacquier, V.D., Wessel, G.M., Wray, G., Zhang, L., Elsik, C.G., Ermolaeva, O., Hlavina, W., Hofmann, G., Kitts, P., Landrum, M.J., Mackey, A.J., Maglott, D., Panopoulou, G., Poustka, A.J., Pruitt, K., Sapojnikov, V., Song, X., Souvorov, A., Solovyev, V., Wei, Z., Whittaker, C.A., Worley, K., Durbin, K.J., Shen, Y., Fedrigo, O., Garfield, D., Haygood, R., Primus, A., Satija, R., Severson, T., Gonzalez-Garay, M.L., Jackson, A.R., Milosavljevic, A., Tong, M., Killian, C.E., Livingston, B.T., Wilt, F.H., Adams, N., Belle, R., Carbonneau, S., Cheung, R., Cormier, P., Cosson, B., Croce, J., Fernandez-Guerra, A., Geneviere, A.M., Goel, M., Kelkar, H., Morales, J., Mulner-Lorillon, O., Robertson, A.J., Goldstone, J.V., Cole, B., Epel, D., Gold, B., Hahn, M.E., Howard-Ashby, M., Scally, M., Stegeman, J.J., Allgood, E.L., Cool, J., Judkins, K.M., McCafferty, S.S., Musante, A.M., Obar, R.A., Rawson, A.P., Rossetti, B.J., Gibbons, I.R., Hoffman, M.P., Leone, A., Istrail, S., Materna, S.C., Samanta, M.P., Stolc, V., Tongprasit, W., Tu, Q., Bergeron, K.F., Brandhorst, B.P., Whittle, J., Berney, K., Bottjer, D.J., Calestani, C., Peterson, K., Chow, E., Yuan, Q.A., Elhaik, E., Graur, D., Reese, J.T., Bosdet, I., Heesun, S., Marra, M.A., Schein, J., Anderson, M.K., Brockton, V., Buckley, K.M., Cohen, A.H., Fugmann, S.D., Hibino, T., Loza-Coll, M., Majeske, A.J., Messier, C., Nair, S.V., Pancer, Z., Terwilliger, D.P., Agca, C., Arboleda, E., Chen, N., Churcher, A.M., Hallbook, F., Humphrey, G.W., Idris, M.M., Kiyama, T., Liang, S., Mellott, D., Mu, X., Murray, G., Olinski, R.P., Raible, F., Rowe, M., Taylor, J.S., Tessmar-Raible, K., Wang, D., Wilson, K.H., Yaguchi, S., Gaasterland, T., Galindo, B.E., Gunaratne, H.J., Juliano, C., Kinukawa, M., Moy, G.W., Neill, A.T., Nomura, M., Raisch, M., Reade, A., Roux, M.M., Song, J.L., Su, Y.H., Townley, I.K., Voronina, E., Wong, J.L., Amore, G., Branno, M., Brown, E.R., Cavalieri, V., Duboc, V., Duloquin, L., Flytzanis, C., Gache, C., Lapraz, F., Lepage, T., Locascio, A., Martinez, P., Matassi, G., Matranga, V., Range, R., Rizzo, F., Rottinger, E., Beane, W., Bradham, C., Byrum, C., Glenn, T., Hussain, S., Manning, F.G., Miranda, E., Thomason, R., Walton, K., Wikramanayke, A., Wu, S.Y., Xu, R., Brown, C.T., Chen, L., Gray, R.F., Lee, P.Y., Nam, J., Oliveri, P., Smith, J., Muzny, D., Bell, S., Chacko, J., Cree, A., Curry, S., Davis, C., Dinh, H., Dugan-Rocha, S., Fowler, J., Gill, R., Hamilton, C., Hernandez, J., Hines, S., Hume, J., Jackson, L., Jolivet, A., Kovar, C., Lee, S., Lewis, L., Miner, G., Morgan, M., Nazareth, L.V., Okwuonu, G., Parker, D., Pu, L.L., Thorn, R. and Wright, R. (2006). The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**, 941–952.

Solovyev, V.V. (1993). Fractal graphical representation and analysis of DNA and Protein sequences. *BioSystems* **30**, 137–160.

Solovyev, V. (1997). Fgenes – Pattern based finding multiple genes in human genome sequences. `http://www.softberry.com/berry.phtml?topic=fgenes&group=programs& subgroup=gfind`.

Solovyev, V., Kolchanov, N. (1994). Search for functional sites using consensus. In *Computer Analysis of Genetic Macromolecules. Structure, Function and Evolution*. N.A. Kolchanov and H.A. Lim, eds. World Scientific, pp. 16–21.

Solovyev, V.V., Korolev, S.V., Tumanyan, V.G. and Lim, H.A. (1991). A new approach to classification of DNA regions based on fractal representation of functionally similar sequences. *Proceedings of the National Academy of Sciences of USSR (Russ) (Biochemistry)* **319**(6), 1496–1500.

Solovyev, V., Kosarev, P., Seledsov, I. and Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biology* **7**(Suppl. 1), 10-1–10-12.

Solovyev, V.V., Lawrence, C.B. (1993a). Identification of Human gene functional regions based on oligonucleotide composition. In *Proceedings of First International Conference on Intelligent System for Molecular Biology*, L. Hunter, D. Searls and J. Shalvic, eds. AAAI Press, Menlo Park, Californiya, pp. 371–379.

Solovyev, V., Lawrence, C. (1993b). Prediction of human gene structure using dynamic programming and oligonucleotide composition. In *Abstracts of the 4th Annual Keck Symposium*, Pittsburgh, PA, p. 47.

Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Research* **22**, 6156–5153.

Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. (1995). Prediction of human gene structure using linear discriminant functions and dynamic programming. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, C. Rawling, D. Clark, R. Altman, L. Hunter, T. Lengauer and S. Wodak, eds. AAAI Press, Cambridge, pp. 367–375.

Solovyev, V.V. and Salamov, A.A. (1997). The Gene-Finder computer tools for analysis of human and model organisms' genome sequences. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, C. Rawling, D. Clark, R. Altman, L. Hunter, T. Lengauer and S. Wodak, eds. AAAI Press, Halkidiki, pp. 294–302.

Solovyev, V.V. and Salamov, A.A. (1999). INFOGENE: a database of known gene structures and predicted genes and proteins in sequences of genome sequencing projects. *Nucleic Acids Research* **27**(1), 248–250.

Solovyev, V., Shahmuradov, I., Akbarova, Y. (2003). The RegsiteDB: A database of transcription regulatory motifs of animal and plant eukaryotic genes: `http://www.softberry.com/ berry.phtml?topic=regsite`.

Staden, R. (1984a). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research* **12**, 505–519.

Staden, R. (1984b). Mesurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucleic Acids Research* **12**, 551–567.

Staden, R. and McLachlan, A. (1982). Codon prefernce and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Research* **10**, 141–156.

Stanke, M., Tzvetkova, A. and Morgenstern, B. (2006). AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biology* **7**(Suppl. 1), S11.

Stark, A., Brennecke, J., Russell, R.B. and Cohen, S.M. (2003). Identification of Drosophila microRNA targets. *PLoS Biology* **1**, 1–13.

Stormo, G.D. and Haussler, D. (1994). Optimally parsing a sequence into different classes based on multiple types of evidence. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 47–55.

Stormo, G.D., Schneider, T.D., Gold, L. and Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *Escherichia coli*. *Nucleic Acids Research* **10**, 2997–3011.

Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Morishita, S., Okubo, K., Sakaki, Y., Nakamura, Y., Suyama, A. and Sugano, S. (2001). Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Reports*, **2**, 388–393.

Suzuki, Y., Yamashita, R., Sugano, S. and Nakai, K. (2004). DBTSS, database of transcriptional start sites: progress report 2004. *Nucleic Acids Research* **32**, D78–D81.

Tarn, W.Y. and Steitz, J.A. (1996a). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell* **84**(5), 801–811.

Tarn, W.Y. and Steitz, J.A. (1996b). Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. *Science* **273**, 1824–1832.

Tarn, W.Y. and Steitz, J.A. (1997). Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. *Trends in Biochemical Sciences* **22**(4), 132–137.

The ENCODE Project Consortium, (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–639.

Tjian, R. (1995). Molecular machines that control genes. *Scientific American* **272**, 54–61.

Tjian, R. and Maniatis, T. (1994). Transcriptional activation: a complex puzzle with few easy pieces. *Cell* **77**, 5–8.

Thanaraj, T.A. (2000). Positional characterization of false positives from computational prediction of human splice sites. *Nucleic Acids Research* **28**, 744–754.

The Honeybee Genome Sequencing Consortium. (2006). Insights into social insects from the genome of the honey bee Apis mellifer. *Nature* **433**(7114), 931–949.

Thomas, A. and Skolnick, M. (1994). A probabilistic model for detecting coding regions in DNA sequences. *Ima Journal of Mathematics Applied in Medicine and Biology* **11**, 149–160.

Tyson, G., Chapman, J., Hugenholtz, P., Allen, E., Ram, R.J., Richardson, P., Solovyev, V., Rubin, E., Rokhsar, D. and Banfield, J.F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43.

Wadman, M. (1998). Rough draft' of human genome wins researchers' backing. *Nature* **393**, 399–400.

Wahle, E. (1995). 3′-end cleavage and polyadelanytion of mRNA precursor. *Biochimica et Biophysica Acta* **1261**, 183–194.

Wahle, E. and Keller, W. (1992). The biochemistry of the 3′-end cleavage and polyadenylation of mRNA precursors. *Annual Review of Biochemistry* **61**, 419–440.

Werner, T. (1999). Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome* **10**, 168–175.

Wieringa, B., Hofer, E. and Weissmann, C. (1984). A minimal intron length but no specific internal sequence is required for splicing the large rabbit Bglobin intron. *Cell* **37**, 915–925.

Wilusz, J., Shenk, T., Takagaki, Y. and Manley, J.L. (1990). A multicomponent complex is required for the AAUAAA-dependent cross-linking of a 64-kilodalton protein to polyadenylation substrates. *Molecular and Cellular Biology* **10**, 1244–1248.

Wingender, E. (1988). Compilation of transcription regulating proteins. *Nucleic Acids Research* **16**, 1879–1902.

Wingender, E., Dietze, P., Karas, H. and Knuppel, R. (1996). TRANSFAC: a database of transcription factors and their binding sites. *Nucleic Acids Research* **24**, 238–241.

Wu, Q. and Krainer, A.R. (1997). Splicing of a divergent subclass of AT-AC introns requires the major spliceosomal snRNAs. *RNA* **3:6**, 586–601.

Xu, Y., Einstein, J.R., Mural, R.J., Shah, M. and Uberbacher, E.C. (1994). An improved system for exon recognition and gene modeling in human DNA sequences. In *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*,

R. Altman, D. Brutlag, P. Karp, R. Lathrop and D. Searls, eds, , Menlo Park, Californiya, pp. 376–383.

Yada, T., Ishikawa, M., Totoki, Y., Okubo, K. (1994). Statistical analysis of human DNA sequences in the vicinity of poly(A) signal. Technical Report TR-876, Institute for New Generation Computer Technology.

Yu, J., Hu, S., Wang, J., Wong, G., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y. and Zhang, X. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296**, 79–92.

Zhang, M. and Marr, T. (1993). A weight array method for splicing signal analysis. *Computer Applications in the Biosciences* **9**, 499–509.

Comparative Genomics

J. Dicks

Department of Computational and Systems Biology, John Innes Centre, Norwich, UK

and

G. Savva

Centre for Environmental and Preventive Medicine, Wolfson Institute of Preventive Medicine, London, UK

Over the last couple of decades, numerous genetic and physical maps have been developed for a wide range of species. These data have led to the development of the field of *comparative genomics*, in which we analyse characteristics of whole genomes, in contrast to the analysis of single genes in comparative genetics. By comparing homologous markers between species, we can get a feel for their relative distributions on the chromosomes of the respective species. Such information also enables us to deduce segments of chromosomes where the gene content, and sometimes also the order of the genes, is similar in two or more species. More recently, DNA sequences of whole genomes have become available, enabling us to compare genomes at both the sequence and the gene levels. Furthermore, these datasets have provided us with more detailed information with which to study the processes involved in genome dynamics. The promise of the post-genome era means that we will see many fully sequenced genomes within the next decade and we should develop analytical methods that are mature when significant quantities of data become available. However, not all species will be sequenced, at least in the near future, and so we should continue to develop methods of analysis that are appropriate for both types of data. In this chapter, we give a flavour of the types of comparative genome analysis that are currently possible. We highlight particular problems such as phylogenetic inference and the use of maps to compare genomes. Finally, we look at problems that should be tackled to enable us to make the most of the emerging genomic data.

5.1   INTRODUCTION

Comparative genomics is pervasive in almost all branches of computational biology. At its broadest, comparative genomics is simply the comparison between two or more

genomic entities from different taxa. These entities can be genomic sequences (either whole sequences or subsets), gene order, gene content, or some other genomic feature such as codon usage or GC content. Furthermore, the way in which we compare these entities across taxa is dependent on the type of data we are analysing.

The growing focus on comparative genomics within computational biology is simple to understand. For example, we would like to be able to look at a particular dataset, say the DNA sequence of a gene, from a single organism and predict the functional characteristics of that dataset. Unfortunately, *ab initio* methods are currently rare or not yet powerful enough to be useful in everyday science. More commonly, we discover information about biological objects that we know little about by comparing them to other objects about which we have been able to gain information, usually by experimental means. Thus, many of the concepts in computational biology with which we are most familiar, such as functional prediction of DNA and protein sequences, have their foundations in comparative biology. Furthermore, we see how methodologies are evolving to take advantage of large quantities of genomic data in related species. For example, in the field of gene prediction, recently developed methods such as Twinscan and N-scan (`http://mblab.wustl.edu/software/`) make use of existing annotated genomes, other than that of the organism on which an analysis is being performed, to improve the results of the analysis. The considerable breadth of computational genomics means that we cannot cover the whole spectrum of methodologies here. Consequently, in this chapter we concentrate on a few areas that most closely embody the spirit of comparative genomics by comparing significant quantities of genomic information from two or more organisms to gain new biological insights. We do not aim to provide an exhaustive bibliography on each of the topics covered, but we introduce many of the key papers that in turn provide the interested reader with a more complete history of the relevant methodology.

In the kinds of analyses we examine in this chapter, we would like to compare the genomes of two or more organisms. But what sort of data do we wish to compare, what is our motivation for comparing them and how do we go about the comparison? To answer these questions, it is simplest to examine in turn each of the datasets. Essentially, different types of dataset have arisen as technological advances have been made. In the early days of comparative genomics, our datasets consisted of mapped genetic markers. Later, as sequencing technologies improved, we began to gain information on the sequences of small genomes, often organellar genomes, leading to the study of gene content and order. More recently, we have seen the sequencing of large eukaryotic organisms and a new series of analytical tools have arisen as a result.

In the next two sections we look at the concepts of homology and genomic mutation, both of which we must understand in order to carry out a comparative genomic analysis. In Section 5.4, we look at comparative mapping and the problems that comparing maps can solve. We refer to the complex nature of chromosomal evolution in Section 5.5 and use the concepts described in Section 5.3 to look at measures of gene order and content difference. We introduce models of chromosomal evolution and show how these models, together with these measures of difference, can be used in evolutionary studies. In Section 5.6, we examine new methodologies emerging for the analysis of genome sequences. These new methods include whole genome alignment, taking the lead from smaller-scale alignment tools, and genomic palaeontology, essentially a combination of sequence and gene order studies that together enable us to understand a genome's evolutionary past through self-comparison. Finally, in Section 5.7, we look at areas of potential future research.

Table 5.1   Examples of databases and software tools for comparative genomic analysis.

Databases	
COGs	`http://www.ncbi.nlm.nih.gov/COG/`
Ensembl	`http://www.ensembl.org/`
Inparanoid	`http://inparanoid.sbc.su.se/`
PhIGs	`http://PhIGs.org/`
Gene order software	
BADGER	`http://badger.duq.edu/`
BPAnalysis, DERANGE II	`http://www.mcb.mcgill.ca/~blanchem/software.html`
CHROMTREE	`http://cbr.jic.ac.uk/dicks/software/CHROMTREE/`
GOTREE	`http://www.mcb.mcgill.ca/~bryant/GoTree/`
GRAPPA	`http://www.cs.unm.edu/~moret/GRAPPA/`
GRIMM	`http://www-cse.ucsd.edu/groups/bioinformatics/GRIMM/`
MGR	`http://www-cse.ucsd.edu/groups/bioinformatics/MGR/`
ParIS server	`http://www.stats.ox.ac.uk/~miklos/`
SHOT	`http://www.bork.embl-heidelberg.de/~korbel/SHOT/`
Gene content software	
GeneContent	`http://xgu.zool.iastate.edu/`
MPP	`http://cbr.jic.ac.uk/dicks/software/mpp/`
SHOT	`http://www.bork.embl-heidelberg.de/~korbel/SHOT/`
Block-finding software	
ADHoRe, i-ADHoRE	`http://bioinformatics.psb.ugent.be/software.php`
CloseUp	`http://contact14.ics.uci.edu/closeup/`
DiagHunter	`http://www.tc.umn.edu/~cann0010/Software.html`
FISH	`http://www.bio.unc.edu/faculty/vision/lab/FISH/`
Gene Teams	`http://www-igm.univ-mlv.fr/~raffinot/geneteam.html`
GRIL	`http://asap.ahabs.wisc.edu/software/gril/`
LineUp	`http://titus.bio.uci.edu/lineup/`
Whole genome alignment software	
AVID, MAVID	`http://baboon.math.berkeley.edu/mavid/`
DIALIGN	`http://bibiserv.techfak.uni-bielefeld.de/dialign/`
LAGAN, MLAGAN, SLAGAN	`http://lagan.stanford.edu/lagan_web/index.shtml`
MAUVE	`http://gel.ahabs.wisc.edu/mauve/`
MUMmer	`http://mummer.sourceforge.net/`
TBA	`http://www.bx.psu.edu/miller_lab/`
WABA	`http://www.soe.ucsc.edu/~kent/xenoAli/`

Throughout the chapter we give examples of databases and software tools pertinent to the subject matter covered in the section. References for these items are given in the text, with URLs provided in Table 5.1.

5.2   HOMOLOGY

Comparative genomics is built upon the concept of *homology*. Haldane (1927) laid the foundations of comparative genetics by examining the coat colour of rodents and carnivores. He discussed genetic homology and stated that 'Structures in two species are said to be homologous when they correspond to the same structure in a common ancestor.' Historically, homology was established on the basis of a large array of experimental evidence (see Searle, 1968). Today, we are more used to seeing sequence similarity with, perhaps, the additional evidence of homologous flanking regions (i.e. being within a larger region of homology) as sufficient to state a putative homology.

For much of this chapter, we refer to homology in the context of genes and the similarity of their sequences. However, we should keep in mind that homology is equally relevant to non-genic objects (e.g. markers in comparative mapping studies are not always gene based). There are different categories of homology. For a given gene in one organism we can often, though not always, find a corresponding gene that carries out a similar role in a closely related organism. If the two genes arose from a common ancestor then they are *orthologues*. However, if they evolved from different origins but have similar sequences or functions as a consequence of convergent evolution, then they are analogous to one another and are known as *analogues*. Relationships between genes are further complicated by gene duplication, where we need to consider *paralogous* genes. Fitch (1970) wrote of the difficulties and importance of correctly distinguishing homologous from paralogous proteins. The recent burgeoning popularity of comparative genomic studies has led to this matter being discussed widely, for example by Remm *et al.* (2001). Here, the term *inparalogue* has been created to indicate paralogues that have arisen through a gene duplication event after speciation, such that a group of such genes may together be co-orthologous to a gene in another organism, while *outparalogues* have arisen following a gene duplication preceding speciation. In general, outparalogues should have a more diversified function than inparalogues. Figure 5.1 gives a depiction of the various types of homology, using the representation of O'Brien *et al.* (2005).

Information on homologous structures, particularly genes, is also now available in many public data resources. For example, the original Ensembl database (Hubbard *et al.*, 2002) was developed for dissemination of the annotated human genome sequence. However, it has since been applied to several other genome sequences and this has led to the development of resources for the linking of homologous genes across species and the alignment of pairs of groups of whole genome sequences. In addition, some databases are developed purely on the basis of homology. For example, the Inparanoid program (Remm *et al.*, 2001) was developed to identify orthologous clusters while differentiating between inparalogues, which are included in clusters, and outparalogues, which are not. The Inparanoid database (O'Brien *et al.*, 2005) provides data on eukaryotic orthologues from 17 genomes, examining over 500 000 proteins derived from Ensembl and UniProt. The Clusters of Orthologous Groups (COGs; Tatusov *et al.*, 1997) database compares the protein sequences of genes from complete genomes representing major phylogenetic lineages, grouping together genes that are co-located in multiple organisms. Each COG consists of individual proteins or groups of paralogues from at least three lineages and thus corresponds to an ancient conserved domain. Phylogenetically Inferred Groups (PhIGs; Dehal and Boore, 2006) is a set of databases and web tools that analyse gene
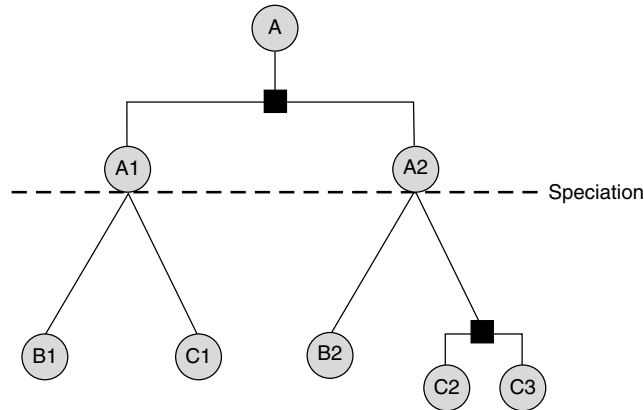
Figure 5.1 Homologous relationships between five contemporary genes in species B and C, all descended from a single gene in the ancestral species A. Black boxes denote gene duplications. Genes C2 and C3 are inparalogues and are co-orthologous to B2. B1 and C1 are orthologues but B1 is an outparalogue of B2 and of C2 and C3.

sets from completely sequenced genomes, building clusters of genes using a graph-based approach, and using maximum likelihood phylogenetic analysis to uncover the evolutionary relationships among all gene families.

5.3   GENOMIC MUTATION

Before we begin to look at some of the main types of comparative genome analysis, we must first understand the ways in which genomes evolve, so that we can understand their relevance to these analyses. It is widely known that DNA sequences evolve via the mutation events substitution, insertion and deletion, which affect single nucleotide sites or perhaps a small number of adjacent sites in the latter two cases. When comparing two relatively small sequences, it is usually sufficient to consider the differences between them in light of these events alone. However, genomes may also change by a series of much larger, but less frequent, mutations known as *chromosomal rearrangements*. When we compare larger genomic datasets such as long DNA sequences or gene orders, we also need to consider these events. Essentially, there are two types of chromosomal rearrangement: those that alter the gene content of the chromosome (the *non-conservative* rearrangements) and those that do not (the *conservative* rearrangements). For ease of understanding, it is simplest to consider chromosomal rearrangements by describing their effect on a series of linear and distinct regions along a genome, such as genes. However, it can be seen easily that these events affect the underlying genomic sequence in a similar way (as we see in later sections of this chapter).

First, let us consider the simple case of a single linear chromosome (this case can be adapted easily to that of a circular chromosome such as the mitochondria or chloroplast). The chromosome contains N genes g_i (for $i = 1$ to N) that are represented as signed integers, such that homologues in other species share the same number. The sign of the number represents the orientation of the gene, denoting the way in which it is *transcribed*,

or read. For example, if we denote a gene as '+1' on our first species then its homologues in all other species must be denoted as '+1' or '−1', depending on their orientations and regardless of their positions on their respective chromosomes. Therefore we can represent the order of genes on the chromosome as a signed permutation:

$$G = \{g_1 g_2, \ldots, g_{N-1} g_N\}.$$

For simplicity, we will define all mutation events as the outcomes of breaks *between genes*, in the intergenic regions. The locations of these breaks are known as *breakpoints*. Although this convention will not always represent biological reality, it is likely to represent the majority of cases.

There are three conservative (*inversion*, *shift* and *inverted shift*) and two non-conservative rearrangements (*tandem duplication* and *deficiency*) that can affect our single chromosome. An inversion (often termed a *reversal* in the computer science literature) is the result of two breaks, with the central segment turning about 180° before rejoining with the two broken ends. Consider e.g.

$$G = \{+1 + 2 + 3 + 4 + 5 + 6 + 7 + 8\} \longrightarrow G' = \{+1 + 2 + 3 - 6 - 5 - 4 + 7 + 8\}.$$

A shift, or *transposition*, is a three-break mutation, where a segment is moved to another part of the same chromosome. Consider e.g.

$$G = \{+1 + 2 + 3 + 4 + 5 + 6 + 7 + 8\} \longrightarrow G' = \{+1 + 2 + 6 + 7 + 3 + 4 + 5 + 8\}.$$

An inverted shift is similar to a shift, but here the segment inverts before it is inserted into its new location. A tandem duplication causes a short chromosomal segment to be duplicated and inserted adjacent to the original copy. Consider e.g.

$$G = \{+1 + 2 + 3 + 4 + 5 + 6 + 7 + 8\} \longrightarrow$$
$$G' = \{+1 + 2 + 3 + 4 + 4 + 5 + 6 + 7 + 8\}.$$

The new segment can remain in its original position or it can be moved to another location by subsequent mutations. For example, the evolution of the haemoglobin locus, from a single gene to a four-gene complex, is thought to be a result of repeated tandem duplication. Finally, a deficiency causes a segment to be removed from a chromosome. Consider e.g.

$$G = \{+1 + 2 + 3 + 4 + 5 + 6 + 7 + 8\} \longrightarrow G' = \{+1 + 2 + 3 + 4 + 5 + 8\}.$$

A genome consisting of two or more chromosomes can be affected by additional mutations (and of course any one of its chromosomes may undergo any of the mutations seen for the single-chromosome case). There are three major additional conservative mutations (*reciprocal translocation*, *centric fusion* and *dissociation*) that alter one or more chromosomes. A reciprocal translocation is a two-break mutation that occurs when parts of two chromosome arms swap with one another. Consider e.g.

$$G = \{+1 + 2 + 3 + 4 + 5\}\{+6 + 7 + 8\} \longrightarrow$$
$$G' = \{+1 + 2 + 3 + 7 + 8\}\{+6 + 4 + 5\}.$$

A centric fusion or *Robertsonian translocation* is a two-break mutation where two distinct chromosomes join at their centromeres (with the very small remaining parts of the chromosomes above the centromeres being lost – these do not usually contain genes). Consider e.g.

$$G = \{+1 + 2 + 3 + 4 + 5\}\{+6 + 7 + 8\} \longrightarrow G' = \{-8 - 7 - 6 + 1 + 2 + 3 + 4 + 5\}.$$

A dissociation (or *fission*) is a one-break mutation, where a single chromosome breaks at its centromere to produce two fully functional chromosomes. Consider e.g.

$$G = \{+1 + 2 + 3 + 4 + 5\} \longrightarrow G' = \{+1 + 2 + 3\}\{+4 + 5\}.$$

Whole genomes may also undergo *polyploidy* events, essentially a genome doubling. There are two kinds of polyploidy: auto-polyploidy and allo-polyploidy. An auto-polyploidy event occurs when a genome gives rise to a new genome with two copies of each of its chromosomes. The organism containing the new genome is an instant species that cannot interbreed with the species from which it derives. An allo-polyploidy event occurs when a new genome receives the genomes of two distinct (but usually closely related) species. Again, the organism containing the new genome belongs to a new instant species that cannot interbreed with either of the species from which it derives.

In addition to the genomic mutations described above, we may wish in future to define other mutations, as we become aware of them. For example, O'Keefe and Eichler (2000) have discussed the possibility of a duplicative transposition event in human evolution, where duplications of large genomic segments are transferred to a new location in the genome. However, the events described above are those most widely considered in comparative genomic analyses at the present time, and so will be sufficient to understand the methodologies introduced in the next sections.

5.4   COMPARATIVE MAPS

Many of the first comparative genome analyses were carried out using comparative maps. Even now, and despite the excitement generated by the human genome mapping program and the genome sequencing projects that preceded and followed it, whole genome sequences are available only for a small minority of organisms. For many species, physical and genetic maps are still highly fruitful resources. Comparative maps focus on the identification of homologous links between markers on two or more maps, where the evidence supporting those links may take a variety of forms, from sequence similarity to experimental evidence (e.g. cross hybridisation of probes). Figure 5.2 shows an example of a comparative map, comparing the locations of homologues between a physical map of *Arabidopsis thaliana* and a genetic map of *Brassica napus*. Note that the comparison appears to indicate an inversion between the two maps.

Comparative maps can be used to extrapolate information from one species, on the location and function of markers and genes, for application in another. For example, imagine that we know about gene A, which is involved in an important function in species 1. We would like to know if there is a 'similar' gene in species 2 that perhaps carries out a similar function. We know that gene A is flanked by genes B and C
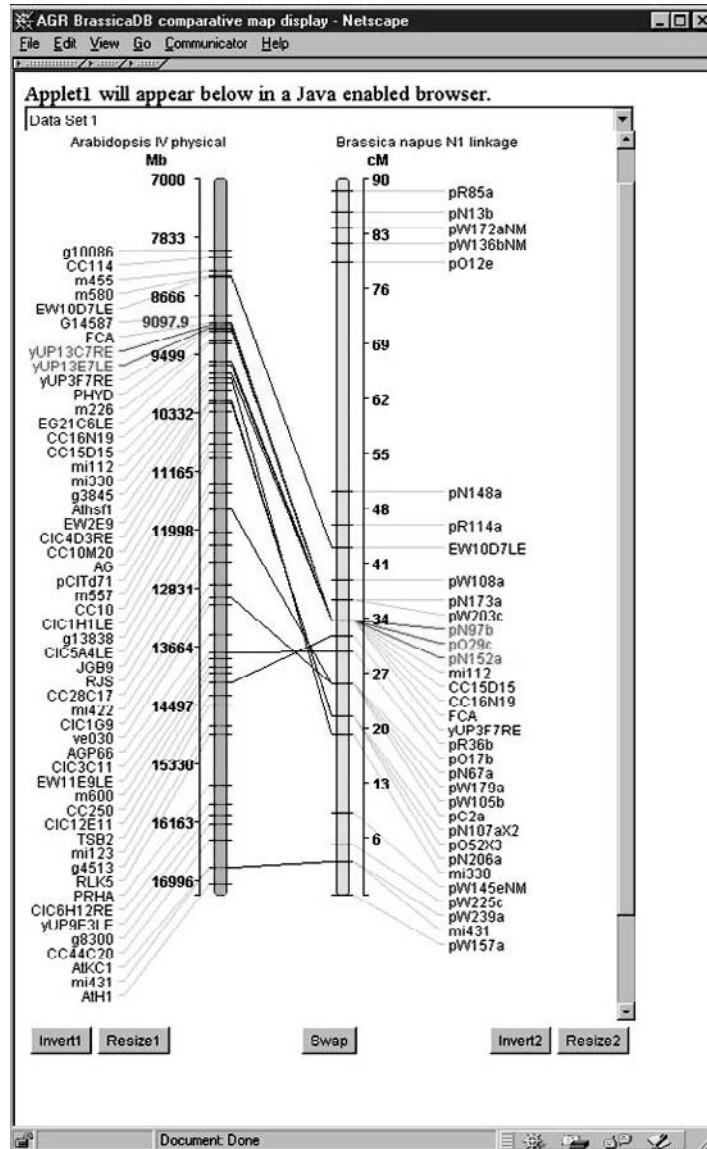
Figure 5.2 An example of a comparative map between two plants, *Arabidopsis thaliana* and *Brassica napus*. [Reproduced from Dicks *et al.*, 2000 UK CropNet: a collection of databases and bioinformatics resources for crop plant genomics, *Nucleic Acids Research* **28**(1), 104–107 by permission of Oxford University Press.]

in species 1. If we know the locations of the homologues of genes B and C in species 2 we can predict the location of a potential homologue in species 2. This provides a starting point for further experimentation to confirm or refute the prediction. Of course, this problem is often made more difficult by gene duplication because we cannot always know if two genes are orthologous or paralogous and we may look for a gene in the wrong part of

the genome. Alternatively, we may be interested in gene D in species 3 but we do not know its function. By looking at the functions of its homologues in other species and perhaps the functions of its close neighbours, we may gain more information on a putative function.

A comparative mapping analysis can help us answer questions such as the following:

- Where in species 2 is the homologue of gene A from species 1?

- What is the likely function of gene D in species 3?

- How many conserved chromosomal segments would we see if we knew about every gene within a genome, rather than the subset we see within a map?

When comparing genomes by their maps, we have information about homologous markers, which may or may not be gene based. There are gaps between these markers and we do not know how many genes lie within these gaps or where their homologues lie in other species. Other datasets may hold even less information, where a marker is known to be located on a particular chromosome (a *chromosome assignment*) but its map position is not known. This 'missing' information will undoubtedly bias the results of any analysis carried out on those data. Despite this, it is possible to make inferences about conserved chromosomal segments and evolution. Nadeau and Taylor (1984) used mapping data from the autosomal chromosomes of human and mouse to estimate the number and lengths of conserved chromosomal segments between these two species. They showed that the length of a segment $\hat{m}$, given the assumption that markers and breakpoints were scattered randomly across a genome (the *random breakage model*), could be found as follows:

$$\hat{m} = \frac{r(n+1)}{(n-1)}, \tag{5.1}$$

where n is the number of markers in a segment and r is the distance between the two outermost markers. They reasoned that segments with fewer than two markers would be missed and they used this to derive the expected sample mean of the transformed segment lengths $\hat{m}$:

$$E[x'] \cong \frac{(L^2 D + 3L)}{(LD + 1)}, \tag{5.2}$$

where L is the mean length of conserved segments measured in centimorgans and D is the density of mapped homologous loci in the genome. They further estimated the number of homology disruptions between human and mouse.

Waddington and colleagues generalised the work of Nadeau and Taylor in order to compare maps of chickens with those of human and mouse (Waddington, 2000; Waddington *et al.*, 2000). They no longer assumed that chromosome lengths were large when compared to segment lengths (important as chicken linkage groups vary considerably in length, with some likely to contain just one segment while others may contain several) and they did not assume chromosome breakage occurred at random. They used the beta distribution to model segment lengths and used a scaling parameter l_k, the length of chromosome k, to derive the density function of segment length y on chromosome k:

$$f(y_k) = \frac{\frac{1}{l_k} \left(\frac{y}{l_k}\right)^{a-1} \left(1 - \frac{y}{l_k}\right)^{b-1}}{B(a, b)}, \tag{5.3}$$

where a and b are the beta distribution parameters and $B(a, b)$ the beta function $\int_0^1 x^{a-1}$ $(1 - x)^{b-1}dx$. Consequently, the expected number of segments on chromosome k is $S_k = (a + b)/a$ and the mean segment length on chromosome k is $l_k(a + b)/a$. The special case $a = 1$ corresponds to the random breakage model. They also derived the distribution of the number of genes n in a syntenic group (count data), on the basis of a conserved segment of length y, and the joint distribution of n and the distance between the two outermost markers in the syntenic group (range data) and used these singly and in combination to analyse the chicken dataset. They noted that the flexibility of the new models made them suitable for the analysis of a wide range of datasets, particularly as their formulation enabled testing of the models and data assumptions.

Ehrlich *et al.* (1997) developed methods to estimate *synteny conservation* (i.e. where two or more homologous markers are located on the same chromosome in two or more species). They used their methods to find the percentage of conserved syntenies in human and mouse that had already been observed. They also showed that rates of chromosomal rearrangement varied considerably between mammalian lineages and that inter-chromosomal mutations had occurred four times more often than intra-chromosomal mutations in the lineages leading to human and mouse, despite the strong selective forces against them.

Sankoff *et al.* (2000b) analysed the Nadeau and Taylor approach, in light of the large number of genes mapped in both human and mouse since 1984 (approximately 1500 genes compared to the 83 analysed by Nadeau and Taylor). They concluded that the results of the Nadeau and Taylor analysis, which are consistent with current evolutionary estimates, were accurate due to the formulation of their model and the data used rather than due to luck alone. They also looked at the effect of using chromosomal assignment only on estimating the number of conserved chromosomal segments. They noted that, for example, small intra-chromosomal mutations (e.g. inversions) could result in two chromosomal segments being regarded as one (see Section 5.5 for further analysis of this phenomenon). For this case, they added a correction to their probability function $P(a, m, n)$:

$$P(a, m, n) = \frac{\binom{m-1}{a-1}\binom{n+1}{a}}{\binom{n+m}{m}}, \tag{5.4}$$

the probability of observing a non-empty segments for m genes and n breakpoints so that:

$$Q(b, m, n, c) = \sum_{a=b}^{n+1} P(a, m, n) \binom{a-1}{a-b} \left(\frac{1}{c}\right)^{a-b} \left(\frac{c-1}{c}\right)^{b-1}, \tag{5.5}$$

is the probability that only b of the a non-empty segments are counted on the c chromosomes being examined. They also calculated the number of segments due to inter- rather than intra-chromosomal rearrangements.

Schoen (2000) looked at the effect of marker density on the estimation of the number of chromosomal breakpoints between pairs of species for the random breakage model. He showed that the estimated number of breakpoints was close to the expected value when marker density was high. However, the amount of rearrangement could be underestimated when marker density was low, particularly when the species being analysed were distantly related. He also showed that underestimation could occur when inversions were common

in the divergence of the species and discussed the results of Ehrlich *et al.* (1997) in light of the bias against detecting inversion events.

5.5   GENE ORDER AND CONTENT

Sequencing of small genomes, such as the mitochondria of organisms from many lineages, led to the first datasets where the content and order of all the genes was known. Unlike comparative mapping datasets, these datasets did not contain gaps between the datapoints. These early whole genome datasets led to new methods for genomic comparison, and indeed the field is still highly active today, with the progression from organellar datasets to gene order and gene content of large nuclear genomes.

5.5.1   Gene Order

Gene orders have been studied computationally since the late 1980s, such that current methodologies enable us to:

1.   Estimate the evolutionary distance between two gene orders.

2.   Find the sequence of changes that could have altered the gene order of one genome to that of another.

3.   Estimate an evolutionary tree based on gene order data.

4.   Hypothesise the gene order and content of an ancestral species.

The problems listed above can be tackled with reference to some measure of *genome difference* between a pair of genomes. Genome difference measures can be broadly grouped into two categories: *distance based* and *path decomposition*. The first of these includes several fairly simple measures of the differences between gene orders. The second category involves analysing likely sequences of mutations between gene orders with identical gene contents. However, as we see later, the two categories are not mutually exclusive and some distance measures are based upon the results of a path decomposition analysis. Here we do not present an exhaustive list of distance measures but rather give a flavour of the types of distances that have been proposed.

5.5.1.1   *Distance Measures*

When two DNA sequences are being compared, the *edit distance* is defined as the smallest number of substitutions that transforms one sequence into another. This is fairly simple to calculate and has been used successfully for many years. An analogous distance for gene order was proposed by Sankoff *et al.* (1992). This edit distance d_E is based on the *insertion/deletion distance* d_D, which is just the number of genes found in one genome that are not present in the other, and the *rearrangement distance* d_R, which is the minimum number of chromosomal rearrangements required to convert one gene order into the other (allowing for any differences in gene content) such that $d_E = d_D + d_R$. This distance measure was applied to the gene orders of a variety of plant, animal, and fungal

mitochondrial genomes. A value of d_E was calculated for each pair of gene orders and a distance matrix constructed from the results. A least-squares algorithm was then applied to the distance matrix to obtain a phylogenetic tree. At the time, practical application of the algorithm was very slow, mainly due to the calculation of the rearrangement distance but, as we see below, considerable advances in algorithmic efficiency have been made since then.

In order to compare gene orders more rapidly, Sankoff and Blanchette (1997; 1998) defined the *breakpoint distance* d_{BP}. This is simply the number of adjacent gene pairs in one gene order that are not adjacent (and in a consistent relative orientation) in a second. For example, the breakpoint distance between the two gene orders below (with G_1 being the *identity gene order* for seven genes) is 3, with the vertical bars between genes in G_2 denoting the locations of the breakpoints:

$$G_1 = \{+1+2+3+4+5+6+7\}$$
$$G_2 = \{+1|-6-5|-7|+2+3+4\}.$$

Sankoff *et al.* (2000a) further developed the *normalised induced breakpoint distance* to allow the comparison of very differently sized genomes. First, genes that are not common to both species are removed before calculating the breakpoint distance on the remaining genes, the result being known as the *induced breakpoint distance*. Second, this distance is normalised by dividing it by the number of the remaining shared genes.

Similar to the analysis of breakpoints, gene adjacency conservation was used as a simple measure by Keogh *et al.* (1998) to compare different fungal genomes. The neighbour pair distance d_{NP} is simply the proportion of adjacent gene pairs in one genome that are adjacent in the second. For G_1 and G_2 above, $d_{NP} = 0.5$. Cosner *et al.* (2000) also looked at neighbouring pairs but used them as binary characters – known as a *binary encoding*. Pairs that were adjacent (and of consistent orientation) were scored as '1' and those that were not as '0'. This encoding, which in addition could be weighted, could then be analysed using a maximum parsimony analysis in order to deduce an evolutionary tree.

The distances above are not suitable for application to genomes containing many paralogous genes, as is the case for most eukaryote genomes. Consequently Sankoff (1999) defined the *exemplar distance* for use in this situation. This distance measure involves reducing the size of each gene family to 1, deleting all family members except for the gene that leads to the smallest distance between the reduced genomes. The exemplar distance can be calculated on the basis of any of the above distances (e.g. the exemplar breakpoint distance or the exemplar edit distance). For example, in the simple case below:

$$\{+1+2+3-4+6+7+2-5\},$$

deleting the first instance of the gene '+2' would lead to a breakpoint distance with the identity gene order of 5, whereas deleting the second instance would lead to a distance of 3. Consequently, we would delete the second instance of the gene in order to calculate the exemplar breakpoint distance. Although the breakpoint distance may be calculated rapidly in linear time, calculation of the exemplar breakpoint distance is an NP-hard problem, as there are many combinations in which the gene family members can be deleted in more complex cases. Sankoff chose to calculate the exemplar distance using a

branch-and-bound technique. However, this approach has been surpassed recently by that of Nguyen *et al.* (2005) who used a divide-and-conquer approach.

5.5.1.2 Path Decomposition

The sequence of chromosomal rearrangements that mutates the gene order G_i to gene order G_j is known as an *evolutionary path* $\rho_{i,j}$. Finding such a path or paths between two genomes is known as a path decomposition analysis. For example, one successful path between G_2 and G_1 above, involving just inversion events, is as follows, with underlined segments indicating the location of the mutation at each step along the path:

$$G_2 = \{+1\underline{-6 - 5 - 7} + 2 + 3 + 4\}$$
$$\{+1\underline{+7 + 5 + 6} + 2 + 3 + 4\}$$
$$\{+1\underline{-4 - 3 - 2} - 6 - 5 - 7\}$$
$$\{+1 + 2 + 3 + 4\underline{-6 - 5} - 7\}$$
$$\{+1 + 2 + 3 + 4 + 5 + 6\underline{-7}\}$$
$$\{+1 + 2 + 3 + 4 + 5 + 6 + 7\} = G_1.$$

Note that the number of mutations on this path is different from the number of breakpoints and, indeed, the relationship between these two quantities is not straightforward, with the former generally thought to be a better indicator of true evolutionary distance.

A simple strategy to find an evolutionary path between two fairly short gene orders is to carry out a tree searching technique. In an unrestricted tree search we would mutate one of our genomes in all possible ways, according to our evolutionary model (i.e. our n_E allowed mutations). We are left with n_E new gene orders. We then mutate the n_E gene orders, each according to our n_E mutations to give us n_E^2 new gene orders. We repeat this process, the result being a branching search tree. At each node of the search tree we check to see if we have a genome identical to our target genome. If this is the case, we have achieved a successful path. Thus, the number of nodes in the search tree of depth l is:

$$\sum_{k=1}^{l} n_E^k = \frac{n_E(n_E^l - 1)}{n_E - 1}. \tag{5.6}$$

Obviously, the number of nodes quickly becomes very large, even for a small number of genes. This is greatly pronounced in multi-chromosome genomes and indeed, n_E is not necessarily a constant in this case and so the search tree can become quite complex. In practice, the tree must be pruned according to some heuristic.

The tree searching approach has been used by Blanchette *et al.* (1996) and Dicks (1999). The authors differed slightly in their pruning strategies and more markedly in the goals of their search. Blanchette *et al.* looked for the shortest path(s) separating the query genome from the target genome, pruning the search tree with d_{BP} and using a depth-first search with limited look ahead and branch and bound for a rapid search. Different weights were used for inversions and shifts, comparing the optimal path to the random case. This method was implemented in the DERANGE

II software. Dicks again used d_{BP} to prune the tree with a depth-first search, but here the search did not stop once the shortest path was achieved. Instead, the computation was stopped once the tree had been completely pruned or once a conservative upper bound on tree depth had been achieved, whichever occurred first. The result was a *set of paths R* separating the two genomes, which were then used to calculate a transition likelihood between G_i and G_j, under some probabilistic model M. We can define this likelihood as follows (setting the constant of proportionality to 1):

$$L(M|G_i, G_j) = \sum_R P(\rho_{i,j}|M), \quad (5.7)$$

which is often approximated by:

$$L(M|G_i, G_j) = \max_R P(\rho_{i,j}|M). \quad (5.8)$$

The former approach uses all paths to calculate the transition likelihood. Note that the real set of paths separating two gene orders is of infinite size and that the real transition likelihood is merely approximated by summing over R. A combined likelihood approach was shown by Thorne *et al.* (1992) to be less biased than a single maximum likelihood in the analysis of DNA sequence alignment and, although not formally tested for chromosomal evolution, it is likely to be the case here. The latter definition uses the maximum likelihood path $\hat{\rho}_{i,j}$, which is often used for computational convenience. Under any sensible model of evolution, where $\ell(\rho)$ is the length of a path ρ:

$$\lim_{\ell(\rho) \to \infty} P(\rho) = 0. \quad (5.9)$$

Ideally, we would like to be able to use this fact to choose just those paths that make a significant contribution to the combined transition likelihood and we could even consider using it as a tree-pruning strategy. Indeed Savva (2001) has shown that, for simple cases, just a few short paths make a very large contribution to the combined transition likelihood. Unfortunately, for large gene orders, a tree-based approach is not computationally feasible.

One alternative to finding the maximum likelihood path between two gene orders is to find the most parsimonious path between them. This research area began in earnest in 1995, with the introduction of the Hannenhalli–Pevzner algorithm (Hannenhalli and Pevzner, 1995a) for sorting signed permutations by reversals (i.e. finding the shortest path between two gene orders while permitting only inversion events to occur). This paper has been extremely influential in gene order studies, defining concepts such as the breakpoint graph that have been used both in other parsimony algorithms and in the statistical approaches introduced below. To create a breakpoint graph for a given gene order of size n, the gene order is first transformed so that, (1) a gene i with a positive sign is replaced by elements $2i - 1$ $2i$, (2) a gene j with a negative sign is replaced by elements $2|j|$ $2|j| - 1$ and (3) the gene order is *framed* by elements 0 and $2n + 1$. Thus, the gene order $\{+1 \ -6 \ -5 \ -7 \ +2 \ +3 \ +4\}$ becomes $\{0 \ 1 \ 2 \ 12 \ 11 \ 10 \ 9 \ 14 \ 13 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 15\}$. A graph is then constructed from this transformed gene order by drawing edges between (1) elements on either side of a breakpoint (e.g. between 2 and 12) and (2) even-numbered elements and those with values one larger when the two elements are not adjacent. The result is a graph where all

elements are either isolated or part of a cycle. In our example, it is easy to discover that the number of cycles c is 2. In the subsequent decade, several polynomial-time algorithms have been developed for other mutations or for combinations of mutations, such as for translocations (e.g. Hannenhalli, 1996) and for inversions, translocations, fissions and fusions (e.g. Hannenhalli and Pevzner, 1995b). For other mutations or combinations, such as transpositions, fissions and fusions (e.g. Dias and Meidanis, 2001), heuristic or approximation algorithms have been developed. Many of the major advances are cited in Yancopoulos *et al.* (2005) so we will not list them here. This paper also describes a new algorithm to sort multi-chromosomal gene orders by inversions, reciprocal translocations, fissions, fusions and block interchanges (an event which swaps two segments on a single chromosome, resulting in a transposition event when the segments are contiguous), thereby including all of the major conservative mutations. Interestingly, all mutations are based on a common operation known as a *double-cut-and-join* and the rearrangement distance d is simply the number of breakpoints b minus the number of cycles c in the corresponding breakpoint graph.

The path decomposition problem has also been studied using Markov chain Monte Carlo (MCMC) techniques. York *et al.* (2002) used a simple Poisson process model with a uniform prior for the mean to estimate the posterior distribution of the number of inversions between two gene orders. This method has an advantage over non-statistical approaches in that credible sets of estimates are made. Importantly, simulation studies showed that the method was capable of providing good estimates of the true numbers of inversion events and that these numbers were often underestimated by parsimony approaches. Miklos (2003) also studied this problem, again using a Poisson process model, but examined the broader problem of inversions, shifts and inverted shifts. The method was further implemented in the ParIS genome rearrangement server (Miklos *et al.*, 2005).

The path or paths resulting from a path decomposition analysis can be used directly to answer evolutionary questions about the number and types of mutation separating two gene orders. Furthermore, path decomposition analyses lead naturally to mutation-based distance measures between two genomes. Firstly, we note that the number of events in the shortest path is simply the rearrangement distance d_R as used above by Sankoff *et al.* (1992). Secondly, we can calculate a probabilistic distance measure based on the path decomposition. For example, for each genome pair, where our data are evolving under some stochastic model M, we can find a maximum likelihood estimate for time, $\hat{t}$, as the distance between our two genomes. Durbin *et al.* (2000) showed that maximum likelihood estimates of distance are additive, lending themselves well to a subsequent phylogenetic analysis. Consequently, a few distances based upon maximum likelihood estimates have been developed. Dicks (1999) modelled the number of breakpoints between two gene orders as a function of the mutations lying on an evolutionary path and used it to calculate the probability of a given path for an observed value of d_{BP}. Savva (2006) also investigated the joint conditional probability distribution of d_{BP} and d_R and used it to calculate path likelihoods under a simple evolutionary model. The latter method is implemented in the CHROMTREE software. Caprara and Lancia (2000) also showed some interesting results, such as the expected number of breakpoints in a random permutation of N genes:

$$E[D_{BP}] = N - 1, \tag{5.10}$$

and the expected number of breakpoints in a random permutation of N genes after a random path of k inversions:

$$E[N(k)] = (N-1)\left(1 - \left(\frac{N-3}{N-1}\right)^k\right), \tag{5.11}$$

using the results to calculate the conditional probability of the number of inversions given an observed value of d_{BP}.

5.5.1.3 *Phylogenetic Analysis*

Chromosomal banding patterns gave rise to what was probably the first comparative genome analysis, carried out by Dobzhansky and Sturtevant (1938), where the orders of bands on the chromosomes of the fruit fly *Drosophila* were used to derive an evolutionary tree. This analysis was carried out visually, made possible by the relative simplicity of the dataset. Today, we wish to carry out similar analyses, but for more complex datasets and in a systematic manner.

Markov chains and, more recently hidden Markov models (HMMs), have been used widely to model the evolution of DNA and amino acid sequences, with these models being used successfully for phylogenetic inference in the Felsenstein likelihood framework (Felsenstein, 1981). For DNA sequences, where each nucleotide site is considered independently, the Markov property lends itself well to modelling the transitions between the four possible states A, C, G and T in time t. Even when considering pairs of adjacent sites, as is done for modelling RNA sequences, we need only consider 16 potential states. Amino acid evolution is a little more complicated, with 20 states to consider. We can see that chromosomal mutations also cause gene orders to move between states and it would be natural to consider an analogous model for their evolution. However, a gene order is a signed permutation, where adjacent genes are not independent. Imagine a small chromosome consisting of just 10 genes. If we wished to use a Markov chain model, we would need to consider $10! \times 2^{10}$, or a little over 3.7 billion states. Obviously, this is computationally infeasible, even for such a small dataset. On the basis of the tree search approach discussed above, Dicks (2000) tried to circumvent this problem by suggesting an approximation to the maximum likelihood tree for gene order data, using a small set of states at internal nodes (i.e. those states that lie along successful pairwise search paths), where those states are thought to make a large overall contribution to the tree likelihood. However, such an approach is slow even for small gene orders.

The simplest way to estimate a gene order phylogeny is to carry out a pairwise analysis, perhaps using one of the distance measures described above. A distance matrix is constructed, where each entry in the matrix is the distance between a pair of gene orders. This matrix is then used as input to a tree-building algorithm. The neighbor-joining algorithm (Saitou and Nei, 1987) is perhaps the best known of these algorithms but others such as the FITCH (Fitch and Margoliash, 1967) and KITSCH algorithms within the PHYLIP phylogenetic package (Felsenstein, 1993) are also applicable. Each of these algorithms has its own advantages and disadvantages, as has been described frequently elsewhere. Two software packages developed specifically for gene order analysis are GOTREE and SHOT. GOTREE calculates breakpoint distances for gene orders with both equivalent and differing gene contents and estimates phylogenetic trees

using neighbor-joining. SHOT takes genome sequences as its data rather than gene orders. It first finds homologous genes and subsequently estimates trees using various pairwise distances. It can also perform tree bootstrapping using the jackknife.

A more difficult problem is to find the maximum parsimony tree, searching for a topology and internal node gene order states that minimise the sum of the distances over all edges of the tree. This problem was studied by Sankoff and Blanchette (1998) and Blanchette *et al.* (1999) where the statistic to be minimised was the number of breakpoints on the tree edges. This method was implemented in the BPAnalysis software. The approach was improved by other researchers, as described in Tang and Moret (2003), with later algorithms also facilitating the search for most parsimonious trees that minimised the reversal distance. Tang and Moret's method was implemented in the GRAPPA software. Bourque and Pevzner (2002) also studied the problem of minimising the reversal distance over a tree topology, calling it the Multiple Genome Rearrangement (MGR) problem. They showed how the approach could be extended to the multi-chromosomal case, including not only inversions but also reciprocal translocations, fusions and fissions. This approach is implemented in the MGR software which itself uses algorithms from the GRIMM software (Tesler, 2002) for calculating shortest distances between pairs of gene orders.

The approaches discussed so far illustrate a growing efficiency in the solution of optimal gene order trees, where the main criteria for success are speed, efficiency and parsimony. However, these approaches can tell us little about the accuracy of such methods, except perhaps through simulation studies. In contrast, statistical approaches have the benefit of assessing the uncertainty of parameter estimates. Bayesian MCMC methods have been used successfully to search through tree space for the topology with the maximum posterior probability in the case of DNA sequences. Larget *et al.* (2002) also showed how MCMC methods could be applied to gene order problems, devising a model for inversion events on single chromosomes that could be used in a Bayesian phylogenetic analysis. The model priors defined a uniform distribution for the tree topology τ over the space of all unrooted trees with ℓ leaves and a gamma distribution $\Gamma(\alpha, \lambda)$ for branch lengths β. Counts of inversions were independent Poisson variables with means equal to the relevant branch lengths. Inversion events were chosen uniformly, at random, along a branch from the set M_n, where the size of the set is $\binom{n+1}{2} = n(n+1)/2$ and the position of the event on the branch was a distance u from its start. Given that D is an indicator variable representing whether the observed data were consistent with the parameters and unobservable variables, and which takes the value 1 when this is the case and 0 when it is not, the joint posterior distribution of all of the parameters was shown to be:

$$p(\tau, \beta, x, r, u|D) \propto p(\tau, \beta, x, r, u)p(D|\tau, \beta, x, r, u)$$

$$= \frac{1}{|T_\ell|} \prod_{i=1}^{2\ell-3} g(\beta_i)h(x_i|\beta_i) \left(\frac{1}{\beta_i}\right)^{x_i} \left(\frac{1}{|M_n|}\right)^{x_i} 1_{\{(\tau,x,r) \hookrightarrow D\}}, \quad (5.12)$$

where

$$g(\beta_i) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \beta_i^{\alpha-1} e^{-\lambda\beta_i}$$

and

$$h(x_i|\beta_i) = \frac{e^{-\beta_i} \beta_i^{x_i}}{x_i!},$$

which, after integrating out branch lengths β and inversion locations u analytically and ignoring factors not depending on τ, x or r, becomes

$$p(\tau, x, r | D) \propto 1_{\{(\tau,x,r) \hookrightarrow D\}} \left(\frac{1}{|M_n|(1+\lambda)} \right)^{\sum_{i=1}^{2\ell-3} x_i} \prod_{i=1}^{2\ell-3} \frac{\Gamma(\alpha + x_i)}{x_i!\,\Gamma(\alpha)}. \qquad (5.13)$$

An MCMC technique is used to sample from this distribution, using an update scheme with three moves. The first two moves affect just the histories (i.e. sequences of inversions) in localised areas of the tree, with the third potentially changing the tree topology as well as the histories. This latter move is based on the concept of the breakpoint graph, as described in Hannenhalli and Pevzner (1995a). Notably, York *et al.* (2002) and Miklos (2003) also used the breakpoint graph in their update schemes. Later computational advances in the update scheme enabled the method to be used on larger real datasets (Larget *et al.*, 2005b) and the model and method have been implemented in the BADGER software. Recently, the performance of the method has been contrasted to the maximum parsimony approach on a number of real datasets. For example, Larget *et al.* (2005a) reanalysed the herpesvirus dataset examined in Bourque and Pevzner (2002). This dataset consisted of the gene orders of Herpes simplex virus (HSV), Epstein–Barr virus (EBV) and Cytomegalovirus (CMV). The problem was to construct the state of the ancestral gene order. Figure 5.3 shows the solution found by Bourque and Pevzner, which involved seven inversions. Larget *et al.* showed that there were two most parsimonious solutions and that one of them (shown in Figure 5.3) had a posterior probability three times as great as the other. The Larget *et al.* analyses showed that the method was capable of finding parsimonious solutions as efficiently as parsimony software and also that, for some datasets, parsimonious solutions had a low probability of being correct. The paper also described a strategy for adopting the method within an empirical Bayes framework, which involved estimating values of the hyperparameters α and λ from the datasets undergoing analysis.

5.5.2 Fragile Breakage versus Random Breakage Models

As we saw above, many of the early algorithmic advances in gene order studies were inspired by organellar genomes, particularly mitochondrial genomes. Here, the order and arrangement of a small number of genes was provided, without the need to carry out a complex pre-processing of a raw dataset. However, the advent of sequencing large nuclear genomes has brought new challenges to gene order studies. Firstly, there has been a need to increase the efficiency of the algorithmic approaches. For example, improvements in the computing time of rearrangement distance algorithms (e.g. Bader *et al.*, 2001; Bergeron, 2001; Tesler, 2002) now enable them to be applied to genomes possessing thousands of genes. Secondly, nuclear genomes must be analysed to identify genes and to establish homologous relationships within and between genomes. This latter step is difficult and prone to error, particularly relatively soon after sequencing has been completed. However, these larger genomes also give us larger datasets with which to test hypotheses of evolutionary scenarios and to develop more realistic models of genomic rearrangement.

Pevzner and Tesler (2003a; 2003b) established the concept of the breakpoint reuse statistic r:

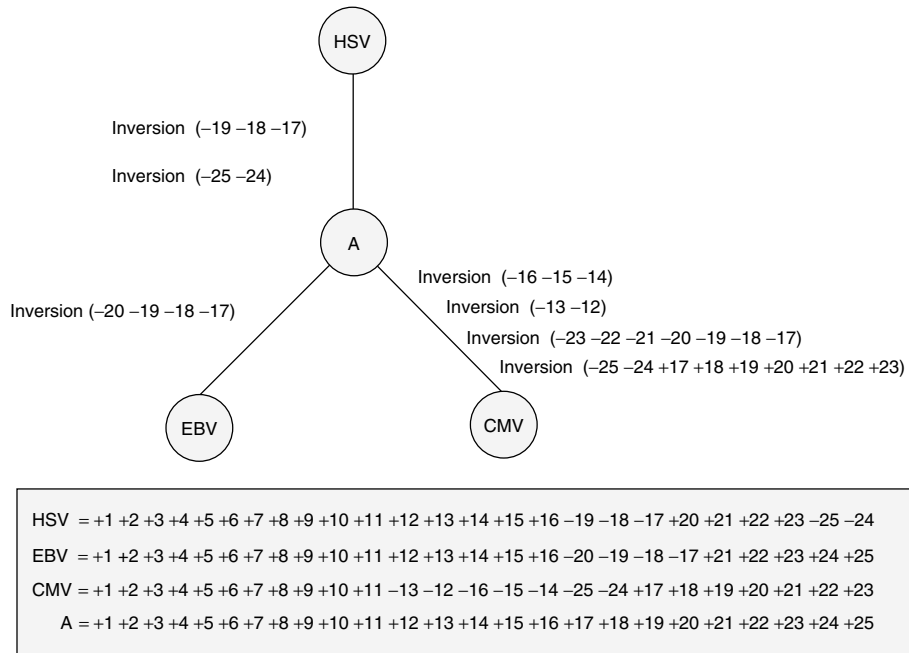$$r = \frac{2d}{b}, \qquad (5.14)$$

Figure 5.3 A gene order phylogeny of three herpesviruses, showing the inferred state A of the ancestral gene order for the most parsimonious tree.

where d is the rearrangement distance between two genomes of size n and b the number of breakpoints separating their gene orders. As it is known that $b/2 \leq d \leq b$ we see that $1 \leq r \leq 2$. A value of $r = 1$ indicates a rearrangement path where each event breaks the genome at two sites unique to that arrangement (i.e. no reuse, thereby supporting random breakage) and $r = 2$ where each event after the first breaks the genome at one new site and one site that is already used. Pevzner and Tesler identified a value of $r = 1.9$ between the human and mouse genomes, suggesting a high level of breakpoint reuse. Indeed, this *fragile breakage* hypothesis, where particular regions are implicated in multiple rearrangement events has been advanced by others (e.g. Kent *et al.*, 2003) and, of course, it had already been hypothesised in the case of comparative maps.

However Sankoff and Trinh (2005), while not rejecting the hypothesis of fragile breakage, argued that the statistic r was not a good estimator of breakpoint reuse and that the high value of r achieved by Pevzner and Tesler may have been due to the method with which they achieved their dataset. Rather than attempting to study large permutations of genes, Pevzner and Tesler analysed almost 600 000 short aligned sequence fragments and used them to build large collinear blocks, ignoring regions of local rearrangement. The result was a set of 281 large blocks. This method was highly useful in that it largely circumvented the problem of establishing gene order from genomic sequence.

Sankoff and Trinh showed through simulation that high values of r could be achieved with no breakpoint reuse, if a proportion of genes θ was deleted at random (essentially attempting to simulate Pevzner and Tesler's smoothing out of small rearrangements). In particular they showed that the greater the rearrangement distance between the two genomes, the more r grew with increasing θ. Furthermore, as θ increased, r initially

increased with a rate depending on d/n and then reached a maximum, before descending sharply. Sankoff and Trinh went on to model the relationship between these statistics, defining d_1 and d_2 as the number of one-breakpoint and two-breakpoint events needed to sort the genome such that $b = 2d_2 + d_1$. Furthermore, they used the model to estimate the rate of change in r:

$$\frac{\mathrm{d}r}{\mathrm{d}\theta} = n\left(\frac{2}{b(t)}\frac{-d_1(t)[d_1(t)-1]^+}{(1-\varepsilon(t))(n-t)(n-t-1)} - \frac{2d(t)}{b(t)2}\frac{-b(t)(b(t)-1)}{(n-t)(n-t-1)}\right), \quad (5.15)$$

where $\varepsilon(t) = d_2(t)/[b(t)(b(t)-1)]$, verifying that $\mathrm{d}r/\mathrm{d}\theta \approx 2d/n$ near to $\theta = 0$. Sankoff and Trinh suggested that, although the reuse statistic r may not be a good estimator of breakpoint reuse, it is a good indicator of a breakdown of genomic structure and therefore could be used to measure the strength of the evolutionary signal in the comparison of genome sequences. Peng *et al.* (2006) countered this argument by showing that the algorithm used by Sankoff and Trinh to calculate collinear blocks was inappropriate and did not mirror the Pevzner and Tesler approach. Peng *et al.* then demonstrated that their method, implemented in the GRIMM-Synteny software, which inferred a greater number of blocks than the Sankoff and Trinh method, did not inflate values of r so significantly in a similar simulation study, and consequently the high value of r from the human/mouse comparison could not be dismissed. They also hypothesised that inhomogeneity of gene distribution and long regulatory regions could in part be responsible for breakpoint reuse. Sankoff and Haque (2006) then went on examine the distribution of cycles in the breakpoint graph for two random genomes. They used their results, together with the relation $d = b - c$ as proved in Yancopoulos *et al.* (2005), to show that r tended to 2 as genomes became randomly ordered with respect to one other, backing up their previous argument. However, whether or not r can be used reliably as an indicator of the fragile breakpoint hypothesis, the model is gaining popularity through the results of empirical studies.

5.5.3 Gene Content

When we compare the *gene contents* of two genomes we are looking to see how many homologues they share and how many genes each contains for which the other does not have a homologue. On average, we would expect closely related species to exhibit roughly the same gene content and more distantly related species to possess species- or group-specific genes, although this does not always hold for individual cases. Data on gene content can be informative in evolutionary studies. For example, the availability of whole genome sequences has enabled many researchers to infer species phylogenies using gene content data (e.g. Herniou *et al.*, 2001; House and Fitz-Gibbon, 2002; Lin and Gerstein, 2000; Montague and Hutchison, 2000; Snel *et al.*, 1999; Tekaia *et al.*, 1999; Whittam and Bumbaugh, 2002; Winzeler *et al.*, 2003). These phylogenetic studies use a variety of approaches, including distance measures derived from fractions of shared genes in genome pairs and parsimony. Gene content data can also be highly useful on a practical level. For example, in a recent analysis of the genomes of three parasites (*Trypanosoma brucei*, *Trypanosoma cruzi* and *Leishmania major*), the authors found a 'core' of 6200 conserved genes (Berriman *et al.*, 2005). It was noted that these common genes provided targets for new drugs that could potentially fight all three parasites. In this section, we will concentrate on recent advances in gene content modelling and analysis for the study of genome evolution.

5.5.3.1   General Approaches

A simple way of analysing gene content data is simply to score gene presence and absence in a number of species, with a '1' representing presence and a '0', absence. The resulting matrix can be further analysed, for example using parsimony algorithms available in many phylogenetic software packages. In addition, a number of distance measures have been developed for gene content datasets. Ferretti *et al.* (1996) proposed the *syntenic distance* for multi-chromosomal genomes. Here, the minimum number of centric fusions, fissions and translocations required to ensure that every chromosome within one genome has the same gene content as a second genome, ignoring gene order, is counted. Snel *et al.* (1999) proposed that the percentage of genes shared by two genomes could be used as a measure of similarity between them:

$$d = 1 - \frac{|G_1 \cap G_2|}{\min\{|G_1|, |G_2|\}}, \qquad (5.16)$$

where $|G_1|$ and $|G_2|$ are the numbers of genes in genomes 1 and 2 respectively and $|G_1 \cap G_2|$ is the number of genes common to both.

Kunin *et al.* (2005) developed a distance measure, which they called *genome conservation*, which takes into account both gene content and sequence similarity within a whole genome context. The authors carried out a BLASTP analysis on the genes of genome pairs, noting hits with an e-value cut-off of e^{-10} and using the 'bit score' as the measure of sequence similarity. The sum of all best hits between genomes A and B was calculated and was denoted by $\sum(A, B)$ (which notably is a non-reciprocal value as reciprocal best hits are not used, so as to filter out paralogous genes) and the conservation between A and B by $S = \min(\sum(A, B), \sum(B, A))$. Two genome conservation measures were then defined as:

$$D_1 = \frac{1 - S}{\min(\sum(A, A), \sum(B, B))}, \qquad (5.17)$$

and

$$D_2 = -\ln\left(S / \frac{\sqrt{2} \times \sum(A, A) \times \sum(B, B)}{\sqrt{(\sum(A, A)^2 + \sum(B, B)^2)}}\right), \qquad (5.18)$$

corresponding to strategies for the transformation and normalisation of self-similarity and adjustment for genome sizes. The new genome conservation distance measure was then used to estimate a phylogeny for 153 sequenced genomes, with a comparison to the results of two other methods. The first of these alternative analyses involved the gene content method as used in Korbel *et al.* (2002) and the second an average orthologue similarity method. In the genome conservation analysis, the pairwise score S was divided by the smallest number of hits between the genomes, with the values being normalised between 0 and 100, for ease of comparison with the other measures. The resulting distance matrices were then analysed using the neighbor-joining tree construction algorithm (Saitou and Nei, 1987). It was noted that all three distance matrices contained a strong phylogenetic signal but that the gene content, and to a lesser extent genome conservation, approaches were affected by significant gene loss in certain lineages. However, genome conservation was found to be capable of producing accurate phylogenetic groupings at a wide range of taxonomic levels and was, in general, superior to both the gene content and average orthologue similarity approaches. Consequently, merging gene content and sequence

information within a single distance measure appears to be a valuable approach for the comparison of whole genome data.

Many gene content analytical methods have been developed with phylogenetic tree inference in mind, but none was developed specifically for more complex scenarios involving genome fusions. In contrast, a recent method known as *conditioned reconstruction* has been developed by Lake and Rivera (2004). The method takes gene presence and absence data conditioned on a reference genome, finds the set of most probable topologies through a bootstrapping technique and tests whether the opposing topologies are consistent with an evolutionary scenario involving genome fusions (i.e. a ring phylogeny). It has been applied by the same authors to a series of bacterial, archaeal and eukaryotic organisms (Rivera and Lake, 2004), which they claim gives evidence for the eukaryotes having evolved through a genome fusion of two diverse prokaryotes.

5.5.3.2 Birth and Death Model–based Approaches

Birth and death processes are proving highly useful in the analysis of gene content, being well-studied mathematical techniques with a natural affinity to this type of dataset. For example Gu (2000) introduced a statistical framework for gene content analysis. The model, which used a death process to describe gene loss, could be adopted to infer a gene content phylogeny using both distance-based and maximum likelihood approaches. Unfortunately, the maximum likelihood approach proved to be very computationally intensive but Zhang and Gu (2004) later introduced a computationally tractable model for gene content evolution. They showed that the transition probabilities between an ancestral genome X_0, possessing r_0 copies of a gene, and a subsequent genomic state X_t, possessing k copies of the same gene, were:

$$P(X_t = k|X_0 = r_0) = \sum_{j=0}^{\min(r_0,k)} \frac{r_0!}{j!(r_0 - j)!}$$

$$\times \frac{(r_0 + k - j - 1)!}{(k - j)!(r_0 - 1)!} \beta^{r_0 - j} \alpha^{k - j} (1 - \alpha - \beta)^j, k \geq 1$$

$$P(X_t = 0|X_0 = r_0) = \beta^{r_0}, \tag{5.19}$$

where

$$\alpha = \lambda \frac{1 - e^{(\lambda - \mu)t}}{\mu - \lambda e^{(\lambda - \mu)t}}$$

$$\beta = \mu \frac{1 - e^{(\lambda - \mu)t}}{\mu - \lambda e^{(\lambda - \mu)t}},$$

simplify to:

$$P(Y = 0|X_0 = r_0) = \beta^{r_0}$$

$$P(Y = 1|X_0 = r_0) = 1 - \beta^{r_0}, \tag{5.20}$$

when gene absence and presence are considered instead of gene family counts. They went on to show that these simplified transition probabilities could be used for maximum likelihood phylogenetic tree estimation on trees with four taxa (using the maximum likelihood phylogenetic framework described by (Felsenstein, 1981) and extensible to

greater numbers of taxa) as the probabilities of all 15 unrooted topologies themselves then greatly simplified to expressions that could be calculated rapidly. Zhang and Gu (2004) went on to evaluate the approach in real and simulated datasets and showed it to possess high accuracy in many simulated situations and to provide a topology consistent with other approaches for the real datasets analysed.

Gu and Zhang (2004) also developed the concept of *extended gene content*. Rather than contrasting the two states of gene presence and gene absence, or a multi-state character denoting the number of members of a gene family, they developed a three-state intermediate character. These three states represented gene families with no members (i.e. gene absence), those with a single member only and those with more than one member. Using this concept, they developed an additive gene content distance measure:

$$G = \sum_{i=1,2} \frac{\alpha_i + \beta_i}{\beta_i - \alpha_i} \ln \frac{1 - \alpha_i}{1 - \beta_i} \tag{5.21}$$

where α_i and β_i represent the gene proliferation and loss parameters within lineage i respectively. Estimates of these parameters may be calculated from an extended gene content matrix using a maximum likelihood-based approach. The authors also introduced a software package, GeneContent (Gu *et al.*, 2004), to implement this and other distance-based methods.

Kunin and Ouzounis (2003) and Karev *et al.* (2003; 2004) also used birth and death processes to model gene content evolution. The latter introduced a series of non-linear stochastic birth, death and innovation models, assessing the suitability of different models to the gene family sizes of several species, but the models have not been used within comparative studies. A recent method developed by Hahn *et al.* (2005) noted that studies of gene presence and absence were not capable of identifying genes involved in adaptive change. They developed a birth–death model of gene family size that could be used to find unusual gene family size change (both expansion and contractions) across a phylogeny of species.

Huson and Steel (2004) developed a maximum likelihood–based distance for gene content. Using a constant-birth, proportional-death model, with gene birth rate λ and gene loss rate μ, they showed that:

$$t_{\mathrm{ML}} = -\frac{1}{\mu} \log \left(\frac{\beta + \sqrt{\beta^2 + 4\alpha_{12}}}{2} \right), \tag{5.22}$$

where $m = \lambda/\mu$, $\alpha_1 = |G_1|/m$, $\alpha_2 = |G_2|/m$, $\alpha_{12} = |G_1 \cap G_2|$ and $\beta = 1 + \alpha_{12} - \alpha_1 - \alpha_2$. The authors then compared the performance of this distance measure to the measure above proposed by Snel *et al.* (1999) and to a Dollo parsimony approach (Le Quesne, 1974), which stipulated that each gene could be gained only once on a phylogeny but could be lost many times. They found that their measure had a similar performance on simulated datasets to the Dollo parsimony approach and that both outperformed the Snel *et al.* measure.

The idea of genome conditioning, as described above, has also been investigated by Savva (2006) modelling gene loss from a reference genome using a death process, while also taking into account rate heterogeneity through a gamma distribution and a 'core' set of genes that cannot be lost. The method, which defines a maximum likelihood–based distance for conditioned gene content, has been applied successfully to datasets arising

from Comparative Genome Hybridisation microarray experiments. A software package, MPP, has been developed to implement the method, along with other steps in the analytical pipeline (Davey, 2005).

5.5.4 Comparison of Gene Order and Gene Content Methods

From the sections above on gene order and gene content, we can see that there are overlaps between the types of analysis that we can perform on each of our datasets. In particular, gene content and gene order have both been used for phylogenetic inference. So how should we decide which type of method is most informative for a particular dataset? This question has been posed by several researchers in the past few years and we are beginning to get a picture of the relative adequacies of each method with regard to different types of dataset.

Herniou *et al.* (2001) examined the phylogeny of nine baculovirus genomes, carrying out three separate analyses. In the first analysis, they applied maximum parsimony algorithms to alignments found for each of the 63 genes common to all 9 genomes and also to the concatenated sequence from these genes. In the second analysis, they calculated the pairwise normalised induced breakpoint distances on their dataset, a method they developed independently of Sankoff *et al.* (2000a) and calculated a phylogeny from the distance matrix using neighbor-joining. Also in this second analysis, they independently developed a binary encoding method, which they applied using maximum parsimony. In the third analysis, they derived the presence/absence gene content matrix from their nine genomes and estimated a phylogeny using maximum parsimony. From their results, Herniou *et al.* concluded that the concatenated sequence phylogeny gave the tree topology most consistent with current knowledge of the evolutionary history of the baculoviruses. Notably, this topology was only identical to 7 of the 63 phylogenies found from analyses of individual genes, highlighting the inadequacy of inferring species trees from genic histories. They also found that the gene content and gene order trees were very similar to the optimal topology, noting that the observed differences could be a consequence of widely different genome sizes and long branch attraction. Interestingly, they mapped gene content gains and losses onto the optimal topology, noting a large number of putative homoplasies (i.e. where two genomes appear erroneously similar because of independent acquisitions of an individual gene, perhaps through horizontal transfer, or because of acquisition and subsequent loss of a gene that was absent from the other genome's lineage). They concluded that gene content and gene order gave useful methods with which to find further supporting evidence for phylogenies derived from concatenated sequences, noting that, 'Gene order and gene content have the advantage of providing independent datasets from the gene sequences, with independent dynamics and rates of evolution.'

Wolf *et al.* (2001) also carried out a comparative analysis of different methods, applying them to bacterial and archaeal genomes. These researchers carried out five different analyses on these genomes. In the first analysis, the presence/absence gene content matrix was derived from the COGs database. A phylogeny was then generated using Dollo parsimony. Crucially, this algorithm reduces the number of homoplasies in the resulting phylogeny and is less sensitive to gene loss (common in prokaryotes) than other parsimony algorithms. In the second analysis, a matrix of conserved pairs of neighbouring genes in genome pairs was constructed. This analysis used a more relaxed definition of a neighbourhood than other algorithms (e.g. the approach used by Herniou *et al.* in which the two genes must be adjacent in both genomes), requiring

adjacency in one genome but allowing a small number of intervening genes in the second. This matrix was again analysed using Dollo parsimony. In the third analysis, pairwise genomic distances were calculated on the basis of the distributions of identity percentages between orthologues (e.g. the mean, median or mode of the percentage values). In the fourth analysis, a maximum likelihood phylogeny was calculated for the concatenated sequence of the ribosomal proteins. In the fifth analysis, maximum likelihood phylogenies were estimated for a subset of carefully selected individual genes. The results indicated that, overall, the sequence-based methods and the third method based on identity percentages of orthologous genes gave trees that were most consistent with the known evolutionary history of the organisms undergoing analysis. While the trees based on local gene order and gene content were partially consistent with these results (and notably included a previously unknown grouping now believed to be correct), the differences were thought to be a consequence of extensive horizontal gene transfer and lineage-specific gene loss.

Looking at these two approaches together, it appears that gene order and gene content methodologies can be valuable in evolutionary studies but that this value varies considerably depending on the type of dataset undergoing analysis. Furthermore, other researchers have noted that gene order and gene content methods may be superior when gene sequences have become saturated for mutations. Of course, if we are looking to find out evolutionary information other than the topology, such as rearrangement histories and ancestral states, then these can only be found using the relevant methodology. However, an alternative strategy would be to map gene content and order changes onto a phylogeny derived from gene sequences, as carried out by Herniou *et al.* above.

We can also compare gene order and gene content methods using the models themselves. In gene order studies, path probabilities tend to a uniform ergodic distribution. This means that, once a particular evolutionary time has passed or a number of events have occurred, the probability of one path is equal to that of any other. Evolutionary analyses under such conditions would be meaningless. Dicks (unpublished) has used a model-based approach to show the time to stationarity under the model in a single chromosome is linear with the number of genes it contains. Savva (2006) has compared simple gene order and gene content models using Fisher's information. Preliminary results show that, for large t, gene content models provide a greater amount of information than gene order models. Consequently, for widely diverged genomes, this would indicate that it is better to compare their gene contents rather than the order of their genes. However, further research is required before we can fully understand the evolutionary boundaries of each method.

5.6   WHOLE GENOME SEQUENCES

Since the first genome, of the bacterium *Haemophilus influenzae*, was sequenced in 1995 we have seen a proliferation of genome sequencing projects. In addition to dozens of bacterial and viral organisms, the genomes of around 50 eukaryotic organisms have now been sequenced. The Institute for Genomic Research (now part of the J. Craig Venter Institute) provides a useful webpage listing the various genome projects underway (`http://www.tigr.org/tdb/`). With the availability of multiple whole genome sequences, we now have a motivation to compare them along their entirety, particularly for relatively close species. How close are two sequences? Do we see good general

correspondence or are there areas that are significantly more closely related than others? In addition, we can delve deeper into problems of their evolutionary histories, looking both within and between organisms. In this section, we examine some of the recently developed methodologies for whole genome alignment of sequences, estimation of conserved chromosomal blocks and dating of large-scale evolutionary events.

5.6.1 Whole Genome Alignment

One simple way to compare two whole genome sequences is to visualise them as a dot plot. Such a plot is based on an underlying matrix derived from a comparison of the two sequences. We 'slide' a window of a certain size (e.g. five bases or residues) along the sequences and, if the sub-sequences within those windows match within a certain range of precision, then the corresponding entry within the matrix is denoted as 1. If there is no such match then an entry 0 is made. By adjusting the window size and tolerance of mismatches, we can change the precision of our matrix. If we then plot the resulting matrix, we can quickly see regions of similarity. If our two sequences align perfectly, then our dot plot will show a perfect diagonal line from the bottom left to the top right corners (or equivalently the top left to the bottom right corners). In a whole genome context, dot plots are useful for spotting large genomic rearrangements such as those described in Section 5.2. For example, Figure 5.4 shows a dot plot in which we see both areas of regional similarity and disjointed regions. Regions that have 'broken away' from the main diagonal are likely to be a consequence of translocations while those that have a differnt orientation become inverted.

In addition to visualising the relationship between two genome sequences, we may wish to align them formally with an alignment algorithm. Over the past couple of decades, it has become commonplace to compare two or more DNA or amino acid sequences by aligning them such that this technique is considered to be part of the biologist's toolkit. If there
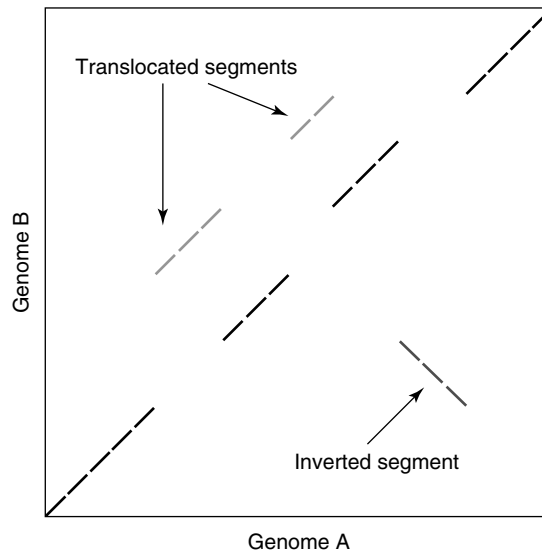
Figure 5.4 Visualising a whole genome comparison with a dot plot.

is a good correspondence between two sequences, then this is often seen as an indication of a common evolutionary origin and perhaps also a structural and functional similarity. There are two main types of sequence alignment algorithms: local and global. Local alignments do not try to match two sequences exactly but rather concentrate on local areas of close similarity separated by more divergent regions. This type of algorithm is useful for sequences without a good overall similarity but which are thought to contain conserved regions within them. The well-known alignment algorithms BLAST (Basic Local Alignment Search Tool; Altschul *et al.*, 1990) and Smith–Waterman (Smith and Waterman, 1981) are both local alignment tools. Conversely, global algorithms attempt to align all sites or residues of a sequence, and are most useful when the sequences undergoing comparison are generally similar and are of equal size. The Needleman–Wunsch algorithm (Needleman and Wunsch, 1970), which is based on dynamic programming, is perhaps the best-known algorithm of this type. For a more detailed look at sequence alignment, see Durbin *et al.* (2000). Global alignments can be used to uncover the shared order of biological features within genomic sequences. Furthermore multiple global alignments, which align more than two sequences, can reveal common features between distantly related species, particularly when intermediate sequences are also included in the alignment.

We saw above that global alignment algorithms have been developed to align large sequences along their length. Unfortunately, for large genomes, using a Needleman–Wunsch algorithm is a very slow process as the algorithm requires time proportional to the products of the lengths of the aligned sequences. Consequently, a number of algorithms have been developed specifically for whole genome alignment. Popular pairwise genome alignment algorithms include MUMmer (Delcher *et al.*, 2002), DIALIGN (Brudno and Morgenstern, 2002), WABA (Kent and Zahler, 2000), AVID (Bray *et al.*, 2003) and LAGAN (Brudno *et al.*, 2003a). Furthermore, some of these algorithms have multiple genome alignment versions, such as MLAGAN (Brudno *et al.*, 2003a) and MAVID (Bray and Pachter, 2004). In general, pairwise algorithms reduce the computational burden of alignment by first finding multiple local similarities, or *anchors*. For example, LAGAN finds local similarities using the CHAOS algorithm (Brudno and Morgenstern, 2002), which identifies short inexact sequence matches. Some other algorithms find longer exact matches and therefore are not as appropriate for alignment of distantly related sequences. LAGAN then selects an ordered subset of the local similarities (the anchors) and chains them together into a *rough global map*. Finally, LAGAN searches for an optimal global alignment using Needleman–Wunch in a small, restricted region bounding the rough global map. Figure 5.5 depicts this process, with the thick black lines representing the anchors and the boxes representing the bounded region within which the Needleman–Wunch algorithm is run.

A multiple alignment version of the LAGAN algorithm, MLAGAN, uses LAGAN within a progressive alignment approach (where sequences are added to the alignment one by one) that also uses a known phylogenetic tree to guide the order in which sequences are added. Owing to the computational difficulties of dealing with very long sequences, full statistical approaches to this problem are also not practical. However some programs, such as MAVID (Bray and Pachter, 2004), use a combination of maximum likelihood and heuristic techniques. MAVID aligns sequences travelling upwards (from the leaves) within a guide phylogenetic tree. At internal nodes, two alignments are merged into a single alignment. First, maximum likelihood ancestral sequences are found for each of the two alignments, and the two ancestral sequences are then aligned using AVID.
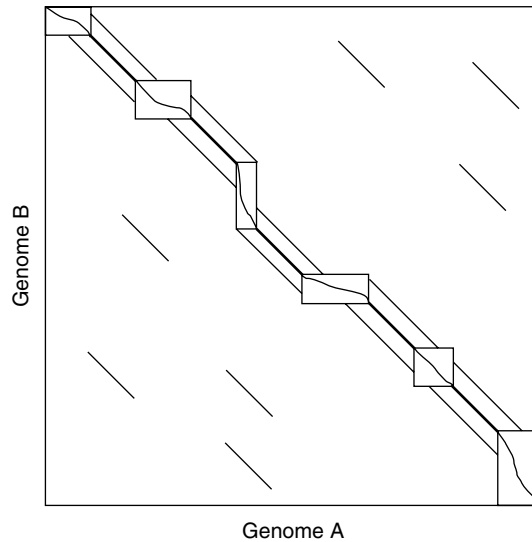
Figure 5.5   Global pairwise whole genome alignment strategy used by LAGAN.

Second, the alignment of the ancestral sequences is used to create a multiple alignment of all sequences in the two alignments being merged. This efficient strategy is capable of aligning whole genomes rapidly.

Unfortunately, most of the whole genome alignment algorithms listed above can only be used reliably within conserved chromosomal regions that have not undergone rearrangement events. To overcome this, new algorithms have been developed specifically to deal with rearrangements. For example, SLAGAN (Brudno *et al.*, 2003b) builds upon the LAGAN tool, using CHAOS local alignments to build a map of the rearrangements between the genome sequences and LAGAN to align conserved regions. The MULTIZ component of the TBA algorithm (Blanchette *et al.*, 2004) is also able to align genomes containing inversions and duplications. The MAUVE tool (Darling *et al.*, 2004b) was also developed specifically for rearranged genomes. It decomposes genomes into locally collinear blocks (LCBs, which are unaffected by rearrangements), which are ordered sets of multi-MUMs (multiple maximum unique matches; sequences that occurred only once in each genome undergoing analysis, that matched exactly and could not be extended in either direction to develop larger unique matches). LCBs are weighted using breakpoint analysis in order to find the most likely set of LCBs in the final alignment. Following a process to lengthen old LCBs and to uncover new LCBs within large gaps, each LCB in the final set is aligned using the CLUSTALW multiple alignment tool (Thompson *et al.*, 1994).

5.6.2   Finding Conserved Blocks

In Section 5.4, we described work on the estimation of conserved chromosomal segments from mapped markers. With whole genome sequences now available and many more promised for the future, a new generation of algorithms must be developed that can derive such segments from these emerging datasets. Indeed, since 2000 many researchers have developed approaches for conserved block estimation from whole genome sequence. However, until recently, the number of sequenced eukaryotic genomes has been low and

those that have been sequenced have not been closely related. This may have led to the early development of methods for *genomic palaeontology*, essentially searching for blocks of duplicated genes *within* genomes such as human, yeast and *Arabidopsis* (e.g. Abi-Rached *et al.*, 2002; Dehal and Boore, 2005; Lynch and Conery, 2000) to discover more about the nature of gene duplication and perhaps to find evidence of large-scale evolutionary events such as polyploidy.

For example, following the analyses of Vision *et al.* (2000) and Simillion *et al.* (2002), Blanc *et al.* (2003) examined the evolutionary history of *Arobidopsis thaliana*. First, duplicated genes were uncovered by carrying out an all-again-all protein similarity search of the genome. To detect conserved blocks, the authors searched for genomic regions in which duplicated genes were in a consistent order. Notably, pairs of duplicated genes were permitted to be separated by up to a fixed number of unduplicated genes. The resulting blocks are also known as *max-gap clusters*, where the *max gap* is this permitted number of unduplicated genes (typically ranging from around 3 in bacterial genomes to 30 in humans; McLysaght *et al.*, 2002). This analysis led to a set of 108 pairs of blocks sharing 6 or more duplicated genes. The statistical significance of the blocks was determined by a permutation test, where the order of the duplicated genes was randomised many times and the block detection algorithm applied to each randomised dataset. Such analyses have also inspired the development of software for conserved block prediction, such as ADHoRe (Automatic Detection of Homologous Regions; Vandepoele *et al.*, 2002) and its successor i-ADHoRe, DiagHunter (Cannon *et al.*, 2003), FISH (Calabrese *et al.*, 2003), LineUp (Hampson *et al.*, 2003) and CloseUp (Hampson *et al.*, 2005). The methods differ in terms of how conserved blocks are defined, the strategies for estimating them and the ways in which the significance of the blocks is determined, making comparisons between them difficult. Indeed, in an analysis of the rice genome, Wang *et al.* (2005) discussed how different methods of assessment of block significance could change the results of an analysis. Furthermore, Vandepoele *et al.* (2003) showed, in their analysis of rice genome, that many block detection methods were not capable of finding all duplicated regions. Instead, more sophisticated approaches were required to detect *hidden* and *ghost* duplications, with hidden blocks being uncovered by examining a third region in the same species while ghost blocks required the third region to be in another species in order to uncover the evolutionary relationships between them.

Recently, work has begun on the evaluation of cluster significance. We noted earlier the concept of a max-gap cluster in conserved block detection methods. Following Durand and Sankoff (2003) on the testing of gene clustering, Hoberman *et al.* (2005) analysed the statistical properties of max-gap clusters. They show that the probability of observing m marked genes (i.e. those identified as being of interest) in a genome of size n, in a max-gap cluster with a maximum gap of size g, and with w_{mg} the maximum length of a cluster of size m (i.e. $m + (m-1)g$), is:

$$P(n, m, g) = \frac{1}{\binom{n}{m}} \begin{cases} (n - w_{mg} + 1)(g+1)^{m-1} + \dfrac{w_{mg} - m}{2}(g+1)^{m-1}, & \text{if } w_{mg} \leq n+1 \\ d_0(m, g, n) & \text{otherwise,} \end{cases}$$

(5.23)

where

$$d_e(k, g, l) = \sum_{i=0}^{(l-k)/(g+1)} (-1)^i \binom{k-1}{i} \binom{l - i(g+1) - e}{k - e}.$$

Gap of size 4

Cluster of size 4
and length 7

Cluster of size 5
and length 8
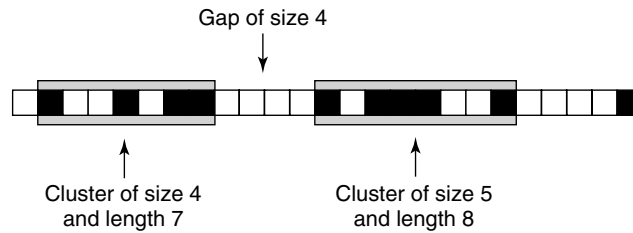
Figure 5.6 Two max-gap clusters for a genome of $n = 25$ genes, $m = 10$ marked genes and a max-gap of $g = 3$.

Figure 5.6 shows a simple example of a max-gap cluster for $n = 25$, $m = 10$ and $g = 3$. Hoberman *et al.* then gave a dynamic programming solution to the problem of finding the probability of observing a cluster with a subset, of size h, of the m marked genes and showed that this has an analytic solution when $h > m/2$. Finally, they tackled the problem of finding the probability of a max-gap cluster of size h from m homologues in a whole genome comparison of two genomes of size n. This was not solved exactly, but rather upper and lower bounds for the probability were established. Simulation experiments using these equations provided interesting results. They showed, for example, that the genome comparison upper and lower bounds are tight when the gap size g is small compared to the genome size n and that the probabilities of observing a cluster of size h are not monotonic with increasing cluster size h but may decrease initially before increasing as h tends to m. The authors also noted that *bottom-up* algorithms to find max-gap clusters (i.e. similar to some of those listed above) may not find the maximal clusters in certain cases and that divide-and-conquer solutions such as that used within the Gene Teams software described in Bergeron *et al.* (2002) should be used instead.

Mau *et al.* (2005) also developed a statistical approach to find LCBs using inexact sequence matching. Within the MAUVE alignment tool introduced above, the same research team (Darling *et al.*, 2004b) had described an approach to develop computational monotypic markers using multi-MUMs. However, the use of exact matches gave the approach a limited evolutionary breadth and so the new approach was developed for larger evolutionary distances. Monotypic markers were chosen to be single genes that possessed a reciprocal best BLAST hit in the other genomes undergoing analysis (and which also occurred just once in those genomes). A pseudo-Gibbs sampler was then used to find the posterior probabilities of markers belonging to particular blocks, thereby defining the number, composition and length of the LCB set. Crucially, a scale parameter allowed the user to choose the resolution of the LCBs. If low resolution was chosen, a few long LCBs would be found (essentially filtering out small rearrangements). At higher resolution, some of the LCBs broke up into more, smaller LCBs. Indeed, this is somewhat reminiscent of the gap length choice in the max-gap clustering described above. The authors then described the application of LCB estimation within four Streptococci genomes and showed how the choice of resolution affected the biological plausibility of a subsequent genome rearrangement analysis. The approach was implemented in the GRIL software (Genome Rearrangement and Inversion Locator; Darling *et al.*, 2004a).

5.6.3 Dating Duplicated Genes and Blocks

The ages of duplicated genes and conserved blocks, once determined by methods such as those outlined above, can be estimated by determining the level of synonymous substitution per site, K_S, (sometime also termed ds), in the underlying nucleotide sequence. For a genome with no large-scale duplication events in its past (or at least in the length of time in which such events are still able to be detected), the distribution of K_S values when plotted exhibits a characteristic 'L' shape (see e.g. Lynch and Conery, 2000). However, genomes in which large-scale evolutionary events have occurred tend to exhibit 'spikes' or bursts overlaid on this curve, as shown in Figure 5.7. In the Blanc *et al.* (2003) study discussed above, values of K_S were estimated using the *codeml* method in the PAML software (with large values discarded due to saturation of mutations). See **Chapter 12** for details of the theory underlying the estimation of rates of synonymous substitution. Blanc *et al.* found that the median K_S values within conserved blocks clustered into two groups: recent blocks and old blocks. The values of the 45 recent blocks showed low variance, suggesting a single, recent, whole genome duplication, while the values of the 63 old blocks were more varied though not wholly inconsistent with a second more distant whole genome event. The blocks were then dated by comparing the K_S values of the two groups to K_S values found by comparing *Arabidopsis* genes with their orthologues in other dicotyledon species. In a later study, Blanc and Wolfe (2004) also applied the method to EST contigs from several plant species, showing that large-scale duplications are likely to have occurred in 9 of 14 species analysed.

Phylogenetic approaches for determining evolutionary histories and dating duplication events have been used by Bowers *et al.* (2003). These authors discussed the potential error involved in dating evolutionary events by K_S values, as such values had been shown to vary considerably between genes belonging to different functional categories. Instead, in their analysis of the *Arabidopsis* genome, rooted phylogenies of gene duplicates together with a dicotyledon orthologue and an orthologue in an outgroup species (the latter of which was used to root the tree) were examined. The relative frequencies of trees where the gene duplicates were each other's closest relative and those where the dicotyledon
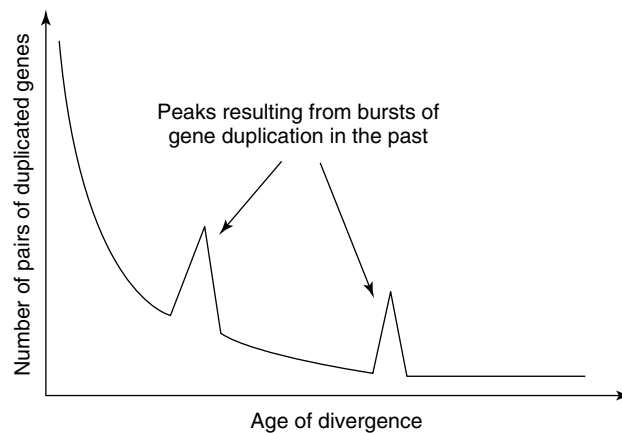
Figure 5.7 Distribution of ages of pairs of duplicated genes, with two ancient bursts of gene duplication.

orthologue was more closely related to one of the duplicate genes than it was to one another were found for a range of dicotyledons. This information was used to assess whether duplicate events were genome wide and when they were most likely to have occurred relative to speciation events. Furthermore, in their analysis of the rice genome, Wang *et al.* (2005) showed that the negative correlation of GC content with K_S values in rice could affect the characteristic K_S distribution and its interpretation. Instead of relying solely on K_S values to date evolutionary events, the authors decided to use a phylogenetic approach as support for the results.

Recently, the use of K_S values as quantities with which to estimate ages of evolutionary events is being examined in more detail. Maere *et al.* (2005) developed a dynamic population model to describe the shape of the K_S distribution as seen, for example, in Figure 5.7. The model described the number of duplicate genes retained in the genome, performing 50 time steps (at K_S intervals of 0.1) for 4 duplication mode categories (0 being a background process of gene gain and 1, 2 and 3 distinct potential whole genome duplications). It possessed three components:

$$D_0(1, t) = \nu \left[\sum_{x'=1}^{\infty} D_{tot}(x', t-1) + G_0 \right]$$

$$D_i(1, t) = \left[\sum_{x'=1}^{\infty} D_{tot}(x', t-1) + G_0 \right] \delta(t, t_i), \quad i = 1, 2 \text{ or } 3 \text{ and}$$

$$D_i(x, t) = D_i(x-1, t-1)[x/(x-1)]^{-\alpha_i}, \quad x > 1, \ i = 0, 1, 2 \text{ or } 3,$$

which were summed to create the overall distribution:

$$D_{tot}(x, t) = \sum_i D_i(x, t). \tag{5.24}$$

The first component represented the background process of single gene gain, where x was the age of the gene, t the time step, $D_i(x, t)$ the number of genes of age x at time t in duplication mode i, G_0 the number of ancestral genes at $K_S = 5$ and ν the gene birth rate. The second component modelled potential whole genome duplications (three in this case, for analysis of the *Arabidopsis* genome) by means of a parameter δ. Finally, the third component modelled the process of gene loss as a power-law curve with decay rates α_i. The power-law curve was thought to be more appropriate than exponential gene loss as it has been hypothesised that gene loss occurs rapidly immediately following genome duplication, gradually evolving to a process preferentially retaining older duplicates. Together, these components modelled the age distribution of duplicated genes which was then transformed to a K_S distribution through a Poisson smoothing technique. The model could then be fitted to real datasets through, for example, simulated annealing optimisation. The authors used the model to lend support for the three whole genome duplications hypothesised in *Arabidopsis* and to show, for example, differences in gene decay parameters for genes of different functional categories.

A different approach was taken by Cui *et al.* (2006) in their analysis of genome duplications in the history of flowering plants, although the basis of the analysis was again to model the K_S distribution. The authors noted that the null model for this distribution followed a constant rate birth–death process, with gene birth following a Poisson process

with rate β and the number of duplicated genes decreasing exponentially with rate δ. The age distribution of retained duplicates $N(t)$ at time t was then:

$$N(t) \sim P_O\left(\gamma \int_0^t \delta \exp(\delta s)\mathrm{d}s\right) = P_O(\gamma \cdot F(t)), \tag{5.25}$$

where $\gamma = \beta/\delta$ and $F(t) = 1 - \exp(-\delta t)$. This null model was then fitted to each dataset, assessing the fit using a χ^2 test. A quantile–quantile (Q–Q) plot was also used to compare the observed data to a dataset simulated under the null model, using a bootstrapped Kolmogorov–Smirnov test. Following the approach of Schleuter *et al.* (2004), datasets that could not be described adequately by the null model were further analysed by a finite mixture model of genome duplications (i.e. the log-transformed K_S values followed a sum of Gaussian distributions), with each component of the mixture representing a duplication class (e.g. two components, where the first represented the background process of gene gain and the second a whole genome duplication event). The authors went on to show, through simulation, the difficulties of determining both very recent and ancient duplications with K_S-based methods but successfully applied this method to several plant datasets, providing support for several large-scale duplication events in their evolutionary histories.

5.7   CONCLUSIONS AND FUTURE RESEARCH

In this chapter, we chose to focus on a few key problems that are paramount in evolutionary studies of whole genomes, using comparative maps, gene orders and contents, and whole genome sequence data. Of course, there are many other ways in which we can compare genomes (see, e.g. Saccone and Pesole, 2003 for a broader overview). We might wish to discover more about GC signatures within genes and codon usage from the genomic sequence. We might be interested in functional data and compare the protein structures or functions of the proteins encoded by the genes. We might not be interested in genes at all and prefer to look at intergenic entities such as *transposable elements*. As new datasets arise, we see methodologies being developed to analyse all these data types, to compare the results of the different methods and perhaps even to integrate them. We have seen that there are many promising strands of research in comparative genome analysis. How do we see research progressing in these areas?

Much of the early work in gene order studies has been of a computational nature and there is still much work to be done in improving, testing and comparing models of evolution. We have already seen that the fragile breakage model is being tested in newly sequenced genomes and we could also model other phenomena such as rapid rearrangement in sub-telomeric chromosomal regions, as has been seen in several eukaryotes. In addition, we must also consider the possibility that a proportion of genes are not randomly distributed along genomes but that their genomic location has some meaning. It has long been understood that genes are clustered in operons within prokaryotic genomes. However, several recent studies have shown that significant quantities of genes in eukaryotic genomes are clustered according to other criteria (e.g. belonging to a common metabolic pathway). A review of the evolutionary dynamics of eukaryotic gene order (Hurst *et al.*, 2004) discusses these criteria and examines potential mechanisms for cluster formation. In an illuminating study Wong and Wolfe (2005) show how several

genome rearrangement events have driven the formation of a secondary metabolite gene cluster in *sensu stricto* yeast species. See also **Chapter 13**, which discusses the implications of gene location (in addition to gene duplication and genetic redundancy) in a section on genome organisation. These papers show that we may not be able to consider all mutation events equally (i.e. those that separate neighbouring co-regulated genes may be less likely to occur than some others) and perhaps we should not consider gene order data in isolation but should rather analyse it in tandem with gene expression data.

In the analysis of whole genome sequences, it is likely that methods for whole genome alignment will continue to improve, both in terms of speed and accuracy and in dealing with problems such as rearrangement events, perhaps with an increase in model-based methods. There is also a need for a thorough comparison of methods for conserved block estimation. Some current methods search for clusters with a strict ordering of the genes, while others allow it to be entirely random and still others fall somewhere between these two extremes. Furthermore, there are essentially (at least) three types of clusters: blocks between genomes (where we may need to filter out micro-rearrangements), blocks within genomes (where we may need to filter out single-copy genes following large-scale duplications) and functional clusters (e.g. those for secondary metabolites, where a neighbourhood of genes from a number of functional classes may be more important than their actual order). We may find that different stratagies are required for each of these situations. Further modelling of dating methodologies (e.g. K_S signatures) would also be illuminating, formalising the goodness-of-fit of models and hypotheses for model comparison.

Overall, it is clear that comparative genomics is an active, flourishing area of computational research and that statistical methods are beginning to make an impact on this field. There is currently a great opportunity for statisticians to become involved in this research and to play an important role in making sense of the enormous quantities of data arising from the genome sequencing projects.

Acknowledgments

REFERENCES

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. and Inoko, H. (2002). Evidence of en bloc duplication in vertebrate genomes. *Nature Genetics* **31**, 100–105.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

Bader, D.A., Moret, B.M. and Yan, M. (2001). A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *Journal of Computational Biology* **8**, 483–491.

Bergeron, A. (2001). A very elementary presentation of the Hannenhalli-Pevzner theory. *Proceedings of the Twelfth Annual Symposium on Combinatorial Pattern Matching*, *Vol. 2089 of Lecture Notes in Computer Science*. Springer-Verlag, New York, pp. 106–117.

Bergeron, A., Corteel, S. and Raffinot, M. (2002). The algorithmic of gene teams. *WABI*, *Vol. 2452 of Lecture Notes in Computer Science*, Springer-Verlag, Berlin pp. 464–476.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B., Böhme, U., Hannick, L., Aslett, M.A., Shallom, J., Marcello, L., Hou, L., Wickstead, B., Alsmark, U.C.M., Arrowsmith, C., Atkin, R.J., Barron, A.J., Bringaud, F., Brooks, K., Carrington, M., Cherevach, I., Chillingworth, T.-J., Churcher, C., Clark, L.N., Corton, C.H., Cronin, A., Davies, R.M., Doggett, J., Djikeng, A., Feldblyum, T., Field, M.C., Fraser, A., Goodhead, I., Hance, Z., Harper, D., Harris, B.R., Hauser, H., Hostetler, J., Ivens, A., Jagels, K., Johnson, D., Johnson, J., Jones, K., Kerhornou, A.X., Koo, H., Larke, N., Landfear, S., Larkin, C., Leech, V., Line, A., Lord, A., MacLeod, A., Mooney, P.J., Moule, S., Martin, D.M.A., Morgan, G.W., Mungall, K., Norbertczak, H., Ormond, D., Pai, G., Peacock, C.S., Peterson, J., Quail, M.A., Rabbinowitsch, E., Rajandream, M.-A., Reitter, C., Salzberg, S.L., Sanders, M., Schobel, S., Sharp, S., Simmonds, M., Simpson, A.J., Tallon, L., Turner, C.M.R., Tait, A., Tivey, A.R., Van Aken, S., Walker, D., Wanless, D., Wang, S., White, B., White, O., Whitehead, S., Woodward, J., Wortman, J., Adams, M.D., Embley, T.M., Gull, K., Ullu, E., Barry, J.D., Fairlamb, A.H., Opperdoes, F., Barrell, B.G., Donelson, J.E., Hall, N., Fraser, C.M., Melville, S.E., and El-Sayed, N.M. (2005). The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309**, 416–422.

Blanc, G., Hokamp, K. and Wolfe, K.H. (2003). A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Research* **13**, 137–144.

Blanc, G. and Wolfe, K.H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions in duplicate genes. *The Plant Cell* **16**, 1667–1678.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D. and Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* **14**, 708–715.

Blanchette, M., Kunisawa, T. and Sankoff, D. (1996). Parametric genome rearrangement. *Gene-Combis* **172**, 11–17.

Blanchette, M., Kunisawa, T. and Sankoff, D. (1999). Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution* **49**, 193–203.

Bourque, G. and Pevzner, P.A. (2002). Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research* **12**, 26–36.

Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438.

Bray, N., Dubchak, I. and Pachter, L. (2003). AVID: a global alignment program. *Genome Research* **13**, 97–102.

Bray, N. and Pachter, L. (2004). MAVID: constrained ancestral alignment of multiple sequences. *Genome Research* **14**, 693–699.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., NISC Comparative Sequencing Program, Green, E.D., Sidow, A. and Batzoglou, S. (2003a). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* **13**, 721–731.

Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I. and Batzoglu, S. (2003b). Glocal alignment: finding rearrangements during alignment. *Bioinformatics* **19**(S1), i54–i62.

Brudno, M. and Morgenstern, B. (2002). Fast and sensitive alignment of large genomic sequences In *Proceedings of the IEEE Computing Society Bioinformatics Conference (CSB)*. Stanford University, Palo Alto, California, USA.

Calabrese, P.P., Chakravarty, S. and Vision, T.J. (2003). Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* **19**(S1), i74–i80.

Cannon, S.B., Kozik, A., Chan, B., Michelmore, R. and Young, N.D. (2003). DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biology* **4**, R68.

Caprara, A. and Lancia, G. (2000). Experimental and statistical analysis of sorting by reversals. In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic Publishers.

Cosner, M.E., Jansen, R.K., Moret, B.M.E., Raubeson, L.A., Wang, L.S., Warnow, T. and Wyman, S. (2000). An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae. In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic Publishers, pp. 99–122.

Cui, L., Wall, P.K., Leebans-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H. and dePamphilis, C.W. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Research* **16**, 738–749.

Darling, A.C.E., Mau, B., Blattner, F.R. and Perna, N.T. (2004a). GRIL: genome rearrangement and inversion locator. *Bioinformatics* **20**, 122–124.

Darling, A.C.E., Mau, B., Blattner, F.R. and Perna, N.T. (2004b). Mauve: multiple alignment of conserved genomic sequences with rearrangements. *Genome Research* **14**, 1394–1403.

Davey, R. (2005). Development and validation of algorithms for gene and marker assignment from microarray experiments. Ph.D. dissertation, University of East Anglia.

Dehal, P. and Boore, J.L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* **3**, e314.

Dehal, P. and Boore, J.L. (2006). A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics* **7**, 201.

Delcher, A.L., Phillippy, A., Carlton, J. and Salzberg, S.L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* **30**, 2478–2483.

Dias, Z. and Meidanis, J. (2001). Genome rearrangements distance by fusion, fission, and transposition is easy. In *Proceedings of SPIRE' 2001–Eighth Symposium on String Processing and Information Retrieval*, Laguna de San Rafael, Chile, pp. 250–253.

Dicks, J.L. (1999). Comparative mapping and phylogeny. DPhil thesis, University of Oxford.

Dicks, J. (2000). CHROMTREE: maximum likelihood estimation of chromosomal phylogenies. In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic Publishers, pp. 333–342.

Dicks, J., Anderson, M., Cardle, L., Cartinhour, S., Couchman, M., Davenport, G., Dickson, J., Gale, M., Marshall, D., May, S., McWilliam, H., O'Malia, A., Ougham, H., Trick, M., Walsh, S. and Waugh, R. (2000). UK CropNet: a collection of databases and bioinformatics resources for crop plant genomics. *Nucleic Acids Research* **28**, 104–107.

Dobzhansky, T. and Sturtevant, A.H. (1938). Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics* **23**, 28–64.

Durand, D. and Sankoff, D. (2003). Tests for gene clustering. *Journal of Computational Biology* **10**, 453–482.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (2000). *Biological Sequence Analysis*. Cambridge University Press.

Ehrlich, J., Sankoff, D. and Nadeau, J.H. (1997). Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics* **147**, 289–296.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Felsenstein, J. (1993). *PHYLIP (Phylogenetic Inference Package) Version 3.5, Documentation*. University of Washington, Seattle, WA.

Ferretti, V., Nadeau, J.H. and Sankoff, D. (1996). Original synteny. In *7th Annual Symposium on Combinatorial Pattern Matching*, Laguna Beach, Palo Alto, California, USA. pp. 159–167.

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology* **19**, 99–113.

Fitch, W.M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155**, 279–284.

Gu, X. (2000). A simple evolutionary model for genome phylogeny based on gene content. In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic Publishers, pp. 515–524.

Gu, X., Huang, W., Xu, D. and Zhang, H. (2004). GeneContent: software for whole-genome phylogenetic analysis. *Bioinformatics* **21**, 1713–1714.

Gu, X. and Zhang, H. (2004). Genome phylogenetic analysis based on extended gene contents. *Molecular Biology and Evolution* **21**, 1401–1408.

Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C. and Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research* **15**, 1153–1160.

Haldane, J.B.S. (1927). The comparative genetics of colour in Rodents and Carnivora. *Biological Reviews* **2**, 199–212.

Hampson, S.E., Gaut, B.S. and Baldi, P. (2005). Statistical detection of chromosomal homology using shared-gene density alone. *Bioinformatics* **21**, 1339–1348.

Hampson, S., McLysaght, A., Gaut, B.S. and Baldi, P.F. (2003). LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Research* **13**, 999–1010.

Hannenhalli, S. (1996). Polynomial-time algorithm for computing translocation distance between genomes. *Discrete Applied Mathematics* **71**, 137–151.

Hannenhalli, S. and Pevzner, P. (1995a). Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on the Theory of Computing*, Las Vegas, Nevada, USA, pp. 178–189.

Hannenhalli, S. and Pevzner, P. (1995b). Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of the Thirty-Sixth Annual IEEE Symposium on Foundations of Computer Science*, Milwaukee, Wisconsin, USA, pp. 581–592.

Herniou, E.A., Luque, T., Chen, X., Vlak, J.M., Winstanley, D., Cory, J.S. and O'Reilly, D.R. (2001). Use of whole genome sequence data to infer baculovirus phylogeny. *Journal of Virology* **75**, 8117–8126.

Hoberman, R., Sankoff, D. and Durand, D. (2005). The statistical analysis of spatially clustered genes under the maximum gap criterion. *Journal of Computational Biology* **12**, 1083–1102.

House, C.H. and Fitz-Gibbon, S.T. (2002). Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *Journal of Molecular Evolution* **54**, 539–547.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Research* **30**, 38–41.

Hurst, L.D., Pal, C. and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nature Reviews Genetics* **5**, 299–310.

Huson, D. and Steel, M. (2004). Phylogenetic trees based on gene content. *Bioinformatics* **20**, 2044–2049.

Karev, G.P., Wolf, Y.I., Berezovskaya, F.S. and Koonin, E.V. (2004). Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evolutionary Biology* **4**, 32.

Karev, G.P., Wolf, Y.I. and Koonin, E.V. (2003). Simple stochastic birth and death models of genome evolution: was there enough time for use to evolve? *Bioinformatics* **19**, 1889–1900.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003). Evolution cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 11484–11489.

Kent, W.J. and Zahler, A.M. (2000). Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. *Genome Research* **10**, 1115–1125.

Keogh, R.S., Seoighe, C. and Wolfe, K.H. (1998). Evolution of gene order and chromosome number in *Saccharomyces, Kluyveromyces* and related fungi. *Yeast* **14**, 443–457.

Korbel, J.O., Snel, B., Huynen, M.A. and Bork, P. (2002). SHOT: a web server for the construction of genome phylogenies. *Trends in Genetics* **18**, 158–162.

Kunin, V., Ahren, D., Goldovsky, L., Janssen, P. and Ouzounis, C.A. (2005). Measuring genome conservation across taxa: divided strains and united kingdoms. *Nucleic Acids Research* **33**, 616–621.

Kunin, V. and Ouzounis, C.A. (2003). GenTRACE-reconstruction of gene content of ancestral species. *Bioinformatics* **19**, 1412–1416.

Lake, J.A. and Rivera, M.C. (2004). Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Molecular Biology and Evolution* **21**, 681–690.

Larget, B., Kadane, J.B. and Simon, D.L. (2005a). A Bayesian approach to the estimation of ancestral genome arrangements. *Molecular Phylogenetics and Evolution* **36**, 214–223.

Larget, B., Simon, D.L., Kadane, J.B. and Sweet, D. (2005b). A Bayesian analysis of metazoan mitochondrial genome arrangements. *Molecular Biology and Evolution* **22**, 486–495.

Larget, B., Simon, D.L. and Kadane, J.B. (2002). Bayesian phylogenetic inference from animal mitochondrial genome arrangements. *Journal of the Royal Statistical Society, Series A* **64**, 1–13.

Le Quesne, W.J. (1974). The uniquely evolved character and its cladistic application. *Systematic Zoology* **23**, 513–517.

Lin, J. and Gerstein, M. (2000). Whole-genome trees based on the occurrence of folds and orthologs: implication for comparing genomes on different levels. *Genome Research* **10**, 808–818.

Lynch, M. and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 5454–5459.

Mau, B., Darling, A.E. and Perna, N.T. (2005). Identifying evolutionarily conserved segments among multiple divergent and rearranged genomes. In *Proceedings of the 2nd RECOMB Comparative Genomics satellite workshop*, *Vol. 3388 of Lecture Notes in Computer Science*. Springer-Verlag, Berlin, pp. 72–84.

McLysaght, A., Hokamp, K. and Wolfe, K.H. (2002). Extensive genomic duplication during early chordate evolution. *Nature Genetics* **31**, 200–204.

Miklos, I. (2003). MCMC genome rearrangement. *Bioinformatics* **19**(S2), ii130–ii137.

Miklos, I., Ittzes, P. and Hein, J. (2005). ParIS genome rearrangement server. *Bioinformatics* **21**, 817–820.

Montague, M.G. and Hutchison, C.A. (2000). Gene content phylogeny of herpesviruses. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5334–5339.

Nadeau, J.H. and Taylor, B.A. (1984). Lengths of chromosomal segments conserved since divergence of mouse and man. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 814–818.

Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453.

Nguyen, C.T., Tay, Y.C. and Zhang, L. (2005). Divide-and-conquer approach for the exemplar breakpoint distance. *Bioinformatics* **21**, 2171–2176.

O'Brien, K.P., Remm, M. and Sonnhammer, E.L.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research* **33**, D476–D480.

O'Keefe, C. and Eichler, E. (2000). The pathological consequences and evolutionary implications of recent human genomic duplications. In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic Publishers, pp. 29–46.

Peng, Q., Revzner, P.A. and Tesler, G. (2006). The fragile breakage versus random breakage models of chromosome evolution. *PLoS Computational Biology* **2**, e14.

Pevzner, P.A. and Tesler, G. (2003a). Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Research* **2**, 37–45.

Pevzner, P.A. and Tesler, G. (2003b). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 7672–7677.

Remm, M., Storm, C.E.V. and Sonnhammer, E.L.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology* **314**, 1041–1052.

Rivera, M.C. and Lake, J.A. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**, 152–155.

Saccone, C. and Pesole, G. (2003). *Handbook of Comparative Genomics: Principles and Methodology*. Wiley Press.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406–425.

Sankoff, D. (1999). Genome rearrangements with gene families. *Bioinformatics* **15**, 909–917.

Sankoff, D. and Blanchette, M. (1997). The median problem for breakpoints in comparative genomics. In *Computing and Combinatorics, Proceedings of COCOON '97, Vol. 1276 of Lecture Notes in Computer Science*, T. Jiang and D.T. Lee, eds. Springer-Verlag, London, pp. 251–263.

Sankoff, D. and Blanchette, M. (1998). Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology* **5**, 555–570.

Sankoff, D., Deneault, M., Bryant, D., Lemieux, C. and Turmel, M. (2000a). Chloroplast gene order and the divergence of plants and algae, from the normalized number of induced breakpoints. In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic Publishers, 89–98.

Sankoff, D. and Haque, L. (2006). The distribution of genomic distance between random genomes. *Journal of Computational Biology* **13**, 1005–1012.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F. and Cedergren, R. (1992). Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 6575–6579.

Sankoff, D., Parent, M.-N. and Byant, D. (2000b). Accuracy and robustness of analyses on numbers of genes in observed segments. In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic Publishers, 299–306.

Sankoff, D. and Trinh, P. (2005). Chromosomal breakpoint reuse in genome sequence rearrangement. *Journal of Computational Biology* **12**, 812–821.

Savva, G. (2001). Estimating transition probabilities using a model of chromosomal evolution. MSc dissertation, University College London.

Savva, G. (2006). Phylogenetic analysis using a model of whole genome evolution. Ph.D. dissertation, University of East Anglia.

Schleuter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J. and Shoemaker, R.C. (2004). Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**, 868–876.

Schoen, D.J. (2000). Marker density and estimates of chromosome rearrangement. In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic Publishers, pp. 307–319.

Searle, A.G. (1968). Coat color genetics and problems of homology. In *Haldane and Modern Biology*, K.R. Dronamraju, ed. Johns Hopkins, pp. 27–41.

Simillion, C., Vandepoele, K., Van Montagu, M.C.E., Zabeau, M. and Van de Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 13627–13632.

Smith, T.F. and Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197.

Snel, B., Bork, P. and Huynen, M.A. (1999). Genome phylogeny based on gene content. *Nature Genetics* **21**, 108–110.

Tang, J. and Moret, B. (2003). Phylogenetic reconstruction from gene rearrangement data with unequal gene contents. In *Proceedings of the 8th Workshop on Algorithms and Data Structures (WADS'03)*, Vol. 2748 of Lecture Notes in Computer Science. Springer-Verlag, Berlin, pp. 37–46.

Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997). A genomic perspective on protein families. *Science* **278**, 631–637.

Tekaia, F., Lazcano, A. and Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Research* **9**, 550–557.

Tesler, G. (2002). GRIMM: genome rearrangements web server. *Bioinformatics* **18**, 492–493.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.

Thorne, J.L., Kishino, H. and Felsenstein, J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution* **34**, 3–16.

Vandepoele, K., Saeys, Y., Simillion, C., Raes, J. and Van de Peer, Y. (2002). The Automatic Detection of Homologous Regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Research* **12**, 1792–1801.

Vandepoele, K., Simillion, C. and Van de Peer, Y. (2003). Evidence that rice and other cereals are ancient aneuploids. *The Plant Cell* **15**, 2192–2202.

Vision, T.D., Brown, D.B. and Tanksley, S.D. (2000). The origins of genomic duplications in *Arabidopsis*. *Science* **290**, 2114–2117.

Waddington, D. (2000). Estimating the number of conserved segments between species using a chromosome based model. In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic Publishers, 321–332.

Waddington, D., Springbett, A.J. and Burt, D. (2000). A chromosome based model for estimating the number of conserved segments between pairs of species from comparative genetic maps. *Genetics* **154**, 323–332.

Wang, X., Shi, X., Hao, B., Ge, S. and Luo, J. (2005). Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytologist* **165**, 937–946.

Whittam, T.S. and Bumbaugh, A.C. (2002). Inferences from whole-genome sequences of bacterial pathogens. *Current Opinion in Genetics and Development* **12**, 719–725.

Winzeler, E.A., Castillo-Davis, C.I., Oshiro, G., Liang, D., Richards, D.R., Zhou, Y. and Hartl, D.L. (2003). Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* **163**, 79–89.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L. and Koonin, E.V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology* **1**, 8.

Wong, S. and Wolfe, K.H. (2005). Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nature Genetics* **37**, 777–782.

Yancopoulos, S., Attie, O. and Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* **21**, 3340–3346.

York, T.L., Durrett, R. and Nielsen, R. (2002). Bayesian estimation of the number of inversions in the history of two chromosomes. *Journal of Computational Biology* **9**, 805–818.

Zhang, H. and Gu, X. (2004). Maximum likelihood for genome phylogeny on gene content. *Statistical Applications in Genetics and Molecular Biology* **3**, 31.

Part 2

Beyond the Genome

Analysis of Microarray Gene Expression Data

W. Huber

Department of Molecular Genome Analysis, German Cancer Research Center, Heidelberg, Germany

A. von Heydebreck and M. Vingron

Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics, Berlin, Germany

This chapter reviews the methods utilized in processing and analysis of gene expression data generated using DNA microarrays. This type of experiment allows relative levels of mRNA abundance in a set of tissue samples or cell populations to be determined for thousands of genes simultaneously. Naturally, such an experiment requires computational and statistical analysis techniques. As processing begins, the computational procedures are largely determined by the technology and experimental setup used. Subsequently, as more reliable intensity values for genes emerge, pattern discovery methods come into play. The most striking peculiarity of this kind of data is that one usually obtains measurements for thousands of genes for a much smaller number of conditions. This is at the root of several of the statistical questions discussed here.

6.1 INTRODUCTION

In the context of the Human Genome Project, new technologies have emerged that facilitate the parallel execution of experiments on a large number of genes simultaneously. The so-called DNA microarrays, or DNA chips, constitute a prominent example. This technology aims to measure mRNA levels in particular cells or tissue samples for many genes at once. To this end, single strands of complementary DNA for the genes of interest – of which there may be many thousands – are immobilized on spots arranged in a grid ('array') on a support which will typically be a glass slide, a quartz wafer, or a nylon

membrane. From a sample of interest, such as a tumor biopsy, the mRNA is extracted, labeled and hybridized to the array. Measuring the quantity of label on each spot then yields an intensity value that should be correlated to the abundance of the corresponding RNA transcript in the sample.

Two schemes of labeling are in common use today. One variant labels a single sample either radioactively or fluorescently. Radioactive labeling is used, for example, in conjunction with hybridization on nylon membranes (Lennon and Lehrach, 1991). The company Affymetrix synthesizes sets of short oligomers on a glass wafer and uses a single fluorescent label (Lipshutz *et al.*, 1999; see also www.affymetrix.com). Alternatively, two samples are labeled with a green and a red fluorescent dye, respectively. The mixture of the two mRNA preparations is then hybridized simultaneously to a common array on a glass slide. This technology is usually referred to as the Stanford technology (Duggan *et al.*, 1999). Quantification utilizes a laser scanner that determines the intensities of each of the two labels over the entire array. Recently, companies like Agilent have immobilized long oligomers 60–70 base pairs in length and used two-color labeling.

The parallelism in this kind of experiment lies in the hybridization of mRNA extracted from a single sample to many genes simultaneously. The measured abundances, though, are not obtained on an absolute scale. This is because they depend on many factors that are hard to control, such as the efficiencies of the various chemical reactions involved in the sample preparation, as well as on the amount of immobilized DNA available for hybridization.

The class of transcripts that is probed by a spot may differ in different applications. Most commonly, each spot is meant to probe a particular gene. The representative sequence of DNA on the spot may be either a carefully selected fragment of cDNA, a more arbitrary polymerase chain reaction (PCR) product amplified from a clone matching the gene, or one of a set of oligonucleotides specific to the gene. Another level of sophistication is reached when a spot represents, for example, a particular transcript of a gene. In this case, or to distinguish mRNA abundances of genes from closely related gene families, careful design and/or selection of the immobilized DNA is required. Likewise, the selection of samples to study and to compare to each other using DNA microarrays requires careful planning, as will become clear upon consideration of the statistical questions arising from this technology (Kerr and Churchill, 2001a; Churchill, 2002; Yang and Speed, 2002).

There are many different ways to plan a microarray experiment. In many cases a development in time is studied, leading to a series of hybridizations following each other. Alternatively, different conditions such as the presence/absence of disease or different disease types may be studied. We generally refer to a time point or a state as a condition, and typically for each condition several replicate hybridizations are performed. The replicates should provide the information necessary to judge the significance of the conclusions one wishes to draw from the comparison of the different conditions. When delving deeper into the subject it soon becomes clear that this simple outline constitutes a rather challenging program.

This chapter is organized along the various steps of analysis of a microarray experiment. Statistical problems arise firstly as a consequence of various technical peculiarities, and their solution is a prerequisite to any meaningful subsequent interpretation of the experiment. Section 6.2 describes some of the issues related to quality control. Visualization methods are introduced because they may greatly help in detecting and

removing obviously failed measurements, as well as in finding more subtle systematic biases associated with variations in experimental conditions.

Microarray measurements are subject to multiple sources of experimental variation, the mathematical treatment of which is discussed in Section 6.3. Some of the variations are *systematic* and may be explicitly corrected for, others are *random* and may be accounted for through an error model. The correction for systematic effects is referred to as calibration or normalization. We will discuss two error models: one involving a constant coefficient of variation, that is, a purely multiplicative noise term; and one allowing for a more general variance-to-mean dependence, with a noise term that has both multiplicative and additive components. From these models we derive *measures of relative abundance* of mRNA.

The goal of many microarray experiments is to identify genes that are differentially transcribed with respect to different biological conditions of cell cultures or tissue samples. Section 6.4 focuses on these issues, paying particular attention to the notoriously low numbers of repeated hybridizations per condition in relation to the high numbers of genes about which one wishes to draw conclusions. Section 6.5 proceeds to highlight some of the issues in pattern discovery in microarray data. Here, again, classical methods of data analysis need to be carefully evaluated with respect to their applicability to the particular type of data at hand. A short summary will be given of the methods that have so far been successfully applied. Emphasis is given to exploratory approaches that allow the subsequent formulation of hypotheses that can be tested through either further analysis or further experiments.

6.2   DATA VISUALIZATION AND QUALITY CONTROL

A microarray experiment consists of the following components: a set of *probes*, an *array* on which these probes are immobilized at specified locations, a *sample* containing a complex mixture of labeled biomolecules that can bind to the probes, and a *detector* that is able to measure the spatially resolved distribution of label after it has bound to the array (see *Nature Genetics* **21**(supplement), 1999). The probes are chosen such that they bind to specific sample molecules; for DNA arrays, this is ensured by the high sequence-specificity of the hybridization reaction between complementary DNA strands. The array is typically a glass slide or a nylon membrane. The sample molecules may be labeled through the incorporation of radioactive markers, such as ^{33}P, or of fluorescent dyes, such as phycoerythrin, Cy3, or Cy5. After exposure of the array to the sample, the abundance of individual species of sample molecules can be quantified through the signal intensity at the matching probe sites. To facilitate direct comparison, the spotted array technology developed in Stanford (Duggan *et al.*, 1999) involves the simultaneous hybridization of two samples labeled with different fluorescent dyes, and detection at the two corresponding wavelengths. Plate 1 shows an example.

6.2.1   Image Quantification

The intensity images are scanned by the detector at a high spatial resolution, such that each probe spot is represented by many pixels. In order to obtain a single overall intensity value for each probe, the corresponding pixels need to be identified (segmentation),

and the intensities need to be summarized (quantification). In addition to the overall probe intensity, further auxiliary quantities may be calculated, such as an estimate of apparent unspecific 'local background' intensity, or a spot quality measure. A variety of segmentation and quantification methods is implemented in available software packages. These differ in their robustness against irregularities and in the amount of human interaction that they require. Different types of irregularities may occur in different types of microarray technology, and a segmentation or quantification algorithm that is good for one platform is not necessarily suitable for another. For instance, the variation of spot shapes and positions that the segmentation has to deal with depends on the properties of the support (e.g. glass or nylon), on the probe delivery mechanism (e.g. quill-pen type, pin and ring systems, ink-jetting), and on the detection method (optical or radioactive). Furthermore, larger variations in the spot positioning from array to array can be expected in home-made arrays than in mass-produced ones. An evaluation of image analysis methods for spotted cDNA arrays was reported by Yang *et al.* (2002).

For a microarray project, the image quantification marks the transition in the work flow from 'wet lab' procedures to computational ones. Hence, this is a convenient time to look at the quality and plausibility of the data. There are several aspects to this: confirmation that positive and negative controls behave as expected; verification that replicates yield measurements close to each other; and checking for the occurrence of artifacts, biases, or errors. In the following we present a number of data exploration and visualization methods that may be useful for these tasks.

6.2.2   Dynamic Range and Spatial Effects

A simple and fundamental property of the data is the dynamic range and the distribution of intensities. Since many experimental problems occur at the level of a whole array or the sample preparation, it is instructive to look at the histogram of intensities from each sample. An example is shown in Figure 6.1. Typically, for arrays that contain quasi-random gene selections, one observes a unimodal distribution with most of its mass at small intensities, corresponding to genes that are not or only weakly transcribed in the sample, and a long tail to the right, corresponding to genes that are transcribed at various levels. In most cases, the occurrence of multiple peaks in the histogram indicates an experimental artifact. To get an overview over multiple arrays, it is instructive to look at the boxplots of the intensities from each sample. Problematic arrays should be excluded from further analysis.

Crude artifacts, such as scratches or spatial inhomogeneities, will usually be observed in the scanner image at the stage of the image quantification. Nevertheless, to get a quick and potentially more sensitive view of spatial effects, a false-color representation of the probe intensities as a function of their spatial coordinates can be useful. There are different options for the intensity scaling, among them the linear, logarithmic, and rank scales. Each will highlight different features of the spatial distribution. Examples are shown in Plate 2.

6.2.3   Scatterplot

Usually, the samples hybridized to a series of arrays are biologically related, such that the transcription levels of a large fraction of genes are approximately the same across the samples. This can be expected, for example, for cell cultures exposed to different conditions or for cells from biopsies of the same tissue type, possibly subject to different
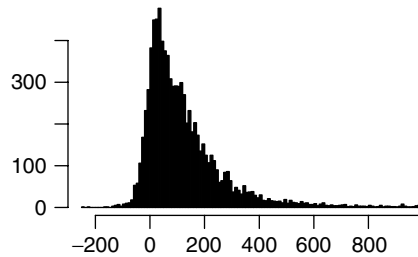
Figure 6.1 Histogram of probe intensities at the green wavelength for a cDNA microarray similar to that depicted in Plate 1. The intensities were determined, in arbitrary units, by an image quantification method, and 'local background' intensities were subtracted. Due to measurement noise, these lead to nonpositive probe intensities for part of the genes with low or zero abundance. The x-axis has been cut off at the 99 % quantile of the distribution. The maximum value is about 4000.
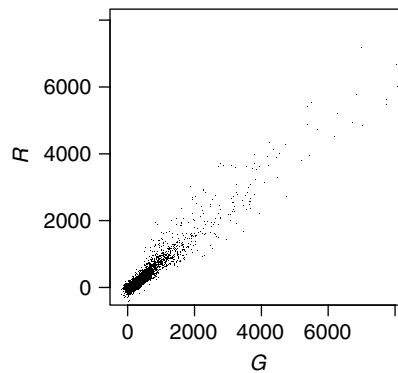
Figure 6.2 Scatterplot of probe intensities in the red and the green color channel from a cDNA array containing 8000 probes.

disease conditions. We call this the *majority of genes unchanged* property. Visually, it can be verified from the scatterplot of the probe intensities for a pair of samples. An example is shown in Figure 6.2.

The scatterplot allows both measurement noise and systematic biases to be assessed. Ideally, the data from the majority of the genes that are unchanged should lie on the bisector of the scatterplot. In reality, there are both systematic and random deviations from this (Schuchhardt *et al.*, 2000). For instance, if the label incorporation rate and photoefficiency of the red dye were systematically lower than that of the green dye by a factor of 0.75, the data would be expected not to lie on the bisector, but rather on the line $y = 0.75x$.

Most of the data in Figure 6.2 is squeezed into a tiny corner in the bottom left of the plot. More informative displays may be obtained from other axis scalings. A frequently used choice is the double-logarithmic scale. An example is shown in Figure 6.3. It is customary to transform to new variables $A = \log R + \log G$, $M = \log R - \log G$ (Dudoit *et al.*, 2002b). Up to a scale factor of $\sqrt{2}$, this corresponds to a clockwise coordinate system rotation by 45°. The horizontal coordinate A is a measure of average transcription
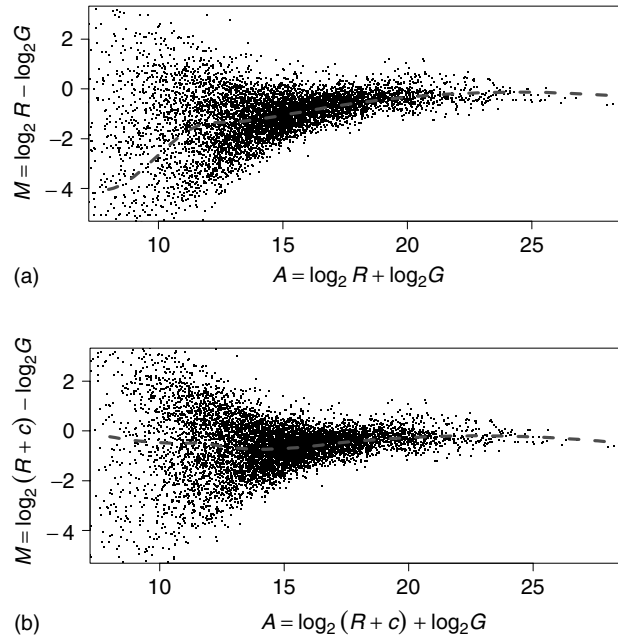
Figure 6.3 (a) The same data as in Figure 6.4, after logarithmic transformation and clockwise rotation by 45°. The dashed line shows a local regression estimate of the systematic effect $M_0(A)$; see text. (b) Similar to (a), but with a constant value $c = 42$ added to the red intensities before log transformation. As a result, the estimated curve for the systematic effect $M_0(A)$ is approximately constant.

level, while the *log ratio M* is a measure of differential transcription. If the majority of genes are not differentially transcribed, the scatter of the data points in the vertical direction may be considered a measure of the random variation. Figure 6.3(a) also shows a systematic deviation of the observed values of M from the line $M = 0$, estimated through a local regression line.[1] There is an apparent dependence $M_0(A)$ of this deviation on the mean intensity A. However, this is most likely an artifact of applying the logarithmic transformation: as shown in Figure 6.3(b), the deviation may be explained sufficiently well through a constant $M_0(A) = M_0$ if an appropriate offset is added to the R values before taking the logarithm. Note that a horizontal line at $M = M_0$ in Figure 6.3(b) corresponds to a straight line with slope 2^{M_0} and intercept c in Figure 6.2.

Figure 6.3 shows the *heteroskedasticity* of log ratios: while the variance of M is relatively small and approximately constant for large average intensities A, it becomes larger as A decreases. Conversely, examination of the differences $R - G$, for example through plots as in Figure 6.2, shows that their variance is smallest for small values of the average intensity $R + G$ and increases with $R + G$. Sometimes one wishes to visualize the data in a manner such that the variance is constant along the whole dynamic range. A data transformation that achieves this goal is called a variance-stabilizing transformation.

---

[1] We used `loess` (Cleveland *et al.*, 1992) with default parameters span $= 0.75$, degree $= 2$.

In fact, *homoskedastic* representations of the data are useful not only for visualization, but also for further statistical analyses. This will be discussed in more detail in Section 6.3.

Two extensions of the scatterplot are shown in Plates 3 and 4. Rather than plotting a symbol for every data point, they use a density representation, which may be useful for larger arrays. For example, Plate 3 shows the scatterplot from the comparison of two tissue samples based on 152 000 probes.[2] The point density in the central region of the plot is estimated by a kernel density estimator. Three-way comparisons may be performed through a projection such as in Plate 4. This uses the fact that the $(1, 1, 1)$ component of a three-way microarray measurement corresponds to average intensity, and hence is not directly informative with respect to differential transcription. Note that if the plotted data were preprocessed through a variance-stabilizing transformation, its variance does not depend on the $(1, 1, 1)$ component.

6.2.4   Batch Effects

Present-day microarray technology measures abundances only in terms of relative probe intensities, and generally provides no calibration to absolute physical units. Hence, the comparison of measurements between different studies is difficult. Moreover, even within a single study, the measurements are highly susceptible to *batch effects* – experimental factors that add systematic biases to the measurements, and may vary between different subsets or stages of an experiment. The following are some examples (Schuchhardt *et al.*, 2000):

1. *Spotting.* To manufacture spotted microarrays, the probe DNA is deposited on the surface through spotting pins. Usually, the robot works with multiple pins in parallel, and the efficiency of their probe delivery may be quite different (see Plate 2d or Dudoit *et al.*, 2002b). Furthermore, the efficiency of a pin may change over time through mechanical wear, and the quality of the spotting process as a whole may be different at different times, due to varying temperature and humidity conditions.

2. *PCR amplification.* For cDNA arrays, the probes are synthesized through PCR, whose yield varies from instance to instance. Typically, the reactions are carried out in parallel in 384-well plates, and probes that have been synthesized in the same plate tend to have highly correlated variations in concentration and quality. An example is shown in Plate 5.

3. *Sample preparation protocols.* The reverse transcription and the labeling are complex biochemical reactions, whose efficiencies are variable and may depend sensitively on a number of circumstances that are hard to control. Furthermore, RNA can quickly degrade, hence the outcome of the experiment can depend sensitively on when and how conditions that prevent RNA degradation are applied to the tissue samples.

4. *Array coating.* Both the efficiency of the probe fixation on the array, and the amount of unspecific background fluorescence strongly depend on the array coating.

5. *Scanner and image analysis.*

These considerations have important consequences for experimental design. First, any variation that can be avoided within an experiment should be avoided. Second, any

---

[2] The arrays used were RZPD Unigene-II arrays (www.rzpd.de).

variation that cannot be avoided should be dealt with so as not to confound the biological question of interest. Clearly, when looking for differences between two tumor types, it would not be wise to have samples of one tumor type processed by one laboratory, and samples of the other type by another laboratory.

Points 1 and 2 are specific to spotted cDNA arrays. For less sensitivity to these variations, the two-color labeling protocol is used, which employs the simultaneous hybridization of two samples to the same array (Duggan *et al.*, 1999). Ideally, if only ratios of intensities between the two-color channels are considered, variations in probe abundance should cancel out. Empirically they do not quite do so, which may, for example, be attributed to the fact that observed intensities are the sum of probe-specific signal and unspecific background (Yue *et al.*, 2001). Furthermore, in the extreme case of total failure of the PCR amplification or the DNA deposition for probes on some, but not all, arrays in an experimental series, artifactual results are hardly avoidable.

If any of the factors 3–5 is changed within an experiment, there is a good chance that this will later show up in the data as one of the most pronounced sources of variation. A simple and instructive visual tool to explore such variations is the correlation plot: given a set of d arrays, each represented by a high-dimensional vector $\vec{Y}_i$ of suitably transformed and filtered probe intensities, calculate the $d \times d$ correlation matrix $\mathrm{corr}(\vec{Y}_i, \vec{Y}_j)$, sort its rows and columns according to different experimental factors, and visualize the resulting false-color images.

6.3 ERROR MODELS, CALIBRATION AND MEASURES OF DIFFERENTIAL EXPRESSION

The relation between a measured intensity y_{ki} of probe k and the true abundance x_{ki} of molecule type k in sample i may be written as

$$y_{ki} = a_{ki} + b_{ki}\, x_{ki}. \tag{6.1}$$

The gain factor b_{ki} represents the net result of the various experimental effects that come between the count of molecules per cell in the sample and the final readout of the probe intensity, such as the number of cells in the sample, the mean number of label molecules attaching to a sample molecule, hybridization efficiency, label efficiency, and detector gain. The additive term a_{ki} accounts for that part of the measured intensity that does not result from x_{ki}, but from effects such as unspecific hybridization, background fluorescence, stray signal from neighboring probes, and detector offset.

The parameters a_{ki} and b_{ki} are different for each probe k and for each hybridization i. It is not practical to determine them exactly, but neither is it necessary. Rather, one is content with obtaining *statistical* statements about *relative* abundances. To this end, one may build stochastic models for the effects a_{ki} and b_{ki}. Different variations on this theme have been proposed, as will be presented below.

First, however, we would like to discuss the functional form of equation (6.1), which says that when the true abundance x_{ki} increases, the measured signal y_{ki} increases proportionally. Might it be necessary to consider more complex nonlinear relationships? Clearly, this cannot be ruled out for all possible experiments or for future technologies. However, a linear operating range over several orders of magnitude has been reported by

a number of authors for current microarray technologies (e.g. Ideker *et al.*, 2000; Ramdas *et al.*, 2001; Irizarry *et al.*, 2003). At the lower end, this range is limited only by the requirement that x_{ki} be nonnegative. At the upper end, the linear range is limited by saturation effects such as quenching, limited probe abundance, and detector saturation. However, for realistic concentrations of sample molecules, the upper limit is not reached in well-conducted experiments.

6.3.1 Multiplicative Calibration and Noise

In a seminal paper, Chen *et al.* (1997) introduced a decomposition of the multiplicative effect (cf. (6.1)),

$$b_{ki} = b_i \beta_k (1 + \varepsilon_{ki}). \tag{6.2}$$

Here, β_k is a probe-specific coefficient, the same for all samples. For each sample i, the normalization factor b_i is applied across all probes. The remaining variation in b_{ki} that cannot be accounted for by β_k and b_i is absorbed by ε_{ki}. Furthermore, since the measured intensities y_{ki} are already 'background-corrected' by the image analysis software's local background estimation, Chen *et al.* assumed the additive effects a_{ki} to be negligibly small. They further simplified the problem in two steps.

First, they noted that one is mainly interested in relative comparisons between the levels of the same gene under different conditions, that is, in the ratios x_{ki}/x_{kj}. Hence the probe-specific effects β_k can be absorbed, $\mu_{ki} = \beta_k x_{ki}$, simply rescaling the units in which molecule abundances are measured, and need not be determined.

Second, they turned to a stochastic description, and modeled ε_{ki} as a normally distributed noise term with mean zero and standard deviation c, independent of i and k. Thus, in the model of Chen *et al.* the measured intensity Y_{ki} is a random variable and depends on the true level μ_{ki} as follows:

$$Y_{ki} = b_i \, \mu_{ki} \, (1 + \varepsilon_{ki}), \qquad \varepsilon_{ki} \sim N(0, c^2). \tag{6.3}$$

Note that Y_{ki} has constant coefficient of variation c.

Chen *et al.* specifically considered two-color cDNA microarrays, where $i = 1, 2$ represents the red and the green color channel, respectively. For a given true ratio μ_{k1}/μ_{k2}, Chen *et al.* derived the distribution of the observed, normalized ratio $M_k = Y_{k2}/Y_{k1} \times b_1/b_2$. This depends only on the values of c and b_1/b_2, and Chen *et al.* gave an algorithm for the estimation of these parameters from the data. Based on this, they were able to formulate a statistical test for differential expression, that is, for the hypothesis $\mu_{k1} = \mu_{k2}$. Hence, the ratios M_k were regarded as a sufficient summary of the results from a single microarray slide, and they, or their logarithms, would then be used as the input for further higher-level analyses of data from multiple slides.

To allow for a more systematic analysis of multiple slide experiments, Kerr *et al.* (2000) proposed an approach based on the ANOVA technique. They modeled the measured intensity Y_{kjlm} of probe k on slide j, in the color channel of dye l, from a sample that received treatment m, as

$$\log Y_{kjlm} = g_k + s_j + d_l + v_m + [gs]_{kj} + [gv]_{km} + \varepsilon_{kjlm}. \tag{6.4}$$

Here g_k, s_j, d_l, v_m are main effects for probe, array, dye, and treatment, respectively. The probe–array interaction $[gs]_{kj}$ accounts for variations of probe quality in the

array manufacture, and the probe–treatment interaction $[gv]_{km}$ for differential levels of transcription of gene g between different treatment groups m. The noise terms ε_{kjlm} account for all other variations and are assumed to be independent and identically distributed. The ANOVA model (6.4) is related to (6.1) by setting

$$a_{ki} = 0, \tag{6.5}$$

$$\log b_{ki} + \log x_{ki} = (s_j + d_l + [gs]_{kj}) + (g_k + v_m + [gv]_{km}) + \varepsilon_{kjlm}, \tag{6.6}$$

where $j \equiv j(i)$, $l \equiv l(i)$ and $m \equiv m(i)$ are slide, dye, and treatment associated with sample i, respectively. The terms in the first set of parentheses on the right-hand side of (6.6) may be attributed to the measurement gain b_{ki}, and the terms in the second set to the actual abundance x_{ki}, but generally such a decomposition is not unique.

Both the models of Chen *et al.* and Kerr *et al.* were formulated with reference to the two-color cDNA array technology. However, (6.4) can be adapted (in fact, simplified) in a straightforward manner to data from one-color array technologies, such as Affymetrix genechips or cDNA membranes with radioactive labeling. Furthermore, to represent more complex experimental designs than simple two-way comparisons, more detailed terms than the single factor v_m can be introduced into (6.4), and the efficiencies of different designs can be compared using standard techniques for linear models (Kerr and Churchill, 2001a).

6.3.2   Limitations

The concepts of Section 6.3.1 have been widely used for microarray data analysis. However, it has also become clear that, for many data sets that are encountered in practice, they are not sufficient. The following points are worth noting:

1. *Robustness.* In order to make model (6.3) identifiable, Chen *et al.* assumed that the transcription levels of all genes were unchanged, and set $\mu_{k1} = \mu_{k2}$ for all k. Thus, their model is misspecified for the part of the data arising from truely differentially transcribed genes, which act as outliers. However, their parameter estimation is based on least-squares criteria, and may be sensitive to the presence of such outliers. In addition, outliers may be caused by technical artifacts.

2. *Heteroskedasticity.* The significance of log ratios depends on the absolute values of the intensities in the numerator and denominator (Beissbarth *et al.*, 2000; Baggerly *et al.*, 2001; Newton *et al.*, 2001; Theilhaber *et al.*, 2001; Kepler *et al.*, 2002). Typically, the variance of log-transformed intensities increases as their mean decreases.

3. *Apparent nonlinearities.* According to the above models, the data from a pair of samples should lie along a straight line in the scatterplot of the log-transformed intensities. However, in real data, several authors have observed data that follows a curved line as in Figure 6.3 (Beissbarth *et al.*, 2000; Dudoit *et al.*, 2002b; Kepler *et al.*, 2002).

4. *Negative values.* While the image quantification's estimates for probe 'foreground' and 'background' intensities are generally positive, this is usually not true for their difference. If a gene is weakly or not expressed, it can happen by chance that the background estimate is larger than the foreground estimate (see Figure 6.1). However, nonpositive values make sense neither for ratios nor for the log-transformation.

To address these problems, various fixes have been proposed. We give a brief and incomplete review.

1. *Robustness.* Robust estimation techniques in the context of microarray data have been described by many authors (Beissbarth *et al.*, 2000; Thomas *et al.*, 2001; Dudoit *et al.*, 2002b; Huber *et al.*, 2002; Kepler *et al.*, 2002). A general overview is given in Rousseuw and Leroy (1987).

2. *Heteroskedasticity.* It is often observed that the variance of the log ratio is a monotonously decreasing function of the mean intensity. One common practice has been to discard the log ratios calculated from intensities below a certain threshold and to treat the rest as if they were homoskedastic.

Newton *et al.* (2001) proposed a *shrinkage estimator*

$$\frac{y'_{k1} + \nu}{y'_{k2} + \nu} \tag{6.7}$$

to replace the naive ratio y'_{k1}/y'_{k2}. Here, $y'_{ki} = y_{ki}/b_i$ are the normalized intensities. Similarly to Chen *et al.* they neglected the additive terms a_{ki} and used a model of the measurement error with a constant coefficient of variation. To arrive at (6.7), they enclosed this in a hierarchical Bayesian model, using a prior distribution for the mRNA abundances, and, in particular, their positivity. The form of this distribution is reflected by the shrinkage parameter ν, which is estimated from the data. To infer differential transcription, they derived 'posterior odds of change', which, however, are not a simple function of the log ratio or of (6.7).

Several authors have addressed the problem of heteroskedasticity by estimating the variance of the log ratios or of log-transformed intensities separately for each gene (e.g. Thomas *et al.*, 2001; Dudoit *et al.*, 2002b). However, in many applications the number of samples available is too small for reliable estimates of gene-specific variance, hence it has been proposed to estimate the variance as a nonparametric smooth function of the mean intensity, through a local regression. The log ratios may then be *studentized* by dividing them by their locally estimated standard deviation (Kepler *et al.*, 2002). Baggerly *et al.* (2001) provided some theoretical foundation for this from models of the measurement error for different levels of replication. According to these, the variance of the log ratio is largest for small intensities and exponentially decreases toward an asymptotic positive value as the intensity increases.

3. *Apparent nonlinearities.* To correct for the curved appearance of the scatterplot of log-transformed data, Dudoit *et al.* (2002b) proposed to replace the normalization factor b_1/b_2 in (6.3) by a smooth function $M_0(A)$ (see Figure 6.3). This is estimated by robust local regression (Cleveland *et al.*, 1992) and, by construction, this correction makes the scatterplot look straight.

A similar correction was proposed by Kepler *et al.* (2002), in the framework of a model similar to (6.4). In their approach, the terms $s_j + d_l$ (slide and dye effects) are replaced by smooth functions of g_k (mean logarithmic abundance of gene k), which are again estimated by robust local regression.

4. *Negative values.* In order to be able to calculate ratios and logarithms from real microarray data, different fixes have been proposed to deal with nonpositive values: marking them as invalid or missing; replacing them by a fixed, small positive value; using an imputation algorithm to replace them by a more acceptable value; adding pseudocounts,

such that the whole set of intensities becomes positive; ignoring the local background estimate (cDNA arrays) or the mismatch probes (Affymetrix genechips) and using only the strictly positive foreground or perfect match intensities.

All of these approaches seem to reflect the common wisdom that molecule abundances are not negative. However, probe intensities are only *measurements* of abundance, and in the presence of an additive component of the measurement noise negative measurements may well be consistent with zero or positive abundance. In any experiment, a certain proportion of genes will have zero or low abundance in some samples but not in others, hence the treatment of the nonpositive intensity measurements may affect a large and potentially informative fraction of the data.

6.3.3 Multiplicative and Additive Calibration and Noise

Interestingly, points 2–4 of Section 6.3.2 can be related to a rather basic assumption of models (6.3) and (6.4), and it appears that in many cases the associated problems can be resolved by using a more general model. Chen *et al.* as well as Kerr *et al.* assumed that the additive terms a_{ki} in (6.1) were negligible, or at least sufficiently accounted for by the image quantification's local background estimation algorithm. One way to arrive at a more realistic model is to set

$$a_{ki} = a_i + b_i \eta_{ki}, \tag{6.8}$$

$$b_{ki} = b_i \beta_k (1 + \varepsilon_{ki}), \tag{6.9}$$

where the decomposition of the multiplicative effect (6.9) is the same as in (6.2), a_i is a sample-specific additive parameter, and η_{ki} are independent and normally distributed random variables with zero mean and common variance. Hence, model (6.3) is replaced by

$$\frac{Y_{ki} - a_i}{b_i} = \mu_{ki}\, e^{\varepsilon_{ki}} + \eta_{ki}, \qquad \varepsilon_{ki} \sim N(0, c^2),\ \eta_{ki} \sim N(0, s^2). \tag{6.10}$$

Model (6.10) was proposed by Rocke and Durbin (2001) and, using different distributional assumptions, by Ideker *et al.* (2000). The latter authors used a more detailed parameterization of the noise terms, allowing for different values of the standard deviations c and s for the red and green color channels $i = 1, 2$ and for correlation between ε_{k1} and ε_{k2}, as well as between η_{k1} and η_{k2}. In both cases, the authors did not try to estimate the calibration parameters a_i, b_i, but rather assumed that a calibration had already been performed by some other means.

6.3.3.1 Consequences

First, the intensities Y_{ki} are no longer supposed to have a constant coefficient of variation. Rather, they obey a variance-to-mean dependence

$$v(u) = c^2(u - a_i)^2 + b_i^2 s^2, \tag{6.11}$$

where, in a slight abuse of notation, $v \equiv \text{var}(Y_{ki})$ and $u \equiv \text{E}(Y_{ki})$, and the equation holds for all probes k for sample i. Recall that a constant coefficient of variation corresponds to

a dependence $v(u) = c^2 u^2$, which is a special case of (6.11) for $a_i = s = 0$. In this case, the logarithm is a variance-stabilizing transformation, that is, the log-transformed data have approximately constant variance. For the more general variance-to-mean dependence (6.11), such a transformation can also be found, as will be explained below.

Second, the ratio of intensities y_{k1}/y_{k2} is no longer the best estimator for the true fold change μ_{k1}/μ_{k2}. This was addressed by Dror *et al* (2002), who estimated $\log(\mu_{k1}/\mu_{k2})$ by the posterior mean of a hierarchical model that consists of (6.10) together with an empirical prior for the distribution of μ_{ki}. Their estimator coincides with the log ratio if both y_{k1} and y_{k2} are large, and remains well behaved for small or nonpositive values of y_{k1} and y_{k2}.

The appropriate variance-stabilizing transformation was described by Huber *et al.* (2002) and Durbin *et al.* (2002). It has the form

$$h_i(y_{ki}) = \operatorname{arcsinh}\left(\frac{c}{s} \cdot \frac{y_{ki} - a_i}{b_i}\right). \tag{6.12}$$

The parameters a_i and b_i may be interpreted as array-specific calibration parameters, while the coefficient of variation c and the background noise level s parameterize the overall error model. The graph of the arcsinh function is depicted in Figure 6.4. The following two relations hold between the arcsinh and the log function:

$$\operatorname{arcsinh}(x) = \log(x + \sqrt{x^2 + 1}), \tag{6.13}$$

$$\lim_{x \to \infty} \{\operatorname{arcsinh}(x) - \log(2x)\} = 0. \tag{6.14}$$

In the framework of Section 6.3.1, log ratios, the differences between the logarithms of normalized intensities, were the appropriate measure of differential transcription to be used in downstream analyses. By analogy, we define (Huber *et al.*, 2002)

$$\Delta h_{k;ij} = h_i(y_{ki}) - h_j(y_{kj}). \tag{6.15}$$
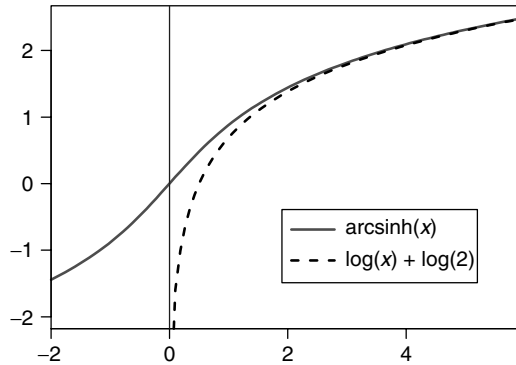
Figure 6.4 Graph of the arcsinh(x) (solid line) and log($2x$) (dashed line). The vertical line marks the singularity of the logarithm function at $x = 0$. The arcsinh function is symmetric, $\operatorname{arcsinh}(x) = -\operatorname{arcsinh}(-x)$; however, most relevant for (6.12) is its behaviour in an x range as depicted here.

For intensities that are much larger than the additive noise level, (6.15) becomes equivalent to the log ratio, as is seen from (6.14). But, in contrast to the log ratio, $\Delta h_{k;ij}$ is well defined and has constant variance c^2 across the whole range of intensities. In fact, $\Delta h_{k;ij}/c$ may be thought of as a 'studentized log ratio'.

To estimate the model and transformation parameters, one could directly fit model (6.10) to the data, using the *majority of genes unchanged* assumption $\mu_{ki} = \mu_k$ for most genes k. A computationally simpler approach is to fit the model

$$h_i(Y_{ki}) = \tilde{\mu}_k + \tilde{\varepsilon}_{ki}, \qquad \tilde{\varepsilon}_{ki} \sim N(0, c^2). \tag{6.16}$$

Up to first and second moments, models (6.10) and (6.16) are equivalent. Parameter estimates can be obtained from a robust variant of the maximum likelihood estimator. A robust estimator with high breakdown point is needed not only because there may be technical outliers, but also because the assumption $\mu_{ki} = \mu_k$ does not hold for a minority of genes that have biologically different transcription levels in different samples (Huber *et al.*, 2002).

The identification of differentially transcribed genes through statistical tests on Δh_k values has been shown to have higher sensitivity and specificity than that through tests on log ratios (Huber *et al.*, 2002). This may be explained by the fact that for nondifferentially transcribed genes the Δh_k values have unimodal distributions with mean zero and variances independent of the genes' mean transcription levels. Hence, within the limits of the error model, all available information with respect to differential transcription of gene k is contained in the values of $\Delta h_{k;ij}$. On the other hand, the distributions of log ratios may have, even for some of the nondifferentially transcribed genes, mean values different from zero due to sensitive dependence on calibration errors, they may have variances that strongly depend on the mean transcription levels, and they may involve missing values, if there are nonpositive net probe intensities. These points are illustrated in Plate 6.

6.3.3.2 Probe Set Summaries

A gene transcript may be represented by multiple probes on an array. To obtain an overall measure of abundance per gene, a straightforward approach is to take the average of the corresponding calibrated and transformed probe intensities (6.12). If additional information on the reliability of the probe measurements is available, a weighted average may be used. This has been investigated most extensively in the context of Affymetrix genechip data (Lipshutz *et al.*, 1999). On these chips, each transcript is represented by 16–20 pairs of oligonucleotide probes referred to as probe sets. Each probe pair consists of an oligonucleotide of 25 bases that exactly matches the target sequence, and of one that has a mismatch in the middle. The mismatch probes are thought to provide estimates of unspecific contributions to the signal measured from the perfect match probes. A good overview, with many further references, was given by Irizarry *et al.* (2003).

6.4   IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

One of the basic goals in the analysis of microarray gene expression data is the identification of differentially expressed genes in the comparison of different types of

cell or tissue samples. In order to control the biological and experimental variability of the measurements, statistical inference has to be based on an adequate number of replicate experiments. Here one may distinguish between cases in which one wishes to make statements on a particular cell population and cases in which one wishes to make statements that hold in the presence of biological variability, such as with biopsy studies of diseased tissues. In the first case independent replications can be obtained at the level of multiple mRNA isolations; in the second they may be provided by samples from different patients.

For the following, we assume that the data are given either as absolute intensities or as relative values with respect to a common reference sample, and have been calibrated (see Section 6.3).

To identify differentially expressed genes with respect to a certain biological question, a suitable statistical test may be performed for each gene (Claverie, 1999). The choice of test statistic depends on the biological question and on the nature of the available experimental data. In the simplest case, one seeks genes that show different transcript abundance between two conditions. In more complex situations, one may look for genes whose abundance is associated with multiple factor levels of one or more sample characteristics. Furthermore, one may consider continuous-valued sample characteristics and test for genes which show nonzero coefficients in a regression model, such as a linear model or a Cox proportional hazards model on patient survival data.

Different test statistics may make more or less strong assumptions on the distributions of replicate measurements. Important questions are whether the distributions are symmetric, how similar or dissimilar they are to normal distributions, what their behavior in the tails is like, and whether or how their variance (or another appropriate measure of scale) varies between different genes or between different conditions. Such differences in the variance may occur for several reasons: in Section 6.3, we discussed the dependence of the variance on the mean, an example of which is given in equation (6.11). There may be other technological effects that can influence the variance of the measurement distributions in a gene- or condition-dependent manner, such as GC content or probe length. Finally, there may be genuine biological differences, such as different tightness of the regulatory control for different genes or for the same gene under different physiological or disease conditions.

Data transformations, such as the logarithmic transformation or a variance-stabilizing transformation like (6.12), may be used to make the distributions more symmetric and possibly close to normal, and to remove the systematic dependence of the variance on the mean (see Section 6.3). In the comparison of two conditions, one might use Student's t-test or the Wilcoxon rank-sum test. Both tests assume that the distributions of the replicate measurements under the two conditions have the same shape and test for differences in the location, with the t-test additionally assuming normal distributions. To account for possibly unequal variances in the two groups, Welch's version of the t-test may be preferred (Welch, 1947; Best and Rayner, 1987). In order to avoid distributional assumptions, Dudoit $et\ al.$ (2002b) proposed to estimate the null distribution of the t-statistic (or, equivalently, of the difference of means) for each gene using permutations of the sample labels. Comparative analyses of different univariate statistical tests in the analysis of gene expression data were presented in Herwig $et\ al.$ (2001), Thomas $et\ al.$ (2001), and Pan (2002), but without a conclusive result.

In addition to standard aspects of hypothesis testing, two specific properties of microarray data have motivated the development of novel strategies:

1. *Variance estimation.* At one extreme, one may try to estimate the variance of the distributions separately for each gene, and possibly for each condition. This requires a large number of repetitions, which are not always available. At the other extreme, one may use a pooled estimate of the variance over all conditions and genes. After the application of a variance-stabilizing transformation such as (6.12), the assumption of constant variance may result in tolerable bias and, due to the large number of genes represented on an array, in very low variance of the estimator. This is the case especially if few repetitions are available. Between the two extremes a number of methods have been proposed that pool the variance estimation over some genes, but also retain some gene dependence.

2. *Multiple testing.* Due to the large number of genes on an array and thus the large number of tests performed, a considerable number of genes may show differential signal intensities simply by chance. Several approaches to assessing the statistical significance of test results obtained from microarray data have been developed.

6.4.1   Regularized t-Statistics

To overcome the instability of the gene-specific variance estimate in the case of few replicate experiments per condition, several authors have proposed methods where a value estimated from a larger set of genes is used to augment the gene-specific standard deviation estimate, thus providing a regularized version of the t-statistic.

Baldi and Long (2001) suggested replacing the within-group empirical variance s_k^2 of gene k in the two-sample t-statistic obtained from d observations by an expression of the form

$$\tilde{\sigma}_k^2 = \frac{v_0\sigma_0^2 + (d-1)s_k^2}{v_0 + d - 2}.$$

This variance estimate results as the posterior mean from a Bayesian hierarchical model for the measurements of each gene under an experimental condition. The measured values are assumed to be normally distributed, and v_0 and σ_0 are hyperparameters of the prior for the parameters of the normal distribution. For practical purposes, the authors recommended choosing σ_0 as the empirical standard deviation obtained from averaging over all genes within a certain intensity range. If a variance-stabilizing transformation has been applied to the data, σ_0 may be obtained from the pooled variance over all genes on the array. The value v_0 is chosen as an integer determining the weight of σ_0 compared to the gene-specific standard deviation. Thus the large number of genes interrogated is exploited to obtain potentially biased but more stable variance estimates for each single gene. The resulting regularized t-statistic, used with a t-distribution with $v_0 + d - 2$ degrees of freedom as null distribution, is shown to perform better than the standard t-test on real and simulated data when there are less than about 5 replications per condition. A similar approach was pursued by Lönnstedt and Speed (2002). Tusher *et al.* (2001), see also Efron *et al.* (2000), also proposed to use a regularized version of the t-statistic, where the empirical standard deviation s_k of gene k is replaced by $\tilde{\sigma}_k = s_k + s_0$, with s_0 determined from the data in a heuristic fashion.

6.4.2 Multiple Testing

Assume that for each gene a statistical test for differential expression has been conducted. If one fixes a genewise significance level of, say, $\alpha = 0.05$, on average one in every 20 genes that are actually not differentially expressed will show a p-value below α just by chance. Due to the large number of genes represented on a microarray, this may lead to a large number of false positive calls. For this reason, Dudoit *et al.* (2002b) suggested choosing a procedure that controls the *familywise error rate* (FWER). The FWER is defined as the probability that the selected set of genes contains at least one false positive. A multiple testing procedure is said to provide *strong control* of the FWER if it controls the FWER for any combination of true and false null hypotheses. If p-values for the test statistics $T_1, \ldots, T_n$ of n genes are available, a simple adjustment that gives strong control of the FWER is the Bonferroni correction, which amounts to multiplying the unadjusted p-values by n. Dudoit *et al.* (2002b) described the use of a stepwise p-value adjustment that is due to Westfall and Young (1993). This procedure is less conservative than the Bonferroni correction and, in contrast to the latter, takes possible dependences between the test statistics into account. The adjusted p-values are estimated by a permutation algorithm.

For many applications, however, control of the FWER is too conservative, with the danger of many interesting genes being missed. As microarrays are often used to screen for candidate genes that may then be validated through further experiments, the researcher may be willing to accept a certain fraction of false positives. This is accommodated by the concept of the *false discovery rate* (FDR; Benjamini and Hochberg, 1995). For a family of hypothesis tests, let R denote the number of rejected null hypotheses, and V the number of falsely rejected null hypotheses. The FDR is defined as

$$FDR = \mathrm{E}\left[\frac{V}{R} \,\middle|\, R > 0\right] \cdot \Pr(R > 0).$$

Benjamini and Hochberg described a procedure to control the FDR under the assumption that the test statistics arising from the true null hypotheses are independent. More precisely, given the set of p-values from all individual hypothesis tests and a desired upper bound q for the FDR, they give a bound p^* such that rejecting all null hypotheses with p-value smaller than p^* guarantees an FDR of at most q for any possible combination of true and false null hypotheses.

Another approach based on the FDR was presented by Storey and Tibshirani (2001); see also Tusher *et al.* (2001). For a given rejection region of the statistical tests, the authors estimated the FDR and the *positive false discovery rate* (pFDR), which is defined as[3]

$$pFDR = \mathrm{E}\left[\frac{V}{R} \,\middle|\, R > 0\right].$$

Rather than computing a rejection region that guarantees an upper bound for the FDR, Storey and Tibshirani assumed that a rejection region was fixed and estimated the FDR on the basis of the distribution of the test statistics. The estimation procedure has been

---

[3] In contrast to the method of Benjamini and Hochberg, where arbitrary but fixed combinations of true and false null hypotheses are allowed, here the null hypotheses are considered as independent and identically distributed Bernoulli random variables that are true with probability π_0.

designed for any kind of dependence between the test statistics and does not require p-values for the single hypothesis tests. The algorithm of Storey and Tibshirani works as follows. They assumed that all null hypotheses were identical and that the same rejection region Γ was used for all test statistics $T_1, \ldots, T_n$, leading to a number $R(\Gamma)$ of rejections. Furthermore, they assumed that the joint null distribution of the test statistics could be simulated by permutations of the sample labels. From this, the authors obtained estimates for the expected number of rejections given that all null hypotheses are true,

$$\widehat{\mathrm{E}}[R^0(\Gamma)],$$

as well as for the probability of at least one rejection,

$$\widehat{\mathrm{Pr}}[R^0(\Gamma) > 0].$$

The pFDR is then estimated by

$$\widehat{\mathrm{pFDR}}(\Gamma) = \frac{\hat{\pi}_0 \cdot \widehat{\mathrm{E}}[R^0(\Gamma)]}{\widehat{\mathrm{Pr}}[R^0(\Gamma) > 0] \cdot \max(R(\Gamma), 1)}, \tag{6.17}$$

and the FDR similarly by

$$\widehat{\mathrm{FDR}}(\Gamma) = \frac{\hat{\pi}_0 \cdot \widehat{\mathrm{E}}[R^0(\Gamma)]}{\max(R(\Gamma), 1)}. \tag{6.18}$$

The expected proportion $\hat{\pi}_0$ of true null hypotheses is estimated as follows. Let Γ' be a rejection region whose complement is likely to be achieved mostly for true null hypotheses. The estimate for π_0 is obtained as

$$\hat{\pi}_0 = \frac{n - R(\Gamma')}{\widehat{\mathrm{E}}[n - R^0(\Gamma')]}.$$

In order to determine how many falsely significant genes may appear with a certain probability, or how likely it is that *all* genes with test statistics in the rejection region are false positives, it is interesting to estimate not only the pFDR, but also quantiles of the distribution of V/R. This is illustrated in Figure 6.5.

Under certain conditions on the dependence structure between the test statistics, it was shown by Storey and Tibshirani (2001) that for all π_0, the estimates are greater than or equal to the true values of the FDR and pFDR in expectation. In Storey (2001) (see also Efron *et al.*, 2001), it is shown that in the case of independent test statistics (and asymptotically also for some forms of dependence) the pFDR can be interpreted in a Bayesian framework as the posterior probability that a gene is not differentially expressed, given its test statistic lies in the rejection region:

$$\mathrm{pFDR}(\Gamma) = \mathrm{Pr}(H = 0 | T \in \Gamma).$$

A special property of the approach of Storey and Tibshirani and Tusher *et al.* is how it makes use of the assumption that the null distributions of the test statistics are identical for all genes. The fact that the estimation procedure is based on the test statistics of *all* genes under permutation of the sample labels gives accurate estimates for relatively few replicate experiments, while at the same time it preserves the dependence structure
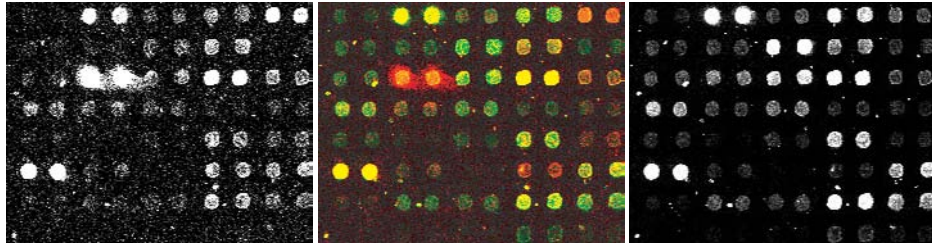
Plate 1. The detected intensity distributions from a cDNA microarray for a region comprising around 80 probes. The total number of probes on an array may range from a few dozen to tens of thousands. Left: gray-scale representation of the detected label fluorescence at 635 nm (red), corresponding to mRNA sample A. Right: label fluorescence at 532 nm (green), corresponding to mRNA sample B. Spots that light up in only one of the two images correspond to genes that are only transcribed in one of the two samples. Center: false-color overlay image from the two intensity distributions. The spots are red, green, or yellow, depending on whether the gene is transcribed only in sample A, sample B, or both.
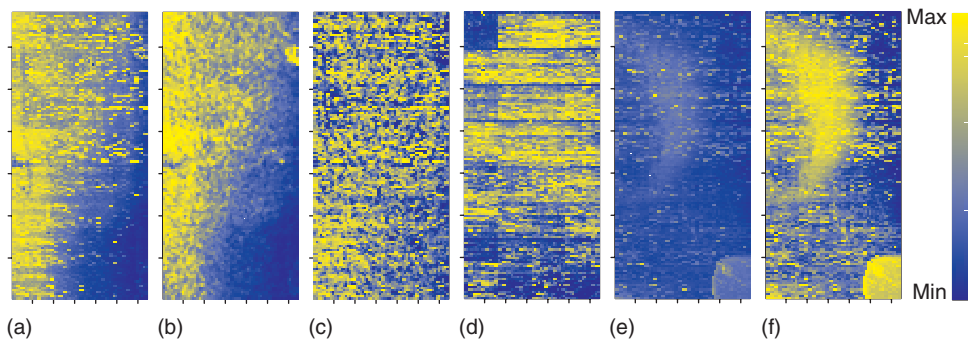
Plate 2. False-color representations of the spatial intensity distributions from three different 64×136 spot cDNA microarrays from one experiment series: (a) Probe intensities in the red color channel; (b) local background intensities; (c) background-subtracted probe intensities. In (a) and (b), there is an artifactual intensity gradient, which is mostly removed in (c). For visualization, the color scale was chosen in each image to be proportional to the ranks of the intensities. (d) For a second array, probe intensities in the green color channel. There is a rectangular region of low intensity in the top left-hand corner, corresponding to one print-tip. Apparently, there was a sporadic failure of the tip for this particular array. (e) and (f) show the probe intensities in the green color channel from a third array. The color scale was chosen proportional to the logarithms of intensities in (e) and proportional to the ranks in (f). Here, the latter provides better contrast. Interestingly, the bright blob in the lower right-hand corner appears only in the green color channel, while the crescent-shaped region appears both in green and red (not shown).
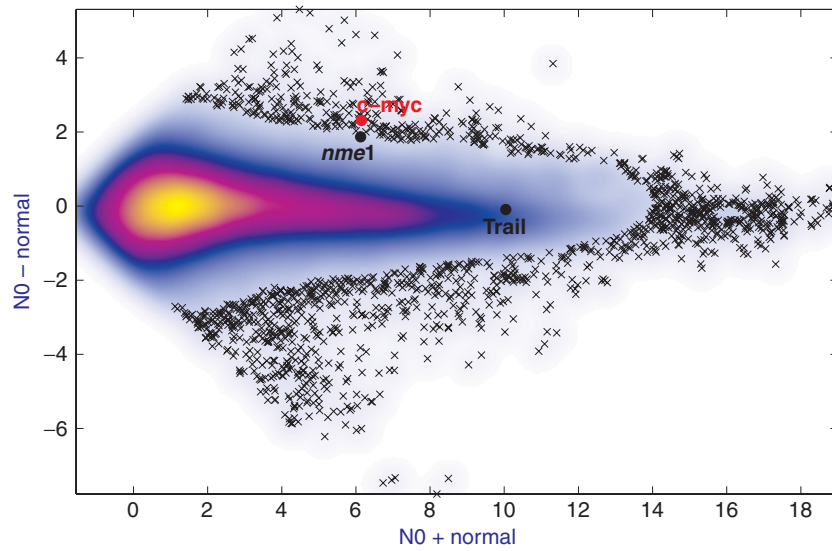
Plate 3. Scatterplot of a pairwise comparison of noncancerous colon tissue and a colorectal tumor. Each individual transcripts is represented by a ×. The *x*-coordinate is the average of the appropriately calibrated and transformed intensities (see Section 6.3). The *y*-coordinate is their difference, and is a measure of differential transcription. The array used in this experiment contained 152 000 probes representing around 70 000 different clones. Since plotting all of these would lead to an uninformative solid black blob in the centre of the plot, the point density is visualized by a color scale, and only 1500 data points in sparser regions are individually plotted.
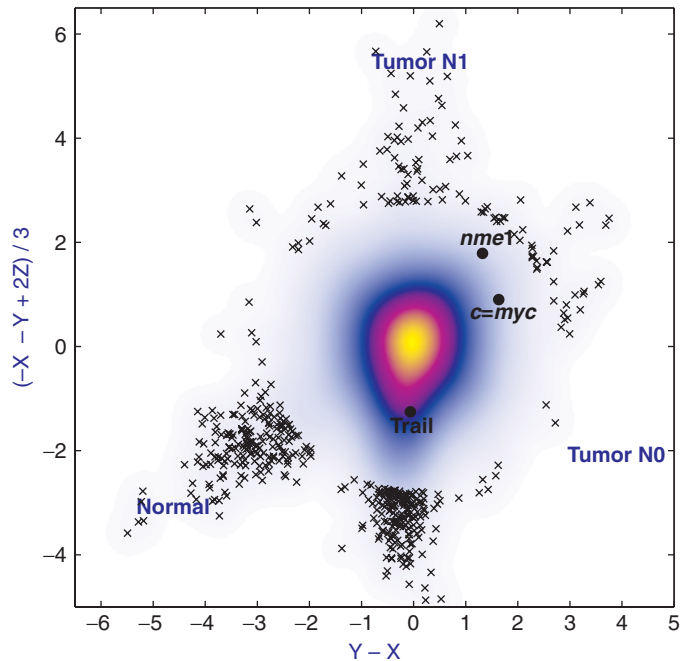
Plate 4. Scatterplot of a triple comparison between noncancerous colon tissue, a lymph-node negative colorectal tumor (N0), and a lymph-node positive tumor (N1). The measurements from each probe correspond to a point in three-dimensional space, and are projected orthogonally on a plane perpendicular to the (1,1,1)-axis. The three coordinate axes of the data space correspond to the vectors from the origin of the plot to the three labels 'normal', 'tumor N0', and 'tumor N1'. The (1,1,1)-axis corresponds to average intensity, while differences between the three tissues are represented by the position of the measurements in the two-dimensional plot plane. For instance, both *c-myc* and *nme1* (nucleoside diphosphate kinase A) are transcribed higher in the N0 and in the N1 tumor, compared to the noncancerous tissue. However, while the increase is approximately balanced for *c-myc* in the two tumors, *nme1* is more upregulated in the N1 tumor than in the N0 tumor, a behavior that is consistent with a gene involved in tumor progression. On the other hand, the apoptosis-inducing receptor *trail-r2* is downregulated specifically in the N1 tumors, while it has about the same intermediate–high transcription level in the noncancerous tissue and the N0 tumor. Similar behavior of these genes was observed over repeated experiments.
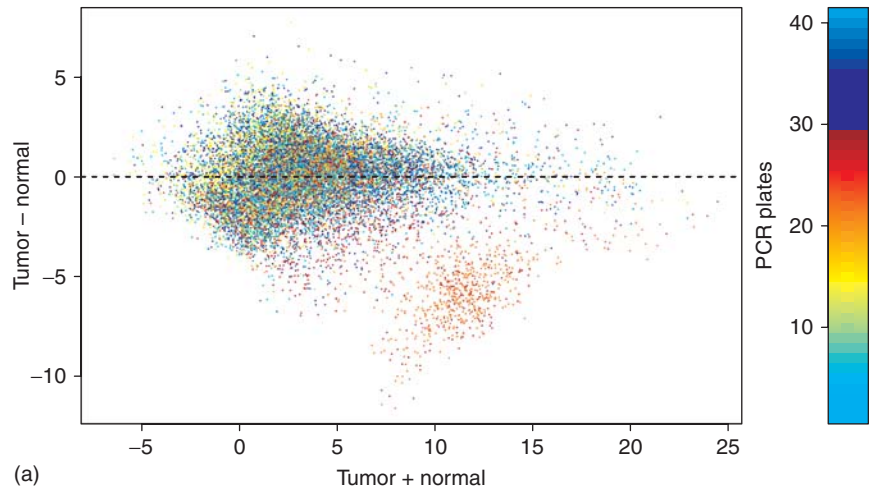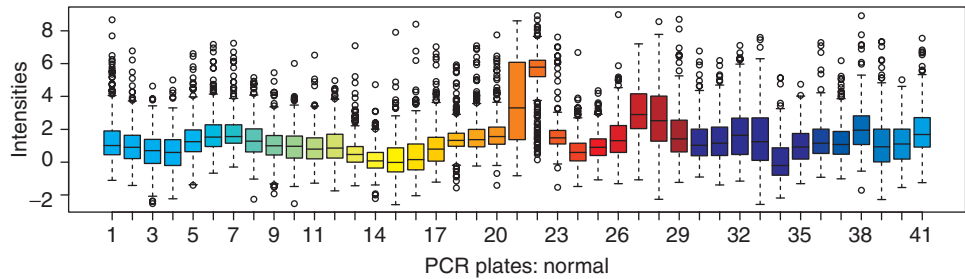
Plate 5. (a) Scatterplot of intensities from a pair of single-color cDNA arrays, comparing renal cell carcinoma to matched noncancerous kidney tissue. Similar to Plate 4, the x-coordinate represents average and the y-coordinate differential signal. At the bottom of the plot, there is a cloud of probes that appear to represent a cluster of strongly downregulated genes. However, closer scrutiny reveals that this is an experimental artifact: (b) shows the boxplots of the intensities for the two arrays, separately for each of the 41 PCR plates (see text). Probes from plates 21, 22, 27, and 28 have extraordinarily high intensities on one of the arrays, but not on the other. Since the clone selection was quasi-random, this points to a defect in the probe synthesis that affected one array, but not the other. The discovery of such artifacts may be helped by coloring the dots in the scatterplot by attributes such as PCR plate of origin or spotting pin. While the example presented here is an extreme one, caution toward batch artifacts is warranted whenever arrays from different manufacturing lots are used in a single study.
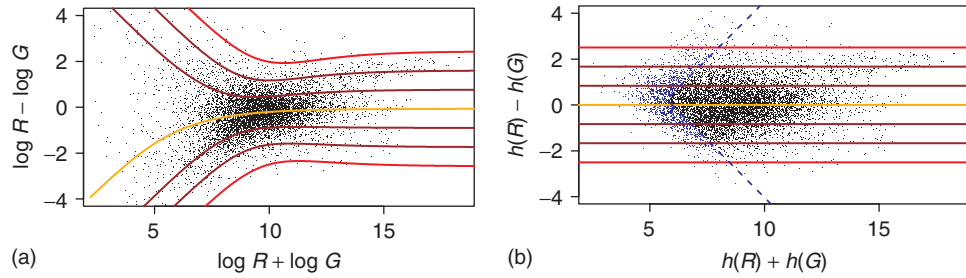
Plate 6. Scatterplot of differential versus total intensities from a two-color cDNA array, using two different transformations: (a) logarithmic; (b) equation (6.12). The horizontal lines correspond to the z-score $\Delta h/\hat{c} = 0, \pm 1, \pm 2, \pm 3$. The z-score of a pair of red and green probe intensities is their difference divided by its expected standard deviation according to the variance-versus-mean function $v(u)$. The z-score is a statistical measure of how strongly an observed pair of intensities is indicative of true differential abundance. While the contours of the z-score are functions of both log ratio and total intensity (a), they are independent of total intensity in the coordinate system of (b). Due to a local background subtraction, this data set contained small and negative net intensities. (a) shows measurements with $R, G > 0$ and $\log(RG) > 2.5$. All data is shown in (b), with the subset of the upper panel to the right of the dashed line.
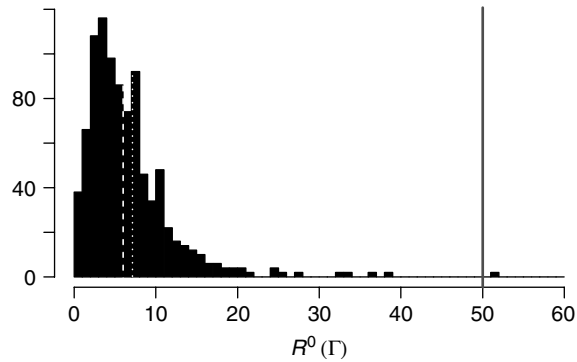
Figure 6.5 Estimation of the false discovery rate. Using 24 arrays with 32 000 cDNA probes each, 12 pairs of matched breast cancer tissue samples dissected before and after neoadjuvant chemotherapy were compared. Differentially transcribed genes were selected according to the absolute value of the one-sample t-statistic. The rejection region Γ was fixed such that 50 genes were selected (solid line). The histogram shows the distribution of $R^0(\Gamma)$, estimated from all 924 balanced sign flips. The dashed and dotted lines show median and mean respectively. The mean may be used as an estimate of $E[R^0(\Gamma)]$ in (6.17) and (6.18). Note the skewness of the distribution of $R^0(\Gamma)$.

between genes. On the other hand, this type of procedure is not able to take possibly unequal variances in the two classes into account.

6.5 PATTERN DISCOVERY

Unsupervised as well as supervised learning methods play a central role in the analysis of microarray gene expression data. Supervised methods aim to infer information from the data with respect to a predefined response variable. For instance, in the context of tumor diagnostics one tries to classify mRNA samples obtained from tumor cells with respect to given tumor types. The application of classification methods to microarray data was discussed, for example, in Golub *et al.* (1999), Ben-Dor *et al.* (2000), Dudoit *et al.* (2002a), and Spang *et al.* (2002). In the following, we focus on unsupervised methods, which aim to detect structures in the data without making use of gene or sample annotations. A primary purpose of such methods is to provide a visualization of the data in which conspicuous structures can easily be recognized. These may be relations among genes, among samples, or between genes and samples. The perception of such structures can lead the researcher to develop new hypotheses: for example, the result of a clustering of genes may indicate the putative involvement of uncharacterized genes in a biological process of interest, whereas a separation of the expression profiles of a set of patient tissue samples into clusters may point to a possible refinement of disease taxonomy. On the other hand, unsupervised methods are often used to confirm known differences between genes or samples on the level of gene expression: if a clustering algorithm groups samples from, say, two different tumor types into distinct clusters without using prior knowledge, this provides evidence that the tumor types do indeed show clearly detectable differences in their global gene expression profiles.

For all of the following methods, we assume that we have a gene expression data matrix of suitably calibrated and transformed expression levels with, say, the rows corresponding to genes and the columns corresponding to cell or tissue samples.

6.5.1 Projection Methods

An important class of unsupervised methods works through dimension reduction. The row or column vectors of a gene expression data matrix are projected onto a low-dimensional space such that some measure of similarity between the vectors is optimally preserved. The projected data may be visualized through one or more scatterplots, in the hope that these convey important information contained in the data.

In *principal component analysis*, mutually orthogonal linear combinations of the row or column vectors (the principal components) are computed, such that the ith principal component has maximal variance among all vectors orthogonal to the first $i - 1$ principal components. In applications, one may hope that the first few principal components carry most of the information contained in the data, which can then be displayed in scatterplots. Alter *et al.* (2000) demonstrated the use of principal component analysis for a gene expression study of the cell cycle in yeast. The first principal component was found to reflect experimental artifacts and was consequently filtered out. After that, the authors found that the first two principal components ('eigengenes') are well described by a sine and cosine function of time, respectively. The interpretation is that these 'eigengenes' reflect oscillating gene expression patterns, while the corresponding 'eigenarrays' define a two-dimensional coordinate system for the cell cycle phases.

In *correspondence analysis*, the rows and the columns of a nonnegative data matrix are simultaneously projected onto a low-dimensional space (Greenacre, 1984). The method decomposes the deviations from homogeneity between rows and columns, as we will now explain. For the data matrix $\mathbf{Y}$, let y_{k+} and y_{+i} be the sum of the kth row and the ith column, respectively, y_{++} the grand total, and $r_k = y_{k+}/y_{++}$ and $c_i = y_{+i}/y_{++}$ the mass of the kth row and the ith column. The matrix $\mathbf{S}$, with elements

$$s_{ki} = (y_{ki}/y_{++} - r_k c_i)/\sqrt{r_k c_i},$$

undergoes singular value decomposition, $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$. Note that the sum of the squared elements of $\mathbf{S}$ is just the χ^2-statistic of $\mathbf{Y}$. Suitably normalized columns of $\mathbf{U}$ and $\mathbf{V}$ provide the principal coordinates for the rows and the columns of $\mathbf{Y}$, respectively:

$$f_{kj} = \frac{\lambda_j u_{kj}}{\sqrt{r_k}}, \qquad g_{ij} = \frac{\lambda_j v_{ij}}{\sqrt{c_i}}.$$

Usually, only the first two or three principal coordinates are used for a simultaneous display of the rows and columns of the data matrix. The corresponding entries of $\mathbf{\Lambda}$ reflect which proportion of the χ^2-statistic of $\mathbf{Y}$ is represented in the low-dimensional projection. The distances between rows and between columns approximate their χ^2 distances, and moreover, the association between rows and columns is reflected in a biplot: a row and a column that are positively (negatively) associated will approximately lie on the same (opposite) half-ray through the origin, with the distance from the origin reflecting the strength of the association. An example is shown in Plate 7.

6.5.2 Cluster Algorithms

Cluster algorithms generally aim to group objects according to some notion of similarity. An overview of issues and methods in cluster analysis is given in Jain and Dubes (1988). For microarray data, clustering may be applied to the genes whose expression levels are measured, with the expectation that functionally related or coregulated genes will show similar expression patterns. On the other hand, one may use clustering to analyse the expression profiles of a set of cell or tissue samples with the hope that samples with similar biological characteristics will be grouped together. Cluster algorithms are explicitly or implicitly based on a quantitative measure of dissimilarity between the objects of interest. In the case of row and column vectors of a gene expression data matrix, typical examples are the Euclidean distance or 1 minus the correlation coefficient. For the clustering of genes based on their expression patterns, the latter is often used because it is invariant under affine-linear transformations of the input vectors and focuses on the pattern of relative changes.

In *hierarchical* clustering algorithms, a tree structure (dendrogram) is computed that contains the objects as leaves. *Agglomerative* methods build the tree starting with the leaves (Eisen *et al.*, 1998), whereas in *divisive* methods (Alon *et al.*, 1999) the set of objects is iteratively partitioned into subsets. To obtain a partition of the set of objects into clusters, the resulting dendrogram may be cut at a certain height. In the analysis of microarray data, however, hierarchical clustering is often simply used to obtain a linear ordering of both the rows and the columns of a gene expression data matrix such that similar rows or columns are located close to each other. The reordered data matrix is then displayed using a color map, which may be a powerful visualization tool. However, many implementations of hierarchical clustering do not try to find an optimal linear order of the n leaves of the obtained dendrogram out of the 2^{n-1} possible orders that are compatible with the tree structure. Bar-Joseph *et al.* (2001) described an efficient algorithm to compute an optimal linear order of the set of leaves of a cluster tree compatible with the tree structure in the sense that the sum of distances of pairs of neighboring leaves is minimized.

Nonhierarchical clustering algorithms directly yield a partition of the set of objects into clusters, the number of which has to be fixed in advance in most methods. Examples are k-means clustering, partitioning around medoids (Dudoit and Fridlyand, 2002) and self-organizing maps (Tamayo *et al.*, 1999). A graph-theoretical clustering method that was developed especially for gene expression data is described in Ben-Dor *et al.* (1999).

An important but difficult question in cluster analysis is that of the validity of the results. Cluster analysis assumes that the data are organized in distinguishable clusters. However, a cluster algorithm will usually also produce a set of clusters if this assumption is not fulfilled. Furthermore, the results may be affected by random fluctuations of the data. Thus, it is of interest to estimate the number of clusters present in a data set (if any), as well as to assess the variability of various features of the result.

Dudoit and Fridlyand (2002) proposed an approach for estimating the number of clusters in a data set. In the first step, they applied a clustering algorithm to a subset of all observations. Then they analyzed how much an assignment of the remaining observations to the clusters by a class prediction method coincided with the partition obtained from clustering these observations. Through comparing the resulting measure of predictability to that obtained under a null model without cluster structure, they arrived at a quality index that can be computed for different numbers of clusters. If none of the predictability values is significant, there is no evidence for clusters in the data, whereas otherwise the

number of clusters yielding the highest quality index is chosen. The performance of this approach was demonstrated using simulated data and real gene expression data, where clustering is applied to the samples. However, one might imagine that real data sometimes lie between the extreme cases of a common distribution for all objects on the one hand and distinct clusters on the other. Furthermore, if the objects are organized in hierarchically nested clusters, such that there are partitions at different levels of granularity, the question of how many clusters are present is not meaningful without further specifications.

Kerr and Churchill (2001b) used the bootstrap in order to assess the reliability of the results of a cluster analysis. Resampling was performed on the residuals of an analysis of variance model, yielding stability values for the assignment of genes to prespecified clusters. In a more general context, Pollard and van der Laan (2002) analyzed the performance of the bootstrap in assessing the variability of the results of a wide class of clustering methods. Assuming that the expression profiles of the biological samples are generated from a mixture of n-variate probability distributions, where n denotes the number of genes, they gave a general definition of clustering methods (for clustering genes or samples, or the simultaneous clustering of genes and samples) as algorithms to estimate certain parameters of the data-generating distribution. Although an underlying probabilistic model is assumed, the cluster algorithms under consideration do not have to be model-based. This framework allows concepts of statistical inference to be applied to clustering algorithms. In a simulation study, they analyzed both the nonparametric (resampling of sample expression profiles) and the parametric bootstrap (based on normal distributions). The results indicated that both bootstrapping methods are able to assess the variability of various quantities describing the output of a clustering algorithm. In addition, the authors proposed to test the statistical significance of a clustering of samples via the comparison with data generated from an appropriate null model. A similar approach for the evaluation of temporal gene expression patterns was presented in Dougherty *et al.* (2002). Assuming the expression pattern of each gene to be generated from one of several multivariate normal distributions, the authors analyzed the error rate of various clustering algorithms in determining the correct cluster membership of genes.

Probabilistic clustering methods assume that each observation belongs to a cluster k with probability π_k, and the observations within each cluster k are generated according to a probability distribution $\mathcal{L}_k$. After the number of clusters and the family of admissible probability distributions have been specified, the model parameters and the most likely cluster assignment of each observation can be estimated by maximum likelihood. This is usually done via the expectation–maximization algorithm (Dempster *et al.*, 1977), starting with some initial clustering. In such a probabilistic framework, it is possible to assess the adequacy of different models – concerning the number of clusters, as well as the allowed parameter space for the component distributions – through the Bayesian information criterion (Fraley and Raftery, 1998). The application of model-based clustering to microarray data is described in Yeung *et al.* (2001), Ghosh and Chinnaiyan (2002), and McLachlan *et al.* (2002). Concerning the clustering of genes, the application of normal mixture models, possibly with constraints on the covariance matrices in order to reduce the number of free parameters, is more or less straightforward. On the other hand, the application of model-based methods to the clustering of samples poses problems, because in typical microarray data sets the number of genes, and thus the number of parameters to be estimated, exceeds by far the number of samples. Ghosh and Chinnaiyan (2002) suggested clustering the samples via a model-based approach using

the first few components obtained from a principal component analysis. McLachlan *et al.* (2002) proposed clustering the samples into a mixture of factor analysis models.

6.5.3   Local Pattern Discovery Methods

One limitation of clustering methods as described in Section 6.5.2 lies in the fact that they are based on a global measure of similarity between the rows or the columns of the data matrix. However, there may be biologically relevant situations where tissue samples share similar expression levels of one particular set of genes, for example, those belonging to a molecular pathway that is active in this group, whereas they differ with respect to the expression of other genes. Also the similarity of the expression levels of a group of genes may be present only under certain biological conditions. We give a brief overview on methods that were developed with the aim of detecting such structures in an unsupervised fashion.

Getz *et al.* (2000) described an algorithm where hierarchical clustering is alternately applied to the rows and columns of submatrices of the data matrix, the rows and columns of which were obtained as stable clusters in previous iterations.

A number of authors have suggested methods to identify interesting submatrices of a gene expression data matrix (Califano *et al.*, 2000; Chen and Church, 2000; Ben-Dor *et al.*, 2002; Lazzeroni and Owen, 2002). The underlying idea is that a set of genes, perhaps belonging to a common molecular pathway, are coregulated only under certain experimental conditions. This notion is quantified in terms of a score function on submatrices of a gene expression matrix. As the number of submatrices is exponential in the number of genes and the number of samples, efficient heuristics are applied in order to find high-scoring submatrices. The submatrices identified can be evaluated in terms of their statistical significance.

For the identification of conspicuous class distinctions among a set of tissue samples based on microarray data, special approaches have been proposed (Ben-Dor *et al.*, 2001; von Heydebreck *et al.*, 2001). These are based on a score function that quantifies the strength of differential gene expression for any possible bipartition of the set of samples. An optimization algorithm is used in order to find high-scoring bipartitions. As the scoring is not based on global properties of the gene expression profiles, but rather on the presence of subsets of genes that are differentially expressed, several independent bipartitions can be obtained, each based on a specific subset of differentially expressed genes. von Heydebreck *et al.* (2001) show that this approach is able to detect biologically meaningful class distinctions that are not identified with cluster algorithms based on a global dissimilarity measure.

6.6   CONCLUSIONS

We have described different aspects of microarray gene expression data analysis, from quality control of the raw probe intensity data, via calibration, error modeling, and the identification of differentially transcribed genes to explorative methods such as clustering or pattern discovery. Yet, statistical analysis is only one part of a microarray experiment. Frequently, data analysis is expected to correct for technological problems or shortcomings in the design of an experiment. Awareness has grown, though, in recent years; interaction

between experimentalists and data analysts is improving and, very importantly, starting early in the planning of an experiment. This raises the hope that statistical analysis will in future be less and less diverted by troubleshooting, and can increase its focus on generating and validating biological hypotheses from the data.

Unlike with, for example, sequence data, it is still extremely difficult to relate different experiments to each other and to quantitatively compare their results. Much stricter standardization of the measurement process, which will have to include significant improvements of present technologies or development of new ones, will be necessary to obtain measurements that would be comparable across laboratories. As a result, currently each experiment has to be large enough to be analyzable by itself because it is still not feasible to view one's own experiment as an incremental addition to an existing knowledge base on gene expression. While this is both a technical problem and a data integration problem, suggestions regarding the data standardization aspect are given in Brazma *et al.* (2001).

Microarray-based experiments are frequently seen as the stronghold of hypothesis-free genome research. While debatable in itself, this assertion simply shifts the responsibility to the computational scientist analyzing the data. In the absence of a clear hypothesis much of the analysis will be of an exploratory nature. Once this leads to a hypothesis further independent verification is needed. This embeds microarray experiments and statistical analysis into a feedback cycle producing new experiments.

Acknowledgments

REFERENCES

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences (USA)* **96**, 6745–6750.

Alter, O., Brown, P. and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences (USA)* **97**, 10,101–10,106.

Baggerly, K.A., Coombes, K.R., Hess, K.R., Stivers, D.N., Abruzzo, L.V. and Zhang, W. (2001). Identifying differentially expressed genes in cDNA microarray experiments. *Journal of Computational Biology* **8**, 639–659.

Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

Bar-Joseph, Z., Gifford, D.K. and Jaakkola, T.S. (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **17**(Suppl. 1), S22–S29.

Beissbarth, T., Fellenberg, K., Brors, B., Arribas-Prat, R., Boer, J.M., Hauser, N.C., Scheideler, M., Hoheisel, J.D., Schütz, G., Poustka, A. and Vingron, M. (2000). Processing and quality control of DNA array hybridization data. *Bioinformatics* **16**, 1014–1022.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. (2000). Tissue classification with gene expression profiles. *Journal of Computational Biology* **7**, 559–583.

Ben-Dor, A., Chor, B., Karp, R. and Yakhini, Z. (2002). *Proceedings of the Sixth Annual International Conference on Computational Molecular Biology (RECOMB)*. ACM Press, New York, pp. 49–57.

Ben-Dor, A., Friedman, N. and Yakhini, Z. (2001). Class discovery in gene expression data. *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB)*. ACM Press, New York, pp. 31–38.

Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology* **6**, 281–297.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series A* **57**, 289–300.

Best, D.I. and Rayner, C.W. (1987). Welch's approximate solution for the Behrens – Fisher problem. *Technometrics* **29**, 205–220.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nature Genetics* **29**, 365–371.

Califano, A., Stolovitzky, G. and Tu, Y. (2000). In *ISMB-2000: Proceedings, Eighth International Conference on Intelligent Systems for Molecular Biology*, P. Bourne, M. Gribskov and R. Altman, *et al.*, eds. AAAI Press, Menlo Park, CA, pp. 75–85.

Chen, Y. and Church, G.M. (2000). In *ISMB-2000: Proceedings, Eighth International Conference on Intelligent Systems for Molecular Biology*, P. Bourne, M. Gribskov and R. Altman, *et al.*, eds. AAAI Press, Menlo Park, CA, pp. 93–103.

Chen, Y., Dougherty, E.R. and Bittner, M.L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* **2**, 364–374.

Churchill, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* **32**(Suppl. 2), 490–495.

Claverie, J.M. (1999). Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics* **8**, 1821–1832.

Cleveland, W.S., Grosse, E. and Shyu, W.M. (1992). *Statistical Models in S*, J.M. Chambers and T.J. Hastie, eds. Wadsworth & Brooks Cole, Pacific Grove, CA, pp. 309–376.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series A* **39**, 1–38.

Dougherty, E.R., Barrera, J., Brun, M., Kim, S., Cesar, R.M., Chen, Y., Bittner, M. and Trent, J.M. (2002). Inference from clustering with application to gene-expression microarrays. *Journal of Computational Biology* **9**, 105–126.

Dror, R.O., Murnick, J.G., Rinaldi, N.J., Marinescu, V.D., Rifkin, R.M. and Young, R.A. (2002). In *Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology (RECOMB2002)*. Association for Computing Machinery, New York, pp. 137–143.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* **3**(7), 0036.1–0036.21.

Dudoit, S., Fridlyand, J. and Speed, T.P. (2002a). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77–87.

Dudoit, S., Yang, Y.H., Speed, T.P. and Callow, M.J. (2002b). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics* **21**(Suppl. 1), 10–14.

Durbin, B., Hardin, J., Hawkins, D. and Rocke, D.M. (2002). A variance-stabilizing transformation from gene-expression microarray data. *Bioinformatics* **18**(Suppl. 1), S105–S110.

Efron, B., Tibshirani, R., Goss, V. and Chu, G. (2000). Microarrays and their use in a comparative experiment. Technical report, Stanford University. http://www-stat.stanford.edu/~tibs/research.
html.

Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J. and Vingron, M. (2001). Correspondence analysis applied to microarray data. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10781–10786.

Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal* **41**, 578–588.

Getz, G., Levine, E. and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 12079–12084.

Ghosh, D. and Chinnaiyan, A.M. (2002). Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* **18**, 275–286.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Herwig, R., Aanstad, P., Clark, M. and Lehrach, H. (2001). Statistical evaluation of differential expression on cDNA nylon arrays with replicated experiments. *Nucleic Acids Research* **29**, e117–e117.

von Heydebreck, A., Huber, W., Poustka, A. and Vingron, M. (2001). Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics* **17**(Suppl. 1), S107–S114.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**(Suppl. 1), S96–S104.

Ideker, T., Thorsson, V., Siegel, A.F. and Hood, L.E. (2000). Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology* **7**, 805–818.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. http://biosun01.biostat.jhsph.edu/~ririzarr/papers/index.html, **4**(2), 249–264.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ.

Kepler, T.B., Crosby, L. and Morgan, K.T. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology* **3**(7), 0037.1–0037.12.

Kerr, M.K. and Churchill, G.A. (2001a). Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77**, 123–128.

Kerr, M.K. and Churchill, G.A. (2001b). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 8961–8965.

Kerr, M.K., Martin, M. and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.

Lazzeroni, L. and Owen, A. (2002). Plaid models for gene expression data. *Statistica Sinica* **12**, 61–86.

Lennon, G.G. and Lehrach, H. (1991). Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics* **10**, 314–317.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. and Lockhart, D.J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics* **21**(Suppl. 1), 20–24.

Lönnstedt, I. and Speed, T.P. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

McLachlan, G.J., Bean, R.W. and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**, 413–422.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**(1), 37–52.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18**, 546–554.

Pollard, K.S. and van der Laan, M.J. (2002). Statistical inference for simultaneous clustering of gene expression data. *Mathematical Biosciences* **176**, 99–121.

Ramdas, L., Coombes, K.R., Baggerly, K., Abruzzo, L., Highsmith, W.E., Krogmann, T., Hamilton, S.R. and Zhang, W. (2001). Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biology* **2**(11), 0047.1–0047.7.

Rocke, D.M. and Durbin, B. (2001). A model for measurement error for gene expression analysis. *Journal of Computational Biology* **8**, 557–569.

Rousseuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H. and Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Research* **28**, e47.

Spang, R., Blanchette, C., Zuzan, H., Marks, J.R., Nevins, J. and West, M. (2002). Prediction and uncertainty in the analysis of gene expression profiles. *In Silico Biology* **2**, 0033.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273–3297.

Storey, J.D. (2001). The positive false discovery rate: a Bayesian interpretation and the q-value. Technical report, Department of Statistics, Stanford University. Annals of Statistics, in press. `http://www.stat.berkeley.edu/~storey/`.

Storey, J.D. and Tibshirani, R. (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical report, Department of Statistics, Stanford University. `http://www.stat.berkeley.edu/~storey/`.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2907–2912.

Theilhaber, J., Bushnell, S., Jackson, A. and Fuchs, R. (2001). Bayesian estimation of fold-changes in the analysis of gene expression: the PFOLD algorithm. *Journal of Computational Biology* **8**, 585–614.

Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, L.P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research* **11**, 1227–1236.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116–5121.

Welch, B.L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika* **34**, 28–35.

Westfall, P.H. and Young, S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York.

Yang, Y.H., Buckley, M.J., Dudoit, S. and Speed, T.P. (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics* **11**, 108–136.

Yang, Y.H. and Speed, T.P. (2002). Design issues for cDNA microarray experiments. *Nature Reviews Genetics* **3**, 579–588.

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987.

Yue, H., Eastman, P.S., Wang, B.B., Minor, J., Doctolero, M.H., Nuttall, R.L., Stack, R., Becker, J.W., Montgomery, J.R., Vainer, M. and Johnston, R. (2001). An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Research* **29**, 1–9, e41.

Statistical Inference for Microarray Studies

S.B. Pounds, C. Cheng and A. Onar

Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA

Microarrays enable investigators to simultaneously measure the expression levels of thousands of genes in a tissue sample. The technology presents exciting avenues for biological research coupled with interesting statistical challenges. Several methods to address these challenges have been proposed and applied in the literature. We review some methods developed for use in what we term *phenotype-association analyses* that aim to identify a set of genes whose expression is associated with a particular phenotype. A phenotype-association analysis typically consists of initial data processing, tests of the association of expression with phenotype, multiple-testing adjustments, annotation analysis, and validation analysis. We briefly describe some methods used for each of these stages of phenotype-association analysis and describe methods used for performing sample-size calculations in planning a microarray study. We also describe the underlying assumptions of these methods and outline some very basic considerations for choosing which methods are best suited for specific applications.

7.1 INTRODUCTION

Microarrays have opened new and exciting possibilities in biomedical research. The technology allows investigators to measure the messenger-RNA levels of thousands of genes simultaneously. More recently, microarrays have been developed that allow researchers to query genomic DNA at thousands of positions simultaneously as well.

Not surprisingly, the analysis of microarray data presents several statistical challenges including, but certainly not limited to, data reduction, normalization, correlation of variables, and multiple testing. None of these are new to statistics; however, in microarray data, the magnitude of these problems is unprecedented. These challenges provide opportunities for development of new statistical methods.

Statistical researchers have responded to these opportunities by proposing new statistical methods to address these challenges. The number of statistical methods proposed for the

analysis of microarray data has been growing exponentially over the past decade (Mehta *et al.*, 2004). In 2003, roughly 90 articles were published that describe new methods for analysis of microarray data (Mehta *et al.*, 2004). Now, there are several books that address the statistical analysis of microarray data (Draghici, 2003; Allison *et al.*, 2006; Do *et al.*, 2006; Lee, 2004; Simon *et al.*, 2003; Gentleman *et al.*, 2005; Speed, 2003; Parmigiani *et al.*, 2003).

Microarray studies often have one or more of the following objectives: class discovery, class prediction, or phenotype association. In class discovery, the objective is to use the microarray data to partition subjects or samples into biologically meaningful classes that have not been previously defined. Clearly, the idea of 'biologically meaningful' is subjective, and hence the analyses are very exploratory in nature. In class prediction, the objective is to use microarray data to accurately assign subjects or samples into a set of predefined classes. Class-prediction analyses typically rely on a variety of machine and statistical-learning methods to achieve their objective. In phenotype-association analyses, the objective is to identify which genes' (or microarray probe sets') expression is associated with a phenotype of interest. For each probe set, a phenotype-association analysis typically tests the null hypothesis that gene expression is not associated with phenotype and then corrects for multiple testing. Additionally, a phenotype-association analysis often performs a bioinformatics follow-up assessment to determine whether genes with specific annotations, such as ontology (`www.geneontology.org`) or pathway membership (`http://www.genome.jp/kegg/pathway.html`), tend to be more highly represented among the 'significant' genes than expected by chance. This chapter outlines some methods for performing phenotype-association analyses; some useful resources for those interested in class discovery or class-prediction analyses are mentioned briefly below. **Chapters 6** and **8** also review statistical methods for the analysis of microarray data.

Class-discovery analyses use clustering algorithms to assign subjects (or samples) into groups that are similar according to a specified distance metric. Class-discovery analyses are often criticized because they do not test a well-specified hypothesis, many of the clustering algorithms utilized will assign samples of any data set (even complete noise) into classes, and it is unclear how to generalize the results of a class-discovery analysis to a population (Mehta *et al.*, 2004). Despite these shortcomings, class-discovery analyses can lead to biologically important discoveries. For example, the results of a class-discovery analysis led Thompson *et al.* (2006) to identify distinct genomic lesions in pediatric medulloblastoma tumor cells. Nevertheless, it is still important that the statistical analysis attempts to address the concerns of Mehta *et al.* (2004). As mentioned above, a major concern is that many clustering algorithms will assign the subjects into groups even if the data set is completely random noise (McShane *et al.*, 2002). Clearly, in such cases, no 'meaningful' subclasses exist. McShane *et al.* (2002) propose a method to test a null hypothesis of unimodality prior to applying clustering algorithms. Rejection of unimodality indicates that the data set has multiple modes that may be attributed to the existence of two or more distinct subgroups. A class-discovery analysis should also assess the reproducibility, stability, or both, of the resulting cluster assignments (Dudoit and Fridlyand, 2003). Dudoit and Fridlyand (2003), and Smolkin and Ghosh (2003) have proposed methods that use resampling or subsampling to assess cluster reproducibility or stability.

Class-prediction analyses apply statistical-learning (Hastie *et al.*, 2001) or machine-learning (Kecman, 2001) algorithms to a 'training' data set to develop a classifier that

assigns samples into predefined classes and then evaluate the performance of the classifier on a 'test' or 'validation' data set in terms of classification accuracy. The prevalence of class-prediction analyses in the literature may give some investigators (and reviewers) the impression that class prediction is the preferred analysis approach for most microarray studies. However, this is not so. Michiels *et al.* (2005) found that each of the seven largest studies that used class-prediction analyses to examine the association of gene expression with relapse published overly optimistic estimates of prediction accuracy. Additionally, they found that five of the seven studies failed to show that their classifiers were significantly more accurate than random class assignment. Some of these overly optimistic estimates may arise because the validation and training phases of the analysis were not performed totally independent of each other (Allison *et al.*, 2005; Quackenbush, 2006). Moreover, these studies were too small to accommodate training and validation data sets of adequate size to achieve the statistical power needed to develop an accurate classifier and demonstrate that the classifier performs significantly better than chance. Nevertheless, preliminary studies indicate that classifiers based on microarray data may be useful diagnostic tools for some clinical applications. Thus, some studies with the objective of developing and validating classifiers based on microarray gene-expression data are still needed. Simon (2005a; 2005b) provides an excellent overview of how to plan and execute microarray studies with the ultimate objective of developing a useful classifier based on gene expression.

A phenotype-association analysis, the primary focus of this chapter, typically involves at least three basic steps: initial data processing (IDP), tests of the association of phenotype with expression (TAPE), and multiple-testing adjustments (MTA). Validation analysis (VA) and annotation analysis (AA) are two additional steps that are becoming common components of phenotype-association analyses. Figure 7.1 illustrates the steps of a phenotype-association analysis in a flowchart. In IDP, image signals are converted, summarized, and normalized into meaningful probe-set or gene-expression signals. Additionally, IDP may filter out probe sets that are believed to represent genes that are not expressed in any study samples or poorly represent the expression of the targeted gene. The second step, TAPE, applies a statistical procedure to describe and test the association of each probe set's expression signals with the phenotype of interest. The phenotype may be assigned experimentally (e.g. treatment vs. control) or collected in an observational study (e.g. tumor subtype in a human study). The third step, MTA, controls or estimates an appropriate multiple-testing error rate such as the false discovery rate (FDR) (Benjamini and Hochberg, 1995) or the generalized family-wise error rate (gFWER) (van der Laan *et al.*, 2004a). The fourth step, VA, explores the reproducibility of results via cross-validation, bootstrapping, or other data perturbation methods. The fifth step, AA, attempts to determine if the most promising genes found in steps 2 and 3 show a significant tendency to be annotated to specific ontologies or pathways.

There are several freely available software tools to implement each component of a phenotype-association analysis. Table 7.1 lists websites where some of these software tools can be obtained. The Bioconductor project (www.bioconductor.org; Gentleman *et al.*, 2004; 2005) represents one of the most comprehensive collections of software tools for the analysis of microarray data. Bioconductor is a collection of R software (www.r-project.org) routine libraries to implement various analysis methods. R is a freely available programming language and implementation platform that includes several functions and software libraries that are useful for statistical computations. Thus,
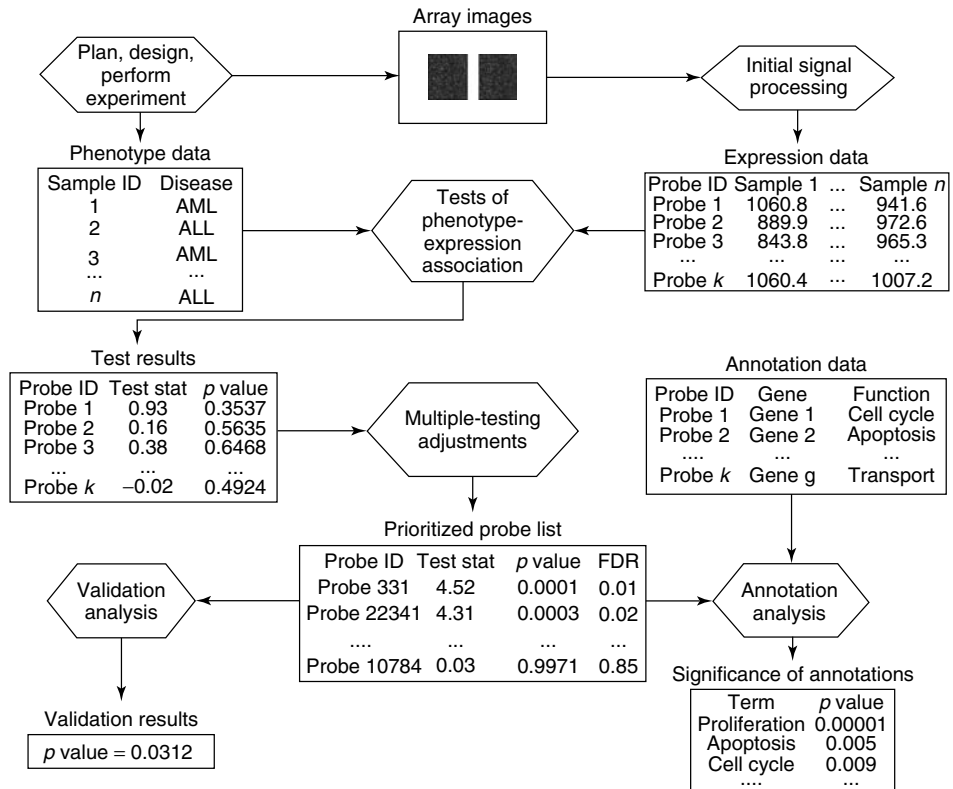
Figure 7.1   Flowchart diagram of a phenotype-association analysis.

effective use of Bioconductor requires some programming (RSP). Other software tools use web-based interfaces (WBI) or are stand-alone packages that offer a graphical user interface (GUI).

The variety of available methods presents the challenge of choosing the best method(s) for a specific application. Several comparisons of methods have been performed via simulation studies, analysis of spike-in data sets, and analysis of case-study applications (Lin and Johnson, 2003; Choe *et al.*, 2005). At present, there is little consensus regarding which methods are best for any of the five components of a phenotype-association analysis and very little work has explored which combination of methods is best for specific applications. Nevertheless, there are some key points to consider when choosing a method. First, the method should be designed to explore the type of phenotype under consideration. For example, it is inappropriate to treat a time-to-event variable, such as duration of remission from cancer or survival time, as a binary outcome without accounting for censoring and follow-up time (Jung *et al.*, 2005a). Additionally, one should consider whether the underlying assumptions of a particular method are reasonable and how robust the method is likely to be against those violations. Many methods have been proposed without explicit or thorough descriptions of their underlying assumptions. The assumptions of a specific method can be ascertained indirectly by envisioning settings that may lead the algorithm to yield misleading results. For example, one might try

Table 7.1 Websites offering freely available software for phenotype-association analysis.

No.	Web address	Available tools⋆	Usage†
1	www.r-project.org	TAPE, MTA	RSP
2	www.bioconductor.org	IDP, TAPE, MTA, VA, AA, SSC	RSP
3	www.stjuderesearch.org/depts/biostats	IDP, MTA, SSC	RSP
4	www.jax.org/staff/churchill/labsite	IDP, TAPE	RSP
5	www.dchip.org	IDP	GUI
6	www.soph.uab.edu/ssg.asp?id=1087	IDP, TAPE, MTA, SSC	WBI
7	visitor.ics.uci.edu/genex/cybert/	TAPE, MTA	WBI
8	www.geocities.com/jg_liao/software	MTA	RSP
9	bioinformatics.skcc.org/webarray/	IDP, TAPE, MTA	WBI
10	www.biostat.harvard.edu/people/ faculty/mltlee/web-front-r.html	SSC	WBI
11	www.biostat.umn.edu/weip/ge.html	SSC	RSP
12	gordonlab.wustl.edu/mills/	IDP	GUI
13	bioinf.wehi.edu.au/folders/ arrayweights/	IDP, TAPE	RSP

⋆IDP = initial data processing; TAPE = tests of the association of phenotype with expression; MTA = multiple-testing adjustment; VA = validation analysis; AA = annotation analysis; SSC = sample-size calculations.
†RSP = requires some programming; GUI = graphical user interface; WBI = web-based interface.

to envision how a method would perform if many or few genes were associated with the phenotype, how it might be affected by outliers, how it may perform under various dependency structures, and how it would manage potential heteroskedasticity. Hopefully, as more comparisons of methods are performed, the conditions under which various methods perform reliably will become more thoroughly understood and more specific recommendations will be possible. In this chapter, we review some methods that have been proposed for phenotype-association analyses. We acknowledge that the review is not exhaustive; an exhaustive review is not feasible in this venue given the vast amount of recent research activity in the area.

7.2 INITIAL DATA PROCESSING

At the most fundamental level, microarray data are image files of the florescent signals on the arrays. In their initial state, these image files are not useful for performing statistical analyses to address questions of practical interest. The information in these image files must be summarized into probe-set expression signals that can be used to perform biologically meaningful comparisons across arrays. A data file with information regarding the location of probe sets on the array is used in conjunction with each image file to produce a file that gives a summarized, florescent-intensity signal for each probe set that targets a known or hypothetical gene. Technically speaking, probes are individual features printed on the array, collections of probes are called probe sets, and a probe set is intended to measure the expression of a specific gene. However, in this chapter, we use the terms *probe*, *probe set*, and *gene* interchangeably. A number of purely technical factors can influence these summarized signals. It is often necessary to adjust the signals for these

technical factors before statistical analyses can meaningfully explore the association of these signals with phenotype. The term *normalization* is used to describe data processing that intends to transform the summarized intensity signals into expression signals that are biologically comparable across arrays. Additionally, sometimes probe sets that are deemed as unreliable or believed to represent genes that are unexpressed across all samples are excluded prior to subsequent analysis. The term *filtering* is used to describe data processing that attempts to identify probe sets that should be excluded from statistical analysis.

In this section, we briefly review some normalization and filtering methods. Table 7.2 lists some methods for normalization and filtering. More detailed reviews on normalization are given by **Chapter 6**, Quackenbush (2002), Reimers (2005), Eckel-Passow *et al.* (2005), and Steinhoff and Vingron (2006).

7.2.1   Normalization

A wide variety of methods to normalize microarray signals have been proposed. From a general perspective, the methods fall into one of a few categories: scaling, quantile normalization, local regression (loess) normalization, and model-based normalization. Scaling methods, such as the normalization algorithm of Microarray Analysis Software, version 5.0 (MAS 5.0, `www.affymetrix.com`), assume that each array should have a similar mean or median signal across probe sets. Quantile-normalization methods assume that each array's probe-set signals should have the same distribution. Model-based methods make assumptions typically used in most general linear models. Loess normalization methods assume that the technical biases are intensity dependent, and fit loess curves to identify and remove those biases.

Table 7.2   Methods for initial processing of microarray data.

Method name or description	Reference(s)	Software[*]
Correction for PM/MM and background	Irizarry *et al.* (2003a)	2
	Bolstad *et al.* (2003)	
Robust multiarray averages (RMA)	Irizarry *et al.* (2003b)	2
	Bolstad *et al.* (2003)	
Preprocessing for two-dye arrays	Irizarry *et al.* (2003b)	2
Sequence-based preprocessing (GCRMA)	Wu *et al.* (2004)	2
Variance stabilization	Huber *et al.* (2002)	2
Normalization by joint modeling of PM/MM signals	Li and Wong (2001)	2
ANOVA-based normalization	Kerr *et al.* (2000)	4
	Wolfinger *et al.* (2001)	
	Cheng *et al.* (2004, Appendix)	
Error model, local regression	Kepler *et al.* (2002)	NA
Quantile normalization	Bolstad *et al.* (2003)	2
Quantile spline normalization	Workman *et al.* (2002)	2
Present/absent *p*-value filter	Pounds and Cheng (2005a)	3
Filter inconsistencies across	Mills and Gordon (2001)	12
technical replicates	Mills *et al.* (2001)	12
Hierarchical quality indices	Hu *et al.* (2006)	NA
Variance-based reweighting	Ritchie *et al.* (2006)	13

[*]Table 7.1 gives full website addresses for software.
NA = not available.
[*Source*: Adapted from Steinhoff and Vingron (2006), used with permission.]

Model-based normalization methods attempt to explicitly account for specific technical sources of variation in the array signals. For example, Wolfinger *et al.* (2001) proposes that the observed initial signals be the dependent variable in a linear model that includes terms for technical factors such as dye, technician, array, spots, pens, and so on, as predictor variables. The residuals from this model are then used as the normalized signals in subsequent analyses. The approach offers essentially the same flexibility as general linear models (Graybill, 1976). Thus, in principle, linear models could be used to normalize data from oligonucleotide arrays or two-color arrays (see **Chapter 6** for more information on the various array platforms). Kerr *et al.* (2000) and Cheng *et al.* (2004, Appendix) also describe normalization methods that utilize linear modeling. Of course, such model-based normalization methods also rely on all the classical assumptions of general linear models, namely that residuals are independent and identically distributed (i.i.d.) normal variables with equal unknown variance. In practice, it may be difficult to determine which specific technical factors should be included in the first model step. Furthermore, the data regarding technical sources of variation may not have been recorded by the laboratory technician or the investigator.

Li and Wong (2001) and Irizarry *et al.* (2003b) propose model-based normalization methods for oligonucleotide arrays. Their models include terms specific to the array platform (e.g. probe effects, etc.), but do not include terms for specific sources of technical variation associated with sample preparation and processing. Thus, the detailed information regarding sample processing and preparation are not required for the algorithms to compute normalized signals. Implicitly, these normalization methods attempt to account for such factors by including an array effect in the model.

Bolstad *et al.* (2003) proposed the quantile-normalization method. Quantile normalization first sorts array signals separately within each array. Next, it computes the average signal for each signal rank. More explicitly, the average of the maximum signal is computed, and then the average of the second largest signal is computed, and so on. Next, for each array, the normalized signals of each probe set are assigned to be the average signals for its rank within the array. For example, the probe set with the highest initial signal is assigned the average of the maximum signals across arrays. Thus, after normalization, a plot of the quantiles of signals from one array against the quantiles of signals from any other array would fall along the line $y = x$.

Yang *et al.* (2002) observed that often the simple scaling methods could not properly normalize two-color cDNA microarray data since the log-transformed signal intensity ratios tended to be intensity dependent. This phenomenon was most prevalent for low intensity spots. Thus they devised an approach based on locally weighted linear regression (lowess) for normalization. This method utilizes what is known as a *ratio-intensity plot* (RI plot) or as a *minus-add plot* (MA-plot), where log ratios of expressions between the two dyes are plotted on the y-axis against log-intensity (average log expression for the two dyes) values on the x-axis. A lowess curve is then fitted to the data which is subsequently used as the basis for normalization by subtracting the best fit from the observed log ratio (see **Chapter 6** for details and examples).

Clearly, normalization methods can have nontrivial effects on the subsequent statistical analyses because the normalized signals are the input data. Thus, any artifacts introduced by the normalization procedure are carried forward into later analyses; as the old adage says, 'garbage in – garbage out'. Hoffmann *et al.* (2002), Irizarry *et al.* (2003b), Choe *et al.* (2005), and Harr and Schlotterer (2006) have explored the impact of normalization

on the inferences of the final statistical analysis. In many cases, the effects are not negligible. At present, there does not appear to be a method that is clearly a good performer across a wide range of data sets. In the analysis of their spike-in experiment, Choe *et al.* (2005) found that lowess renormalization of MAS 5.0-normalized signals gave the most accurate final statistical inferences. In the analysis of the Affymetrix Latin Square spike-in experiment, Irizarry *et al.* (2003b) noted that their normalization methods led to more accurate inferences than did normalization by the Affymetrix algorithm. As mentioned in the introduction, the selection of a normalization method should be guided by considering method with the assumptions that are most reasonable for a specific application.

7.2.2   Filtering

After the microarray signals have been normalized, it is a common practice to remove the data for some probe sets prior to statistical analysis. The term *filtering* is used to refer to this process of excluding selected probe sets from the statistical analysis. A *filter* is a specific filtering method. There are several reasons for which filtering may be useful in practice. The technical aspects of handling and processing samples and arrays may introduce variation or bias so that the probe set does not reliably measure the expression of its intended target gene. Additionally, for some applications, it is expected that many of the probe sets are unexpressed across all experimental conditions. Furthermore, it is difficult to verify differential expression of small magnitudes by polymerase chain reaction (PCR) in follow-up laboratory studies. If these unexpressed, unreliable, or invariant probe sets can be accurately identified, then it seems reasonable to exclude them from the primary analysis so that they do not clutter the final results.

Several filtering methods are available. Some are widely used even though their statistical properties have not been thoroughly explored. A common filter is to remove probe sets with a foreground to background ratio that is lower than a specified threshold. In the analysis of Affymetrix data, it is common to remove probe sets that are called *present* in less than a specified number of samples. Other filters remove probe sets with small mean or variance in expression across all samples. In practice, several filters may be used simultaneously (see e.g. Lampron *et al.*, 2006; Tan *et al.*, 2005).

Ideally, technical replicates will be available to provide useful information to identify unreliable probe sets. Mills *et al.* (2001), Yang *et al.* (2002), and Wang *et al.* (2003) describe methods that use the information in technical replicates for purposes of filtering. The method proposed by Mills *et al.* (2001) utilizes duplicate cRNA generated from a single preparation of sample RNA which is independently hybridized to a pair of chips. These two chips are then compared to each other and genes with expression levels declared as 'increased' or 'decreased' are defined as false positives. A three-dimensional plot is then used to plot the signal intensities of the two chips along with the prevalence of each intensity value combination from the two chips as measured by the fraction of the total number of false positives. For details, see Mills and Gordon (2001). The location of the false positive signals as identified by a grid system is subsequently used to rank the combination of signal intensities based on their likelihood of being false positives. Based on this ranking, lookup tables are constructed, which are used to filter noise from the biologically distinct RNAs. The authors report that their approach eliminated 90 % of false positives in the cases they investigated and that their approach was more reproducible than fold-change based filtering.

For some experiments, technical replication is too expensive or infeasible. Statistical filtering methods have been developed for experiments that do not include technical replication as well. For applications involving Affymetrix arrays, Pounds and Cheng (2005a) proposed two filters that utilize the information in the p-values that are used to make the present/absent calls. The method assumes the present/absent p-values from different samples for an unexpressed or 'absent' probe set are i.i.d. uniform (0,1) observations. Following Fisher (1932), they summarize the present/absent p-values for a probe set across samples into a single, pooled p-value by comparing the negative sum of the log-transformed p-values to a gamma distribution. The *pooled p-value filter* includes all probe sets with a pooled p-value less than a specified α threshold; whereas the *error-minimizing pooled p-value filter* includes all probe sets with a pooled p-value less than the threshold α_{MTE} that minimizes an estimate of the total error risk (Genovese and Wasserman, 2002; Cheng *et al.*, 2004). Additionally, they show that the α-level pooled p-value filter corresponds to the uniformly most powerful test under a β model for the individual present/absent p-values. Furthermore, in simulation studies, they observe that their filters are superior to the widely used filters based on the number of present calls seen across samples.

The rationale for filtering seems reasonable; however, its value in practice is questionable. In their example application, Pounds and Cheng (2005a) observed that the final FDR estimate for the 100 most significant probe sets was greater when their statistically-derived filters were used than when no filter was applied. Similarly, Larsson *et al.* (2005) noted that using a fold-change based filter before running the statistical analysis influences the number of genes that are reported as significant in an unpredictable manner, changes the q values assigned to genes, and affects the results of gene ontology (GO) classifications which may be utilized in an effort to understand the role of the molecular processes in the biological system under study. Therefore, filtering methods should be applied very judiciously.

The intent of filtering is to reduce or eliminate the influence of probe sets that are unexpressed or that measure the expression of their target gene poorly. Recently, several methods have been developed that adhere to this intent without resorting to actually excluding probe sets. Hu *et al.* (2006) introduced hierarchical quality indices that measure the reliability of each probe set, and are used as weights in subsequent test-statistic and p-value calculations. The test statistic for each gene is multiplied by its quality measure and the associated p-value is based on a t-distribution whose degrees of freedom are estimated by Satterthwaite's approximation. The authors consider a variety of approaches for the quality weights which include criteria based on present/absent calls as well as detection p-values. They report that quality weights based on detection p-values generally perform better than weights based on present/absent calls but note that the performance of the various weighting statistics in the presence of MTA appears to be influenced by the magnitude of the treatment effect.

Ritchie *et al.* (2006) proposed a similar approach that uses a heteroskedastic linear model with shared variance terms to assign weights to each microarray based on a residual maximum likelihood (REML) method which requires replicated data at the array level. In addition to the full scale REML, the authors also provide a faster and more computationally efficient method, which updates the weights gene-by-gene in a single pass through the data to estimate the array variances. The array variances are then used as inverse weights in subsequent calculations to detect differential expression. With respect

to detecting differential expression, their results based on both simulations and case studies indicate that their approach is superior to treating all arrays as the same or to removing the worst-quality arrays from further consideration.

7.3   TESTING THE ASSOCIATION OF PHENOTYPE WITH EXPRESSION

The primary objective of a phenotype-association analysis is to identify which probe sets' expressions are associated with the phenotype of interest. A straightforward approach to phenotype-association analysis would be to apply an appropriate classical statistical method to the expression data for each probe set and then adjust for multiple testing. For example, if the phenotype is characterized by membership in one of two groups, then one could apply the two-sample t-test or the rank-sum test to the expression data for each probe set. This approach has a number of advantages. It is conceptually and computationally easy to implement: the data analyst needs only to determine and apply (repeatedly) an appropriate classical method to explore the association of a continuous variable with the phenotype of interest. Most classical methods are implemented in a variety of software packages, including the freely available R software.

Additionally, several methods have been developed specifically for the analysis of microarray data. Many of these more recently developed strategies attempt to borrow information across genes in a way that improves variance estimation or statistical power. As mentioned in the introduction, some of these newer methods have been proposed without a thorough expository on the underlying assumptions. Certainly, information borrowing assumes some form of exchangeability across genes or there are groups of genes that have similar statistical parameters, such as variance of expression that can be identified and pooled to improve estimation of those parameters. The validity of these types of assumptions is difficult to assess for any specific application. However, the inability to assess the validity of assumptions should not necessarily be considered a reason to avoid using a method. It is doubtful that the assumptions of any method hold for all the genes; rather, it is important to consider whether the assumptions hold for most genes and how robust the method is to violations that are likely to occur. In this section, we mention classical methods and briefly review some recently developed methods for some of the most common kinds of analyses, namely two-group comparisons, k-group comparisons, and association with quantitative or time-to-event phenotypes.

7.3.1   Two-group and k-group Comparisons

One of the most common goals for a microarray experiment is to compare expression between two groups. For example, one may wish to assess the effects of a treatment on gene expression by comparing a treated group to an untreated group. Clearly, one approach to the two-group comparison would be to apply a classical method, such as the t-test or the rank-sum test, to the expression of each probe set and then adjust for multiple testing. In recent years, many other methods have been developed specifically for applications involving the analysis of microarray gene-expression data. Most of these newer methods borrow information across genes to modify the t-statistics or their Bayesian analogs. A few of these methods are briefly described below; Pan (2002), and Cui and Churchill (2003b)

Table 7.3 Some methods for phenotype-expression association testing.

Method name and purpose	Reference(s)	Software*	Comments
Local pooled error test Two-group comparison	Jain *et al.* (2003)	2	Assumes genes with similar mean expression have equal variance of expression
CyberT Two-group comparison	Baldi and Long (2001)	7	Uses hierarchical Bayesian model to compute a revised variance estimate for *t*-stat.
Shrinkage estimator K-group comparison	Cui *et al.* (2005)	4	Proposes a shrinkage estimate for error variance and compares it to other existing methods
Rank-based correlation Time-to-event Phenotype	Jung *et al.* (2005a) Jung *et al.* (2005b)	Available from author	Uses a rank-based correlation stat for censored survival times; null dist. derived by permutation
Prop. hazards model Time-to-event phenotype	Cox (1972)	1	Semi-parametric survival time models with/without covariates
Competing risks model Time-to-event phenotype	Fine and Gray (1999)	1	

* See Table 7.1 for web addresses.
[*Source*: Adapted from Steinhoff and Vingron (2006), used with permission.]

provide more specific reviews of additional procedures. Table 7.3 provides software availability information for some of the methods we review in this section.

Baldi and Long (2001) describe a method that computes Bayesian point estimates of variance for each group and then substitutes those estimates into the usual two-sample *t*-statistic. The Bayesian model used to obtain the variance estimates assumes that the mean and variance of each gene's expression arise from a common joint prior distribution. The method can be implemented in a fully Bayesian manner or in an empirical Bayesian manner by using the data to estimate values for hyperparameters.

Jain *et al.* (2003) proposed a local pooled error test that combines information across genes to compute a variance estimate for a statistic similar to the two-sample *t*-statistic. For each pair of arrays testing the same condition, the mean and difference in log signals are computed for each probe set. Then, the difference in log intensity is plotted against the mean log intensity. A nonparametric regression procedure is used to fit a curve estimating the expected variance in log intensity as a function of the mean log intensity. Then, for each gene, a *z*-statistic is computed as the ratio of the difference of two groups' medians to the square root of the sum of the two groups' standard error estimates. The standard error estimates are obtained from the results of the nonparametric regression. Clearly, genes with similar median signals are assumed to have the same signal variance. Under this assumption, the variance estimate obtained from the nonparametric regression should be more precise than that obtained for each gene separately.

Many times, a microarray study is conducted to compare expression levels across several groups. One objective of such a study would be to identify genes that have unequal mean or median expression across the groups. For example, Pounds and Morris

(2003) compared expressions across four treatment groups. As before, the usual classical methods such as one-way ANOVA or the Kruskal–Wallis test could be applied separately for each gene followed by MTA. As with the two-group comparison, a number of methods have been developed specifically for the analysis of microarray data. We briefly describe one of these methods below.

Cui *et al.* (2005) derived a shrinkage estimator for gene-specific variance that pools across all genes. The purpose of the shrinkage estimator is to avoid having large test statistics due to underestimation of the variance that is likely to occur in applications with small sample sizes. They subsequently use this shrinkage estimator as the denominator of an F-like statistic and assess the significance via permutation. They show that their shrinkage estimator converges to the gene-specific variance estimator as the sample size increases. Thus, in some sense, their procedure uses shrinkage only if the sample size is small enough to warrant it. In several simulation studies, they observe that their method has greater power than several other approaches based on F-statistics.

As with other statistical methods, these analytical methods are based on some underlying assumptions. These assumptions are not always explicitly stated in the literature. Certainly, information-borrowing strategies assume that data can be pooled across genes to improve the variance estimates for t-type statistics for all genes. Thus, there is an implicit assumption of some type of exchangeability across all genes or subsets of genes. Furthermore, any specific method that employs information borrowing implicitly assumes that its particular technique for information borrowing improves the variance estimates for most or all genes. These types of assumptions are sometimes difficult to formulate mathematically. Therefore, it is often impossible to statistically test the validity of such assumptions in a data analysis. Furthermore, it is difficult, if not impossible, to biologically assess the validity of such assumptions. In practice, these assumptions can make a considerable impact on the ordering of genes by significance. It may be useful to explore the data with several methods and assess the consistency across analyses. Furthermore, the conclusions drawn from a microarray study should be confirmed in follow-up laboratory studies.

7.3.2   Association with a Quantitative Phenotype

Another possible objective of a microarray study is to explore the association of gene expression with a quantitative phenotype. For example, the diagnostic white blood count is a measure of disease burden and known to be prognostically relevant among pediatric acute myeloid leukemia patients (Ribeiro *et al.*, 2005). Therefore, it may be interesting to study the association of gene expression in tumor cells with white blood count at diagnosis. As before, phenotype-association analyses can be performed by applying a classical procedure for each gene separately and then adjusting for multiple testing. For instance, Spearman's or Pearson's correlation coefficients and corresponding hypothesis testing procedures may be used to examine the association of each gene's expression with the quantitative phenotype. Regression methods may be used to explore the association of expression after adjustment for other covariates.

7.3.3   Association with a Time-to-event Endpoint

One objective of collecting microarray gene-expression data in a clinical study is to explore the association of gene expression with a measure of outcome. Many times,

outcome is measured in terms of a survival-type endpoint. For example, Ross *et al.* (2004) and Bullinger *et al.* (2004) explored the association of the duration of remission from acute myeloid leukemia with the gene expression of the disease cells at diagnosis. Several methods to examine the association of expression with survival have been described (Shoemaker and Lin, 2005). Unfortunately, many of these methods may be subject to severe biases because they treat an outcome such as relapse as a binary variable (remission vs relapse) rather than as a time-to-event variable that may be censored (Jung *et al.*, 2005b). In other words, it is important that the analyses account for the possibility that some patients may not have been observed for a sufficiently long period for a relapse to occur.

This subsection describes three approaches to explore the association of gene expression with a survival endpoint that accounts for follow up and censoring. The first approach is to fit a separate, proportional, hazards regression model (Cox, 1972) for each gene to describe and test the association of expression with survival and then adjust for multiple testing. The second approach uses a nonparametric method developed by Jung *et al.* (2005a; 2005b). Finally, the third approach accommodates various types of failure for time-to-failure phenotypes (Gray, 1988).

Morris *et al.* (2005) examined the association of gene expression in tumor cells taken at diagnosis with the survival of lung cancer patients. They fit a separate, proportional, hazards model for each microarray probe set to test the association of expression with survival. The fitted models included terms for previously known factors of prognostic relevance so that the analysis would identify genes that may provide prognostic information in addition to that available from the known factors. Thus, they obtained a p-value for each probe set and fit a β-uniform mixture model (Pounds and Morris, 2003) to the p-values to estimate the FDR (Benjamini and Hochberg, 1995). Using this approach, at an estimated FDR of 20 %, they identified 26 genes that provide prognostic information beyond that available from previously known factors.

Jung *et al.* (2005b) used their nonparametric method (Jung *et al.*, 2005a) to explore the association of gene expression in diagnostic tumor samples with the survival time of lung cancer patients. They defined a rank-based correlation coefficient that measures the association of a single continuous variable with a censored time-to-event variable. For each gene, they computed this correlation coefficient to measure the association of expression with survival time. Then, they computed these statistics for 10 000 permuted data sets in which the gene-expression profile was paired at random to a set of follow-up data. Controlling the gFWER at 0.10, they identified two genes associated with survival. Their analysis did not account for the influence of previously known prognostic factors.

In some applications, it may be important to distinguish between various types of failure. For example, treatment for leukemia can fail because of death due to therapy-related complications, failure to achieve remission, relapse of the leukemia, or (rarely) development of additional therapy-related malignancies (Bogni *et al.*, 2006). A particular analysis may want to focus on a specific type of failure, such as the development of a secondary malignancy. Clearly, a therapy-related malignancy cannot be observed in any patient who dies of complications during the early phases of therapy. In this situation, there is no opportunity to observe the time-to-development of a therapy-related malignancy. Thus, death due to complications of remission–induction therapy is a *competing risk* for the development of a secondary therapy-related malignancy. See Gray (1988), Fine and Gray (1999), and Kalbfleisch and Prentice (2002) for more details on the theory and

methods to account for competing risks in the analysis of time-to-event data. See Bogni *et al.* (2006) for an example of how to use the methods of Gray (1988) and Fine and Gray (1999) in exploring the association of gene-expression data with the time-to-development of therapy-related leukemia in the context of competing risks.

Each of the three approaches described above has its strengths and limitations. Proportional hazards models that offer the flexibility to include known prognostic factors are readily implemented in standard statistical packages (e.g. R, S-plus, SAS), and are already familiar to readers in the academic medical community (Peto *et al.*, 1976; 1977). However, the results may be biased because it is quite likely that some genes may violate the model assumptions. Thus, it may be necessary to assess significance via permutation or bootstrap methods. In their application, Morris *et al.* (2005) computed p-values in three different ways (likelihood ratio test, permutation, and bootstrap) and found that the results from the three approaches were qualitatively very similar. Nevertheless, this may not be the case in all applications. The rank-based method of Jung *et al.* (2005a) does not rely on the proportional hazards assumption but does not adjust for other known variables. The rank-based method also offers several practical computational advantages over proportional hazards modeling. The method of Fine and Gray (1999) is likelihood based. In our experience, the model-based approaches can be computationally difficult because the rank ordering of the expression of some genes relative to that of the follow-up times leads to a likelihood that is monotone in one or more parameters. For this monotone likelihood, the maximum likelihood estimates are undefined (Bryson and Johnson, 1981), and software packages terminate calculations after reaching a maximum number of iterations for optimization routines.

7.3.4   Computing p Values

For any given statistical method or test statistic, there are often several ways to compute a p-value. For example, a p-value for the two-sample t-statistic may be computed via permutation, exact methods, or by comparison with the t-distribution having the appropriate degrees of freedom. In the context of microarray data analysis, there are thousands of such t-statistics, one per gene. Considering the possibility of combining information across genes, there are several more ways to compute p-values than the classical methods.

Reiner *et al.* (2003) proposed pooling across genes to compute permutation-based p-values. Briefly, under the assumption that the test statistic for each gene has (approximately) the same null distribution, the p-value for a particular gene is computed by comparing its observed test statistic to all genes' test statistics obtained by permuting the assignments of phenotype to gene-expression profiles. This approach allows one to compute p-values with high resolution using relatively few permutations. The principle can be extended to exact testing when the number of possible assignments is very small.

Many analysts choose permutation methods to compute p-values. Certainly, permutation methods are appealing because they are robust and often do not rely as heavily on distributional assumptions, unlike comparison to a mathematically derived null distribution such as the t-distribution. However, some analysts mistakenly believe that permutation methods do not require *any* assumptions and thus advocate the use of permutation in *all* applications. It is important to recall that permutation methods are based on the assumption of exchangeability under the null hypothesis and this assumption does not hold in all

applications. For example, a permutation approach is an invalid way to compare two groups' means when the variances are unequal (Romano, 1990). Additionally, Huang *et al.* (2006) caution that permutation testing can have inflated Type I error rates in the analysis of microarray data because the assumption of equal variance is often violated.

7.4 MULTIPLE TESTING

After IDP, and testing the association of phenotype with each probe-set's expression, there is a test statistic (and usually a *p*-value) for each probe set. At this stage in the analysis, it is important to adjust for multiple testing in interpreting the results. Although microarray gene-expression studies are exploratory in nature, it is still desirable to control or estimate the prevalence of Type I errors among a set of results considered 'significant'. Follow-up and confirmatory experiments can be costly and time-consuming; therefore, it is advisable to keep the Type I error rate low. Clearly, multiple-testing issues are intrinsic to the analysis of microarray data. Dudoit *et al.* (2003) and Ge *et al.* (2003) provide extensive reviews of several approaches to adjust for multiple testing in the analysis of microarray data. Pounds (2006) reviews some methods that use *p*-values to estimate or control the FDR (Benjamini and Hochberg, 1995) and related error rates. Below, we describe some methods that estimate or control the gFWER, the FDR, and other significance criteria. Table 7.4 lists some of these methods and indicates where the related software can be obtained.

7.4.1 Family-wise Error Rate

Traditional approaches which adjust for multiple-testing attempt to control the gFWER. The gFWER is the probability that one or more Type I errors occur in the analysis. These approaches were originally developed for applications that performed a very small number of tests relative to the number of tests performed in the analysis of microarray data. gFWER methods are typically very conservative for microarray applications. For example, it is essentially impossible to have any rejections when using the well-known Bonferroni correction in the context of thousands of tests.

The (gFWER) has been proposed as a more practical metric of Type I errors for the analysis of microarray data (van der Laan *et al.*, 2004a). The gFWER is defined as the probability that *k* or more Type I errors occur in the analysis, where *k* is determined by the investigator. Clearly, the gFWER is a less stringent measure of the incidence of Type I errors and thus leads to increased statistical power. The theory and methods associated with gFWER have been developed by van der Laan *et al.* (2004a; 2004b) and Dudoit *et al.* (2004). Briefly, they proposed bootstrap-based approaches to control the gFWER. Further details can be found in their articles.

7.4.2 The False Discovery Rate

The FDR (Benjamini and Hochberg, 1995) is a widely used measure of statistical significance that accounts for multiple testing in the analysis of microarray gene-expression data (Storey and Tibshirani, 2003b). Benjamini and Hochberg (1995) defined the FDR as the expected value of the ratio Q of the number V of Type I errors to the

Table 7.4 Methods to adjust for multiple testing.

Acronym	Method name or description	Reference(s)	Software*
BH95	FDR-control procedure	Benjamini and Hochberg (1995)	2
BH00	Adaptive FDR control	Benjamini and Hochberg (2000)	2
YB99	Resampling-based FDR control under dependency	Yekutieli and Benjamini (1999)	2
BY01	FDR Control under any dependency structure	Benjamini and Yekutieli (2001)	2,3
St02	Estimation of pFDR Calculation of q value	Storey (2002)	2,3
Al02	FDR estimation via elaborate p-value modeling	Allison *et al.* (2002)	6
PM03	Fit a β-uniform mixture to p-values to estimate FDR	Pounds and Morris (2003)	3
PC04	Estimate FDR by smoothing the p-value histogram	Pounds and Cheng (2004)	3
L04	Estimate local FDR p-value modeling (L04)	Liao *et al.* (2004)	8
Ch04	Estimate significance criteria	Cheng *et al.* (2004)	3
G05	FDR Control for discrete p-values	Gilbert (2005)	NA
PC06	FDR estimation for discrete p-values or one-sided tests	Pounds and Cheng (2006)	3

* See Table 7.1 for full web addresses.
[*Source*: Adapted from Pounds (2006), used with permission.]

number R of rejections (where $Q \equiv 0$ for $R = 0$). They also developed a simple procedure that operates on a set of m p-values to control the FDR at a prespecified value τ. In the phenotype-association approach to microarray data analysis, m is the number of features included in the analysis after initial processing and filtering. First, the procedure places the m p-values in ascending order, i.e. $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$. Next, the procedure computes

$$r_{(i)} = \frac{p_{(i)}}{i/m}, \qquad (7.1)$$

and

$$q_{(i)} = \min_{j \geq i} r_{(j)}, \qquad (7.2)$$

for $i = 1, 2, \ldots, m$. Each null hypothesis (i) with $q_{(i)} \leq \tau$ is rejected.

Under the assumption that p-values testing true null hypotheses are i.i.d. uniform $(0,1)$ observations, Benjamini and Hochberg (1995) prove that their procedure controls the FDR at $m_0 \tau / m$, where m_0 is the number of tests with a true null hypothesis. In our applications, m_0 would be the number of microarray probe sets that do not exhibit an expression-phenotype association. This observation prompted many groups to develop methods that incorporate an estimate of the null proportion $\pi_0 = m_0/m$ in hopes of improving statistical power while maintaining the actual FDR control at or below the nominal level τ. Pounds (2006) provides a review of some of those methods.

Storey (2002) argued that the positive FDR, the expected value of the ratio Q given that $R > 0$, is a more reasonable error rate than the FDR in exploratory settings such as the analysis of microarray gene-expression data. Subsequently, he developed a procedure that computes a quantity he calls the q value that is useful for controlling the positive false discovery rate (pFDR) at a at a prespecified level in applications with at least one false null hypothesis. Storey's (2002) procedure is identical to the procedure of Benjamini and Hochberg (1995) except that it incorporates an estimate of the null proportion π_0 into the numerator of $r_{(i)}$ for each i and is developed to control the pFDR instead of the FDR.

Several other approaches have been proposed as well. Allison *et al.* (2002), and Pounds and Morris (2003) have proposed methods that model *p*-values as independent observations arising from a mixture of a uniform distribution and one or more β distributions. The fitted model is used to compute values for the numerator and denominator of $r_{(i)}$. Pounds and Morris (2003) observed that the fitted model can be partitioned into four regions, one for each of the four classical hypothesis testing outcomes (Figure 7.2). The partition can be used to estimate several error rates such as the FDR, the empirical Bayes posterior (Efron *et al.*, 2001) that the null hypothesis is true, and the total number of Type I and Type II errors (Genovese and Wasserman, 2002; Cheng *et al.*, 2004). Additionally, (Pounds and Cheng 2004; 2006; Cheng *et al.* 2004) have proposed estimators of these and other error rates by using nonparametric density estimators to describe the distribution of the observed *p*-values. The density estimates are used to compute values
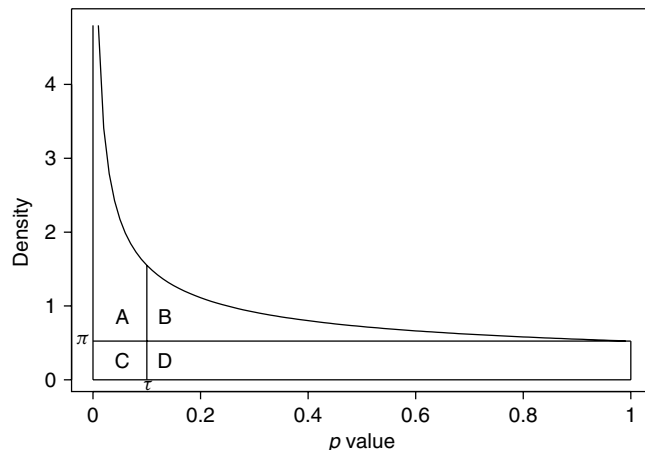
Figure 7.2 Graphical illustration of error-control quantities. Region A corresponds to the occurrence of true positives because it lies above the horizontal line (the alternative component) and to the left of the vertical line (declared significant). Region B corresponds to the occurrence of false negatives because it lies above the horizontal line (the null component) and to the right of the vertical line (not declared significant). Region C corresponds to the occurrence of false positives because it lies below the horizontal line (the null component) and to the left of the vertical line (declared significant). Region D corresponds to the occurrence of true negatives because it lies below the horizontal line (the null component) and to the right of the vertical line (declared significant). [Reprinted from Pounds, S. and Morris S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19**, 1236–1242 by permission of Oxford University Press.]

for the numerator and denominator of the FDR estimate and estimates of various other
error rates.

7.4.3   Significance Criteria for Multiple Hypothesis Tests

When testing a single pair of (null vs alternative) hypotheses, one reports a single p-value.
A decision to reject the null hypothesis is usually made by comparing the p-value to the
customary and subjective levels of Type I error $\alpha = 0.01$, 0.05, or 0.10. It is important
to realize, as seen in the previous section, that in FDR-based multiple hypothesis testing,
5 % FDR does not correspond to the 0.05 significance level (p-value cutoff). In general,
$p = \alpha$ does not correspond to FDR $= \alpha$. Thus, the traditional significance levels are not
necessarily appropriate to use as FDR-control levels. In practice, it is not always clear
how to select an FDR-control level or a p-value threshold α to define the set of results that
will be considered statistically significant, i.e. those having p-value less than or equal to
α. More specifically, it is difficult to know how to balance the trade-off between incurring
too many false positives (Type I errors) against incurring too many false negatives (Type
II errors). Cheng *et al.* (2004) propose two significance criteria to assist in choosing
the p-value significance threshold to balance the number of false positives against the
number of false negatives, namely the profile information criterion (Ip) and the total error
criterion (Er). These significance criteria were developed with microarray applications
in mind; however, they can be used in other applications that involve extensive multiple
testing. Cheng *et al.* (2004) also propose a guide-gene driven criterion for microarray
gene-expression analyses, suitable for applications where a reliable list of differentially
expressed genes is available a priori from existing biological knowledge. This section
briefly reviews Ip and Er.

Let m denote the total number of tests (genes). The Ip criterion is written as follows:

$$Ip(\alpha) = R(\alpha) + \lambda m \pi_0 \alpha. \tag{7.3}$$

The first term, $R(\alpha)$, measures the deviation of the p-values from the uniform (0,1)
distribution in such a way that for each fixed α, $R(\alpha)$ decreases as the p-values distribute
more and more densely around zero than the uniform (0,1) distribution. Furthermore,
when the p-values distribute more heavily around zero than uniform (0,1), $R(\alpha)$ decreases
as α increases. A relatively large number of small p-values indicates the existence of
true alternative hypothesis (discoveries); in this setting, increasing α helps to reduce the
number of false negatives. Thus for the sake of making 'true discoveries' one would
like to reduce $R(\alpha)$ by increasing α. On the other hand, $m\pi_0\alpha$ in the second term is the
expected number of false rejections (false positives/discoveries) and λ is a penalty factor;
this term increases linearly in α. For the sake of avoiding 'false discoveries', one does
not want to make α too large. When the p-values are more heavily distributed around
zero than uniform (0,1), there is a unique α value that minimizes $Ip(\alpha)$. This minimizing
α is the 'optimal' significance threshold (p-value cutoff) according to the Ip criterion.
This α value strikes a balance between the levels of false positives and false negatives,
through the competition of the two terms.

Cheng (2006) formalizes the idea of the profile information criterion Ip and develops
the adaptive profile information (API) criterion: $API(\alpha) = R_\gamma(\alpha) + \lambda m \pi_0 \alpha$. The impor-
tant modification is the inclusion of a parameter γ in the first term. In practice, γ
is estimated from the p-values. This parameter reflects the deviation of the p-value

percentiles from the uniform (0,1) percentiles in a vicinity of zero. The p-value percentiles are directly related to the concentration of small p-values around zero – the higher the concentration, the smaller are the percentiles. If the p-values exactly follow uniform (0,1) then $\gamma = 1$. Also, γ increases as the p-value percentiles become increasingly less than the corresponding uniform (0,1) percentiles. $R_\gamma(\alpha)$ is so constructed that it decreases as the p-values distribute more and more densely around zero than the uniform (0,1) distribution, and decreases in α with the rate of decrement dictated by γ. The larger γ is, the slower is the decrement of $R_\gamma(\alpha)$ in α, in a vicinity of zero. So the γ parameter dictates the emphasis of the deviation from the uniform (0,1) distribution: the higher the concentration of small p-values around zero, the larger γ is, the slower is the decrement of $R_\gamma(\alpha)$ in α; consequently, $API(\alpha)$ tends to minimize at a relatively large α. On the other hand, if the concentration of small p-values around zero is not much higher than that of uniform (0,1), then $\gamma \approx 1$, $R_\gamma(\alpha)$ decreases fast in α, and $API(\alpha)$ tends to minimize at a relatively small α.

A practical issue is the choice of the penalty factor λ. Theoretical investigations and simulation studies in Cheng (2006) and Cheng *et al.* (2004) suggest a few conservative choices. These choices result in 'optimal' α's that guarantee to control the FDR (or equivalently the gFWER) when all null hypotheses are true (i.e. $\pi_0 = 1$, cf. previous section) at the comparable level, as does the Bonferroni procedure. These choices may be too conservative for general microarray applications. For exploratory analyses it may be appropriate to set $\lambda = 1$. This issue remains an open research problem.

Now, we consider the total error criterion, which is written as

$$Er(\alpha) = \pi_0 \alpha + [1 - F_P(\alpha) - \pi_0(1 - \alpha)], \qquad (7.4)$$

where F_P denotes the CDF of the p-values. The Er criterion can be interpreted in terms of the p-value partition (Figure 7.2) recognized by Pounds and Morris (2003). The first term, $\pi_0 \alpha$, is the proportion of all tests resulting in a Type I error when significance is based on comparing p-values to α. In terms of Figure 7.2, $\pi_0 \alpha$ is the area of region A (below $y = \pi_0$ and left of $p = \alpha$). The second term, $[1 - F_P(\alpha) - \pi_0(1 - \alpha)]$, is the proportion of all tests resulting in a Type II error (corresponding to the upper-right region in Figure 7.2). Theoretically, $Er(\alpha)$ is equivalent to the 'total misclassification risk' (Genovese and Wasserman, 2002) in the limiting scenario when the number of tests $m \to \infty$. Thus, choosing α to minimize an estimate $Er(\alpha)$ achieves an optimal balance of Type I and Type II errors in terms of the Er criterion. Cheng *et al.* (2004) develop estimates of π_0, F_P, and $Er(\alpha)$ for each α; the estimate of this optimal α is obtained by minimizing the Er estimate.

7.4.4 Significance Analysis of Microarrays

Tusher *et al.* (2001) proposed the significance analysis of microarrays (SAM) algorithm as a method that incorporates the MTA into the analysis. SAM has become quite popular because it was one of the first statistically developed methods with accompanying user-friendly software made widely available (`www-stat.stanford.edu/~tibs/SAM`). SAM first computes a test statistic of specified form for each of the m probe sets. For a two-group comparison of the expression of probe set i, Tusher *et al.* (2001) proposed a 'relative difference' statistic given by

$$d_i = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{s_i + s_0}, \qquad (7.5)$$

where $\bar{x}_{1i}$ and $\bar{x}_{2i}$ are the mean expression of probe set i in groups 1 and 2, respectively; s_i is a pooled estimate of the standard deviation of expression of probe i; and s_0 is a shrinkage parameter. Tusher *et al.* (2001) provide more details on how the value of the shrinkage parameter is computed. Next, these relative difference statistics are ordered to obtain $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(m)}$. SAM then uses permutation to compute values of the ordered statistics under the null hypothesis. More specifically, let $d^{\star}_{(i)j}$ be the value of the i_{th} ordered statistic in the j_{th} permutation. After B permutations,

$$d_{E(i)} = \frac{1}{B} \sum_{j=1}^{B} d_{(i)j}, \tag{7.6}$$

is used as an estimate of the expected value of $d_{(i)}$ under the complete null hypothesis for $i = 1, \ldots, m$. The significance of $d_{(i)}$ is measured by the deviation $\delta_{(i)} = d_{(i)} - d_{E(i)}$. Next, given a user-specified threshold Δ for the δ's, SAM defines $\tilde{j}_+$ as the smallest j such that $\delta_j \geq \Delta$ and $\tilde{j}_-$ as the smallest j such that $\delta_j \leq -\Delta$. Then, the null hypothesis is rejected for all $j \geq \tilde{j}_+$ and all $j \leq \tilde{j}_-$. The algorithm can be generalized by using other test statistics. Given the form of the test statistic T, the algorithm is very easy to program. Thus, it is conceptually straightforward to generalize the algorithm to various types of applications. However, it is not immediately clear how to calibrate the threshold δ to any specific measure of the Type I error rate. Storey and Tibshirani (2003a) describe how to use information from the individual permutation rounds to calibrate the threshold parameter to an estimate of the FDR.

SAM as described by Tusher *et al.* (2001) has been criticized for a variety of reasons. The example application involved a two-group comparison. The relative difference statistic proposed by Tusher *et al.* (2001) is a modified *t*-statistic. They modified the denominator of the *t*-statistic by incorporating a shrinkage parameter into the denominator. This extra term was included with the intention of preventing Type I errors due to underestimation of variance. However, in the analysis of their spike-in experiment, Choe *et al.* (2005) observed that SAM performed more poorly than the usual *t*-statistic with significance assessed by permutation. They attributed the poor performance of SAM in that application to over shrinkage of the *t*-statistic. Additionally, Dudoit *et al.* (2003) caution users that the developers of SAM no longer recommend using some versions of the algorithm.

7.4.5   Selecting an MTA Method for a Specific Application

The number of available methods to perform MTA raises the practical issue of choosing an appropriate method for a specific application. Currently, the literature provides very limited guidance on this issue. Typically, each newly proposed method is shown to be better than some of the previously proposed methods for a certain application. However, later work may show that the newer method is inferior to other methods for a different setting. Recently, Pounds (2006) outlined some recommendations for choosing from among the methods that use *p*-values, a method to estimate or control the FDR and similar error rates. Subsequently, Pounds and Cheng (2006) updated those recommendations in light of the performance of their robust method of estimating the FDR. These recommendations are briefly summarized below and outlined in Table 7.5.

Table 7.5 Recommendations for FDR Methods.

Setting			Objective	
Correlation[*]	Tests	p-values[†]	Estimate (p)FDR	Control (p)FDR
MoL	two-sided	cont.	Al02,PC03,PC04 L04,Ch04,PC06	BH95, BH00, St02
MoL	two-sided	disc.	PC06	BH95, G05
MoL	one-sided	cont.	PC06	BH95, G05
MoL	one-sided	disc.	PC06	BH95, G05
SaE	two-sided	cont.	BY01[††]	YB99, BY01
SaE	two-sided	disc.	BY01[††]	YB99, BY01
SaE	one-sided	cont.	BY01[††]	YB99, BY01
SaE	one-sided	disc.	BY01[††]	YB99, BY01

Note: Acronyms and full references for methods are listed in Table 7.4.
[*]MoL = mild or limited; SaE = Strong and Extensive.
[†]cont. = continuous; disc. = discrete.
[††]Results should be interpreted cautiously in these cases. The FDR-adjusted p-values should not be interpreted as an estimate of the proportion of significant findings that are false discoveries.
[*Source*: Adapted from Pounds (2006), used with permission.]

MTA used in the analysis of microarray data typically have one of two objectives: estimation or control of a multiple-testing error rate. Control is the classical approach to performing MTA. Control methods attempt to select the set of results considered significant in such a way that the error rate of interest is maintained at or below a prespecified threshold of tolerance. Storey (2002) argues that because microarray studies are very exploratory, it may not be necessary or possible to prespecify an appropriate control level. Thus, Storey (2002) suggests estimation of the error rate as a function of the p-value cutoff (possibly chosen after p-values are computed) used to define the set of results to report in a research publication or examine in greater detail through follow-up experiments. Clearly, the theory of error-rate estimation has not been developed as thoroughly as the theory of error-rate control. Nevertheless, one item of special importance is quite clear: interpreting control quantities such as the q value (Storey, 2002) or FDR-adjusted p-values (Reiner *et al.*, 2003) as estimates of the corresponding error rates can grossly underrepresent the actual prevalence of false discoveries (Pounds and Cheng, 2004). More specifically, because the ratios in equation (7.1) can be interpreted as local estimates of the FDR (Benjamini and Hochberg, 2000), applying the minimization operation in equation (7.2) to the ratios prior to smoothing clearly introduces downward bias into the final 'estimates' of the FDR (Pounds and Cheng, 2004). In particular, the q's in equation (7.2) are not conservative estimates of the FDR.

Thus, the first major consideration in choosing an MTA method is determining whether the application's objective is error-rate estimation or error-rate control. As noted above, interpreting control quantities such as the q values or FDR-adjusted p-values may grossly understate the actual prevalence of Type I errors. Subsequently, estimation methods should be used for applications with error-rate estimation as the objective. However, if error-rate control is the objective, then control methods will tend to have greater power than estimation methods (Pounds, 2006). Unless the experimental design is based on statistical sample size and power calculations, Pounds (2006) suggests that estimation methods

be used. When it is unclear whether error-rate control or error-rate estimation is more appropriate, adaptive significance criteria coupled with FDR estimation (Cheng *et al.*, 2004; Cheng, 2006) can be considered.

The properties of the *p*-values should also be considered in choosing an FDR method. Important properties to consider include the sidedness of the *p*-values (i.e. are the *p*-values based on one-sided or two-sided tests?), the discreteness of the *p*-values, and the correlation of *p*-values with one another. Pounds and Cheng (2005a; 2006) have observed that FDR methods that estimate π_0 can be very unstable when applied to *p*-values arising from one-sided tests. Briefly, *p*-values from one-sided tests may be stochastically greater than uniform (0,1) because the 'untested alternative' is true. For example, suppose the tests are of the form $H_0 : \mu \leq 0$ vs $H_A : \mu > 0$. For genes with $\mu < 0$, the test statistic may be stochastically less than the null distribution derived under the assumption $\mu = 0$. Subsequently, the *p*-value from the right-sided test will be stochastically greater than uniform. This violates the assumption that *p*-values testing true null hypothesis are uniformly distributed. The impact of this violation can be unpredictable. Therefore, Pounds and Cheng (2006) developed a method to estimate the FDR in applications that involve one-sided statistical testing.

The discreteness of the *p*-value distribution can affect the performance of the various FDR procedures as well. In proving the control properties of their FDR-control procedure, Benjamini and Hochberg (1995) assumed that the *p*-values testing true null hypotheses were i.i.d. observations from a continuous uniform (0,1) distribution. Pounds (2006) notes that other methods implicitly rely on an assumption of continuity via the use of modeling or smoothing to compute an estimate of the null proportion, i.e. the proportion of tests with a true null hypothesis. In extensive simulation studies, Pounds and Cheng (2006) observed that discreteness of *p*-values introduced substantial instability in the performance of methods that estimate the null proportion. The instability was more pronounced for the nonparametric methods they explored than for the model-based method (Pounds and Morris, 2003) they examined.

The correlation of *p*-values can affect the performance of several FDR methods. The correlation of *p*-values arises from the correlation of genes. Benjamini and Yekutieli (2001) showed that Benjamini and Hochberg's (1995) procedure maintained the prespecified level of control under a specific dependence structure among test statistics. Storey *et al.* (2004) proved a similar result for Storey's (2002) procedure. Storey and Tibshirani (2003b) argue that genome-wide expression studies naturally result in a dependency structure that satisfies the conditions for Storey's (2002) method to maintain its control properties due to the way that genes work together in pathways. Pounds (2006) notes that most FDR methods perform operations similar to Storey's (2002) procedure or Benjamini and Hochberg's (1995) procedure, and thus should have reasonable performance in this setting as well; however rigorous proofs have not been outlined for many of these additional procedures, and data collected in studies which use custom arrays that focus on particular pathways or diseases may not satisfy the positive regressive dependence structure. To our knowledge, there is no readily implemented procedure to test whether a particular data set satisfies these requirements for dependency among *p*-values. Therefore, we currently recommend that biological knowledge be used to suggest whether the data set satisfies the assumption of positive regressive dependency structure. Benjamini and Yekutieli (2001) developed a simple modification of Benjamini and Hochberg's (1995)

method that maintains its control properties for any dependency structure; however this method is extremely conservative.

7.5   ANNOTATION ANALYSIS

There are several databases that summarize the biological knowledge about genes. For example, the GO Consortium (2000) has defined a nomenclature to describe the biological process, molecular function, and cellular component of each gene. Furthermore, the GO consortium maintains and provides free access to databases that describe each gene in terms of this well-defined nomenclature (www.geneontology.org). Additionally, the Kyoto Encyclopedia of Genes and Genomes (KEGG; www.genome.jp/kegg/pathway.html) databases maintain information about known pathways. Incorporating the information from these databases can provide useful hints to understand and further explore the biology driving the associations identified by the statistical analysis. *AA* incorporates these *annotations* of genes available from freely accessible public databases. Table 7.6 provides software availability information for some of the methods we review in this section.

Several approaches to AA have been used in the literature. A very common approach is to construct a two-by-two table for each annotation term of interest. Each microarray probe set is characterized as 'significant' or otherwise, and as having the annotation or not. Then, the hypergeometric distribution is used to assign a *p*-value to the table. This approach has been criticized because it relies heavily on the assumptions of the hypergeometric distribution, which implies that genes are selected as significant in an independent fashion. In reality, genes operate in pathways, and are therefore correlated, thus violating this critical assumption of the hypergeometric-based approach.

Barry *et al.* (2005) propose the significance analysis of function and expression (SAFE) algorithm for performing AA. First, test statistics are computed for each probe set in the usual way and a test statistic is computed for each annotation term. The test statistics of the probe sets characterize the association of expression with the phenotype. The statistic for a particular annotation (such as the term *apoptosis* or participation in a specific pathway) summarizes how strongly the probe sets with the annotation are associated with phenotype. For example, the annotation statistic could be the average of the *p*-values of the probe sets with the annotation. Then, the significance of the test statistics and annotation statistics are assessed via permutation. The permutation is performed by randomly assigning expression vectors to the phenotype data or labels. For instance, in applications involving two-group or k-group comparisons, the permutation is performed by randomly assigning the group

Table 7.6   Some methods for annotation analysis.

Method name or description	Reference(s)	Software*
Significance analysis of function and expression (SAFE)	Barry *et al.* (2005)	2
Omnibus permutation tests	Potter (2006)	NA
U-statistics	Schaid *et al.* (2005)	NA

* See Table 7.1 for full address of software websites.

labels to the expression vectors. The observed test statistics and annotation statistics are compared to those obtained via permutation to compute p-values and FDR estimates.

A key issue is how the annotation statistics should be defined. For association or differential expression analyses that use linear models (t-test, ANOVA, etc.) at each probe/probe set, Barry *et al.* (2005) propose to define GO annotation statistics by the sum of the F-statistics or squared (or absolute value) t-statistics of the probes/probe sets annotated to the GO category. This is quite a general approach because most of the statistical tests are modified t-statistics or F-statistics. In a similar context, Potter (2006) proposes to use p-values as the basic test statistics and considers several aggregation schemes. Another approach to aggregation can be found in Schaid *et al.* (2005). A complicating issue is how to know whether a particular formula for the annotation statistic actually measures biological significance. For instance, the alteration of one gene may impair an entire pathway, but the overall expression of other genes in the affected pathway may not be substantially altered because of their participation in other pathways. This type of effect may not stand out in an AA that aggregates test statistics across all genes with a particular annotation, but may nevertheless have dramatic biological effects.

7.6   VALIDATION ANALYSIS

Certainly, the results of a microarray study are not definitive if they rely solely on statistical analyses because there are too many possible factors that could cause the statistical analyses to yield misleading results. Even when the 'best' statistical methods are employed, there are too many uncertainties regarding assumption validity and multiple testing for the findings of a microarray study to be considered definitive. Ideally, the findings would be confirmed in a similar, separate, and adequately powered study. Alternatively, auxiliary laboratory experiments may confirm the associations identified in the statistical analysis. However, these studies can be very expensive; hence, it is desirable to avoid committing resources in an attempt to 'validate' a false positive. Thus, computational methods that can provide insights regarding the probability of success for such follow-up studies would prove quite valuable. VA attempts to characterize the 'validity' of the study findings via examination of data sets acquired by resampling, subsampling, permuting, or otherwise perturbing the original study data set.

The specific meaning of the validation depends on the type of analysis performed (recall that the introduction outlines three analysis approaches: class discovery, class prediction, and phenotype association). Although our primary focus has been to describe methods used in phenotype-association analyses, we briefly mention some approaches of VA used in class-prediction analysis. Class-prediction analyses attempt to validate their findings by estimating the misclassification error rate on an independent data set. It is important to note that it is the classifiers, not necessarily the probe sets, that are validated in this analysis. Furthermore, prognostic classifiers must be built and validated in a specific therapeutic context in order to be clinically applicable (Simon, 2005a). Thus, class-prediction analysis and the corresponding VA may not be ideal for the objective of identifying genes to study in greater detail in follow-up laboratory experiments. The typical class-prediction analysis splits the study data set into training and validation sets for an internal validation (Simon, 2005b). Figure 7.3 of Bullinger *et al.* (2004) gives a good illustration of this approach. The power to identify genes is diminished by reducing the sample size in this manner.
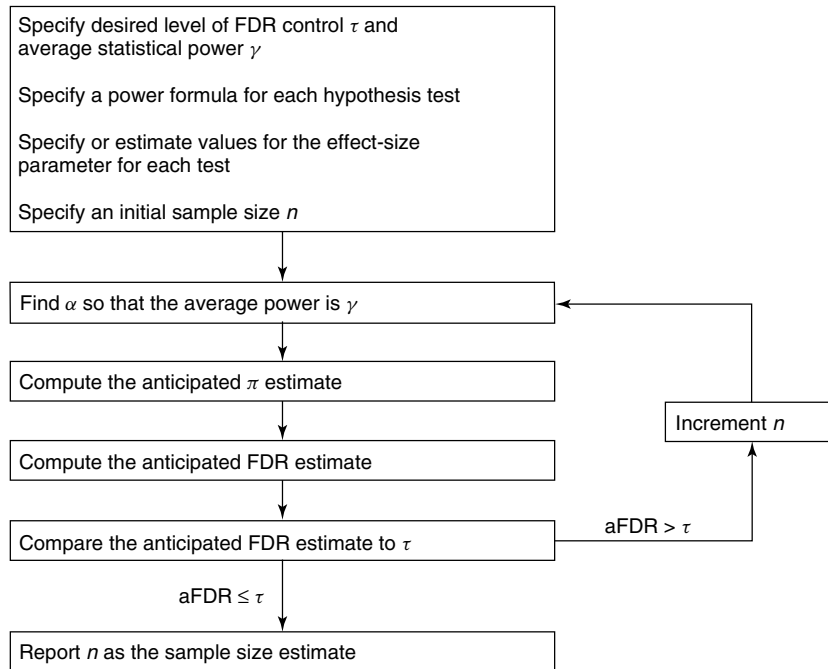
Figure 7.3 Flow diagram of Pounds and Cheng (2005b) method of sample-size calculation.

It is our perception that many investigators choose class-prediction approaches because they wish to demonstrate that their findings have somehow been validated. However, randomly splitting the study data set into two parts just once is often not adequate either to provide an accurate estimate of the classification accuracy, or to demonstrate association on the validation set, due to reduced statistical power. More importantly, Simon (2005b) notes that 'internal validation' by computational methods is no substitute for the 'external validation' provided by a separate, prospective, and adequately powered study. If the purpose is to accurately identify genes and subsequently understand the biology, then a phenotype association is preferred because it uses all the data and thus has greater power than an analysis using only part of the data to identify interesting probe sets.

   In some studies, the phenotype of interest is simply too infrequent to warrant splitting the study data set. Bogni *et al.* (2006) utilize a combination of cluster analysis, multiple-group failure time comparison, and permutation test to internally validate an expression profile associated with the development of therapy-related myeloid leukemia (tML). First, a hazard regression model (Fine and Gray, 1999; Section 7.3.3) is fitted separately for each probe set to measure the association of expression with the development of tML. An expression profile is formed by choosing the probe sets with *p*-value less than or equal to 0.01 from the fitted Fine–Gray model. Next, hierarchical clustering (Gordon, 1999) is performed to assign subjects into three groups, using probe sets in the expression profile as features. Then, Gray's (1988) test is used to compare the cumulative incidence of tML across the clusters. Finally, the entire process is repeated across 1000 data sets obtained by randomly permuting the assignment of expression to outcome data. The χ-square statistics of Gray's test obtained by permutation are compared to the one computed on

the original data set to yield a *p*-value for the expression profile. A similar permutation-based approach to VA was used by Edick *et al.* (2005) to study the association of gene expression with the development of treatment-related brain tumors.

There are several open questions regarding VA in the context of a phenotype-association analysis. Potentially, one could perform VA using bootstrap, jackknife, cross-validation, or permutation techniques. Another open question is how to represent the association of a particular expression profile with the phenotype. In the examples above, a clustering algorithm was applied to the expression data of the selected profile to separate the subjects into groups and then the groups were compared according to phenotype. This raises the question of choosing the number of clusters and the clustering algorithm. Bair *et al.* (2006) have proposed finding the principal components of the expression of the selected probe sets and then testing the association of the first (few) principal component(s) with the phenotype. Certainly, other methods will be proposed, and some theoretical work and analysis of case studies will help to clarify what methods are best suited for specific applications.

7.7   STUDY DESIGN AND SAMPLE SIZE

Experimental design is the most important phase of a microarray gene-expression study. Decisions regarding sample pooling, replication, batching, and reference samples will strongly impact the complexity and interpretation of the eventual data analyses. Poor designs can lead to uninterpretable results due to inappropriate sample pooling, confounding of batches with phenotypes or treatments, confounding of dyes with phenotypes, comparison to a biologically meaningless reference, or inadequate statistical power. These unfortunate outcomes can be avoided with careful planning and study design.

As with other scientific investigations, the design of a microarray study should be directed by practical and statistical considerations. First, the study should be designed to capture the most relevant biological information. For example, while it may be of interest to compare expression across many groups, it may be more practical and efficient to focus the study on the comparisons that may provide the most important biological insights. More specifically, if resources limit one to study only 12 subjects, it may be preferred to study two groups of size six than to examine four groups of size three. If this is the case, then the investigator should choose which two groups to compare to obtain the biological insights of greatest interest. Also, many times, it is not possible to process all samples and arrays on the same day. Subtle variation in sample preparation and handling can have very noticeable effects on the expression signals obtained (see Section 7.2). Hence, the samples must be processed in batches. Batches should be blocked in the study design to avoid confounding batch effects with phenotype effects. Additionally, studies that use two-color arrays should be designed to avoid confounding dye effects with comparisons of interest (Churchill, 2002). Sometimes, each subject cannot yield enough RNA from the relevant tissues required for one array. Then, it becomes necessary to combine samples from multiple subjects into one RNA pool. The RNA pool is then hybridized to the microarray. Then, it becomes necessary to replicate the RNA pools within each condition or phenotype that is studied (Kerr, 2003). As always, studies should be designed with statistical power kept in mind.

Several methods have been proposed to perform power and sample-size calculations for microarray studies (Table 7.7). Pan *et al.* (2002) describe how to compute the sample size

Table 7.7   Methods for sample size and power calculations.

Sample-size objective	Reference(s)	Software
Control number of Type I errors and achieve desired average power	Lee and Whitmore (2002)	10
Control FDR and achieve desired average power	Pounds and Cheng (2005b)	3
Control FDR and achieve desired average power	Jung (2005)	Available from author
Control FDR and achieve desired average power	Gadbury *et al.* (2004)	NA
Control comparison-wise Type I error and per-comparison power	Pan *et al.* (2002)	11
Sample-size estimation with a variety of objectives	Simon *et al.* (2002)	NA
Balancing tech. and biological replicates via cost constraints	Cui and Churchill (2003a)	NA
Sample size needed for a classifier with given accuracy	Mukherjee *et al.* (2003)	NA
Control FDR to achieve desired average power	Hu *et al.* (2005)	1
Sample-size calculation based on Bayesian loss criteria	Müller *et al.* (2004)	NA

required to control the comparison-wise Type I error rate and achieve a specified level of per-comparison power. Lee and Whitmore (2002) describe how to compute the sample size required to achieve a desired level of power while controlling the expected number of Type I errors. Simon *et al.* (2002) describe how to perform sample-size calculations for a variety of objectives. Cui and Churchill (2003a) describe how to balance the level of technical and biological replication to optimize the power under a cost constraint. Mukherjee *et al.* (2003) describe how to determine the number of replicates needed to develop a classifier with a specified level of accuracy. Gadbury *et al.* (2004) and Hu *et al.* (2005) propose methods to determine the sample size for a two-group comparison to achieve desired levels of power while controlling the FDR or pFDR. Müller *et al.* (2004) describe a computationally demanding approach to determine the sample size needed to achieve desired values for specified Bayesian loss functions. Jung (2005) and Pounds and Cheng (2005b) describe general methods to determine the sample size needed to achieve a desired level of power while controlling the FDR at a specified level; these two approaches are described below briefly.

A key element of statistical sample-size calculations is defining an appropriate analog to statistical power in the multiple-testing setting. The average power (Gadbury *et al.*, 2004) is widely used for this purpose. The average power of a procedure is the expected proportion of true associations that are declared statistically significant. This is clearly equal to the arithmetic average of the powers of the individual tests with true alternative hypotheses. Given the formulae for the powers of the individual tests, the power of each test can be estimated given the values for the effect-size parameters. These formulas can be used either to estimate the average power of an experiment with a specified sample size or to estimate the sample size required to achieve a desired average power while controlling the FDR at a specified level. Of note are the methods proposed by Jung (2005), and Pounds and Cheng (2005b), each of which measure the power of a microarray study by the average power.

Jung (2005) notes that the FDR, incurred by performing all tests at a common level α, can be approximated by

$$f \approx \frac{m_0 \alpha}{m_0 \alpha + r_1}, \tag{7.7}$$

where f is the approximate FDR, m is the total number of tests, m_0 is the number of tests with a true null hypothesis, and r_1 is the expected number of rejections among those tests with a true alternative hypothesis. Solving (7.7) for α yields

$$\alpha \approx \frac{f r_1}{m_0 (1 - f)}. \tag{7.8}$$

Substituting a desired level of FDR control for f and a desired number of true discoveries for r_1 into (7.8) gives a value for α. Then, the sample size is determined such that performing all tests at a common level α has an average power $r_1/(m - m_0)$. Jung (2005) outlines more details on how to perform this calculation when the two-sample t-test is used to perform a two-group comparison.

Equation (7.8) highlights the important factors that affect the sample size required to achieve the desired power while controlling the FDR at a specified level. Clearly, the required sample size increases as α in (7.8) decreases. Thus, the required sample size increases when m_0 increases, f decreases, or r_1 increases. In fact, it is clear that some values of m_0, f, and r_1 give values of α that are less than or equal to 0 or greater than 1. More specifically, for any $m_0 > 0$, Jung's (2005) derivations imply that there are some combinations of f and r_1 which cannot be achieved with *any* sample size.

Jung's (2005) approach overlooks some subtle, yet important, issues regarding sample-size estimation. Most FDR control or estimation procedures compute an estimate of the null proportion. The estimate of the null proportion is typically conservatively biased, i.e. $\mathrm{E}(\widehat{\pi}) \geq \pi$, so as to ensure the conservativeness of the control or estimation procedure (Storey, 2002). Therefore, in practice, the specified average power may not be achieved by the sample size computed using Jung's (2005) approach, because the control procedures will tend to choose a smaller p-value threshold than that determined by (7.8). Additionally, Jung (2005) does not consider how to adjust effect-size estimates obtained from pilot data for multiplicity. Failure to adjust effect-size estimates for multiplicity can lead to inflated estimates of power and eventual underestimation of the sample size required to achieve the desired average power.

Pounds and Cheng (2005b) developed a general iterative procedure to perform sample-size calculations to achieve a specified average power while keeping the reported level of FDR (or related measures) controlled at a desired level. The procedure accounts for the dependence of the π_0-estimator on the sample size and is designed so that the computed sample size achieves the desired average power while being able to report controlling the FDR at a prespecified level. The method relies on a quantity called the *anticipated* false discovery rate (aFDR), which is the expected value of the FDR *estimate* given the sample size, effect-sizes of the tests, power formula for the tests, the FDR procedure to be used to compute FDR estimates, and a common level α for all tests. Thus, the power calculations are performed so that one has the desired average power while *reporting* control of the FDR at a desired level. Given all the necessary background information (effect-sizes, power formula, FDR method for final analysis, desired level of FDR control, and desired average power) and a starting value for the sample size, the procedure finds the level α

that achieves the desired average power and then determines if the anticipated FDR is less than or equal to the desired level of FDR control. If the anticipated FDR does not satisfy the specified level for FDR control, then the sample size is increased and the steps described above are repeated. The process is iterated until a maximum possible sample size is reached or the specified requirements for average power and FDR control are satisfied. Figure 7.3 illustrates the method in a flowchart. Furthermore, Pounds and Cheng (2005b) described a method to adjust effect-size estimates obtained from background data for multiplicity. This adjustment is necessary to prevent the sample-size calculations from being overly optimistic. Pounds and Cheng (2005b) provided details on how to perform the effect-size estimation and sample-size calculations for k-group comparisons based on one-factor ANOVA.

7.8 DISCUSSION

Clearly, there are essentially an infinite number of ways to approach the analysis of any microarray data set. Not surprisingly, numerous methods to explore microarray data have been proposed in the literature. As mentioned in the introduction, most microarray analyses can be characterized in terms of having class discovery, class prediction, or phenotype association as an objective. This chapter has provided a very brief overview of some methods that may be useful in phenotype-association analysis. Certainly, the number of methods will continue to increase dramatically during the coming years. Additionally, practical experience with available methods, simulation studies, and theoretical insights will help to clarify how to select the best methods for specific applications.

Given the plethora of possibilities, choosing methods for a specific application can be a daunting task. Nevertheless, there are some sensible considerations that can help in choosing reasonable methods for specific applications. First, the selected method should be designed to address the particular objective of the analysis. This chapter has described some methods for phenotype-association analyses. Second, the method should be designed to accommodate the data type (e.g. continuous, categorical, censored time-to-event) of the phenotype of interest. For example, the analysis should not treat a censored time-to-event phenotype as a binary outcome (Jung *et al.*, 2005a). Third, the assumptions of the selected method should be reasonable for the specific data set to be analyzed. Many authors do not explicitly describe the assumptions of the methods they propose; therefore, evaluation of assumptions can be difficult. Nevertheless, the validity of underlying implicit assumptions will certainly affect the performance of the method and impact the reliability of the obtained inferences. Finally, one must consider whether the study provides adequate power to apply a specific method. For example, class-prediction analyses typically require very large sample size so that each of the training and validation phases of analysis has adequate statistical power.

There are many open research questions pertaining to the statistical analysis of microarray data. An important question involves how to determine the extent to which two or more studies of the same disease confirm or contradict each other's findings. Simply examining reported lists of most significant probe sets is certainly a flawed approach. It is quite plausible that no genes would appear in two or more of the lists simply because the requirements of being a reportable finding are so stringent that it is virtually impossible for any gene to be reported in two or more studies. Additionally, individual microarray studies

may be underpowered, so combining information across studies may guide investigators to important biological insights. An early attempt at meta-analysis of microarray studies has already been published (Rhodes *et al.*, 2004). Many issues can arise in the meta-analysis of microarray studies, including matching probe sets and genes across multiple platforms and attempting to account for possible study-specific biases. Ghosh *et al.* (2003) elaborate on some of the challenges that arise in the meta-analysis of microarray studies and propose some approaches to address those challenges. This will be an exciting area for future research.

Another opportunity for future research is the development of tests that can help determine when permutation- or resampling-based approaches are necessary or preferred. Some researchers assert that permutation approaches are preferred for microarray studies. However, Huang *et al.* (2006) caution that permutation methods can lead to inflated Type I error rates. Additionally, in their application, Morris *et al.* (2005) observed that computationally demanding methods for *p*-value calculation gave qualitatively similar results as the likelihood ratio test. As microarrays with a very large number of features, such as exon arrays and tiling arrays, are more commonly used, the cost and complexity of implementing computationally intensive methods will increase dramatically. Statistical methods that could quickly assess the value of such methods for a specific data set may prove quite valuable in those applications.

There are several other issues worthy of exploration as well. AA and VA are exciting and relatively new areas to explore. Improvements in pathway analysis may help elucidate biological insights by determining when specific pathways are associated with the phenotype of interest. Additionally, methods to incorporate genotype data collected via SNP arrays into the analysis will need to be developed as investigators perform studies that collect genotype and gene-expression data on a common set of subjects. Furthermore, controversies still exist regarding normalization methods (Choe *et al.*, 2005); so even the seemingly more mature areas within the field still need additional research to clarify fundamental issues.

Related Chapters

Chapter 6; **Chapter 8**; and **Chapter 9**.

REFERENCES

Allison, D.B., Cui, C., Page, G.P. and Sabripour, M. (2005). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* **7**, 55–65.

Allison, D.B., Gadbury, G.L. Heo, M., Fernandez, J.R., Lee, C-K., Prolla, T.A. and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* **39**, 1–20.

Allison, D.B., Page, G.P., Beasely, T.M. and Edwards, J.W. (eds) (2006). *DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments.* Taylor & Francis Group, LLC, Boca Raton, FL.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association* **101**, 119–137.

Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

Barry, W.T., Nobel, A.B. and Wright, F.A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* **21**, 1943–1949.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25**, 60–83.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188.

Bogni, A., Cheng, C., Liu, W., Yang, W., Pfeffer, J., Mukatira, S., French, D., Downing, J.R., Pui, C.-H. and Relling, M.V. (2006). Genome-wide approach to identify risk factors for therapy-related myeloid leukemia. *Leukemia* **20**, 239–246.

Bolstad, B.M., Irizarray, R.A., Astrand, M. and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193.

Bryson, M.C. and Johnson, M.E. (1981). The incidence of monotone likelihood in the Cox model. *Technometrics* **23**, 381–383.

Bullinger, L., Döhner, K., Bair, E., Fröhling, S., Schlenk, R.F., Tibshirani, R., Döhner, H. and Pollack, J.R. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *The New England Journal of Medicine* **350**, 1605–1616.

Cheng, C. (2006). An adaptive significance threshold criterion for massive multiple hypothesis testing. In *Optimality: The Second Erich L. Lehmann Symposium*, *IMS Lecture Notes – Monograph Series, Volume 49*, J. Rojo, ed., Institute of Mathematical Statistics, Beachwood, OH, USA, pp. 51–76.

Cheng, C., Pounds, S.B., Boyett, J.M., Pei, D. Kuo, M.L. and Roussel, M.F. (2004). Statistical significance threshold criteria for analysis of microarray gene expression data. *Statistical Applications in Genetics and Molecular Biology* [electronic journal] **3**, 36.

Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M. and Halfon, M.S. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology* **6**, R16.

Churchill, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* **32**, 490–495.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Cui, X., Hwang, J.T.G., Qui, J., Blades, N.J. and Churchill, G.A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59–75.

Cui, X.Q. and Churchill, G.A. (2003a). How many mice and how many arrays? Replication in mouse cDNA microarray experiments. In *Methods of Microarray Data Analysis III*, K.F. Johnson and S.M. Lin, eds. Springer.

Cui, X. and Churchill, G.A. (2003b). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4**, 210.

Do, K.A., Muller, P. and Vannuci, M. (eds) (2006). *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, New York.

Draghici, S. (2003). *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC, New York.

Dudoit, S. and Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* **19**, 1090–1099.

Dudoit, S., van der Laan, M.J. and Pollard, K.S. (2004). Multiple testing. Part I. Single-step procedures for control of general type i error rates. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 13. //www.bepress.com/sagmb/vol3/iss1/art13.

Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103.

Eckel-Passow, J.E., Hoering, A., Therneau, T.M. and Ghobrial, I. (2005). Experimental design and analysis of antibody microarrays: applying methods from cDNA arrays. *Cancer Research* **65**, 2985–2989.

Edick, M.J., Cheng, C., Yang, W., Cheok, M., Wilkinson, M.R., Pei, D., Evans, W.E., Kun, L.E., Pui, C.-H. and Relling, M.V. (2005). Lymphoid gene expression as a predictor of risk of secondary brain tumors. *Genes, Chromosomes and Cancer* **42**, 107–116.

Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Fine, J.P. and Gray, R.J. (1999). A proportional Hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**, 496–509.

Fisher, R.A. (1932). *Statistical Methods for Research Workers*, 4th edition. Oliver & Boyd, London.

Gadbury, G.L., Page, G.P., Edwards, J., Prolla, T.A., Weindruch, R., Permana, P.A., Mountz, J.D. and Allison, D.B. (2004). Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research* **14**, 325–338.

Ge, Y., Dudoit, S. and Speed, T.P. (2003). Resampling-based multiple testing for microarray data analysis. *Test* **12**, 1–77.

The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B* **64**, 499–517.

Gentleman, R., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarray, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H. and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80. http://genomebiology.com/2004/5/10/R80.

Gentleman, R., Carey, V.J., Huber, W., Irizarray, R.A. and Dudoit, S. (eds) (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.

Ghosh, D., Barette, T.R., Rhodes, D. and Chinnaiyan, A.M. (2003). Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Functional and Integrative Genomics* **3**, 180–188.

Gilbert, P.B. (2005). A modified false discovery rate multiple comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Applied Statistics* **54**, 143–158.

Gordon, A.D. (1999). *Classification*, 2nd edition. Chapman & Hall/CRC, Boca Raton, FL.

Gray, R.J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics* **16**, 1141–1154.

Graybill, F.A. (1976). *Theory and Application of the Linear Model*. Wadsworth & Brooks Cole, Pacific Grove, CA.

Harr, B. and Schlotterer, C. (2006). Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Research* **34**, e8.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Hoffmann, R., Seidl, T. and Dugas, M. (2002). Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biology* **3**, 7.

Huang, Y., Xu, H., Calian, V. and Hsu, J.C. (2006). To permute or not to permute. *Bioinformatics* **22**, 2244–2248.

Hu, P., Beyene, J. and Greenwood, C.M.T. (2006). Tests for differential gene expression using weights in oligonucleotide microarray experiments. *BMC Genomics* **7**, 33.

Hu, J., Zou, F. and Wright, F.A. (2005). Practical FDR-based sample size calculations in microarray experiments. *Bioinformatics* **21**, 3264–3272.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003a). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**, e15.

Irizarry, R.A., Hobbs, B., Colin, F., Beazer-Barclay, Y.D., Antonellis, K., Scherf, U. and Speed, T.P. (2003b). Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Jain, N., Thatte, J., Braciale, T., Ley, K., O'Connell, M. and Lee, J.K. (2003). Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **19**, 1945–1951.

Jung, S.-H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics* **21**, 3097–3104.

Jung, S.H., Owzar, K. and George, S.L. (2005a). A multiple testing procedure to associate gene expression levels with survival. *Statistics in Medicine* **24**, 3077–3088.

Jung, S.H., Owzar, K. and George, S.L. (2005b). Associating microarray data with a survival endpoint. In *Methods of Microarray Data Analysis IV*, J.S. Shoemaker and S.M. Lin, eds. Springer, New York.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. John Wiley & Sons, Hoboken, NJ.

Kecman, V. (2001). *Learning and Soft Computing*. MIT Press, Cambridge, MA.

Kerr, M.K. (2003). Design considerations for efficient and effective microarray studies. *Biometrics* **59**, 822–828.

Kerr, M.K., Martin, M. and Churchill, G.A. (2000). Analysis of variances from gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.

Kepler, T.B., Crosby, L. and Morgan, K.T. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology* **3**, research0037.1–0037.12.

Lampron, A., Bourdeau, I., Hamet, P., Tremblay, J. and Lacroix, A. (2006). Whole genome expression profiling of GIP- and ACTH-dependent adrenal hyperplasias reveals novel targets for the study of GIP-dependent Cushing's syndrome. *The Journal of Clinical Endocrinology and Metabolism*, **91**(9), 3611–3618.

Larsson, O., Wahlestedt, C. and Timmons, J.A. (2005). Considerations when using the significance analysis of microarrays (SAM) algorithm. *BMC Bioinformatics* **6**, 129.

Lee, M.-L.T. (2004). *Analysis of Microarray Gene Expression Data*. Kluwer Academic Publishers, Boston, MA.

Lee, M.-L.T. and Whitmore, G. (2002). Power and sample size for microarray studies. *Statistics in Medicine* **11**, 3543–3570.

Li, C. and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 31–36.

Liao, J.G., Lin, Y., Selvanayagam, Z.E. and Shih, W.J. (2004). A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics* **20**, 2694–2701.

Lin, S.M. and Johnson, K. (2003). *Methods of Microarray Data Analysis III*. Kluwer Academic Publishers, Norwell, MA.

McShane, L.M., Radmacher, M.D., Freidlin, B., Yu, R., Li, M-C and Simon, R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* **18**, 1462–1469.

Mehta, T., Tanik, M. and Allison, D.B. (2004). Towards sound epistemological foundation of statistical methods for high-dimensional biology. *Nature Genetics* **36**, 943–947.

Michiels, S., Koscielny, S. and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488–492.

Mills, J.D. and Gordon, J.I. (2001). A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Research* **29**, e72.

Mills, J.C., Roth, K.A., Cagan, R.L. and Gordon, J.I. (2001). DNA microarrays and beyond: completing the journey from tissue to cell. *Nature Cell Biology* **3**, E175–E178.

Morris, J.S., Yin, G., Baggerly, K.A., Wu, C. and Li, Z. (2005). Pooling information across different studies and oligonucleotide chip types to identify prognostic genes for lung cancer. In *Methods of Microarray Data Analysis IV*, J.S. Shoemaker and S.M. Lin, eds. Springer, New York.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T.R. and Mesirov, J.P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology* **10**, 119–142.

Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association* **99**, 990–1001.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18**, 546–554.

Pan, W., Lin, J. and Le, C.T. (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology* [electronic journal] **3**, 5.

Parmigiani, G., Garrett, E., Irizarry, R.A. and Zeger, S.L. (eds) (2003). *The Analysis of Gene Expression Data*. Springer, New York.

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. and Smith, P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: introduction and design. *British Journal of Cancer* **34**, 585–612.

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. and Smith, P.G. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: analysis and examples. *British Journal of Cancer* **35**, 1–39.

Potter, D.M. (2006). Omnibus permutation tests of the association of an ensemble of genetic markers with disease in case-control studies. *Genetic Epidemiology* **30**, 438–446.

Pounds, S. (2006). Estimation and control of multiple testing error rates for microarray studies. *Briefings in Bioinformatics* **7**, 25–36.

Pounds, S. and Cheng, C. (2006). Robust estimation of the false discovery rate. *Bioinformatics* **22**, 1979–1987.

Pounds, S. and Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics* **20**, 1737–1745.

Pounds, S. and Cheng, C. (2005a). Statistical development and evaluation of gene expression data filters. *Journal of Computational Biology* **12**, 482–495.

Pounds, S. and Cheng, C. (2005b). Sample size determination for the false discovery rate. *Bioinformatics* **21**, 4263–4271.

Pounds, S. and Morris, S.W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19**, 1236–1242.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics* **32**(Suppl.), 496–501.

Quackenbush, J. (2006). Microarray analysis and tumor classification. *The New England Journal of Medicine* **354**, 2463–2472.

Reimers, M. (2005). Statistical analysis of microarray data. *Addiction Biology* **10**, 23–35.

Reiner, A., Yekutieli, D. and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368–375.

Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A.M. (2004). Large-scale meta-analysis of cancer microarray data identifies

common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America* **25**, 9309–9314.

Ribeiro, R.C., Razzouk, B.I., Pounds, S., Hijiya, N., Pui, C.H. and Rubnitz, J.E. (2005). Successive clinical trials for childhood acute myeloid leukemia at St. Jude Children's Research Hospital, 1980 through 2000. *Leukemia*, **19**, 2125–2129.

Ritchie, M.E., Diyagamma, D., Neilson, J., van Laar, R., Dobrovic, A., Holloway, A. and Smyth, G.K. (2006). Empirical array quality weights in the analysis of microarray data. *BMC Bioinformatics* **7**, 261.

Romano, J. (1990). On the behavior of randomization tests without a group − symmetry assumption. *Journal of the American Statistical Association* **85**, 686–692.

Ross, M.E., Mahfouz, R., Onciu, M., Liu, H.-C., Zhou, X., Song, G., Shurtleff, S., Pounds, S., Cheng, C., Ma, J., Ribeiro, R.C., Rubnitz, J.E., Girtman, K., Williams, W.K., Raimondi, S.C., Liang, D.-C., Shih, L.-Y., Pui, C.-H. and Downing, J.R. (2004). Gene expression profiling of pediatric acute myeogenous leukemia. *Blood*, **104**, 3679–3687.

Schaid, D.J., McDonnell, S.K., Hebbring, S.J., Cunningham, J.M. and Thibodeau, S.N. (2005). Nonparametric tests of association of multiple genes with human disease. *American Journal of Human Genetics* **76**, 780–793.

Shoemaker, J.S. and Lin, S.M. (eds) (2005). *Methods of Microarray Data Analysis IV*. Springer, New York.

Simon, R., Radmacher, M.D. and Dobbin, K. (2002). Design of studies using DNA microarrays. *Genetic Epidemiology* **23**, 21–36.

Simon, R., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W. and Zhao, Y. (2003). *Design and Analysis of DNA Microarray Investigations*. Springer, New York.

Simon, R. (2005a). Development and validation of therapeutically relevant multi-gene biomarker classifiers. *Journal of the National Cancer Institute* **97**, 866–867.

Simon, R. (2005b). Roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of Clinical Oncology* **23**, 7332–7341.

Smolkin, M. and Ghosh, D. (2003). Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics* **4**, 36.

Speed, T. (ed) (2003). *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/ CRC, New York.

Steinhoff, C. and Vingron, M. (2006). Normalization and quantification of differential expression in gene expression microarrays. *Briefings in Bioinformatics* **7**, 166–177.

Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

Storey, J.D. and Tibshirani, R. (2003a). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data*, G. Parmigiani, E.S. Garrett, R.A. Irizarry and S.L. Zeger eds. Springer, New York.

Storey, J.D. and Tibshirani, R. (2003b). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–9445.

Storey, J.D., Taylor, J.E. and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society, Series B* **66**, 187–205.

Tan, F.K., Hildebrand, B.A., Lester, M.S., Stivers, D.N., Pounds, S., Zhou, X., Wallis, D.D., Milewicz, D.M., Reveille, J.D., Meyes, M.D., Jin, L. and Arnett, F.C. Jr. (2005). Classification analysis of the transcriptosome of nonlesional cultured dermal fibroblasts from systemic sclerosis patients with early disease. *Arthritis and Rheumatism* **52**, 865–876.

Thompson, M.C., Fuller, C., Hogg, T.L., Dalton, J., Finkelstein, D., Lau, C.C., Chintagumpala, M., Adesina, A., Ashley, D.M., Kellie, S.J., Taylor, M.D., Curran, T., Gajjar, A. and Gilberson, R.J. (2006). Genomics Identifies Medulloblastoma Subgroups That Are Enriched for Specific Genetic Alterations. *Journal of Clinical Oncology* **24**, 1924–1931.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116–5121.

van der Laan, M., Dudoit, S. and Pollard, K.S. (2004a). Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 14. `//www.bepress.com/sagmb/vol3/iss1/art14`.

van der Laan, M., Dudoit, S. and Pollard, K.S. (2004b). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 15. `//www.bepress.com/sagmb/vol3/iss1/art15`.

Wang, X., Hessner, M.J., Wu, Y., Pati, N. and Ghosh, S. (2003). Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics* **19**, 1341–1347.

Wu, Z., Irizarry, R.A., Gentleman, R., Murillo, F.M. and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* **99**, 909–917.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R.S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**, 625–637.

Workman, C., Jensen, L.J., Jarmer, H., Berka, R., Gautier, L., Nielser, H.B., Saxild, H-H., Nielsen, C., Brunak, S. and Knudsen, S. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology* **3**(9), research 0048.1–0048.16.

Yang, I., Chen, E., Hasseman, J., Liang, W., Frank, B., Wang, S., Sharov, V., Saeed, A., White, J., Li, J., Lee, N., Yeatman, T. and Quackenbush, J. (2002). Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biology* **3**, 1–12.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* **82**, 171–196.

<div style="text-align: right">

8

</div>

Bayesian Methods for Microarray Data

A. Lewin and S. Richardson

Division of Epidemiology, Public Health and Primary Care, Imperial College, London, UK

In this article, we review the use of Bayesian methods for analyzing gene expression data. We focus on methods that select groups of genes on the basis of their expression in RNA samples derived under different experimental conditions. First, we describe Bayesian methods for estimating gene expression level from the intensity measurements obtained from the analysis of microarray images. Next, we discuss the issues involved in assessing differential gene expression between two conditions at a time, including models for classifying the genes as differentially expressed or not. In the last two sections, we present models for grouping gene expression profiles over different experimental conditions, in order to find coexpressed genes, and multivariate models for finding gene signatures, i.e. for selecting a parsimonious group of genes that discriminate between entities such as subtypes of disease.

8.1 INTRODUCTION

High-throughput technologies such as DNA microarrays have emerged over the last $5-10$ years as one of the key source of information for functional genomics. Microarrays permit researchers to capture one of the fundamental processes in molecular biology, the transcription process from genes into messenger ribonucleic acid (mRNA) that will be subsequently translated to form proteins. This process is called *gene expression*. By quantifying the amount of transcription, microarrays allow the identification of the genes that are expressed in different types of cells and different tissues and help to understand the cellular processes in which they intervene, thus giving a unique insight into the function of genes. However, transforming the huge quantity of data that is currently produced in experiments that involve microarrays into useful knowledge for system biology is not trivial and research into ways of interpreting this rich body of data has become an active area, involving statisticians, machine learning and computer scientists.

Microarrays generally contain thousands of spots (or probes) at each of which a particular gene or sequence is represented. In effect, a microarray experiment represents data comparable to that obtained by performing tens of thousands of 'experiments' of a similar type in parallel. The 'experiments' on a given array will share certain characteristics related to the manufacturing process of the particular array used and the extraction and handling of the biological sample hybridized to the array. The interest is in comparing expression levels between arrays with samples from different biological conditions of interest (e.g. cancerous against noncancerous cells) and the challenge is identifying differences that are related to the biology of the samples rather than to technical experimental variation.

Many of the characteristic features of experiments involving microarrays render them particularly well suited to a flexible modeling strategy within the Bayesian framework. The aim of this chapter is to focus on the unique contribution that Bayesian methods offer and highlight this by discussing in detail the steps taken for modeling the variability in gene expression data at several levels.

The framework of Bayesian hierarchical modeling refers to a generic model building strategy in which unobserved quantities are organized into a small number of discrete levels with logically distinct and scientifically interpretable functions and probabilistic relationships between them that capture inherent features of the data. It is of course important to perform some basic exploration and visualization of the data before formulating complex models; see **Chapter 6**, for examples. The hierarchy of levels makes it particularly suitable for modeling gene expression data, which arises from a number of processes and is affected by many sources of variability. In the next sections, we look at an approach to modeling these different sources of variability using fixed effects, random effects and distributional assumptions.

One of the most important aspects of Bayesian hierarchical modeling as regards microarray data is the sharing of information across parallel units. For example, gene expression experiments used by biologists to study fundamental processes of activation/suppression frequently involve genetically modified animals or specific cell lines, and such experiments are typically carried out only with a small number of biological samples. It is clear that this amount of replication makes standard estimates of gene variability unstable. By assuming exchangeability across the genes, inference is strengthened by borrowing information from comparable units.

Another strength of the Bayesian framework is the propagation of uncertainty through the model. Owing to the many sources of systematic variation between arrays and samples, gene expression data is often processed through a series of steps, each time estimating and subtracting effects in order to make the arrays comparable. The end result of this process can be overconfident inference, as the uncertainty associated with each step is ignored. In a Bayesian model, it is straightforward to include each of these effects simultaneously, thus retaining the correct level of uncertainty on the final estimates. Further, when including structured priors that are associated with classification, e.g. mixture priors, in the model, estimates of uncertainty of the classification are obtained along with the fit of the model.

The field of microarray data analysis is very large, and it is not possible to cover all aspects in this chapter. In particular, we do not discuss methods for estimating graphical models and Bayesian networks that are aimed at understanding regulatory networks or metabolic pathways. There is a large literature on this subject; see, e.g. a number of

chapters in Do *et al.*, 2006, and references therein. See also **Chapter 9**, **Chapter 6** and **Chapter 7** for non-Bayesian approaches.

We focus on methods that select groups of genes on the basis of their expression in RNA samples derived under different experimental conditions. In the next section, we look at two Bayesian methods for estimating gene expression level from the intensity measurements obtained from analysis of microarray images. This section includes some discussion of the steps involved in a microarray experiment. Section 8.3 discusses the issues involved in assessing differential gene expression between two conditions at a time. There is an extensive literature on this topic. Our presentation is divided into sections on normalization, gene variability and models for classifying the genes as those that are differentially expressed and those that are not. We include a brief explanation of mixture models and some discussion of different decision rules used to choose the lists of genes that are considered to be differentially expressed. In Section 8.4, we present a range of models for grouping gene expression profiles, which are vectors of gene expression over several different experimental conditions, for ordered samples (e.g. time-course data) and for samples with no ordering. Finally, Section 8.5 reviews the current work on multivariate methods for finding subsets of genes that can predict and classify phenotypes. We focus on variable selection methods that have been used to find, e.g. the so-called gene signature of different subtypes of disease, as well as Bayesian shrinkage methods. In both cases, the emphasis is on parsimony of the multivariate model in order to enhance interpretation. In Bayesian models, inference is made in either an empirical or fully Bayesian framework. In the case of fully Bayesian models, Markov Chain Monte Carlo (MCMC) is usually used to estimate the posterior distribution of the model. We do not go into details of these procedures, except in the case of nonstandard algorithms. Table 8.1 gives URLs for software for the models discussed in this chapter.

A note on notation: we use x_g for gene expression measures used as data in Sections 8.3 and onward, rather than the more usual y_g. This is to allow the standard formulation for the variable selection models in Section 8.5, where y stands for the outcome and x stands for the variables (here genes). Throughout the chapter, p stands for the number of genes and n stands for the number of samples or experimental conditions.

8.2   EXTRACTING SIGNAL FROM OBSERVED INTENSITIES

The output from a microarray experiment is an the image of an array (see **Chapter 6**). This image must be gridded and segmented into spots, which are the basis for inference about the sequences present in the sample of interest. There has been some work on Bayesian methods for image analysis (see, e.g. Ceccarelli and Antoniol, 2006; Gottardo *et al.*, 2006a and references therein). This work is beyond the scope of this chapter. The methods we describe in this section first deal with an intensity measurement for each pixel on an array, found from the image analysis, and these are used to construct a summary measure of the amount of RNA present in the sample for each gene of interest.

There are two main types of microarrays in use, spotted or cDNA arrays, which are usually two color, and oligonucleotide arrays, which are one color. Spotted arrays are microscopic slides onto which long strands of cDNA are fixed in a regular grid layout. Each 'spot' on the array will then contain millions of copies of the same (known) sequence

Table 8.1   Web pages containing software or code for the models described in this chapter.

Extracting signal from observed intensities

Frigessi *et al.*, 2005	TransCount	`http://alba.uio.no/base/local/` `demotranscount/index.html`
Hein *et al.*, 2005	BGX	`http://www.bgx.org.uk/` `software.html`

Differential expression

Baldi and Long, 2001	Cyber-T	`http://visitor.ics.uci.edu/` `genex/cybert/`
Broët *et al.*, 2002	nmix	`http://www.bgx.org.uk/` `software.html`
Broët *et al.*, 2004	gmix	`http://www.bgx.org.uk/` `software.html`
Do *et al.*, 2005	BayesMIX	`http://odin.mdacc.` `tmc.edu/~kim/` `bayesmix/`
Efron *et al.*, 2001	EBAM	`http://bioconductor.fhcrc.` `org/packages/2.0/` `bioc/html/` `siggenes.html`
Gottardo *et al.*, 2006b	rama	`http://www.stat.ubc.ca/` `~raph/Software/` `BiocRPackages/` `BiocRPackages.html`
Gottardo *et al.*, 2006c	bridge	`http://www.stat.ubc.ca/` `~raph/Software/` `BiocRPackages/` `BiocRPackages.html`
Ibrahim *et al.*, 2002	code available from author: `mhchen@stat.` `uconn.edu`	
Ishwaran and Rao (2003; 2005a; 2005b)	BAM	`http://www.bamarray.com/`
Lewin *et al.*, 2006	BayesDE	`http://www.bgx.org.uk/` `software.html`
Lönnstedt and Speed, 2003	SMA	`http://www.stat.berkeley.` `edu/~terry/zarray/` `Software/smacode.html`
Newton *et al.* (2001; 2004)	EBarrays	`http://www.stat.wisc.` `edu/%7Enewton/` `research/arrays.html`
Parmigiani *et al.*, 2002	POE	`http://astor.som.jhmi.` `edu/poe/`
Reilly *et al.*, 2003		`http://www.biostat.umn.` `edu/cavanr/` `geneNormRepHier.txt`

(continued overleaf)

Table 8.1 (*continued*).

Smyth, 2004	limma	`http://bioinf.wehi.edu.au/limma/`
Clustering profiles		

Gottardo *et al.*, 2006c	bridge	`http://www.stat.ubc.ca/~raph/`
		 `Software/BiocRPackages/`
		 `BiocRPackages.html`

Heard *et al.* (2006a; 2006b)		`http://stats.ma.ic.ac.` `uk/naheard/public_html/`

Kendziorski *et al.*, 2003	EBarrays	`http://www.stat.wisc.`
		 `edu/%7Enewton/`
		 `research/arrays.html`

Vogl *et al.*, 2005		`http://genome.tugraz.at/` `BayesianClustering/`
Zhou and Wakefield, 2006		`http://faculty.washington.` `edu/jonno/cv.html`

of cDNA (called *probes*). One sequence corresponds to one gene, or expressed sequence tag (EST). In order to find out what sequences are present in a sample of mRNA, a sample of cDNA (the target) is produced from the mRNA by reverse transcription and fluorescently labeled. This sample is introduced onto the array, where hybridization reactions take place between sequences of cDNA that match in the sample and on the array. The array is then washed to remove target cDNA that has not hybridized to the array, and scanned to detect the fluorescent labels of the cDNA strands that have hybridized. For two-color arrays, two samples of cDNA, labeled with dyes of two different frequencies (Cy3 and Cy5), are placed on the array. The two samples are usually from different mRNA samples, enabling the concentrations of particular sequences to be compared between the two samples.

There are two particularly important statistical issues arising from the process of the microarray experiment. First, the Cy3 dye tends to appear brighter than the Cy5, owing to differences in the reaction with the cDNA and different responses to the laser used in the scanning process. This leads to the so-called dye effect. In addition, the known cDNA sequences are printed onto the array using a number of spotting pins. The different pins may deliver slightly different amounts of cDNA to the array, and thus there can be a systematic effect between spots printed with different pins. This is known as the *print-tip effect*.

Oligonucleotide arrays work in a similar way. There are three main differences from spotted arrays (from a data analysis point of view), the first being that just one sample is hybridized to each array, and thus only one dye is used, so there is no dye effect. The second is that the same printing head is used for all spots, and thus there is no print-tip effect. The third difference is that the probe sequences fixed to oligonucleotide arrays are shorter than those used in spotted arrays. For this reason, several probes (of different sequences) are used to detect one gene or EST. Spots on the array come in pairs: one containing the 'perfect match' (PM) probe and an adjacent spot containing the 'mismatch' (MM) probe. The PM probe is a strand of cDNA that has the sequence of interest. The MM probe has the same sequence except for the central nucleotide, which is different.

The reason for this is to provide a measure of cross-hybridization: target cDNA having a similar but not identical sequence to the PM probe may hybridize to the PM, contaminating the signal. The idea behind the MM probe is that these mismatched target cDNAs would also hybridize to the MM probe, but the true matches to the PM would only hybridize to the PM and not to the MM. Thus the amount of cDNA with the exact same sequence as the PM could be estimated by subtracting the MM signal from the PM signal. In reality, target cDNA with exact match to PM also hybridizes partially to the MM, so estimating the correct amount is a complicated process (see Section 8.2.2).

8.2.1   Spotted cDNA Arrays

Most work with data from spotted arrays takes the ratio of the Cy3 and Cy5 intensities as a measure of the *relative* expression of each gene in the two RNA samples. These can be used in a fairly straightforward way to compare gene expression under different experimental conditions. Care must be taken to account for the dye and print-tip effects. These effects are often included as part of the model for differential expression, as seen in Section 8.3.1.

Here we discuss a Bayesian model developed by Frigessi *et al.* (2005) for obtaining estimates of concentrations of RNA from two-color arrays. This model includes the dye and print-tip effects, along with other aspects of the experimental process, usually absorbed into empirical normalization methods. The idea is to follow through the process that the RNA molecules undergo in order to be detected as hybridized to the array. The steps in this process are modeled with a hierarchical model.

The principal data used in the model are the intensity measurements in each pixel j on each array a. These are denoted by $L_{j,s}^{i,a}$, where s labels spots to which the pixel belongs and i labels the particular RNA sample hybridized to the array. The background intensity is assumed to have been subtracted as part of the image analysis. The quantity of interest to estimate is the concentration of RNA (in molecules per unit weight) for gene g in sample i, denoted by K_g^i. The main steps relating this concentration to the observed intensities are hybridization, washing, and scanning.

In order to model the scanning process, consider the number of molecules $J_s^{i,a}$ from sample i left on spot s after hybridization and washing. The observations that contribute directly to $J_s^{i,a}$ are the intensities for the pixels in the spots corresponding to that gene; for a given sample and array, $L_{j,s}^{i,a} \propto J_s^{i,a}/n_s^a$, where n_s^a is the number of pixels in spot s. The constant of proportionality is $2^{f_{\text{dye}} \cdot PMT^{i,a}} \alpha_{\text{dye}}$, where $PMT^{i,a}$ is the voltage used in scanning and f_{dye} is the scanner amplification factor, both of which are known. The factor α_{dye} accounts for the dye effect (this is estimated as part of the model). With this expected relation between intensity and number of molecules on a spot in place, the intensity measurements are modeled as coming from a normal distribution,

$$L_{j,s}^{i,a} \sim N(2^{f_{\text{dye}} \cdot PMT^{i,a}} \alpha_{\text{dye}} J_s^{i,a}/n_s^a, (\sigma_s^{i,a})^2). \tag{8.1}$$

The variance $(\sigma_s^{i,a})^2$ is estimated from the sample variance of the intensities, and treated as fixed in the analysis.

In the second level of the hierarchical model, the prior for $J_s^{i,a}$ depends on the concentrations $K_{g(s)}^i$, where $g(s)$ is the gene corresponding to spot s, and also on other parameters for various effects encountered in the hybridization step. This step is treated as

a selection process where each molecule of gene $g(s)$ has an equal chance of hybridizing to and remaining on spot s. Thus the number of molecules has a binomial distribution:

$$J_s^{i,a} \sim \text{Bin}(cn_s^a q^{i,a} K_{g(s)}^i, p_s^{i,a}), \qquad (8.2)$$

where $q^{i,a}$ is the total weight of sample i and c is a hybridization factor (estimated in a calibration experiment). The probability of hybridization $p_s^{i,a}$ has several contributions:

$$p_s^{i,a} = L^{-1}(\gamma_0 + \gamma_{g(s,a)} + \gamma P^i + \beta X_s^a). \qquad (8.3)$$

P^i is a measure of the purity of sample i, and X_s^a contains the covariates for spot s on array a: probe length, probe quality, print tip and array. Various link functions are used for L. To ensure identifiability, several arrays must be analyzed together. The main object of inference is K_g^i and a purposely designed MCMC algorithm is used to get posterior samples.

8.2.2 Oligonucleotide Arrays

In contrast to cDNA arrays, the intensity measurements for spots on oligonucleotide arrays cannot be combined in a simple manner to form gene expression measurements. The simplest way to use the PM and MM measurements would be to use PM–MM as a measure of expression. However, there is a problem with this, as very often the MM intensity is larger than the PM intensity (see Figure 8.1, top row). There has been much work in the microarray literature on methods to deal with this phenomenon. Here, we present Bayesian models developed to model the PM and MM intensities in order to produce measures of gene expression.

Hein *et al.* (2005) present a fully Bayesian hierarchical model, estimated by MCMC, for obtaining gene expression measures for each gene in each experimental condition. If there are replicate samples for a condition (including biological replicates), the model produces an estimate for that condition, rather than separate estimates for each replicate. On the other hand, by using the variability of the probe sets for each gene, the model can be used with a single array for each condition and meaningful comparison between conditions without any replicates can be obtained (Hein and Richardson, 2006).

The data used for the model in Hein *et al.* (2005) are the PM and MM intensities for probe pair j of gene g in replicate r of condition c, denoted by PM_{gjcr} and MM_{gjcr}. Each c, r pair corresponds to one physical array. The intensity observed at a PM probe is assumed to be the result of hybridization, partly of fragments that perfectly match the probe (specific hybridization, signal: S_{gjcr}) and partly of fragments that do not perfectly match the probe (nonspecific hybridization: H_{gjcr}). A similar pattern is assumed for the MM probe, with only a fraction ϕ of the perfectly matching fragments undergoing binding. Both specific and nonspecific hybridization are estimated separately for each gene and probe. To account for the possibility of the MM being bigger than the PM, the model includes an additive error on the normal scale.

$$PM_{gjcr} \sim N(S_{gjcr} + H_{gjcr}, \tau_{cr}^2),$$
$$MM_{gjcr} \sim N(\phi S_{gjcr} + H_{gjcr}, \tau_{cr}^2). \qquad (8.4)$$
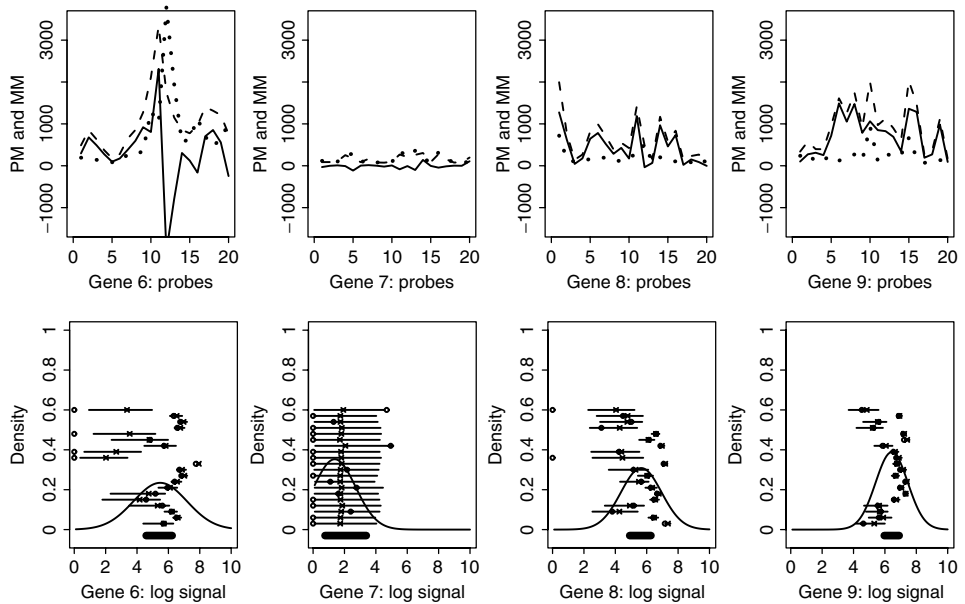
Figure 8.1 Upper panel: probe set response for four genes from an oligonucleotide microarray. Each probe set consists of 20 probe pairs. Solid lines show PM−MM, dashed lines show PM, dotted lines show MM. Lower panel: summaries of posterior distributions related to expression of the four genes, from the model in Hein *et al.* (2005). The 95 % equal-tailed credibility intervals of the S_{gjcr} are shown as horizontal lines (shifted vertically) and should be read off the x axis. The bold line shows the 95 % equal-tailed credibility interval for μ_{gc}. Circles show the observed log(PM−MM) values (plotted at zero when MM > PM). Curves show $TN(\hat{\mu}_{gc}, \hat{\sigma}_{gc}^2)$, with $\hat{\mu}_{gc}$ and $\hat{\sigma}_{gc}^2$ equal to the posterior means. [Reprinted from Hein, A.-M. K., Richardson, S., Causton, H. C., Ambler, G. K., and Green, P. J. (2005). BGX: a fully Bayesian gene expression index for A ymetrix GeneChip data. *Biostatistics* **6**(3), 349–373, by permission of Oxford University Press.]

At the next level of the model, estimates for the specific hybridization for each gene in each condition μ_{gc} are obtained, averaging across probes and replicates. These are the final measures of interest. The nonspecific hybridization is modeled with an arraywide distribution (indexed by c and r).

$$\log(S_{gjcr} + 1) \sim TN(\mu_{gc}, \sigma_{gc}^2),$$
$$\log(H_{gjcr} + 1) \sim TN(\lambda_{cr}, \eta_{cr}^2). \tag{8.5}$$

Here TN stands for the normal distribution truncated at zero on the left. This and the shifted log function allow the hybridization signals to be zero.

Array-specific parameters (those indexed by c, r) are given independent priors. The gene-specific variances σ_{gc}^2 are modeled exchangeably, to share information across the genes and stabilize the variance estimates.

Hein *et al.* (2005) have fit this model to the GeneLogic spike-in data set at `http://www.genelogic.com/media/studies/index.cfm`. This is a widely used data set consisting of gene expression measurements for replicate samples of cRNA

from an acute myeloid leukemia (AML) tumor cell line, with 11 exogenous cRNAs spiked into each sample at a different known concentration in each sample. Each sample was hybridized on one array, thus all measurements for spike-in genes on a particular array correspond to the same cRNA concentration. The top row of Figure 8.1 shows the PM and MM measurements for four of the spiked-in genes (all from the same array). It can be seen that the measurements of different probes within a gene vary widely, owing to the different sequences being detected.

The lower panel shows results for the same four genes when the model is fit to the single array. The plots show the posterior estimates of S_{gjcr} compared with $\log(PM_{gjcr} - MM_{gjcr})$. It can be seen that the posterior estimates get more precise for larger PM–MM, i.e. for probes with high specific hybridization, and consequently that the posterior credibility intervals for the μ_c are reduced. For probes with large MM, the estimates of S_{gjcr} are drawn toward those for the rest of the probes for that gene.

8.3 DIFFERENTIAL EXPRESSION

One of the most widely studied problems in microarray analysis is that of differential gene expression between two experimental conditions, e.g. between knockout and wild-type animals, or between cases and controls. Most work in this area starts with the gene expression measures for each gene on each array, or the log ratios of expression under two conditions. Expression measures have been observed to have increasing variability with increasing value (Schadt *et al.*, 2000), so they are often modeled on the log scale. Sometimes a shifted log transform is used, as, e.g. in Gottardo *et al.* (2006b). **Chapter 6** discusses many transformations used in the literature.

Many models that have been developed for differential expression can be written as a linear model for the log expression level x_{gcr} for gene g, condition $c = 1, 2$ and replicate array r:

$$x_{gcr} = \mu_{gc} + \gamma_{cr} + \varepsilon_{gcr}, \qquad (8.6)$$

where μ_{gc} represents the level of expression of gene g for condition c, γ_{cr} is a normalization term for the array containing the replicate r sample of condition c, and ε_{gcr} is the residual.

Not all models we discuss can be fitted exactly into this format, e.g. Newton *et al.* (2004) and Kendziorski *et al.* (2003) use the gamma distribution to model gene variability and so their models do not quite fit into the linear framework. However, they still involve parameters corresponding to the same biological quantities. In addition, the vast majority of models can be fitted into the linear framework, and thus it is useful to give these equations in an attempt to clarify where models differ or otherwise. We will indicate in the text where models do not use the linear formulation (8.6).

The parameters of interest are the μ_{gc}. Before we discuss how these are modeled, we look at the normalization and error terms.

8.3.1 Normalization

Microarray data show systematic differences between expression levels found on different arrays (e.g. Schadt *et al.*, 2000). Some of these differences are due to dye and print-tip

effects discussed in Section 8.2.1. This may be taken into account when analyzing data at lower levels, but generally empirical differences between arrays are still found for the gene expression values. Often the systematic effect is such that there is nonlinear relationship between the expression levels on different arrays.

Much work has been done in the classical statistical literature on different methods of accounting for these systematic nonlinear differences (normalizing). These usually involve a transformation of the data before it is analyzed with another method. Most work on Bayesian models for gene expression has also assumed that this process has been done beforehand. Bayesian models incorporating normalization include those proposed by Parmigiani *et al.* (2002) and Gottardo *et al.* (2006b). Both of these include a constant term in a linear model, estimated in a fully Bayesian manner.

Bhattacharjee *et al.* (2004) and Lewin *et al.* (2006) model normalization as a nonlinear function of expression level. Bhattacharjee *et al.* (2004) use a normalization term γ_{gcr}, which is modeled as a piecewise linear function of gene expression level. Owing to marginalization over the joint posterior, posterior estimates of γ_{gcr} will be reasonably smooth functions of expression level.

Lewin *et al.* (2006) propose a model starting with that given in (8.6), but for which the normalization term has an additional gene index: $\gamma_{gcr} = f(\mu_{gc})$, where the function f is a quadratic spline. They show that transforming the data first rather than modeling the normalization simultaneously with the other unknown quantities can introduce bias, as the gene expression levels μ_{gc} have to be estimated and thus have variability, as in measurement error problems (Carroll *et al.*, 1995). Figure 8.2 shows the posterior mean array effects γ_{gcr} as a function of expression level for a group of three arrays hybridized to cDNA from wild-type mice, as presented in Lewin *et al.* (2006).

8.3.2 Gene Variability

There are many sources of variation in gene expression data. It is possible to put replicate RNA samples from the same individual on different microarrays, but this is usually considered unnecessary as it has been observed that these so-called technical replicates show very high correlation. More usually different arrays are hybridized with samples taken from different individuals. Thus the variability incorporated in the error term ε_{gcr} in (8.6) represents the biological variability. It is generally accepted that different genes
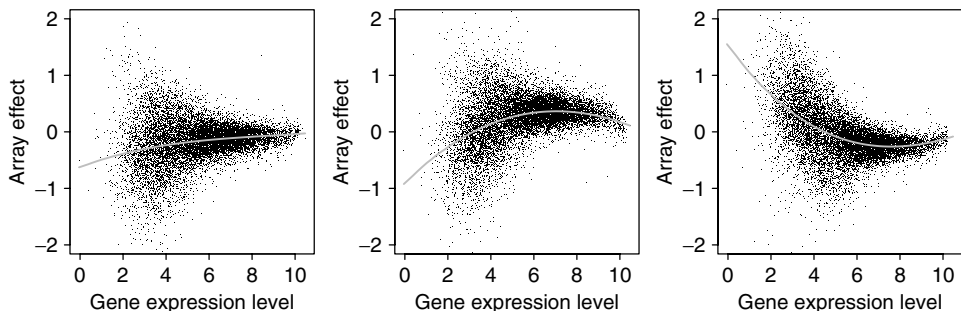
Figure 8.2 Array effects as a function of expression level for a wild-type mouse expression data set of three arrays as presented in Lewin *et al.* (2006). Each plot shows the array effect from one array (curve) with the data residuals from the mean across arrays (points).

show different levels of biological variability, and thus parameters in the distributions for the errors will depend on the gene index.

Several Bayesian models in the literature assume normal errors (Lönnstedt and Speed, 2003; Baldi and Long, 2001; Bhattacharjee *et al.*, 2004; Lewin *et al.*, 2006). Gottardo *et al.* (2006c) use a *t*-distribution (bivariate for cDNA data) to accommodate more outlying data points. Newton *et al.* (2001; 2004) give the data a gamma likelihood rather than the lognormal implied by (8.6). Simple model-checking techniques suggest that the gamma and lognormal families are equally suitable for gene expression data.

Since the number of individuals for each experimental condition is often small, independent estimates of gene variance parameters would be unstable. Therefore, gene variances σ_{gc}^2 are usually shrunk, by assuming exchangeability across genes (and sometimes conditions). Both empirical Bayes (Lönnstedt and Speed, 2003) and fully Bayesian methods (Lewin *et al.*, 2006; Gottardo *et al.*, 2006c) relying on MCMC algorithms for inference have been used. Rather than allowing a separate variance for each gene, Bhattacharjee *et al.* (2004) allow gene variances to take one of three values, estimated as part of the model, as an alternative way of sharing information across genes. Baldi and Long (2001) allow gene variances to depend on expression level, by making the variances exchangeable among genes with similar expression levels (defined by a window on the expression level) and estimating these using Empirical Bayes methods.

8.3.3 Expression Levels

It is useful to write the expression levels in two experimental conditions as

$$
\begin{aligned}
\mu_{g1} &= \alpha_g - \delta_g/2, \\
\mu_{g2} &= \alpha_g + \delta_g/2,
\end{aligned}
\tag{8.7}
$$

where α_g represents the overall expression level for gene g and δ_g represents the log differential expression. For two-color arrays, the data can be given as log fold changes between the conditions (the data is paired) and in that case there is no α_g parameter. When the data is given separately for the two conditions, α_g must be modeled. It is usually treated as a fixed effect, so no information is shared between genes for this parameter.

The fold change parameter δ_g can also be given an unstructured prior (e.g Baldi and Long, 2001; Bhattacharjee *et al.*, 2004; Lewin *et al.*, 2006); however, many people choose to use mixture models to classify genes as those differentially expressed and those that are not. Usually, this implies putting a mixture prior on some measure of the difference between expression levels in the two experimental conditions. The mixture models can be classified into two groups: those that put a mixture prior on the model parameter δ_g, and those that model the data directly as a mixture. These are not intrinsically different, as the parameters δ_g could be integrated out to give a mixture model on the data, but it is convenient to describe the models separately.

A finite mixture distribution for a quantity Δ_g is a weighted sum of probability distributions,

$$
\Delta_g \sim \sum_{k=0}^{K-1} w_k f_k(\phi_k),
\tag{8.8}
$$

where the weights sum to one ($\sum_{k=0}^{K-1} w_k = 1$). Each mixture component has a certain distribution f_k, with parameters ϕ_k. The weight w_k represents the probability of Δ_g being assigned to mixture component k. In the context of differential expression, most mixture models used consist of two components ($K = 2$), one of which (f_0) can be thought of as representing the 'null hypothesis' that there is no differential expression. The second component corresponds to the alternative hypothesis that there is differential expression. Of course, it is not necessary to see the model in terms of hypothesis testing; in the Bayesian framework, it is a straightforward procedure to classify each gene into one or other of the mixture components. This is usually done using the posterior probability of component membership (see Section 8.3.4 for details).

One of the earliest mixture models used in gene expression analysis was that of Efron *et al.* (2001). In this model, Δ_g in (8.8) is a regularized *t*-statistic t_g, one for each gene.

$$t_g \sim w_0 f_0 + w_1 f_1. \tag{8.9}$$

The densities of the mixture components are estimated nonparametrically using standard kernel density procedures. Regularized *t*-statistics are calculated using expression data from the same experimental condition, to provide an estimate of the null component f_0. An estimate of w_0 (which represents the proportion of genes in the null, or not differentially expressed) is obtained using empirical Bayes methods. The whole mixture distribution ($w_0 f_0 + w_1 f_1$) can be estimated using all the t_g. Thus the second component f_1 can be inferred.

A fully Bayesian version of this model has been discussed by Do *et al.* (2005). In this work, the framework of Dirichlet Process Mixtures (DPMs) is used to formulate a prior probability model for the distributions f_0 and f_1. A DPM model, characterized by a base measure G^*, a scalar parameter α and a mixing kernel to be specified, is one of the most popular nonparametric Bayesian models in view of the simplicity of its representation and MCMC implementation (Escobar and West, 1995; Walker *et al.*, 1999). Do *et al.* (2005) choose base measures $G_0^* \sim N(0, \tau^2)$ and $G_1^* \sim \frac{1}{2}N(-b, \tau^2) + \frac{1}{2}N(b, \tau^2)$ for f_0 and f_1, respectively, and Gaussian mixing kernels with common variance parameter σ^2. The specification of G_1^* reflects the prior belief that DE in either direction is equally likely, in the absence of more specific prior information. Use of the stick-breaking construction of DP (Sethurman, 1994) leads to a useful representation of $f_k, k = 0, 1$ as an infinite mixture of normals:

$$f_k = \sum_{h=1}^{\infty} p_{hk} N(\mu_{hk}, \sigma^2),$$

with $\mu_{hk} \stackrel{\text{i.i.d}}{\sim} G_k^*, k = 0, 1$ and the weights following the stick-breaking structure: $p_{hk} = U_h \prod_{j<h}(1 - U_j)$ with $U_h \stackrel{\text{i.i.d}}{\sim} \beta(1, \alpha)$. In Do *et al.* (2005), all the model parameters are given hyperprior distributions, conjugate inverse gamma for τ^2 and σ^2 and conjugate normal for b, α is fixed at 1 and w_0 is given either a beta prior or a uniform prior away from 0. As in Efron *et al.* (2001), within-condition data differences are used to estimate f_0, while between-condition differences are modeled as arising from the mixture defined in (8.9). Figure 8.3 shows posterior estimates of f_0, f_1 and $f \equiv w_0 f_0 + w_1 f_1$ for the Alon colon cancer data set (Alon *et al.*, 1999) as analyzed in Do *et al.* (2005). This is a data set of gene expression measurements for 2000 genes in 62 tissue samples (40 tumors and 22 normal samples). The density f_1 is bimodal, showing that there are genes that
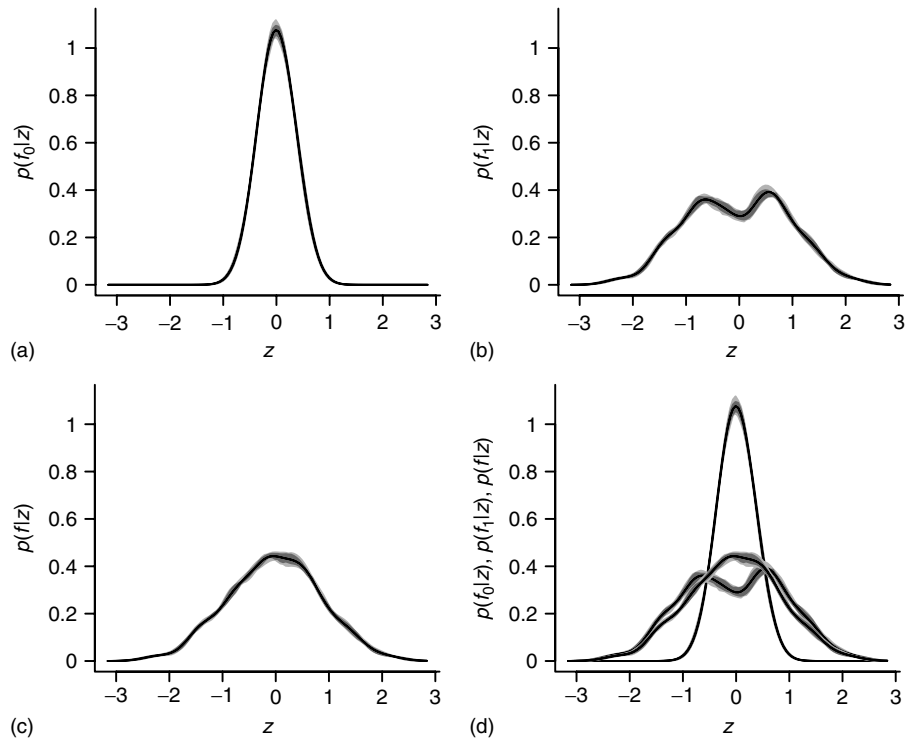
Figure 8.3 Illustration of the posterior distributions for the mixture densities in the Do *et al.* (2005) model, found for the Alon *et al.* (1999) cancer data set. (a–c) show f_0, f_1 and f; (d) shows all three. Each curve is a draw from the posterior distribution of the relevant mixture component. [Figure courtesy of Peter Müller.]

are expressed more in tumors than in normal samples, and genes that are expressed more in normal samples. The estimate of the proportion of differentially expressed genes was around 1 % in this data set. The performance of this model will depend on the number of within-replicate differences that are used to calibrate f_0 and on the information introduced in the hyperprior specification. When there are only a few replicates, the mixture might be close to nonidentifiability.

Broët *et al.* (2002) suggest another model using a mixture at the data level to classify genes. Here the data is first transformed with a linear model to produce normalized log fold changes d_g. The d_g are modeled using a fully Bayesian mixture of normals, which includes estimation of the proportion of differentially expressed genes (the weights in the mixture). The number of components in the mixture K is not restricted to 2. There is still just one component representing the null, but several representing differentially expressed genes. This allows grouping of genes into different levels of differential expression. In fact K is not fixed in this model, but estimated, in a fully Bayesian way, using the split and merge algorithm for mixtures with an unknown number of components introduced in Richardson and Green (1997).

When mixture distributions are put on parameters of the model (prior) rather than on the data (likelihood), care must be taken to ensure identifiability of the parameters

of the mixture components. A common choice is to make the null component a point mass. This corresponds to testing the null hypothesis $\delta_g = 0$ versus the two-sided alternative.

Lönnstedt and Speed (2003), Lin *et al.* (2003) and Smyth (2004) use mixture priors on the parameter δ_g representing difference between conditions. Lönnstedt and Speed (2003) use a mixture of a point mass at zero and a conjugate normal prior on the δ_g. Smyth (2004) uses the same mixture model, but on data that has first been transformed using a linear model similar to that in (8.6) but using a robust estimation method, to obtain log fold changes. These two models are estimated using empirical Bayes methods. The proportion of true nulls is not estimated. Thus these methods produce a ranking of the genes rather than an actual estimate of how many genes are differentially expressed.

Rather than putting the mixture directly on the δ_g, Newton *et al.* (2004) propose a mixture prior on the pair of parameters μ_{g1}, μ_{g2}. Their likelihood is gamma, but the μ_{gc} still represent mean expression in the two conditions. One component of the mixture has $\mu_{g1} = \mu_{g2}$ drawn from one distribution, and the other has μ_{g1}, μ_{g2} drawn from two separate distributions. These distributions are estimated nonparametrically, using an expectation maximization (EM) algorithm. Gottardo *et al.* (2006c) has a similar mixture on the pair μ_{g1}, μ_{g2}, this time using normal priors, and a fully Bayesian estimation method, including estimating the proportion of differentially expressed genes. Reilly *et al.* (2003) has a model with a similar structure, but in addition incorporates prior information about certain genes being controls (and therefore not differentially expressed).

An early model for differential expression, which does not employ a mixture model on the difference between the two conditions, is that of Ibrahim *et al.* (2002). They model the data *in each condition* as that from a mixture of a point mass and a lognormal distribution, the point mass representing the threshold for genes to be unexpressed. A measure for differential expression is formed from the ratio of expectation of expression in the two conditions.

As a final comment, note that finding differentially expressed genes can also be cast in a multivariate framework. This approach was adopted by Ishwaran and Rao (2003) who use multivariate shrinkage effected via a continuous version of the spike and slab variable selection model (see Section 8.5.1 for a discussion of variable selection approaches). They propose to detect differentially expressed genes by formulating the problem as a linear regression. They then use a multivariate shrinkage approach to find posterior means of the differentially expressed parameters and finally they compare these values to percentiles of a standard normal distribution (with a scaling coefficient) in order to select differentially expressed genes.

8.3.4   Classifying Genes as Differentially Expressed

In differential expression problems, the aim is to produce a list of genes that are considered to be differentially expressed between the different experimental conditions. A decision rule is used to classify genes as either differentially expressed (DE) or not differentially expressed (non-DE). In Bayesian models, this will either be based on the value of some model parameter (usually the posterior mean), or on posterior probabilities of some criterion in the model, e.g. of being classified into a certain mixture component or of some parameter being above a certain threshold.

8.3.4.1 Models with Mixture Priors for Fold Changes

In the fully Bayesian mixture models described above, decisions are usually made using the posterior probabilities of a gene being allocated to the different mixture components. The mixture example given in (8.8) can also be written as

$$
\begin{aligned}
\Delta_g | z_g &\sim w_{z_g} f_{z_g}(\phi_{z_g}), \\
\mathbb{P}(z_g = k) &= w_k,
\end{aligned} \tag{8.10}
$$

where the z_g are allocation parameters, which label the mixture component to which gene g is assigned. The posterior probability of gene g being in component k is $\mathbb{P}(z_g = k|\mathbf{x})$.

Defining a loss function enables one to form the decision rule. First, denote the set of genes declared to be DE by S_1 and the set of genes called *non-DE* by S_0. In the two-component mixture models, since there are two possible classifications for each gene, there are two possible penalties for misclassification, one for false positives and one for false negatives. If the ratio of these two penalties is λ, the same for all genes, the loss function is proportional to

$$
L \propto \sum_{g \in S_0} \mathbb{P}(z_g \neq 0|\mathbf{x}) + \lambda \sum_{g \in S_1} \mathbb{P}(z_g = 0|\mathbf{x}). \tag{8.11}
$$

This is minimized by defining S_0 as the set of genes for which $\mathbb{P}(z_g = 0|\mathbf{x}) \geq 1/(1 + \lambda)$, i.e. genes are classified using a threshold on the posterior probabilities of classification in the mixture. Müller *et al.* (2007) discuss different possible loss functions and the decision rules they lead to.

The posterior probabilities can also be used to obtain an estimate of the false discovery rate, which is the ratio of false positives to total declared positives:

$$
FDR = \frac{1}{|S_1|} \sum_{g \in S_1} \mathbb{P}(z_g = 0|\mathbf{x}), \tag{8.12}
$$

(see Newton *et al.*, 2004; Broët *et al.*, 2004; Müller *et al.*, 2007). An estimate of the false nondiscovery rate (ratio of false negatives to total negatives) can be similarly defined. The false discovery rate is widely used in classical statistical analysis of gene expression data (see **Chapter 6** and **Chapter 7**). It is useful to be able to give this estimate when comparing with different analysis methods and it has generally be found in simulation studies that (8.12) gives quite accurate estimates of the true FDR.

For mixtures of more than two components, one may consider different rules. The most obvious would be to assign genes to the component with highest probability, i.e. gene g is assigned to component $k = \max_{k'} \mathbb{P}(z_g = k'|\mathbf{x})$. However, when there are more than two components, this can lead to genes being declared DE (in a particular component) when their posterior probability of being classified into that component is low. For example, with four components, a gene that has almost equal probability of being classified in all components can be declared DE (into the best component for that gene) with posterior probability of 0.26 of being in that component. An alternative, but more conservative, suggestion would be to classify into one of the components representing DE only those

genes for which the corresponding posterior probability of belonging to that component is above a set threshold, e.g. 50 % or higher. Otherwise the genes are classified into the null. Again, evaluating the associated FDR of such rules will guide the choice of appropriate thresholds. Such a rule was used in a related context of modeling DNA copy number changes (gains or losses) in comparative genomic hybridization experiments by a spatially structured mixture model with three components (gain, loss, normal) in Broët and Richardson (2006). For typical noise-to-signal ratio, the authors found that classifying DNA sequences with a posterior probability above 0.8 into the gain or loss components gave good operational characteristics in this context, whereas the Bayes rule had poorer performance.

When mixture models are estimated using empirical Bayes methods, without an estimate of the number of genes in the null, the posterior probability of being allocated to the null can only be estimated up to a constant. In this situation, the posterior odds ratio can be used to rank genes:

$$Odds_g = \frac{\mathbb{P}(z_g = 0|\mathbf{x})}{\mathbb{P}(z_g \neq 0|\mathbf{x})}, \tag{8.13}$$

(Lönnstedt and Speed, 2003; Smyth, 2004).

8.3.4.2 Models with Nonstructured Priors on Fold Change Parameters

In the nonmixture methods of the previous section, a variety of measures of differential expression are used to classify genes. Baldi and Long (2001) and Smyth (2004) propose the so-called regularized or moderated *t*-statistics. These consist of the Bayesian posterior mean estimate of the log fold change parameter, divided by a shrunken estimate of standard deviation. This shrunken estimate is the square root of the posterior mean of the variance parameter, shrinkage being provided by the exchangeable prior on the variances estimated in an expectation maximization (EM) framework.

The model used by Bhattacharjee *et al.* (2004) allows gene variances to take one of three values (a mixture on gene variability). These can be used to classify genes into groups based on their variability within and between tissues.

With a noninformative prior on the δ_g, Lewin *et al.* (2006) proposed a decision rule based on a threshold δ_{cut} set according to a biologically interesting level of differential expression. Differential expression is defined as δ_g being greater than δ_{cut}, corresponding to an interval null hypothesis with the interval fixed *a priori*. The decision rule is that genes are declared to be differentially expressed if the posterior probability $\mathbb{P}(|\delta_g| > \delta_{cut}|\mathbf{x})$ is greater than some threshold probability (e.g. 0.5). This rule combines statistical and biological significance.

When the interval for an interval null hypothesis is required not to be fixed *a priori*, Bochkina and Richardson (2006) suggest two types of decision rule based on tail posterior probabilities. They define a loss function

$$L \propto \sum_{g \in S_0} I[|\delta_g| > \theta(\sigma_g)|\mathbf{x}] + \lambda \sum_{g \in S_1} I[|\delta_g| \leq \theta(\sigma_g)|\mathbf{x}]. \tag{8.14}$$

They consider two possibilities for θ: first, $\theta \propto \sigma_g$, which leads to a decision rule where S_0 is defined as the group of genes with $\mathbb{P}(|\delta_g/\sigma_g| \leq T^{\alpha}|\mathbf{x}) \geq 1/(1 + \lambda)$, which is an analog

of a t-statistic procedure. The second choice is with constant θ, in which case the decision rule defines S_0 as genes with $\mathbb{P}(|\delta_g| \leq \delta_g^\alpha|\mathbf{x}) \geq 1/(1 + \lambda)$. A heuristic argument is used to choose the thresholds T^α and δ_g^α; these are defined as the percentiles of the distribution of δ_g/σ_g or δ_g found by hypothetically conditioning on $\bar{x}_{g2} - \bar{x}_{g1} = 0$ in the model, e.g. $f(\delta_g|\bar{x}_{g2} - \bar{x}_{g1} = 0, s_g^2)$. Bochkina and Richardson (2006) also consider a one-sided rule with threshold zero. This is shown to be equivalent to the moderated t-statistic of Smyth (2004) (when the same variance model is used).

8.3.5 Multiclass Data

A number of models that extend the methods used for differential expression in two conditions to compare expression in several conditions or classes have been proposed. These might be used, e.g. to compare the actions of several drugs and a control sample simultaneously, or to compare different tumor samples. As with the mixture models described previously, it can be useful to describe the classification of genes in terms of null and alternative hypotheses. There are a number of different choices of alternative hypothesis for multiclass data. Here we discuss models that use hypotheses of the type 'the gene is differentially expressed (or not) in at least one condition', without distinguishing what the condition is. Section 8.4 deals with models that classify genes by clustering them according to the pattern of expression across the experimental conditions, thus distinguishing between being differentially expressed in condition 4 only and being differentially expressed in condition 2 only, for example. An intermediate approach is taken by Ishwaran and Rao (2005b) who, similarly to their work on differential expression, formulate the multiclass analysis as a multivariate regression problem, use variable selection and shrinkage to output lists of significant genes between any two conditions and then use these lists to highlight patterns of interest between the conditions.

A common formulation is the classical analysis of variance (ANOVA) model, which tests the null hypothesis 'the gene has the same expression in all conditions' versus the alternative 'the gene has differential expression in at least one condition'. This is used in a Bayesian framework by Broët *et al.* (2004), who start with an F-statistic for each gene. These F-statistics are transformed to the normal scale and modeled with a two-component mixture to classify genes as differentially expressed or not differentially expressed. The null component is a standard normal, while the alternative component is modeled semiparametrically with a mixture of normals. This type of formulation is also suggested by Smyth (2004), who uses moderated F-statistics, using shrunken estimates of variances as with the moderated t-statistics (see Section 8.3.4).

A slightly different formulation is considered by Bochkina and Richardson (2006), who use alternative hypotheses such as 'the gene is DE in a set of pairwise comparisons of interest' versus a compound null, which is the opposite of this, i.e. genes are only selected as being of interest if they show changes in a predefined set of comparisons of interest.

8.4 CLUSTERING GENE EXPRESSION PROFILES

When gene expression is measured in several conditions simultaneously, the data is in the form of a matrix x_{gc} with the index c taking more than two values, $c = 1, \ldots, n$. In

this case, the interest focuses on finding groups of genes that have the same pattern of coexpression across the conditions. These patterns are called *expression profiles*.

8.4.1   Unordered Samples

The models dealt with in this section are designed for experiments in which there is no special ordering of the different conditions, e.g. several tumor samples. Most commonly, this sort of data has no replicate measurements, or at any rate samples from different individuals are not treated as replicates. When replicates under different conditions have been measured, the modeling of the profiles can take this into account to improve the classification (Medvedovic *et al.*, 2004; Dahl, 2006). Even when there are no replicates, the different samples can however be used together to estimate gene variances, even though the samples have different expression levels.

Most of these models are based on a similar linear model to that in (8.6). It is convenient to write the gene profile as a vector:

$$\vec{x}_g = \vec{\mu}_g + \varepsilon_g, \tag{8.15}$$

where the vectors $\vec{x}_g$ and $\vec{\mu}_g$ are of length n. The normalization term is omitted here, as most published Bayesian models do not include this term (instead requiring the data to have been transformed in a suitable way beforehand).

Choices for the distribution used to model gene variability are similar to those discussed in Section 8.3.2. Here our focus is on modeling the mean expression levels in the different groups. As with the differential expression models, a mixture model can be put on the $\vec{\mu}_g$ parameters, or on the data directly. The correspondence with null and alternative hypotheses can also be carried forward to this type of model, though now there may be several alternatives. The null is 'no difference in expression across conditions'. The alternatives are usually all possible patterns showing some difference, though Kendziorski *et al.* (2003) allow the number of alternative patterns to be restricted, which is a useful feature for large numbers of experimental conditions.

The probability of expression (POE) model of Parmigiani *et al.* (2002) and Garrett-Mayer and Scharpf (2006) is a simple three-component mixture model on the data. Its aim is to estimate allocation probabilities for each gene and condition to one of three groups: reference, under-, and overexpressed. In its simplest form, biological information is available to classify some indices c as giving 'normal', i.e. reference values. Reference values are modeled as arising from a normal distribution with additive gene effects plus global condition effects, whereas the under and over components are assumed to be uniformly left or right shifted from the reference mean with a range to be estimated. All unknown parameters are given prior distribution and the mixture is estimated in a fully Bayesian way via MCMC algorithms, (Parmigiani *et al.*, 2002). The output of this mixture model is a simple transformation of the data matrix into a probability scale based on an underlying assumption that the information in the expression values is essentially categorical. In contrast to other work described below, clustering of the probabilities to define interesting subgroups of genes is not attempted within the model, but the authors suggest that data mining tools may be used at a second stage.

Kendziorski *et al.* (2003), Gottardo *et al.* (2006c) and House *et al.* (2006) suggest models for clustering gene expression profiles using a mixture model on the parameters $\vec{\mu}_g$. These models are extensions of those used in differential expression, and are

formulated via combinations of point masses and continuous distributions for the various hypotheses/clusters. Kendziorski *et al.* (2003) extend the model of Newton *et al.* (2004), described in the differential expression section, to a mixture model on gene expression profiles. As an example, when there are three experimental conditions, the gene expression parameters are $\mu_{g1}, \mu_{g2}, \mu_{g3}$. There are five possible patterns for a profile over three conditions: one pattern with $\mu_{g1} = \mu_{g2} = \mu_{g3}$, three patterns with two of the μ_{gc} equal to each other and the third drawn from a separate distribution, and one pattern where all three μ_{gc} are drawn from different distributions. Their implementation allows the restriction to a few interesting patterns, as the number of possible patterns increases rapidly with the number of conditions. One version of the model of Kendziorski *et al.* (2003) is an extension of that used in Newton *et al.* (2004) where the likelihood is a gamma and the mean expression parameters have gamma priors. They also look at a version with lognormal likelihood and normal priors. Gottardo *et al.* (2006c) propose a similar model for profiles, implemented for three experimental conditions. This model is an extension of their differential expression model mentioned in Section 8.3.3, using log normal likelihood and normal priors. They automatically consider all possible patterns. House *et al.* (2006) give another similar model, this time implemented in five conditions.

Vogl *et al.* (2005) gives an example of a mixture model on the data (see Section 8.3.3 for more explanation on mixture models). In this case $\vec{\mu}_g$ is replaced by $\vec{\mu}_k$ in (8.15), where k labels the mixture component and the mixture allocation parameter z_g is equal to k. This model assumes a normal distribution for each mixture component, with variance σ_k^2, i.e. equal variance for all genes in the same component. Note that gene variances integrated over different allocations will not be equal for all genes, as different genes will be allocated to different combinations of mixture components. The prior used in Vogl *et al.* (2005) for the $\vec{\mu}_k$ is a conjugate prior, with independence between different experimental conditions. The number of clusters $k = 1, \ldots, K$ is estimated in the model along the lines of Richardson and Green (1997). This model does not in fact automatically include the null cluster of equal expression in all conditions. Figure 8.4 shows some of the gene profiles found by Vogl *et al.* (2005) for the Spellman cell-cycle data (Spellman *et al.*, 1998). Different clusters of genes peak at different phases of the cell cycle.

Rather than using finite mixture models for clustering profiles, a number of authors (Medvedovic *et al.*, 2004; Dahl, 2006; Lau and Green, 2006a; 2006b) have recently developed fully Bayesian profile clustering, based on DPMs. In this setup, it is assumed that gene profiles characterized by parameters $\vec{\mu}_g$, $g = 1, \ldots, p$ are clustered according to a tractable distribution on partitions corresponding to the Dirichlet process and that, within each cluster, the profiles follow the same distribution. DPM is a popular formulation for implementing clustering and partitions models. Indeed, besides their representation as infinite mixtures (see Section 8.3.3), DP models with baseline distribution G^* and scalar α can be equivalently defined via a prior structure on the space $\mathbf{C}$ of partitions of p items (here the genes) into K clusters, $\{1, \ldots, p\} = \bigcup_{k=1}^{K} C_k$, with p_k items in cluster C_k, a joint distribution on the partition given by:

$$p(C_1, C_2, \ldots, C_K) = \frac{\alpha^K \Gamma(\alpha) \prod_{k=1}^{K} (p_k - 1)!}{\Gamma(\alpha + p)},$$

associated i.i.d. draws of $\vec{\mu}_k^*$ from G^*, $k = 1, \ldots, K$, and setting $\vec{\mu}_g = \vec{\mu}_k^*$ if $g \in C_k$.

Authors differ in their choice of specification of the base measure distribution, and in whether they choose a fully conjugate specification between the mixture kernels for the
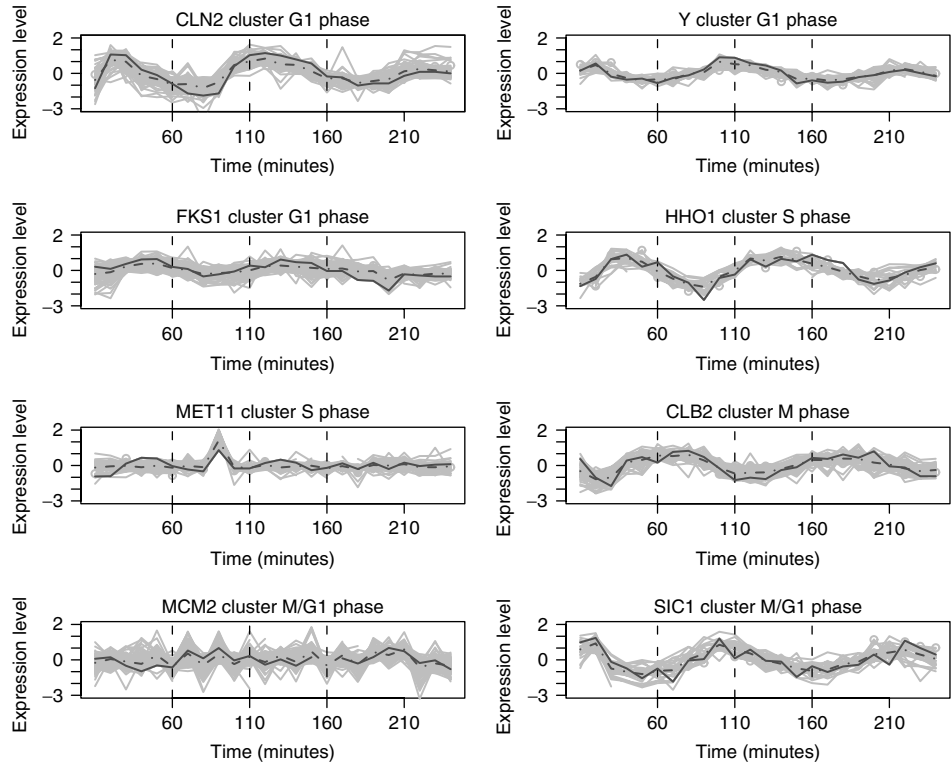
Figure 8.4   Clusters of cell-cycle-regulated genes found in the Spellman *et al.*, 1998 data using the model of Vogl *et al.*, 2005. [Reprinted from Vogl, C., Sanchez-Cabo, F., Stocker, G., Hubbard, S., and Wolkenhauer, O. (2005). A fully Bayesian model to cluster gene-expression pro les. *Bioinformatics* **21**, ii 130–ii 136, by permission of Oxford University Press.]

data part of the model and the base measure. In Lau and Green (2006b), the standard DPM approach is extended by replacing the DP model by a variant in which there is a background cluster not exchangeable with the others and in which there is a different prior distribution of the cluster-specific parameters. Fully conjugate specification (Dahl, 2006; Lau and Green, 2006a; 2006b) is computationally advantageous as all the cluster parameters can be integrated out and efficient MCMC algorithms that update solely the partition can be used. Within cluster, posterior distributions for the parameters are then sampled, conditional on the partition.

In general, drawing inference from the complex posterior clustering distribution is not straightforward and, as the number of partitions grows rapidly with increasing p, recording the partition with the highest posterior probability does not guarantee that it is close to the posterior mode. Medvedovic *et al.* (2004) suggest computing the pairwise posterior probabilities for two genes to be in the same cluster and then postprocessing this output by traditional hierarchical clustering algorithms. Dahl (2006) proposes that, among all the observed clustering, the clustering that minimizes, in the least square sense, the distance between its 0-1 association matrix and the estimated pairwise posterior probabilities be chosen. Lau and Green (2006a) have formulated a Bayesian solution to

define the optimal clustering that optimizes a posterior expected loss function. This loss function penalizes pairs that are clustered together when they should not be and vice versa. They derived an efficient approximation to define the optimal clustering. Heard *et al.* (2006a) propose a hierarchical algorithm that approximates the posterior mode in a Bayesian clustering model without requiring MCMC computations (see next section for details).

It is possible to apply the models presented above to time-course or dose−response data, where there is an ordering of the samples, e.g. in Medvedovic *et al.* (2004) and Vogl *et al.* (2005). In these examples, the unordered sample models work well. However, for data with less pronounced patterns, it is better to use models that take into account the ordering information, such as those presented in the next section.

8.4.2 Ordered Samples

Models for ordered samples also usually start with a linear model,

$$x_{gt} = \mu_{gt} + \varepsilon_{gt}, \qquad (8.16)$$

where again we omit the normalization term, as this is usually assumed to be already taken care of. We use the index t for the ordered data points. For convenience we refer to these as *time points*, though they could be any ordered data. This type of data tends not to have repeated measurements for the same time point. Because of this, it is not possible to estimate separate μ_{gt} for each gene and perform clustering on them. The mixture prior for clustering must be at the data level, i.e. μ_{gt} is replaced by μ_{kt}, where k labels the mixture component to which gene g is allocated.

Two broad classes of models for dependence between time points have been proposed in the literature. One class models the parameter at any given time t in terms of the previous time point or several time points. The other class uses parametric basis functions to give a shape to the parameters across the time points. Note also that the formulation presented in Lau and Green (2006a; 2006b) allows the structuring of $\vec{\mu}_g$ as a linear function of a fixed set of covariates, in particular time (or function thereof), and can thus be used effectively for both ordered and unordered samples.

Ramoni *et al.* (2002) implement the first class of model, using an autoregressive (AR) prior on the μ_{kt} (thus μ_{kt} is regressed on $\mu_{k,t-1}, \ldots, \mu_{k,t-q}$, where q is the order of the AR prior). The AR prior assumes a stationary time series, so it will not be appropriate for many types of microarray data, especially as data is often measured at irregular time points. The clustering used in this method is hierarchical and agglomerative, that is, the posterior space of clusters is not explored fully, but a path is taken through the space in a similar way to classical hierarchical clustering methods. The scoring function used to decide which clusters are to be merged is based on ratios of posterior probabilities of the original and merged partitions.

A more flexible model in this class is given in Wakefield *et al.* (2003) and Zhou and Wakefield (2006), who use a random walk model on the μ_{kt}:

$$\mu_{kt} = \mu_{k,t-1}, +u_t, \qquad (8.17)$$

where $u_t \sim N(0, |X_t - X_{t-1}|\tau^2)$ and X_t is the value of time at the tth time point. Thus the closer adjacent time points are, the more dependent they are. This model is fitted in a fully Bayesian way, with the number of clusters estimated as part of the model. Inference is

made on the basis of posterior probability of cluster membership, with particular attention given to finding pairs of genes with high probability of being allocated to the same cluster. This is to find genes that are coexpressed.

A model in the second class is proposed by Heard *et al.* (2006a). This uses splines (with degree to be chosen) as basis functions for the trajectories of the genes over time,

$$\mu_{kt} = \sum_h X_{th} \beta_{hk}, \tag{8.18}$$

where X_{th} is the (fixed) value of the hth basis function at time t and β_{hk} is the coefficient of the hth basis function for cluster k. By using a fully conjugate specification, the joint distribution of the data conditional on any partition can be computed in closed form. Hence the posterior probability of any partition can also be evaluated. The clustering method exploits this and proceeds by an agglomerative algorithm similar to that used by Ramoni *et al.* (2002), in order to find the partition that approximates the posterior mode. Heard *et al.* (2006b) extend this model to time series taken in a number of different conditions, and estimate covariance between time series in different conditions.

Wakefield *et al.* (2003) also fit a basis function model, for periodic data

$$\mu_{gt} = A_g \sin(wX_t) + B_g \cos(wX_t), \tag{8.19}$$

applied to a cell-cycle data set.

8.5   MULTIVARIATE GENE SELECTION MODELS

In the previous sections, we have discussed *gene expression association studies* where the aim is to find gene expression changes that relate to biological outcomes by comparing, for each gene, their differential expression under different conditions, and *profile clustering* where the interest is to find patterns of coexpressed genes across different experimental conditions, in order to understand pathways. In this section, we are concerned with a different, but related, problem, which is that of using gene expression for phenotype prediction. Our aim here is to build multivariate molecular profiles based on combination of the expression of a subset of genes that can characterize different phenotypes (e.g. clinical outcomes). We are thus in the framework of multivariate regression and classification models. The specific difficulty of genomic applications is that there are typically many more covariates than samples: the so-called large p (thousands of genes), small n (50–100 samples) regression paradigm, and consequently standard regression/discrimination techniques do not apply. Further, the interest is in finding parsimonious regression models that include only small subsets of genes so that biological interpretation and validation can be attempted.

Bayesian approaches to multivariate gene selection have broadly followed two related lines of development: (1) regression methods with covariate selection, (2) multivariate regression with shrinkage priors that favor sparsity. We shall review these in turn. Mostly, we shall discuss so-called supervised classification situations where the characteristic of the samples that one wants to predict are known. Variable selection can also be performed simultaneously with the task of uncovering clustering patterns of the samples in an unsupervised manner.

8.5.1 Variable Selection Approach

Suppose that we have potentially p predictor variables, each measured on a set of n samples: x_{gc} with $g = 1, \ldots, p$ and $c = 1, \ldots, n$. Thus, for each predictor variable g, we have a vector of n measurements $\vec{x}_g$. For the present, the outcome variable, $y_c, c = 1, \ldots, n$, can be continuous (e.g. measuring a biomarker) or categorical (e.g. encoding a cancer subtype) and we denote the vector of regression parameters linking $\mathbf{X}$ and $\mathbf{Y}$ as $\boldsymbol{\beta}$, (these being the matrices for predictors and outcomes, respectively). Thus β_g is the regression parameter corresponding to the covariate $\vec{x}_g$.

Bayesian variable selection (BVS) is usually implemented through a hierarchical model, where all possible 2^p models are represented by a p-dimensional indicator variable $\boldsymbol{\gamma}$:

$$\gamma_g = \begin{cases} 1 & \text{variable (gene) } g \text{ is included} \\ 0 & \text{variable (gene) } g \text{ is excluded} \end{cases}.$$

A prior on the model space can be specified via a prior $p(\boldsymbol{\gamma})$. A common choice is $p(\boldsymbol{\gamma}) = \prod_{g=1}^{p} \pi^{\gamma_g} (1 - \pi)^{1-\gamma_g}$, and by choosing small π, the number of variables selected can be controlled. Alternatively, a beta prior distribution can be assumed for π and the sparsity of the regression only controlled by the choice of prior for the βs.

This generic approach to variable selection, often referred to as the *spike and slab approach*, was taken in Mitchell and Beauchamp (1988), George and McCulloch (1993; 1997) and in many subsequent papers (Clyde, 1999; Brown *et al.*, 1998; 2002). Much of the work on BVS was developed for linear models where y_c is continuous. In this context, authors differ in the choice of the prior distribution for $\boldsymbol{\beta}$, in particular whether the components of $\boldsymbol{\beta}$ are treated as independent, and whether a conjugate formulation is chosen so that the prior on $\boldsymbol{\beta}$ includes the noise parameter of the linear model. Typically, the prior for $\boldsymbol{\beta}$ is formulated via a mixture. Most models define a point mass at zero for β_g when $\gamma_g = 0$, while when $\gamma_g = 1$, large variances are favored with a distribution to be specified. Ishwaran and Rao (2003; 2005a; 2005b) use a modified spike and slab model that assumes a continuous bimodal prior for β_g, a scale mixture of two centered normals, one having a small variance. They show that this prior is useful in gene expression; see Sections 8.3.3 and 8.3.5. In the common case of a prior for β_g with a point mass at zero, for ease of notation, we shall denote all nonzero elements of $\boldsymbol{\beta}$ by β_γ, and correspondingly, we denote the columns of X corresponding to those elements of $\boldsymbol{\gamma}$ equal to 1 by $\mathbf{X}_\gamma$.

A standard choice of prior for β_γ is the so-called g prior, where $\beta_\gamma \sim N(0, c(X_\gamma^T X_\gamma)^{-1})$, where c is a positive scale factor to be chosen. Similarly, if independent normals are specified for the components of $\boldsymbol{\beta}$, again a scalar has to be chosen. These choices influence the sparsity of the final regression model (Chipman *et al.*, 2001) and a full understanding of this aspect is the object of current research.

In the microarray context, because we are in a 'large p, small n' situation, the posterior distribution over the model space of variable dimensions is multimodal. Moreover, full posterior inference for the entire model space of size 2^p is not feasible if p is larger than about 20. Hence, MCMC methods are rather used as stochastic search algorithms with the aim to quickly find many regions of high posterior probability. The Markov chain needs to move quickly around the support of the posterior distribution and, as usual, it is useful to integrate out as many parameters as possible. For this reason, conjugate settings have been favored in the linear model. When proposing changes to $\boldsymbol{\gamma}$, a key question is how to propose sensible changes to the regression vector $\boldsymbol{\beta}$. The current parameter values

may be of little relevance in this case and joint moves that simultaneously update γ and β produce an improvement (Holmes and Held, 2006).

Much of the application of variable selection in microarrays has concerned binary or categorical variables rather than the linear model. Typically, samples are classified as good or poor prognosis or linked to different clinical subentities, like subtypes of cancer. There is no immediate conjugate formulation of Bayesian categorical regression, but following the approach of Albert and Chib (1993), probit regression can be efficiently implemented through the use of latent auxiliary variables, which allows integration of the regression coefficients in the full conditional distribution of the indicator variables γ. This approach was taken by Lee *et al.* (2003) and Sha *et al.* (2004). These authors have implemented different MCMC schemes for updating γ (Gibbs sampling in Lee *et al.* (2003), which will tend to be slow mixing, Metropolis with add/delete/swap moves in Sha *et al.* (2004)). In a recent paper, Holmes and Held (2006) have proposed an auxiliary variable formulation of the logistic model, which allows, in a similar way to the probit model, integration of the regression coefficients when updating the indicator variables, thus improving mixing. For both probit and logistic regression models, the calibration of the prior distribution of the regression coefficients again influences the outcome of the variable selection process. In this respect, the logistic regression, which is more commonly used for binary regression, is easier to calibrate than the probit model as it has heavier tails and so exhibits less sensitivity.

In general, MCMC variable selection algorithms in high dimension are difficult to implement owing to slow convergence. Recent developments in stochastic simulation algorithms, such as population-based reversible jump MCMC and the use of parallel tempered chains seem promising (Jasra *et al.*, 2006). An alternative search algorithm, the *shotgun stochastic search* method, which is close in spirit to MCMC but aims to search rapidly for the most probable models, has been recently proposed by Hans *et al.* (2007). Note that it is a discussion point whether to report models with high posterior probabilities and their associated variables, or to extract marginal information about the selected variables by looking at their marginal posterior probabilities of inclusion (Sha *et al.*, 2004).

We end this section by referring to the recent work of Tadesse *et al.* (2005) and Kim *et al.* (2006) where variable selection and clustering of the samples are performed simultaneously. This joint modeling is motivated by the remark that using a high-dimensional vector of gene expression to uncover clusters among the samples might not be effective and on the contrary can tend to mask existing structure, while a more parsimonious model that selects a only a small subset of covariates to inform the clustering is more easily interpretable. Such joint modeling was discussed in a Bayesian context in two related papers, which differ in their model for the clustering structure. Tadesse *et al.* (2005) formulate their clustering structure using a finite mixture of multivariate normals with a variable number of components and use reversible jump techniques to explore different structures, while Kim *et al.* (2006) exploit the computational benefits of DP mixtures.

8.5.2 Bayesian Shrinkage with Sparsity Priors

An alternative approach to BVS for selecting a small number of regressors is to use a hierarchical formulation of the regression problem with a prior on the regression coefficients that favors sparsity. Effectively, a large number of regression coefficients are essentially set to zero by having very small posterior values. Note that different choices of prior and hierarchical structures can be interpreted as different choice of penalties if one

adopts the point of view of penalized likelihood, framework into which ridge regression and other shrinkage methods can be cast.

A general formulation that encompasses many of the models that have been proposed is that of scale mixture of normals. In this formulation, the regression coefficients β_g are given independent normal priors: $\beta_g \sim N(0, \tau_g)$, and the variances τ_g are themselves given a hyperprior distribution: $\tau_g \sim p(\tau_g)$. Choice of this prior distribution leads to a different kind of sparsity for the βs, but all priors used have in common the desirable feature that the integrated prior for β is a heavy tail distribution with a peak around zero, thus favoring only a small number to be substantially different from zero. Note that a Laplace prior for β_g, which corresponds to a Lasso type penalization, can also be written as a scale mixture of normals with $p(\tau_g)$ being a one-parameter exponential distribution, (Griffin and Brown, 2005).

Bae and Mallick (2004) implement three different choices of prior for τ_g in the context of gene expression studies: an inverse γ with two hyper parameters that are chosen in order to favor large variances, a Laplace prior with one parameter and a Jeffreys improper prior (which implies an improper prior of the form $\frac{1}{\beta_g}$ for β_g). They found that Jeffreys prior induces more sparseness than the Laplace prior and yields good performance. There is currently a lot of interest in using general families of scale mixtures and in calibrating them for efficient inference in high-dimensional setups (Griffin and Brown, 2005).

Sparsity priors can also be used in the context of latent factor models (West, 2003; Lucas *et al.*, 2006). Modeling high-dimensional data via latent factor models is a powerful dimension reduction technique that allows the identification of patterns of covariation among genes. In their application of factor models to the analysis of gene expression data, West (2003) and Lucas *et al.* (2006) further structure the factor loading matrix to encourage sparsity via a mixture prior with point mass at zero. A biological interpretation of the factors as potentially representing biological pathways is then derived by examining the list of genes most weighted on each factor.

Acknowledgments

Related Chapters

Chapters 6; **Chapters 7**.

REFERENCES

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon

tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences* **96**, 6745–6750.

Bae, K. and Mallick, B. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* **20**, 3423–3430.

Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of micro-array data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

Bhattacharjee, M., Pritchard, C.C., Nelson, P.S. and Arjas, E. (2004). Bayesian integrated functional analysis of microarray data. *Bioinformatics* **20**, 2943–2953.

Bochkina, N. and Richardson, S. (2006). Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics*, in press.

Broët, P., Lewin, A., Richardson, S., Dalmasso, C. and Magdelenat, H. (2004). A mixture model based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics* **20**(16), 2562–2571.

Broët, P. and Richardson, S. (2006). Bayesian hierarchical model for identifying change in gene expression from microarray experiments. *Bioinformatics* **9**, 671–683.

Broët, P., Richardson, S. and Radvanyi, F. (2002). Bayesian hierarchical model for identifying change in gene expression from microarray experiments. *Journal of Computational Biology* **9**, 671–683.

Brown, P., Vannucci, M. and Fearn, T. (1998). Multivariate Bayes variable selection and prediction. *Journal of the Royal Statistical Society, Series B* **60**(3), 627–641.

Brown, P., Vannucci, M. and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B* **64**(3), 519–536.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC.

Ceccarelli, M. and Antoniol, G. (2006). A deformable grid-matching approach for microarray images. *IEEE Transactions on Image Processing* **15**, 3178–3188.

Chipman, H., George, E. and McCulloch, R. (2001). The practical implementation of Bayesian model selection (discussion). *Institute of Mathematical Statistics. Lecture Notes - Monograph Series* **38**, 67–134.

Clyde, M. (1999). Bayesian model averaging and model search strategies. In *Bayesian Statistics 6*, J. Bernardo, J. Berger, A. Dawid and A. Smith, eds, Proceedings of the Sixth Valencia International Meeting. Oxford University Press, pp. 157–185.

Dahl, D. (2006). Model-based clustering for expression data via a dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics*, K.-A. Do, P. Müller and M. Vannucci, eds. Cambridge University Press, pp. 201–218.

Do, K., Müller, P. and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society, Series C* **54**, 627–644.

Do, K.-A., Müller, P. and Vannucci, M. (2006). *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Frigessi, A., van de Wiel, M., Holden, M., Svendsrud, D., Glad, I. and Lyng, H. (2005). Genome-wide estimation of transcript concentrations from spotted cDNA microarray data. *Nucleic Acids Research* **33**(17), e143.

Garrett-Mayer, E. and Scharpf, R. (2006). Models for probability of under-and overexpression: the POE scale. In *Bayesian Inference for Gene Expression and Proteomics*, K.-A. Do, P. Müller and M. Vannucci, eds. Cambridge University Press, pp. 137–154.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

George, E. and McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Gottardo, R., Besag, J., Stephens, M. and Murua, A. (2006a). Probabilistic segmentation and intensity estimation for microarray images. *Biostatistics* **7**(1), 85–99.

Gottardo, R., Raftery, A.E., Yeung, K.Y. and Bumgarner, R.E. (2006b). Quality control and robust estimation for cDNA microarrays with replicates. *Journal of the American Statistical Association* **101**, DOI 10.1198.

Gottardo, R., Raftery, A.E., Yeung, K.Y. and Bumgarner, R.E. (2006c). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics* **62**, 10–18.

Griffin, J. and Brown, P. (2005). Alternative prior distributions for variable selection with very many more variables than observations. Technical report. `http://www2.warwick.ac.uk/fac/sci/statistics/staff/academic/griffin/personal/vspaper.pdf`.

Hans, C., Dobra, A. and West, M. (2007). Shotgun stochastic search in regression with many predictors. *Journal of the American Statistical Society*, to appear.

Heard, N.A., Holmes, C.C. and Stephens, D.A. (2006a). A quantitative study of gene regulation involved in the immune response of anopheles mosquitoes: an application of Bayesian hierarchical clustering. *Journal of the American Statistical Society* **101**, 18–29.

Heard, N.A., Holmes, C.C., Stephens, D.A., Hand, D.J. and Dimopoulos, G. (2006b). Bayesian coclustering of anopheles gene expression time series: study of immune defense response to multiple experimental challenges. *Proceedings of the National Academy of Sciences* **102**, 16939–16944.

Hein, A.-M.K., Richardson, S., Causton, H.C., Ambler, G.K. and Green, P.J. (2005). BGX: a fully Bayesian gene expression index for Affymetrix GeneChip data. *Biostatistics* **6**(3), 349–373.

Hein, A.-M.K. and Richardson, S. (2006). A powerful method for detecting differentially expressed genes from GeneChip arrays with no replicates. *BMC Bioinformatics* **7**, 353.

Holmes, C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145–168.

House, L., Clyde, M. and Huang, Y.-C.T. (2006). Bayesian identification of differential gene expression induced by metals in human bronchial epithelial cells. *Bayesian Analysis* **1**, 105–120.

Ibrahim, J.G., Chen, M.-H.C. and Gray, R.J. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association* **97**, 88–99.

Ishwaran, H. and Rao, J. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association* **98**, 438–455.

Ishwaran, H. and Rao, J. (2005a). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics* **33**, 730–773.

Ishwaran, H. and Rao, J. (2005b). Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association* **100**, 438–455.

Jasra, A., Stephens, D. and Holmes, C. (2006). Population based reversible jump Markov chain Monte Carlo. Technical Report, Imperial College.

Kendziorski, C., Newton, M., Lan, H. and Gould, M.N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**, 3899–3914.

Kim, S., Tadesse, M. and Vanucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93**(4), 877–893.

Lau, J. and Green, P. (2006a). Bayesian model based clustering procedures. *Journal of Computational and Graphical Statistics*, in press.

Lau, J. and Green, P. (2006b). Bayesian clustering using a heterogeneous Dirichlet process, with application to parametric gene expression profiles. Technical report, University of Bristol.

Lee, K., Sha, N., Dougherty, E., Vannucci, M. and Mallick, B. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics* **19**(1), 90–97.

Lewin, A., Richardson, S., Marshall, C., Glazier, A. and Aitman, T. (2006). Bayesian modelling of differential gene expression. *Biometrics* **62**, 1–9.

Lin, Y., Reynolds, P. and Feingold, E. (2003). An empirical Bayesian method for differential expression studies using one-channel microarray data. *Statistical Applications in Genetics and Molecular Biology* **2**, Article 8.

Lönnstedt, I. and Speed, T. (2003). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. and West, M. (2006). Sparse statistical modelling in gene expression genomics. In *Bayesian Inference for Gene Expression and Proteomics*, K.-A. Do, P. Müller and M. Vannucci, eds. Cambridge University Press, pp. 155–176.

Medvedovic, M., Yeung, K. and Bumgarner, R. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics* **20**, 1222–1232.

Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**, 1023–1032.

Müller, P., Parmigiani, G. and Rice, K. (2007). FDR and Bayesian multiple comparison rules. In *Bayesian Statistics 8*, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, eds. Oxford University Press.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F. and Tsui, K. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

Newton, M., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics* **5**, 155–176.

Parmigiani, G., Garrett, E., Anbazhagan, R. and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society, Series B* **64**, 717–736.

Ramoni, M.F., Sebastiani, P. and Kohane, I.S. (2002). Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences* **99**, 9121–9126.

Reilly, C., Wang, C. and Rutherford, M. (2003). A method for normalizing microarrays using genes that are not differentially expressed. *Journal of the American Statistical Association* **98**, 868–878.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B* **59**, 731–792.

Schadt, E., Li, C., Su, C. and Wong, W. (2000). Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry* **80**, 192–202.

Sethurman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica* **4**, 639–650.

Sha, N., Vannucci, M., Tadesse, M., Brown, P., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, N., Buckley, C. and Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular singatures of disease stage. *Biometrics* **60**, 812–819.

Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* Article 3.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273–3297.

Tadesse, M., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* **100**, 602–617.

Vogl, C., Sanchez-Cabo, F., Stocker, G., Hubbard, S. and Wolkenhauer, O. (2005). A fully Bayesian model to cluster gene-expression profiles. *Bioinformatics* **21**, ii130–ii136.

Wakefield, J.C., Zhou, C. and Self, S.G. (2003). Modelling gene expression data over time: curve clustering with informative prior distributions. In *Bayesian Statistics 7*, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, eds. Oxford University Press, pp. 721–732.

Walker, S., Damine, P., Laud, P. and Smith, A. (1999). Bayesian nonparametric inference for distributions and related functions (with discussion). *Journal of the Royal Statistical Society, Series B* **61**, 485–527.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics 7*, J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West, eds, Proceedings of the Seventh Valencia International Meeting. Oxford University Press, pp. 733–742.

Zhou, C. and Wakefield, J.C. (2006). A Bayesian mixture model for partitioning gene expression data. *Biometrics* **62**, 515–525.

Inferring Causal Associations between Genes and Disease via the Mapping of Expression Quantitative Trait Loci

S.K. Sieberts and E.E. Schadt

Rosetta Inpharmatics LLC, Seattle, WA, USA

The ability to monitor transcript levels in a comprehensive fashion allows for a more general characterization of transcriptional networks and their relationship to disease and other complex physiological processes. Information on how variations in DNA impact complex physiological processes flows through the transcriptional networks. Therefore, integrating DNA variation, transcription, and phenotypic data has the potential to enhance identification of the associations between DNA variation and disease, as well as characterize those parts of the molecular networks that drive disease. Toward that end, we discuss mapping expression quantitative trait loci (eQTL) for gene expression traits and then detail a method for integrating eQTL and expression and clinical data to infer causal relationships among gene expression traits and between expression and clinical traits. We further describe methods to integrate these data in a more comprehensive manner by constructing coexpression gene networks, which leverage pair-wise gene interaction data to represent more general relationships. To infer gene networks that capture causal information, we describe a Bayesian algorithm that further integrates eQTL and expression and clinical phenotype data to reconstruct whole gene networks capable of representing causal relationships among genes and traits in the network. These emerging high-dimensional data analysis approaches, that integrate large-scale data from multiple sources, represent the first steps in statistical genetics moving away from considering one trait at a time and toward operating in a network context. Evolving statistical procedures that operate on networks will be critical to extracting information related to complex phenotypes like disease, as research goes beyond the single-gene focus. The early successes achieved with some of the methods described herein suggest that these more integrative genomics approaches to dissecting disease traits will significantly enhance the identification of key drivers of disease beyond what could be achieved by genetics alone.

9.1   INTRODUCTION

DNA microarrays have revolutionized the way we study genes and the role they play in everything from the regulation of normal cellular processes to complex diseases like obesity and cancer. High-throughput DNA microarray technologies are capable of simultaneously providing quantitative measures of RNA present in cells or tissue samples for tens of thousands of protein-coding genes, in addition to other noncoding transcribed sequences. The quantitative measures of transcript abundances in cells are often referred to as *gene expression traits*. In their typical use, microarrays allow researchers to screen thousands of genes for differences in expression or differences in how genes are connected in the network (Schadt and Lum, 2006) between experimental conditions of interest (see **Chapter 6**, Sections 4 and 5, and also **Chapters 7** and **8**). These data are often used to discover genes that differ between normal and diseased tissue, to model and predict continuous or binary measures, to predict patient survival, and to classify disease or tumor subtypes. Because gene expression levels in a given sample are measured simultaneously, researchers are able to identify genes whose expression levels are correlated, implying an association under specific conditions or more generally.

Causal associations among genes or between genes and traits have also been investigated using time series experiments, gene knockouts or transgenics that overexpress a gene of interest, RNAi-based knockdown or viral-mediated overexpression of genes of interest, and chemical activation or inhibition of genes of interest. However, recently a number of studies have demonstrated that naturally occurring DNA polymorphism can be used to help establish causal associations, since gene expression and other molecular phenotypes in a number of species have been shown to be significantly heritable and at least partially under the control of specific genetic loci (Jin *et al.*, 2001; Brem *et al.*, 2002; Klose *et al.*, 2002; Oleksiak *et al.*, 2002; Schadt *et al.*, 2003; Monks *et al.*, 2004; Morley *et al.*, 2004; Hubner *et al.*, 2005; Stranger *et al.*, 2005; DeCook *et al.*, 2006). By examining the effects that naturally occurring variations in DNA have on variations in gene expression traits in human or experimental populations, other phenotypes (including disease) can then be examined with respect to these same DNA variations and ultimately ordered with respect to genes to infer causal control (Figure 9.1) (Mehrabian *et al.*, 2005; Schadt *et al.*, 2005; Kulp and Jagalur, 2006; Lum *et al.*, 2006). The power of this integrative genomics strategy rests in the molecular processes that transcribe DNA into RNA and then RNA into protein, so that information on how variations in DNA impact complex physiological processes often flows directly through transcriptional networks. As a result, integrating DNA variation, transcription, and phenotypic data has the potential to enhance identification of the associations between DNA variation and disease, as well as characterize those parts of the molecular networks that drive disease.

A number of groups have now published on approaches for identifying key drivers of complex traits by examining genes located in regions of the genome genetically linked to a complex phenotype of interest, and then looking for colocalization of cis-acting expression quantitative trait loci (eQTL) for those genes residing in a genomic region linked to the phenotype (Jansen and Nap, 2001; Brem *et al.*, 2002; Schadt *et al.*, 2003; Monks *et al.*, 2004; Morley *et al.*, 2004; Alberts *et al.*, 2005; Chesler *et al.*, 2005; Cheung *et al.*, 2005; Schadt *et al.*, 2005; Petretto *et al.*, 2006a; 2006b). Those genes with (1) expression values that are significantly correlated with the complex phenotype of interest (including disease), (2) transcript abundances controlled by QTL that colocalize with the phenotype QTL, and
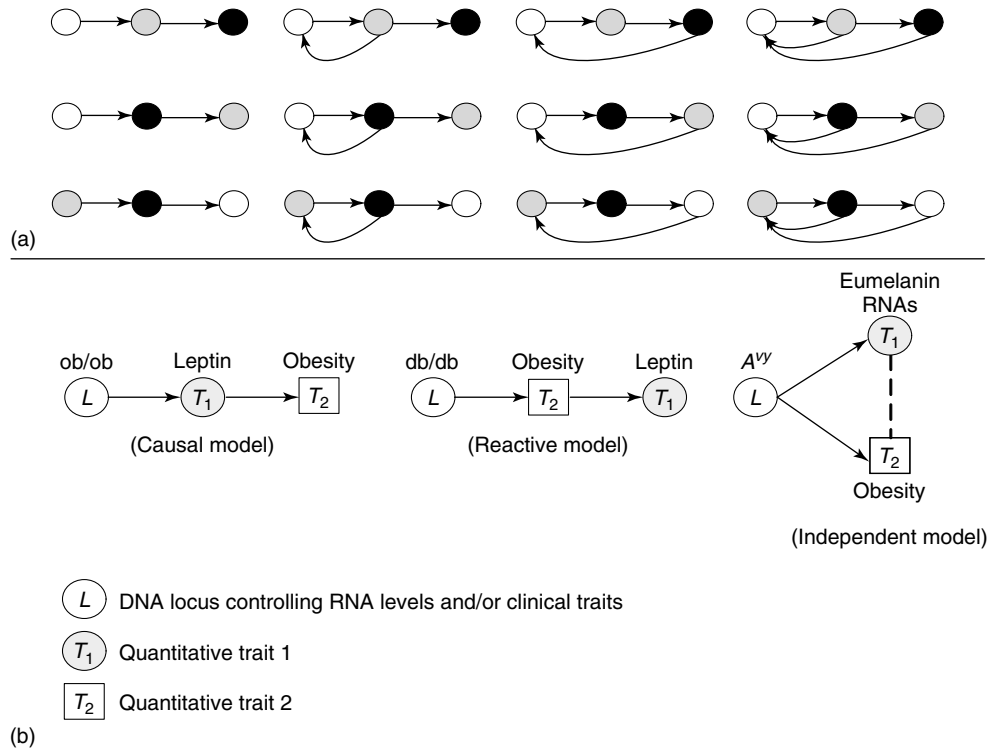
Figure 9.1 Possible relationships between phenotypes with and without genetic information. Edges between nodes in each of the graphs represent an association between the nodes. A directed edge indicates a causal association between the nodes. (a) A subset of the number of possible relationships between three variables. (b) The set of possible relationships between two traits and a controlling genetic locus, when feedback mechanisms are ignored.

(3) physical locations supported by the phenotype and expression QTL are natural causal candidates for the complex phenotype of interest. Since DNA variation leads to changes in transcription and other molecular trait activity, it can be used to partition the thousands of gene expression traits that may be correlated with a given phenotype into sets of genes that are supported as causal for, reacting to, or independent of the given phenotype. The key to the success of this approach is the unambiguous flow of information from changes in DNA to changes in RNA and protein function (Figure 9.1). That is, given that two traits are linked to the same DNA locus, and a few simplifying assumptions discussed below, there are a limited number of ways in which such traits can be related with respect to that locus (Schadt, 2005; 2006; Schadt *et al.*, 2005), whereas in the absence of such genetic information, many indistinguishable relationships would be possible, so that additional data would be required to establish the correct relationships.

Here we discuss mapping eQTL for gene expression traits and then detail a method for integrating eQTL and expression and clinical data to infer causal relationships among gene expression traits and between expression and clinical traits. We further describe methods to integrate these data in a more comprehensive manner by constructing coexpression gene networks, which leverage pair-wise gene interaction data to represent more general

relationships. This type of network provides a useful construct for characterizing the topological properties of biological networks and for partitioning such networks into functional units (modules) that underlie complex phenotypes like disease. However, these networks are, by design, undirected and so do not capture causal relationships among genes. To infer gene networks that capture causal information, we describe a Bayesian algorithm that, like the methods operating on only two or three expression traits and/or clinical traits mentioned above, integrates eQTL and expression and clinical phenotype data to reconstruct whole gene networks capable of representing direction along the edges of the network. Here, directionality among the edges corresponds to causal relationships among genes and between genes and clinical phenotypes. These emerging high-dimensional data analysis approaches, which integrate large-scale data from multiple sources, represent the first steps in statistical genetics moving away from considering one trait at a time and toward operating in a network context. Evolving statistical procedures that operate on networks will be critical to extracting information related to complex phenotypes like disease, as research goes beyond the single-gene focus. The early successes achieved with some of the methods described herein suggest that these more integrative genomics approaches to dissecting disease traits will significantly enhance the identification of key drivers of disease beyond what could be achieved by genetics alone.

9.2 AN OVERVIEW OF TRANSCRIPTION AS A COMPLEX PROCESS

Messenger ribonucleic acid (mRNA) is an intermediate product of a process that generates protein starting with DNA, where DNA is transcribed to produce RNA, and then RNA is translated to produce protein. The nucleus is where DNA is transcribed into its RNA analog, and in the case where RNA corresponds to a protein-coding gene, the RNA is processed into a mature mRNA form. The mRNA then migrates to the cytoplasm where it is translated, via tRNA, into a protein, which is often considered the functional product of the gene itself. Thus, gene expression measures corresponding to protein-coding genes are surrogates for either the state of a protein or the amount of protein that is being produced by the corresponding gene. Therefore, the amount of mRNA detected in the cytoplasm of a cell is seen to be the result of a number of different molecular processes, including the transcription process itself, the rate of RNA degradation (which is sequence specific and often sensitive to different functional forms of a protein), transport from the nucleus, and alternative splicing (a microarray probe may target a part of the transcript that is alternatively spliced).

The overall regulation of transcription is a complex process, especially in eukaryotic cells. Transcription in all types of cells begins when RNA polymerase binds to specific DNA sequences known as *promoters*, which are generally located close to the 5′ end of the gene. In order to prevent the binding of the polymerase, and thus regulate the otherwise constant transcription of a given gene sequence, many genes have negative control elements called *operators*, which when bound by specific proteins called *repressor proteins* prevent the binding of the polymerase to the promoter region. Using these simple positive and negative controllers, cells can regulate the transcription of DNA sequences in response to specific conditions or, as appropriate, for a given tissue type.

In eukaryotes, promoters and operators are usually complemented by other regulatory sequences that can operate at much longer distances, often tens of kilobytes or even longer. Some elements, like enhancers, encourage and speed up transcription, while others, like silencers, discourage and slow down transcription. While promoters and operators cause binary, on/off regulation, enhancers and silencers can have much more subtle effects on transcription. Nevertheless, mutations in any of these sequence elements can alter gene expression at the basal level or affect a gene's ability to react under varied cellular conditions. These regulatory elements, which occur in or near the gene transcription region, are referred to as *cis-acting elements* (Figure 9.2).

Complementary to the sequence-based regulatory elements are the protein elements that interact with the sequence elements to regulate transcription. Repressor proteins are one example of regulatory proteins that directly bind to DNA sequences. Other proteins may bind to the promoter region to increase the affinity of RNA polymerase to the region and, thus, are positive controllers of transcription. In eukaryotes, in order for polymerase to bind the DNA and begin transcription, a number of additional molecules must first bind the promoter region. These transcription factor binding–associated proteins (TAFs) form transcription factor complexes that help position the polymerase in the promoter region. This complex alone provides basal regulation of transcription but may also be enhanced by additional ancillary proteins called *activators*, which bind to enhancer sequence elements to speed transcription. Repressor proteins may bind to silencer sequence elements in the DNA as well to slow transcription. Of course, these regulatory proteins may themselves be regulated, where feedback control or other such processes are known to regulate the quantities of these proteins in the cell. Other regulatory proteins may require that allosteric effectors be in their active form or that they become inactivated by effector proteins. Other proteins may inhibit the formation of the transcription factor complexes. These elements,
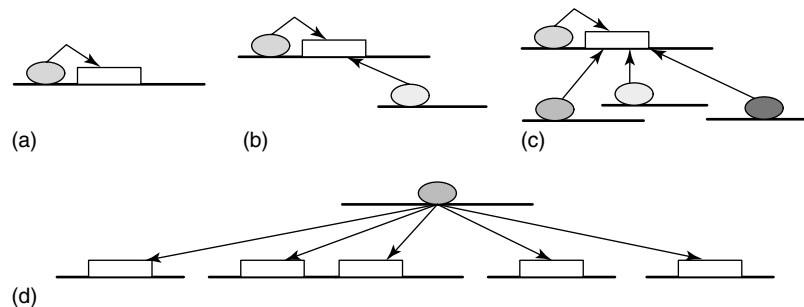
Figure 9.2 Mapping proximal and distal eQTL for gene expression traits. The white rectangles represent genes that are controlled by transcriptional units. The ellipses represent the transcriptional control units, which could be transcription regulatory sites, other genes that control the expression of the indicated gene, and so on. (a) Cis-acting control unit acting on a gene. DNA variations in this control unit that affected the gene's expression would lead to a cis-acting (proximal) eQTL. (b) cis and trans control units regulating the indicated gene. DNA variations in these control units that affected the gene's expression would lead to proximal and distal eQTL. (c) cis control unit and multiple trans control units regulating the indicated genes. DNA variations in these control units would lead to a complex eQTL signature for the gene. (d) A single control unit regulating multiple genes. DNA variations in this single control unit could lead to a cluster of distal eQTL (an eQTL hot spot).

which are usually coded for far from the genes they regulate, are called *trans-acting elements* (Figure 9.2).

The view of transcription just discussed, although necessarily simplified for this presentation, highlights that it is indeed a complex process, and with the recent discovery of large numbers of noncoding ribonucleic acids (ncRNAs), many of which have already been shown to regulate transcription, the complexity of how cells regulate transcription will no doubt be shown to be more complex than we can appreciate at this time. Variations in transcript abundances can be caused in a number of ways, many of which are ultimately due to variation in genetic sequences, where altering the sequence in the regulatory region or altering the proteins or quantity of the proteins that bind to these regions results in variations in transcript abundances. Additionally, the rate of transcription of degradation of gene transcripts can also be affected by ncRNAs. One class of ncRNAs currently under intense study because of their regulation of transcription is micro-RNAs. These small molecules bind to specific sequences in the mRNA, so that mutations in the DNA encoding either the micro-RNA or the gene transcript can alter the binding affinity and thus the rate of degradation. Whether variations in gene expression are due to changes in the rate of transcription, degradation, transcript transport from the nucleus, alternative splicing, or other such processes, these encoding or micro-RNA variations can be comprised of genetic and environmental components, which when properly identified and integrated with other information can lead to the identification of key drivers of complex phenotypes like disease.

9.3   HUMAN VERSUS EXPERIMENTAL MODELS

eQTL analysis can be done using either designed crosses of experimental organisms or natural segregating populations, like humans, using methods appropriate for the study design (**Chapters 18**, **19**, **20**, **32**, **33**, **34**, **35**, **36**, **37** and **38**). eQTL mapping has been successful both in animal crosses and in humans using both linkage and association. The choice of animal model versus human should depend on many factors that include the goal of the study and the availability of tissue samples in the chosen population.

In humans, there are two primary difficulties that make these studies less common. First is the availability of tissue samples. Preliminary studies in humans were done using lymphoblastoid cell lines, and with the exception of cell types that can be extracted from blood, the procurement of most tissue types of interest is relatively invasive and often prohibitive in most circumstances. Another barrier to performing these types of studies in humans is that of power, which is always an issue when choosing between human and designed animal experiments, due to the lack of full informativeness in human data. In the situation of eQTL mapping, the issue of power is exacerbated by the multiple testing issues that are obtained in examining thousands of expression traits rather than one or few physical phenotypes. Studies must be appropriately powered to distinguish signal from noise given both the dimentionality of the expression data and the size of the genome.

Crosses of inbred lines in animal models are more informative in terms of identifying the grandparental source of alleles; therefore, fewer individuals are required to detect QTL. Additionally, tissue samples of all types are more readily available in animal experiments than they are from human. Still, questions exist about the validity

of extending the results from animal to human. Additional experiments are always required to validate findings in human. Crosses of inbred lines are also limited by the genetic differences between the two parental lines. Because all inbred strains are related in some manner, between any two strains there is always some portion of the genome that is nonvarying (identical by descent) between the two, and, thus, identifiable QTL are limited to those at loci segregating polymorphisms. These are likely to be fewer in number than in a natural population. As a result, no one experiment can fully identify all eQTL and findings from one cross may not replicate in another.

9.4   HERITABILITY OF EXPRESSION TRAITS

As a quantitative measure, gene expression can be treated as any other quantitative trait in a genetic analysis. A variety of linkage methods for the analysis of quantitative traits are described in **Chapters 18**, **19** and **20**, which are just as useful for identifying genetic controllers of gene expression as they are for, say, regulation of plasma LDL cholesterol levels. This premise, of course, is only valid if expression is, indeed, a trait under genetic control. Thus, characterizing the proportion of the trait variance that is attributed to genetics (i.e. the heritability, see **Chapter 32**, Section 3) over a large set of gene expression traits provides important information about the landscape of genetic control, and ultimately determines how useful eQTL mapping strategies are. For instance, if the genetic component of expression control is vastly outweighed by environmental factors and measurement error, then strategies to map genetic contributions and ultimately correlate these contributions with complex disease will fail.

Equally important is identifying the types of expression traits that are heritable. Many studies have identified gene transcripts that (1) are associated with complex disease phenotypes (Karp *et al.*, 2000; Schadt *et al.*, 2003), (2) are alternatively spliced (Johnson *et al.*, 2003), (3) elucidate novel gene structures (Shoemaker *et al.*, 2001; Mural *et al.*, 2002; Schadt *et al.*, 2004), (4) can serve as biomarkers of disease or drug response (DePrimo *et al.*, 2003), (5) lead to the identification of disease subtypes (van 't Veer *et al.*, 2002; Mootha *et al.*, 2003; Schadt *et al.*, 2003), and (6) elucidate mechanisms of drug toxicity (Waring *et al.*, 2001). However, identifying the heritable traits and the extent of their genetic variability provides insight about the evolutionary forces contributing to the changes in expression that associate with biological processes that underlie complex traits like disease, beyond what can be gained by looking at the transcript abundance data alone.

Estimates of genetic heritability can be made with or without the aid of genetic marker data. In the absence of genetic marker data, the correlation between the trait values of related individuals provides information about the heritability of a trait by identifying the extent to which an individual's trait values can be predicted from their relatives. In outbred populations, traditional methods involve regressing an offspring's trait value on those for the parents, comparing the within-family variance for groups of siblings to the among-family variance in an ANOVA, or contrasting the correlation between pairs of monozygotic and dizygotic twins. In general outbred pedigrees, the heritability can be estimated using linear mixed models as in Section 3.2 of **Chapter 19**. In the absence of

genetic marker information, however, specific QTLs are not modeled. Instead, the mixed model contains only the fixed covariate effects and the random polygenic effect

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{9.1}$$

where

$$u \sim N(0, \mathbf{A}\sigma_u^2),$$

Here $\mathbf{A}$ is twice the kinship matrix and σ_u^2 is the additive genetic variance of the trait, so that σ_u^2/σ_y^2 is the narrow-sense heritability. This polygenic model assumes many loci with small effects (i.e. assumes the trait is complex), but these assumptions may be violated in the presence of one or few loci with strong effects.

When a major gene contributes to the trait and genotype data is available, better estimates of the genetic variance can be made by including allelic or genotypic effects at a QTL locus or loci using the models in Sections 3.2 and 3.3 of **Chapter 19**. This is identical to performing linkage analysis through variance components methods. Here, summing over the appropriate variances for the QTL and polygenic components provides an estimate for the genetic variance. Bayesian MCMC (Markov chain Monte Carlo) methods (see Section 4.3 of **Chapter 19**) can provide more accurate estimates of the genetic variance, as well as estimates of the number of major genes, by sampling over the space of genetic models conditional on the data. The drawback to this type of approach, when the number of traits being mapped is high, is that it is computationally intensive, and the interpretations are not as clear-cut as point estimates obtained from frequentist methods. However, even when computation time is a constraint, appropriate filters can be applied to reduce the number of expression traits to consider in such an analysis. For example, one may choose to restrict attention to those genes with at least one eQTL at a relatively loose significance threshold. Brem and Kruglyak (2005) found 132 genes with QTLs significant at the 5×10^{-4} level, which provides a much smaller set of genes to consider, compared to the tens of thousands of genes that were profiled in this study.

Linkage approaches can also be used to estimate the genetic variance contribution in crosses of inbred lines (Brem and Kruglyak, 2005). However, in this case, care should be taken in the extrapolations made. Conclusions in these analyses apply only to populations founded by the same parental strains and are limited by the extent and pattern of identity by descent (IBD) between them. Here, estimates should be made in the presence of genetic marker data. The power of segregation approaches is limited because, in common inbred line designs such as F2 intercrosses or backcrosses, the subjects are genetically equivalent to one large sibship. Because the relationship between each pair of individuals is identical, the kinship coefficient among all pairs is also identical, resulting in no power to estimate the polygenic component. In order to estimate the genetic variance of a trait, each genetic contribution must be modeled explicitly. Further discussion of multilocus models is given in the following section.

9.5 JOINT eQTL MAPPING

Ultimately, expression traits are quantitative measures that can be analyzed like any other quantitative trait. The difficulty in analysis and interpretation comes with the large

number of traits examined. Whatever method is ultimately chosen must be computationally tractable enough to be performed hundreds of thousands of times. In addition, significance thresholds must be adjusted for multiple testing. Multiple testing issues relate not only to the number of transcripts tested but also to the number of markers or proportion of the genome tested. Further, single-trait analyses lose information by ignoring the correlation between associated expression traits. Appropriately chosen joint mapping methods can leverage trait correlation and more elegantly handle multiple testing.

Typical approaches to the joint analysis of genetic traits involve mapping each gene expression trait individually and inferring the genetic correlation between pairs or sets of expression traits based on pair-wise Pearson correlation, eQTL overlaps, and/or tests for pleiotropy. Using a family-based sample, Monks *et al.* (2004) estimated the genetic correlation between pairs of traits using a bivariate variance-component-based segregation analysis, and showed that the genetic correlation was better able to distinguish clusters of genes in pathways than correlations based on the observed expression traits. This method is similar to single-trait variance-component-based segregation analysis as in (9.1). For two trait vectors $\mathbf{v}$ and $\mathbf{y}$, let $\mathbf{t}$ denote the bivariate trait vector

$$\mathbf{t} = \begin{bmatrix} \mathbf{v} \\ \mathbf{y} \end{bmatrix},$$

with piecewise mean vector

$$\mu_t = \begin{bmatrix} \mu_v \\ \mu_y \end{bmatrix}.$$

The partitioned covariance matrix can be written as

$$\mathbf{V}_t = \begin{bmatrix} \mathbf{V}_v & \mathbf{V}_{vy} \\ \mathbf{V}_{vy} & \mathbf{V}_y \end{bmatrix},$$

where $\mathbf{V}_v$ and $\mathbf{V}_y$ are the univariate covariance matrices for traits v and y respectively. More details are available in Section 3.3 of **Chapter 19**. The trait covariance matrix $\mathbf{V}_{vy}$ is modeled as

$$\mathbf{V}_{vy} = \mathbf{A}\sigma_{uvy}^2 + \mathbf{I}\sigma_{evy}^2,$$

where $\mathbf{A}$ is twice the kinship matrix, $\mathbf{I}$ is the identity matrix, σ_{uvy}^2 is the genetic covariance between the two traits, and σ_{evy}^2 is the nongenetic covariance. The genetic and nongenetic covariances can be expressed in terms of the genetic and nongenetic correlation as

$$\sigma_{uvy}^2 = \sigma_{uv}\sigma_{uy}\rho_{uvy},$$

and

$$\sigma_{evy}^2 = \sigma_{ev}\sigma_{ey}\rho_{evy},$$

where σ_{uv} and σ_{uy} are the square root of the genetic variances for traits v and y, σ_{ev} and σ_{ey} are the square root of the environmental variances, and ρ_{uvy} and ρ_{evy} are the genetic and nongenetic correlations respectively.

These methods can be extended to perform bivariate and multivariate linkage analysis, which can be more highly powered to detect linkage when traits are correlated. Clusters of correlated gene expression traits can often contain hundreds or thousands of genes,

which would be computationally prohibitive in a joint analysis. Kendziorski *et al.* (2006) approached this problem in a different way by employing a Bayesian mixture model to exploit the increased information from the joint mapping of correlated gene expression traits, which is computationally tractable for large sets of genes. Instead of doing a linkage scan by computing LOD scores at positions along the genome, Kendziorski *et al.* (2006) compute the posterior probability that a particular gene expression trait maps to marker *m* for each marker, as well as the posterior probability that the trait maps nowhere in the genome. More specifically, for a particular gene expression trait, *k*, from a correlated gene cluster of interest, the marginal distribution of the data, $\mathbf{y}_k$, is

$$p_0 f_0(\mathbf{y}_k) + \sum_m p_m f_m(\mathbf{y}_k),$$

where p_m is the probability that the transcript maps to marker *m* and f_m is the distribution of the data if it does, p_0 is the probability that the transcript does not map to the genome, and f_0 is the data distribution in this case. Given appropriate choices for the distributions, model parameters can be estimated via the EM algorithm, and posterior estimates of $\{p_m\}$ can be obtained. Nonlinkage is declared for a transcript if the posterior probability of nonlinkage exceeds a threshold that bounds the posterior expected false discovery rate (FDR). One benefit to this approach is that it controls false discovery for the number of expression traits being tested, whereas assessing the appropriate significance cutoffs in single-transcript linkage analysis often requires data permutation analyses. The drawback is that this method assumes that linkage occurs at either one or none of the markers tested, and lacks a well-defined method for the case when multiple eQTL control an expression trait.

9.6   MULTILOCUS MODELS AND FDR

In the study of inbred strain crosses, the only valid way of estimating the extent of genetic control of a given trait is to explicitly model each eQTL, including any epistatic interactions if they exist. Brem *et al.* (2005) showed that epistatic interactions were prevalent in the gene expression levels in yeast, and similar suggestions have been made in other species as well (Schadt *et al.*, 2005), but more definitive studies are needed to characterize the extent of epistasis among eQTL in these other species. In the absence of epistasis, the genetic contribution for each transcript can be estimated by summing contributions for each individual eQTL, assuming that little or no allelic association exists between the eQTL loci. In the presence of epistasis, however, this practice does not yield a valid estimate, and multilocus models are required to obtain valid estimates. Additionally, multilocus modeling can identify loci contributing to expression traits that would have been missed in single-locus eQTL scans (Brem and Kruglyak, 2005; Storey *et al.*, 2005).

While forward selection methods for general regression cases, as with any greedy algorithm, do not typically choose the best model, they have been shown to be consistent in the linkage modeling framework (Broman, 1997). For further discussion of model selection techniques, see Section 4.4 of **Chapter 18**. In the eQTL framework, given the large number of tests being performed, particular care needs to be taken to limit the FDR. Storey *et al.* (2005) suggest controlling the FDR at each stage of a forward model selection

procedure, and they employ a Bayesian technique to estimate false positive probabilities. In other words, the posterior probability of linkage at K loci is calculated as

$$Pr(\text{loci } 1, \ldots, K \text{ linked|data}) = Pr(\text{locus 1 linked|data})$$
$$\times Pr(\text{locus 2 linked|locus 1 linked, data}) \times \cdots$$
$$\times Pr(\text{locus } K \text{ linked|loci } 1, \ldots, K-1 \text{ linked, data}).$$

At each stage of selection, the posterior probability of the new locus, conditional on the data and the linkage of the previously selected loci is calculated. All traits with $Pr(\text{loci } 1, \ldots, K \text{ linked|data}) \geq T$, for some threshold T, are called *significant*. If, at any stage, the model for a trait fails to be significant, the last significant model for that trait is chosen and the selection procedure stops.

The posterior probability of linkage is estimated empirically by comparing the distribution of conditional test statistics to the null distribution, which is estimated using permutation methods. Thus, the high dimensionality of the data is exploited to estimate empirically based on the data, instead of assuming a fixed prior. More specifically, let F_{iK} be the test of the model containing all K loci versus the model containing $K-1$ loci at stage K in the model selection procedure. The null distribution for this test statistic can be obtained by permuting the data conditional on the first $K-1$ loci, by permuting the traits within each of the $(K-1)$-locus genotype classes. This maintains the relationship between the trait and the previously selected loci, but randomizes the trait with respect to the remaining loci. For some number of permutations, let F_{iK}^b be the test statistic for the ith trait and the bth permutation. The set of true test statistics $\{F_{iK}\}$ can be assumed to be from a mixture of true tests and false tests, and the set of permuted test statistics $\{F_{iK}^b\}$ provides an estimate of the null distribution of test statistics. For a trait with test statistic equal to some value F at the Kth selection step, the posterior probability of no linkage, conditional on previous linkages, can be estimated as

$$\frac{\left|\{F_{iK} \leq F\}\right|}{\left|\{F_{iK}^b \leq F; b = 1, \ldots, B\}\right|/B},$$

where $|\cdot|$ denotes the cardinality of the set. Note that, since this method relies on empirical estimation of test-statistic distributions over the set of 'active' traits that were significant at the previous step, the number of 'active' traits and the number of permutations must be high enough to give reasonable estimates. When proceeding through the model selection procedure, at the point in which the number of significant traits becomes sufficiently low, this approach to error rate estimation becomes inaccurate, and the model selection should stop.

This Bayesian model selection criteria lends itself well to the FDR concept because the posterior probability of linkage can be used to estimate the number of false models, where a false model is defined to be one in which at least one of the K selected loci is not truly linked. In other words, a false discovery is one in which any of the selected linkages is false. At any stage of the model selection, K, the FDR is defined as the ratio of expected number of false discoveries to the number of significant K-locus selections. The expected number of false discoveries can be estimated as $\sum(1 - Pr(\text{loci } 1, \ldots, K \text{ linked|data}))$, where the summation is over the K-locus models called significant based on the threshold, T.

9.7 eQTL AND CLINICAL TRAIT LINKAGE MAPPING TO INFER CAUSAL ASSOCIATIONS

While understanding the mechanisms of RNA expression is in itself important for understanding biological processes, the ultimate use of this information is identifying the relationship between variation in expression levels and disease phenotypes in an organism of interest. Microarray experiments are commonly used to explore differential expression between disease and normal tissue samples or between samples from different disease subtypes (see **Chapter 6** Sections 4–5, **Chapters 7** and **8**). These studies are designed to detect association between gene expression and disease-associated traits, which in turn can lead to the identification of biomarkers of disease or disease subtypes. However, in the absence of supporting experimental data, these data alone are not able to distinguish genes that drive disease from those that respond. As discussed above, eQTL mapping can aid traditional clinical quantitative trait loci (cQTL) mapping by narrowing the set of candidate genes underlying a given cQTL peak and by identifying expression traits that are causally associated with the clinical traits.

Expression traits detected as significantly correlated with a clinical phenotype may reflect a causal relationship between the traits, either because the expression trait contributes to, or is causal for, the clinical phenotype, or because the expression trait is reactive to, or a marker of, the clinical phenotype. However, correlation may also exist in cases when the two traits are not causally associated. Two traits may appear correlated due to confounding factors such as tight linkage of causal mutations (Schadt *et al.*, 2005) or may arise independently from a common genetic source. The A^y mouse provides an example of correlations between eumelanin RNA levels and obesity phenotypes induced by an allele that acts independently on these different traits, causing both decreased levels of eumelanin RNA and an obesity phenotype. More generally, a clinical and expression trait for a particular gene may depend on the activity of a second gene, in such a manner that, conditional on the second gene, the clinical and expression traits are independent.

Correlation data alone cannot indicate which of the possible relationships between gene expression traits and a clinical trait are true. For example, given two expression traits and a clinical trait detected as correlated in a population of interest, there are 112 ways to order the traits with respect to one another. That is, for each pair of nodes, there are five possible ways the nodes can be connected: (1) connected by an undirected edge, (2) connected by a directed edge moving left to right, (3) connected by a directed edge moving right to left, (4) connected by a directed edge moving right to left and a directed edge moving left to right, (5) not connected by an edge. Since there are three pairs of nodes, there are $5 \times 5 \times 5 = 125$ possible graphs. However, since we start with the assumption that all traits are correlated, we exclude 12 of the 125 graphs in which one node is not connected to either of the other two nodes, and we exclude the graph in which none of the nodes are connected, leaving us with 112 possible graphs, some of which are illustrated in Figure 9.1(a). The joint trait distributions induced by these different graphs are often statistically indistinguishable from one another (i.e. they are Markov equivalent, so that their distributions are identical), making it nearly impossible in most cases to infer the true relationship. On the other hand, when the two traits are at least partially controlled by the same genetic locus and when more complicated methods of control (e.g. feedback loops) are ignored, the

number of relationships between the QTL and the two traits of interest can be reduced to the three models illustrated graphically in Figure 9.1(b). The dramatic reduction in the number of possible graphs to consider (from 112 to 3) is mainly driven by fact that changes in DNA drive changes in phenotypes and not vice versa (it is extremely unlikely that changes in RNA or protein lead to changes in DNA at a high enough frequency to detect associations between germ-line transmitted polymorphisms and phenotype).

It is important to note here that, when we use the term causality, it is perhaps meant in a more nonstandard sense than most researchers in the life sciences may be accustomed to. In the molecular biology or biochemistry setting, claiming a causal relationship between, say, two proteins usually means that one protein has been determined experimentally to physically interact with or to induce processes that directly affect another protein, and that in turn leads to a phenotypic change of interest. In such instances, an understanding of the causal factors relevant to this activity are known, and careful experimental manipulation of these factors subsequently allows for the identification of genuine causal relationships. However, in the present setting, the term 'causal' is used from the standpoint of statistical inference, where statistical associations between changes in DNA, changes in expression (or other molecular phenotypes), and changes in complex phenotypes like disease are examined for patterns of statistical dependency among these variables that allows directionality to be inferred among them, where the directionality then provides the source of causal information (highlighting putative regulatory control as opposed to physical interaction). The graphical models (networks) described here, therefore, are necessarily probabilistic structures that use the available data to infer the correct structure of relationships among genes and between genes and clinical phenotypes. In a single experiment, with one-time point measurement, these methods cannot easily model more complex regulatory structures that are known to exist, like negative feedback control. However, the methods can be useful in providing a broad picture of correlation and causative relationships, and, while the more complex structures may not be explicitly represented in this setting, they are captured nevertheless given they represent observed states that are reached as a result of more complicated processes like feedback control. A mathematical theory of causal inferences from observed dependency patterns from raw data has been established, and Judea Pearl, a pioneer of mathematical and computational methods for this purpose, provides an excellent description and treatment of this underlying theory (Pearl, 1988; 2000).

While in principle the techniques we describe can be applied to any pair of traits, continuous or binary, clinical or molecular (e.g. RNA, protein, or metabolite), we focus on the problem of inferring causality among gene expression traits and between gene expression and a continuous clinical phenotypes. We let T_1 and T_2 denote RNA expression traits or an expression trait and a clinical trait, and L the locus genotype. In the first model, the genetic locus is causal for T_2 only through T_1 (causal model). In the second model, the T_1 is a reaction to T_2 (reactive model). In the final model, the genetic locus affects both T_1 and T_2 independently (independence model). With the inclusion of genetic marker information, the data models are distinguishable due to the conditional independence structure of the variables.

9.7.1 A Simple Model for Inferring Causal Relationships

Assuming the conditional independence structure implied in the graphical models in Figure 9.1(b), the following simplifications can be made for the joint probability distributions for the causal, reactive, and independence models, respectively:

$$
\begin{aligned}
P_1(L, T_1, T_2) &= P_{\theta_L}(L) P_{\theta_{T_1 L}}(T_1|L) P_{\theta_{T_1 T_2}}(T_2|T_1), \\
P_2(L, T_1, T_2) &= P_{\theta_L}(L) P_{\theta_{T_2 L}}(T_2|L) P_{\theta_{T_1 T_2}}(T_1|T_2), \\
P_3(L, T_1, T_2) &= P_{\theta_L}(L) P_{\theta_{T_1 L}}(T_1|L) P_{\theta_{T_2 L}^*}(T_2|L, T_1) \\
&= P_{\theta_L}(L) P_{\theta_{T_2 L}}(T_2|L) P_{\theta_{T_1 L}^*}(T_1|L, T_2).
\end{aligned}
\tag{9.2}
$$

In the causal model depicted in Figure 9.1(b), the clinical trait (T_2) is independent of the genetic locus conditional on the gene expression trait (T_1). In other words, the locus genotype lends no additional information about the clinical phenotype when the gene expression measurement is known. Similarly, in the reactive model, the expression trait is independent of the underlying genetics conditional on the clinical trait. In the independence model, the gene expression and clinical trait are not assumed to be conditionally independent to allow for correlation due to other shared genetic and environmental influences. Failure to account for this correlation, when it is of moderate size, can result in falsely choosing one of the two other models because the two traits contain information, in addition to that provided by the genetic locus, about the other, due to unmeasured common influences.

The modeling framework in (2) is quite general and can accommodate a wide range of genetic and trait dependence models. Genetic models for dichotomous and quantitative traits have been developed for crosses of inbred lines (**Chapter 18**) and for outbred pedigrees (**Chapter 19**). For continuous traits, a variety of trait models have been developed for the purposes of testing for linkage of a QTL to a single locus. These same models can be used to model an expression or clinical trait conditional on the genetic locus. For example, when the data come from an intercross of two inbred lines, it is common to model the trait using a normal distribution with a mean that depends on the genotype at the locus (Section 3.1 of **Chapter 18**).

Given appropriate choices for the conditional distributions in the three models, $P(T_1|L)$, $P(T_2|L)$, $P(T_1|T_2)$, $P(T_2|T_1)$, and $P(T_1|L, T_2)$, likelihoods for the three different models can be maximized with respect to the model parameters and the likelihoods can be subsequently compared. Note that the log-likelihood of each model is the sum of log-likelihoods for each of the three variables with no common parameters. For example,

$$
\begin{aligned}
\log P_1(L, T_1, T_2) &= \log P_{\theta_L}(L) + \log P_{\theta_{T_1 L}}(T_1|L) + \log P_{\theta_{T_1 T_2}}(T_2|T_1) \\
&= \log L(\theta_L|L) + \log L(\theta_{T_1 L}|T_1, L) + \log L(\theta_{T_2 T_1}|T_2, T_1) \\
&= \log L_1(\boldsymbol{\theta}_1|L, T_1, T_2).
\end{aligned}
$$

Thus MLEs (Maximum likelihood estimates) for $\theta_L, \theta_{T_1 L}$, and $\theta_{T_2 T_1}$ can be obtained by separately maximizing each corresponding term. The likelihoods are then compared among the different models in order to find the most likely of the three. When the number of model parameters among the models differs, a penalized function of the likelihood

is used to avoid the bias against parsimony: the model with the smallest value of the penalized statistic

$$-2 \log L_i(\hat{\boldsymbol{\theta}}_i | L, T_1, T_2) + k \times p_i,$$

is chosen. Here, $L_i(\hat{\boldsymbol{\theta}}_i | L, T_1, T_2)$ is the MLE for the ith model, p_i is the number of parameters in the ith model, and k is a constant. Common choices for k include $k = 2$, in which case the statistic is known as the *Akaike information criteria* (AIC), and $k = \log(n)$, where n is the number of observations, which is called the *Bayesian information criteria* (BIC).

An alternate approach, which is often taken in causal inference, is to use a significance tests to identify a model consistent with the data. For example, in the causal model depicted in Figure 9.1(b), the clinical trait is conditionally independent of the genetic locus, L. Thus, a rejected test of conditional independence between T_2 and L implies that the data is not consistent with the causal model. In order to infer among the three models, the steps are as follows:

1. Obtain the p value, $p_{T_2 L | T_1}$, for the test of conditional independence between T_2 and L.

2. Obtain the p value, $p_{T_1 L | T_2}$, for the test of conditional independence between T_1 and L.

3. Infer the model based on the following logic

    i.   If $p_{T_2 L | T_1} \geq \alpha$ and $p_{T_1 L | T_2} < \alpha$, then the data is consistent with the causal model.

    ii. If $p_{T_1 L | T_2} \geq \alpha$ and $p_{T_2 L | T_1} < \alpha$, then the data is consistent with the reactive model.

    iii. If $p_{T_2 L | T_1} < \alpha$ and $p_{T_1 L | T_2} < \alpha$, then the data is consistent with the independence model (neither the causal nor reactive model).

    iv. If $p_{T_2 L | T_1} \geq \alpha$ and $p_{T_1 L | T_2} \geq \alpha$, then the data is consistent with both the causal and reactive models, the test is inconclusive.

In cases in which the clinical and gene expression trait are very highly correlated, this approach may result in an inconclusive test because, conditional on the other trait, the genetic data provides almost no additional information.

The model selection procedure described above does not lend itself very well to a testing paradigm because, when choosing among these three models, there is no standard null distribution that can be used in this context to carry out a test of hypothesis for the 'best' model. As a result, significance cannot be assessed by attaching a p value to a hypothesis test in the usual way. Instead, when the scientific aim is to identify whether a particular gene expression trait is causal, reactive, or independent to a second gene expression trait or clinical trait, with respect to a given locus, a confidence measure for the chosen model can be constructed using resampling methods. In such cases, a large number of bootstrap samples is drawn from the empirical distribution of (L, T_1, T_2), and, for each resample, the model selection procedure is carried out to estimate the proportion of times each model is chosen. This proportion can then be considered as a measure of confidence for the selected model.

A simpler case occurs when only one direction of causality is of interest. For example, in order to identify expression traits that are specifically causal for a clinical trait of

interest, with respect to a particular locus, an appropriate test can be constructed whose alternative hypothesis is

$$P(T_2|T_1, L) = P(T_2|T_1),$$

where, for this case, T_1 denotes an expression trait and T_2 a clinical trait. For example, given the linear model for the clinical trait

$$E[T_{2_i}] = \mu + \gamma f(T_{1_i}) + \beta g(L_i) + \lambda h(T_{1_i}, L_i),$$

where f and g are functions or parameterizations of the RNA expression and the genotype respectively and h is a function of the interaction, we can perform a test with null hypothesis $\beta \neq 0$ and $\lambda \neq 0$. This is a test for the independence of the clinical trait and locus genotype conditional on the gene expression, which would occur when the genetics affects the clinical trait only through the expression trait. In order to test hypotheses of this nature, a bioequivalence paradigm is required, so that hypotheses are actually of the form $|\beta_i| > \varepsilon_i$ and $|\lambda_i| > \delta_i$, which requires the selection of appropriate values for the ε and δ. Dixon and Pechmann (2005) suggest selecting bounds based on biological knowledge and informed judgment. One criterion for the choice of bound might be based on the proportion of variance explained by the genetic component in this model. For example, in the case of an F2 intercross population constructed from inbred lines of mice and where we assume the genetic model contains only additive effects and no genotype-by-gene expression interaction, a common parameterization of the genetic model is $E[T_{2_i}] = \mu + \gamma f(T_{1_i}) + \beta_A g(L_i)$, where

$$g(L) = \begin{cases} -1 & L = BB \\ 0 & L = Bb \\ 1 & L = bb. \end{cases}$$

In this case, the expected additive genetic variance is $\frac{1}{2}\beta_A^2$. Bounding the proportion of variance explained by the genetic component results in a bound of the form $|\beta_A| < \sqrt{2\varphi_A \sigma_{T_2}^2}$ for some value of φ_A, and where $\sigma_{T_2}^2$ is the variance for trait T_2. More complex models can be managed in a similar way, including a bound on each component of the genetic variance, including dominance and interaction terms when appropriate.

Given a single parameter of interest, hypothesis testing in the bioequivalence framework requires testing two one-sided subhypotheses, $H_{01} : \beta < -\varepsilon$ and $H_{02} : \beta > \varepsilon$. Because both null hypotheses cannot simultaneously be true, testing each of these tests at level α results in a test that is also of level α for the composite hypothesis, $H_0 : |\beta| > \varepsilon$. Thus, H_{01} is rejected if the t-statistic $T^{(1)} = (\hat{\beta} + \varepsilon)/\hat{\sigma}_\beta$ is greater than the one-sided t-distribution level-α critical value. Similarly, H_{02} is rejected if the t-statistic $T^{(2)} = (-\hat{\beta} - \varepsilon)/\hat{\sigma}_\beta$ is greater than the same level-α critical value. Confidence intervals can also be used to infer the bioequivalence; however, care needs to be taken to ensure the proper significance level. Specifically, H_0 is rejected at level α if the $(1 - 2\alpha)\,100\,\%$ confidence interval lies entirely within the alternative hypothesis region, $|\beta| \leq \varepsilon$.

Significance under the null hypothesis of the independence model can also be assessed via permutation. In other words, permuting the relationship between the clinical trait and the gene expression trait within each genotype results in a conditionally independent relationship between the two traits.

9.7.2   Distinguishing Proximal eQTL Effects from Distal

As discussed in previous sections, all genes expressed in living systems are cis regulated at some level and so are under the control of various cis-acting elements such as promoters and TATA boxes. In this context, expression as a quantitative trait for eQTL mapping presents a unique situation in quantitative trait genetics because the expression trait corresponds to a physical location in the genome (the structural gene that is transcribed, giving rise to the expression trait). The transcription process operates on the structural gene, and so DNA variations in the structural gene that affect transcription are identified as eQTL in the mapping process. In such cases, we would identify eQTL as cis acting, given that the most reasonable explanation for seeing an eQTL coincident with the physical location of the gene is that variations within the gene region itself give rise to variations in its expression (Doss *et al.*, 2005). However, because we cannot guarantee that the eQTL is truly cis acting (i.e. it could arise from variation in a gene that is closely linked to the gene expression trait in question), it is more accurate to refer to such eQTL as *proximal*, given they are close to the gene corresponding to the expression trait. Because the cis-regulated components of expression traits are among the most proximal traits in a biological system with respect to the DNA (Plate 8), we might expect that true cis-acting genetic variance components of expression traits are among the easiest components to detect via QTL analysis if they exist. This indeed has been observed in a number of studies, where proximal (presumably cis-acting) eQTL have been identified that explain unprecedented proportions of a trait's overall variance (several published studies highlight examples where greater than 90 % of the overall variation was explained by a single cis-acting eQTL) (Brem *et al.*, 2002; Schadt *et al.*, 2003; Monks *et al.*, 2004; Cervino *et al.*, 2005; Cheung *et al.*, 2005; Lum *et al.*, 2006).

Variations in expression levels induced by DNA variations in or near the gene itself may in turn induce changes in the expression levels of other genes. Each of these genes in a population of interest may not harbor any DNA variation in their structural gene, so that they do not give rise to true cis-acting eQTL, but they would give rise to eQTL nevertheless that link to the gene region inducing changes in their expression. Therefore, we see that the individual variation in gene expression can be of two fundamental types. The first, termed *proximal*, often results from DNA variations of a gene that directly influence transcript levels of that gene. The second, termed *trans acting* or *distal*, does not involve DNA variations of the gene in question but rather is secondary to alterations of other true cis-acting genetic variations (Figure 9.2). In reality, variation in expression traits may be due to variation in cis-acting elements and/or one or multiple trans-acting elements. Additionally, master regulators of transcription, which affect the expression of many traits in trans, may exist, though the evidence on this is mixed (Figure 9.2).

Because, in many cases, it is not possible to infer the true regulatory effects (i.e. cis vs trans) of an eQTL without complex bioinformatics study and experimental validation, we instead categorize eQTL into proximal and distal types based on the distance between the eQTL and the location of the structural genes. Obviously, if these are on different chromosomes the eQTL is distal, but if they fall on the same chromosome then we require the distance between the structural gene and the eQTL to not exceed some threshold. The exact threshold is a function of the number of meioses and extent of recombination in a given population dataset. In a completely outbred population, where LD (Linkage Disequilibrium) mapping has been used to finemap the eQTL, it is reasonable to require

the distance between the proximal eQTL and structural gene to be less than 1 MB (Cheung *et al.*, 2005). However, in an F2 intercross population constructed from two inbred lines of mice, the extent of LD will be extreme, given all animals are descended from a single F1 founder, with only two meiotic events separating any two mice in the population. In such cases, the resolution of linkage peaks is quite low, requiring the threshold of peak-to-physical gene distance to be more relaxed, so that eQTL that are within 20 or 30 MB could be considered cis acting (Schadt *et al.*, 2003; Doss *et al.*, 2005). While the proximal eQTL provide an easy path to making causal inference, given the larger effect sizes commonly associated with proximal eQTL make them easier to detect (Brem *et al.*, 2002; Schadt *et al.*, 2003; Monks *et al.*, 2004; Cervino *et al.*, 2005; Cheung *et al.*, 2005; Lum *et al.*, 2006), the methods discussed above work for distal as well as proximal eQTL.

9.8 USING eQTL DATA TO RECONSTRUCT COEXPRESSION NETWORKS

Networks provide a convenient framework for representing high-dimensional data in which relationships among the many variables making up such data are the key to understanding the properties that emerge from the complex systems they represent. Networks are simply graphical models comprised of nodes and edges. For gene networks associated with biological systems, the nodes in the network typically represent genes, and edges (links) between any two nodes indicate a relationship between the two corresponding genes. For example, an edge between two genes may indicate that the corresponding expression traits are correlated in a given population of interest (Zhu *et al.*, 2004), that the corresponding proteins interact (Kim *et al.* 2005), or that changes in the activity of one gene lead to changes in the activity of the other gene (Schadt *et al.*, 2005). Interaction or association networks have recently gained more widespread use in the biological community, where networks are formed by considering only pair-wise relationships between genes, including protein interaction relationships (Han *et al.*, 2004), coexpression relationships (Gargalovic *et al.*, 2006; Ghazalpour *et al.*, 2006), as well as other straightforward measures that may indicate association between two genes.

Forming association networks from expression data based purely on correlations between genes in a set of experiments of interest can give rise to links in the network driven by correlated noise structures between array-based experiments or other such artifacts. The eQTL data can be simply leveraged in this case by filtering out gene–gene correlations in which the expression traits are not at least partially explained by common genetic effects. For example, we may connect two genes with an edge in a coexpression network if (1) the p value for the Pearson correlation coefficient between the two genes was less than some prespecified threshold, and (2) the two genes had at least one eQTL in common. One intuitive way to establish whether two genes share at least one eQTL is to carry out single-trait eQTL mapping for each expression trait and then consider eQTL for each trait overlapping if the corresponding LOD for the eQTLs are above some threshold and if the eQTL are in close proximity to one another.

The p-value threshold for considering two genes linked in the coexpression networks is chosen such that the resulting network exhibits the scale-free property (Barabasi and Albert, 1999; Ghazalpour *et al.*, 2006; Lum *et al.*, 2006) and the FDR for the gene–gene
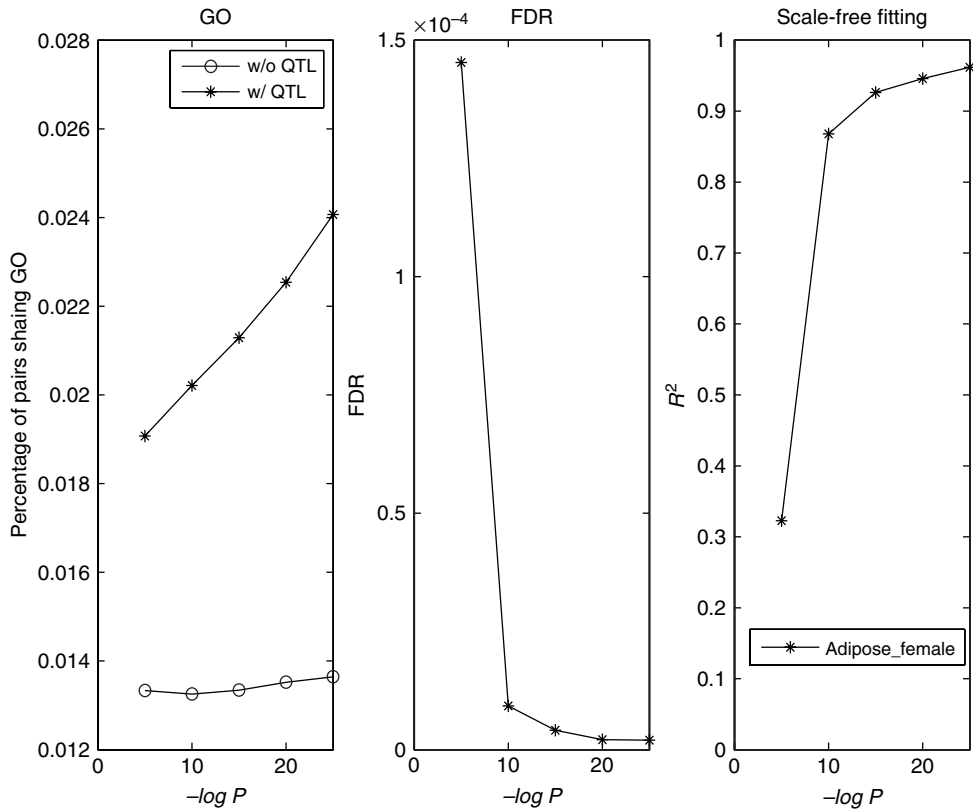
Figure 9.3 Variation in key parameters over different p values provides an objective way to select p-value thresholds for reconstructing coexpression networks. The first box of this figure plots the percent of gene pairs connected in the network that share GO biological process category terms, as a function of $-\log$ of the p-value threshold for the correlation between the gene pairs used to construct the network. The dark gray curve represents genes in the network connected by an edge because the corresponding correlation coefficient is significant at the indicated p-value threshold and the genes share at least one common eQTL (as described in the main text). The light gray curve represents edges based on correlation significance only (no genetics). The second box plots the false discovery rate (FDR) for edges in the coexpression network constructed as a function of p-value threshold. The third box plots the coefficient of determination corresponding to how well the degree distribution for the network nodes fits the inverse power law.

pairs represented in the network is constrained (Figure 9.3). Filtering correlations based on eQTL overlap can be seen to improve the quality of the coexpression networks by examining whether networks reconstructed without the eQTL data are more coherent than networks reconstructed with the eQTL data with respect to the GO (Gene Ontology) database biological process categories (see also 'annotation analysis' in **Chapter 7**). We say one network is more coherent than another with respect to GO biological process categories according to its percentage of gene–gene pairs sharing a common GO biological process category. The pathways represented in GO are independently determined and so provide an independent source of information in testing how much of the known information is captured in the network. As an example, consider a previously described

F2 intercross population constructed from the B6 and C3H inbred lines of mice (referred to throughout as the *BXH cross*) (Lum *et al.*, 2006; Yang *et al.*, 2006). Figure 9.3 plots the percentage of gene–gene pairs sharing common GO biological process categories as a function of minus log of the correlation p value for the coexpression network reconstructed from the BXH adipose data, both with and without the eQTL data. Over a range of p values, the network reconstructed with the eQTL data is seen to be significantly more coherent than the network reconstructed without the eQTL data. This provides direct experimental evidence that incorporating eQTL information into network reconstruction can enhance the accuracy of the network.

9.8.1 More Formally Assessing eQTL Overlaps in Reconstructing Coexpression Networks

While testing whether gene–gene pairs that are significantly correlated are partially controlled by common genetic loci using the overlap method discussed above is intuitively appealing, it fails to make full use of the data to infer whether overlapping eQTLs are really the same eQTL or closely linked eQTL, and such an approach does not lend itself to statistically robust hypothesis testing. One way to test more formally whether two overlapping eQTL represent a single eQTL or closely linked eQTL is to employ a pleiotropy effects test (PET) based on a pleiotropy model initially described by Jiang and Zeng (1995) and Zeng *et al.* (2000). While we discuss this method for considering only two traits simultaneously, the method can be easily extended to consider more traits. The statistical model for PET is an extension of the single-trait model as defined in the following equation:

$$\begin{pmatrix} y_{11} \cdots y_{n1} \\ y_{12} \cdots y_{n2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} (x_1 \cdots x_n) + \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} (z_1 \cdots z_n) + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix},$$

where y_i is the vector of trait values for individual $i (i = 1, \ldots, n), a_j$ and d_j are the additive and dominance effects for trait $j (j = 1, 2), x_i$ is as defined above, and e_j is the residual effect for trait j. In this case, we are assuming an F2 intercross population constructed from two inbred lines of mice, but the model is easily generalizable to other experimental cross populations.

From this statistical model, a series of tests of hypotheses can be performed to test whether the two traits are supported as being driven by a single QTL at a given position. The first test involves testing whether a given region is linked to the joint trait vector for the traits under study:

$$H_0 : a_1 = 0, d_1 = 0, a_2 = 0, d_2 = 0,$$

$$H_1 : \text{at least one of the above terms is not 0.}$$

To test the above null hypothesis of no linkage against the alternative linkage hypothesis, likelihoods associated with the null and alternative hypothesis are maximized with respect to the model parameters. From the maximum likelihoods, the log-likelihood ratio statistic is formed and used to test whether the alternative hypothesis (H_1) is supported by the data. With this model, the log-likelihood ratio statistic under the null hypothesis is chi-square distributed with 4 degrees of freedom. If the null hypothesis (H_0) is rejected, the implication is that trait 1 and/or trait 2 have a QTL at the given test locus.

Subsequent to the test just described, resulting in a rejection of the null hypothesis, second and third tests of hypotheses can be performed to establish whether the detected

QTL affects both traits. For a given QTL test position,

$$H_{10} : a_1 = 0, d_1 = 0, a_2 \neq 0, d_2 \neq 0,$$
$$H_{11} : a_1 \neq 0, d_1 \neq 0, a_2 \neq 0, d_2 \neq 0,$$

assesses whether the first trait has a QTL at the test position, and

$$H_{20} : a_1 \neq 0, d_1 \neq 0, a_2 = 0, d_2 = 0,$$
$$H_{21} : a_1 \neq 0, d_1 \neq 0, a_2 \neq 0, d_2 \neq 0,$$

assesses whether the second trait has a QTL at the test position. As above, the log-likelihood ratio statistics are formed for each of these tests, where under the null hypotheses these statistics are chi-square distributed with 2 degrees of freedom. If both null hypotheses H_{10} and H_{20} are rejected, the QTL is supported as having pleiotropic effects on the two traits under study.

At the 2.8 LOD score threshold indicating suggestive linkage in an F2 intercross population, the expected number of QTL detected is 1 for the conditional linkage tests just described (Lander and Kruglyak, 1995). Therefore, an intuitive way to compute the probability that both of these tests for independent traits give rise to a QTL at a given location (so rejection of the null hypothesis for both tests at the same chromosome position) is to consider the probability that the two QTLs expected by chance to be identified for each test happen to be detected at the same chromosome location. This probability can be approximated by the fraction 1/655, where 655 represents the effective number of tests carried out in searching the entire genome at the 2.8 LOD threshold (Lander and Kruglyak, 1995). However, because such an estimate is based on theoretical arguments relying on assumptions that may not hold exactly, and because in practice the traits under consideration will not be independent (especially not in the coexpression network setting where we are interested in highly interconnected sets of genes), permutation testing can also be used to assess the significance of both tests leading to a rejection of the null hypothesis at the 2.8 LOD threshold. In the adipose BXH cross expression data described above, the significance level associated with the 2.8 LOD score threshold was estimated to be 0.004 using permutation methods, only slightly larger than the theoretical estimate of 0.002 given above. Therefore, using the suggestive 2.8 LOD score threshold established for single traits seems reasonable in this situation as it corresponds to a genomewide significance level of 0.004 when considering two traits at a time.

9.8.2 Identifying Modules of Highly Interconnected Genes in Coexpression Networks

Given the scale-free and hierarchical nature of coexpression networks (Barabasi and Oltvai, 2004; Ghazalpour et al., 2006; Lum et al., 2006), one of the problems is to identify the key network modules in the network, representing those hub nodes (nodes that are significantly correlated with many other nodes) that are highly interconnected with one another, but that are not as highly connected with other hub nodes. Plate 9(a) illustrates a topological connectivity map for the most highly connected genes in the adipose tissue of the BXH cross (Chen et al., 2007). After hierarchically clustering both dimensions of this plot, the network is seen to break out into clearly identifiable modules. Gene–gene

coexpression networks are highly connected, and the clustering results shown in Plate 9 illustrate that there are gene modules arranged hierarchically within these networks.

Ravasz *et al.* (2002) used manually selected height cutoff to separate tree branches after hierarchical clustering, in contrast to Lee *et al.* (2004), who formed maximally coherent gene modules with respect to GO functional categories. Another strategy is to employ a measure similar to that used by Lee *et al.* (2004), but without the dependence on the GO functional annotations, given it is of interest to determine independently whether coexpression modules are enriched for GO functional annotations. A gene module in the coexpression network is defined as a maximum set of interconnected genes. A coherence measure for a given gene module can be defined as

$$\text{Coherence} = \frac{GP_{\text{obs}}}{GP_{\text{tot}}},$$

where GP_{obs} is the number of gene pairs that are connected and GP_{tot} is the total number of possible gene pairs in the module. The efficiency of a gene module can then be defined as

$$\text{Efficiency} = \frac{\text{Coherence} \times G_{\text{mod}}}{G_{\text{net}}},$$

where G_{mod} is the number of genes in the module and G_{net} is the number of genes in the network. Given these definitions, we define a process to iteratively construct gene modules by the following steps:

1. Order genes in the gene–gene connectivity matrix according to an agglomerative hierarchical clustering algorithm as previously described (Hughes *et al.*, 2000).

2. Calculate the efficiency $e_{i,j}$ for every possible module, including genes from i to j as given in the ordered connectivity matrix, where $j \geq i + 9$ (i.e. minimum module size is 10), using a dynamic programming algorithm.

3. Determine the maximum $e_{i,j}$.

4. Set $e_{i \ldots j, 1 \ldots G_{\text{net}}} = 0$ and $e_{1 \ldots G_{\text{net}}, i \ldots j} = 0$.

5. Go to step 3 until no additional modules can be found.

The modules identified in this way are informative for identifying the functional components of the network that may underlie complex traits like disease (Lum *et al.*, 2006). It has been demonstrated that the types of modules depicted in Plate 9(a) are enriched for known biological pathways, enriched for genes that are associated with disease traits, and enriched for genes that are linked to common genetic loci (Ghazalpour *et al.*, 2006; Lum *et al.*, 2006).

9.9 USING eQTL DATA TO RECONSTRUCT PROBABILISTIC NETWORKS

Coexpression networks are informative as discussed for gross characterizations of the properties of biological networks, identification of highly connected (hub) nodes, and

identification of functional modules that aid in the characterization of subnetworks associated with disease. Despite these and other advantages, coexpression networks do not provide explicit details on the connectivity structure among genes in the network, including the representation of causal associations among genes and between genes and phenotypes. As detailed in previous sections, naturally occurring variations in DNA can be leveraged as a systematic source of perturbations to infer causal associations among gene expression traits and between gene expression and clinical traits. By examining the effects naturally occurring DNA have on variations in gene expression traits in human or experimental populations, other phenotypes (including disease) can then be examined with respect to these same DNA variations and ultimately ordered with respect to genes to infer causal control (Plate 8) (Mehrabian *et al.*, 2005; Schadt *et al.*, 2005; Kulp and Jagalur, 2006; Lum *et al.*, 2006).

While previous sections highlighted inferring causal associations by integrating eQTL and quantitative trait data for two traits, Zhu *et al.* were among the first to formally incorporate genetic data into the reconstruction of whole gene networks using Bayesian network reconstruction methods (Zhu *et al.*, 2004). Bayesian networks are directed acyclic graphs that, while limited with respect to representing temporal information or feedback loops, allow for the explicit representation of causal associations among nodes in the network. With Bayesian network reconstruction methods taking gene expression data as the only source of input, many relationships between genes in such a setting are Markov equivalent (symmetric), similar to what was discussed for three-node graphs in Figure 9.1(a). This means one cannot statistically distinguish whether a given gene causes another gene to change or vice versa. To break this symmetry, Zhu *et al.* incorporated eQTL data as prior information to establish more reliably the correct direction among expression traits.

Bayesian network methods have been applied previously to reconstruct networks comprised only of expression traits, as well as to networks comprised of both expression and disease traits, where the aim has been to identify those portions of the network that are driving a given disease trait. Forming candidate relationships among genes was carried out using an extension of standard Bayesian network reconstruction methods (Chiellini *et al.*, 2002). In the first approach to extend this method using genetic data, QTL information for the transcript abundances of each gene considered in the network was incorporated into the reconstruction process. It is well known that searching for the best possible network linking a moderately sized set of genes is an NP-hard problem. Exhaustively searching for the optimal network with hundreds of genes is presently a computationally intractable problem. Therefore, various simplifications are typically applied to reduce the size of the search space and to reduce the number of parameters that need to be estimated from the data. Two simplifying assumptions to achieve such reductions are commonly employed. First, while any gene in a biological system can control many other genes, a given gene can be restricted so that it is allowed to be controlled by a reduced set of genes. Second, the set of genes that can be considered as possible causal drivers (parent nodes) for a given gene can be restricted using the type of causality arguments discussed in previous sections, as opposed to allowing for the possibility of any gene in the complete gene set to serve as a parent node.

One method to select potential parents for each gene is to assess the extent of genetic overlap between any two gene expression traits. If RNA levels for two genes are tightly associated, or if such levels are genetically controlled by a similar set of loci, then their

eQTL should overlap. The extent of QTL overlap can be carried out using the PET procedure described in a previous section, or, to consider simultaneously more of the eQTL data for any two traits, the correlation coefficient between vectors of LOD scores associated with eQTL identified over entire chromosomes for each gene can be computed. If we assume no epistasis between eQTL for a given expression trait, then the eQTL can be considered independent between chromosomes, and a measure of genetic relatedness over all chromosomes for any two traits can be subsequently computed as a weighted average of correlations for each individual chromosome:

$$r = \sum_c w(c) \times r(c), \tag{9.3}$$

where $r(c)$ is the correlation coefficient for chromosome c and $w(c)$ is the chromosome-specific weight. The chromosome-specific correlation coefficient is given by

$$r(c) = \frac{\sum_l x_c(l) \times y_c(l) \times I_c(l)}{\sum_l x_c(l) \times x_c(l) + \sum_l y_c(l) \times y_c(l)}. \tag{9.4}$$

In this equation, $x_c(l)$ and $y_c(l)$ are LOD scores at locus l on chromosome c and $I_c(l)$ is an indicator function defined as

$$I_c(l) = \begin{cases} 1, & \text{if } x_c(l) > 1.5, y_c(l) > 1.5 \\ 0, & \text{otherwise} \end{cases}, \tag{9.5}$$

which is incorporated to eliminate low LOD scores that would have likely only contributed noise to the correlation measures. The chromosome-specific weight terms are given by

$$w(c) = \max_l(\min(x_c(l), 10) \times \min(y_c(l), 10)). \tag{9.6}$$

In effect, this equation provides for those chromosomes with high LOD scores to more significantly influence the overall correlation measure. This heuristic weight is intuitively appealing since gene expression traits with common significant eQTL have a larger percentage of their overall variation explained by these common genetic effects. For a particular data set, the weighted correlations can be computed for all gene pairs, and the resulting list rank ordered, so that those genes in the upper percentiles can be chosen as candidate parental nodes (Zhu *et al.* (2004) used the 80th percentile as the threshold).

To differentiate between colocalization of QTL for a pair of traits due to common genetic effects (pleiotropy) versus multiple closely linked QTL, the extent of phenotypic (RNA levels) association can be measured by the mutual information measure

$$mi(A, B) = \sum_{i,j} p(a_i, b_j) \log \frac{p(a_i, b_j)}{p(a_i)p(b_j)}, \tag{9.7}$$

where $p(x)$ is the probability density function for the expression of gene X in the system of interest. For a particular data set, the mutual information measure can be computed for all gene pairs, and the resulting list rank ordered on this measure, so that those genes in the upper percentile can be chosen as candidate parental nodes (Zhu *et al.* (2004) again used the 80th percentile as the threshold).

The selection of the genes in the 80th percentile of the rank-ordered lists generated from the correlation and mutual information measures provides the prior evidence that two genes may be causally related. The QTL data provide the causal anchors that allow this type of inference to be made. That is, by definition, a QTL controlling for the expression of two gene expression traits implies that DNA variations in the QTL lead to variations in the expression of the associated gene traits. Therefore, it must be the case that any gene expression trait pair controlled by a common QTL is either independently driven by the same QTL, or causally associated in that one of the two traits is driven by the QTL, while the other trait responds to the trait driven by the QTL.

In addition to utilizing the QTL information as prior information to restrict the types of relationships that can be established among genes, the QTL information can be more intimately integrated into the network reconstruction process. As indicated in previous sections, correlation measures are symmetric and so can indicate association but not causality. However, QTL mapping information for the gene expression traits can be used to help sort out causal relationships. The different tests described in the section on making causal inferences between pairs of traits provide one way to explicitly sort out such relationships. Zhu *et al.* (2004) leveraged the eQTL data in a different way to make similar types of inferences in their network reconstruction algorithm. Suppose gene expression trait X has two high confidence eQTL at loci L_1 and L_2, while gene expression trait Y has a single eQTL at L_1 that is more significant than the eQTL for X at L_1. In this instance, it is reasonable to infer that Y may control X (or is 'causal' for X), since if X were causal for Y we would expect Y to have an eQTL at L_2 in addition to the eQTL at L_1, given X has an eQTL at L_2. Further, the asymmetry in the significance of the eQTL at L_1 for X and Y also favors Y as being causal for X. Thus, the eQTL overlap information can be used in this more heuristic way to infer causality by defining the prior for a candidate relationship as

$$p(X \to Y) = r(X, Y) \frac{N(Y)}{N(X) + N(Y)}, \qquad (9.8)$$

where N is a gene expression trait's complexity as measured by the number of significant eQTL mapped for the given gene expression trait. If X and Y have coincident QTL, but the overall complexity of Y is greater than the overall complexity of X, then the relationship X causes $Y (X \to Y)$, as opposed to $Y \to X$, is weighted as the more likely scenario in the prior defined above. Implicit in this weighting scheme is the assumption that traits driven by common QTL are causally related (i.e. one trait drives the other), even though it is possible that the traits could be independently driven by the same set of QTL. However, in cases where multiple traits are independently driven by a common set of QTL, the correlation between the traits is smaller than when the traits are causally related, so the prior in this case carries less weight. Further, the conditional mutual information measure discussed below also serves to prevent, in at least some cases, causal links from being made between genes that are independently driven by a common set of QTL.

The causality relationships between gene expression traits can be further assessed by considering whether the QTL for a particular gene expression trait is proximal. As discussed previously, genes that give rise to eQTL that are coincident with the gene's physical location most likely harbor DNA variations that at least partially explain variations in the gene's observed RNA levels (proximal or cis-acting eQTL) (Doss *et al.*,

2005). Therefore, it is reasonable to infer that a gene with a proximal eQTL is at least partially under the control of the gene itself. In this particular situation, where another gene expression trait Y has an eQTL mapping to the physical location of gene X, and $(X \rightarrow Y)$ is inferred as the most likely relationship using a test for causality such as that described in previous sections, we can infer that Y is likely not causal for X and set $p(Y \rightarrow X)$ equal to 0.

With the various constraints and measures defined above, the goal in reconstructing whole gene networks is to find a graphical model M (a gene network) that best represents the relationships between genes, given a gene expression data set, D, of interest. That is, given data D, we seek to find the model M with the highest posterior probability $P(M|D)$. The prior probability $P(M)$ of model M is

$$p(M) = \prod_{X \rightarrow Y} p(X \rightarrow Y), \tag{9.9}$$

where the product is taken over all paths in the network (M) under consideration. The algorithm Zhu *et al.* (2004) employed to search through all possible models to find the network that best fits the data is similar to the local maximum search algorithm implemented by Friedman *et al.* (2000).

The Bayesian network reconstruction algorithm can be employed to elucidate the module connectivity structure depicted in Plate 9(a). Because reconstruction of Bayesian networks is an NP-hard problem (Garey and Johnson, 1979), the number of nodes that can be considered in the network and the extent of connections (edges) among these nodes must be reduced (over what can be considered in reconstructing coexpression networks) in order to make the problem tractable, thereby making such networks more sparse compared to coexpression networks. Toward this end, Plate 9(b) shows the result of the Bayesian network reconstruction algorithm discussed above applied to module 2 depicted in Plate 9(a) for the BXH cross. Further highlighted in Plate 9(c) is a subnetwork containing the gene *Lpl*, a gene recently identified as causal for obesity in the BXH cross (Schadt *et al.*, Submitted). The more detailed structure provided in Plate 9(b) and (c) allows for the examination of the context in which specific genes like *Lpl* operate, providing insights into which parts of the network may impact a given gene's function, as well as what other parts of the network may be impacted by the gene's function.

9.10   CONCLUSIONS

The eQTL mapping methods and network reconstruction methods, including the causality test, discussed here provide a convenient framework for moving beyond examining genes one at a time to understand complex phenotypes like common human diseases. Whether considering the relationship between two traits with respect to common QTL driving each of the traits, interaction networks, Bayesian networks, or other types of networks reconstructed by integrating genetic and molecular profiling data, the advantage the network view affords is the ability to consider more of the raw data informing on complex phenotypes simultaneously, taking into account dependency structures between all of the different fundamental components of the system, and providing a framework to

integrate a diversity of data types (something Bayesian network reconstruction methods are particularly good at). Researchers in the life and biomedical sciences will have few other options available to them in considering the vast amounts of data being generated to elucidate how the hundreds of thousands or even millions of fundamental components within the cell interact to give rise to complex phenotypes. Networks provide one of the only frameworks of which we are aware to systematically and simultaneously take all of the fundamental components into account. Statistical inferences on networks will come to define much of the future research needed in this field to adequately leverage what network models can provide.

Of course, the types of approaches reviewed here represent only the first steps being taken to reconstruct meaningful gene networks, and even in this chapter we have largely restricted attention to genetic, gene expression, and disease phenotype data. Ultimately, it will be necessary to integrate many different lines of experimental data simultaneously. Protein–protein interactions, protein–DNA interactions, protein–RNA interactions, RNA–RNA interactions, protein state, methylation state, and especially differential methylation states that have now been shown to act transgenerationally (Anway *et al.*, 2005), and interactions with metabolites, among other interactions, are all important components that define complex phenotypes that emerge in living systems. What a given protein and RNA does will give way to what a network of protein, RNA, DNA, and metabolite interactions do, where such networks of interaction are defined by the context in which they operate, with environment playing a critical role. A particular network state that drives disease (or other complex phenotypes that define living systems) will not only require knowledge of DNA and environmental variation and the changes these variation components induce in the network but also information on the previous states of the network that led to the current state, where environmental stresses interacting in complex ways with genetic background not only influence the current state of the network but can also lead to longer lasting effects on the network that act transgenerationally.

While this more comprehensive reconstruction of biological networks is still outside the scope of what is presently doable, the types of approaches discussed here represent useful first steps toward this ultimate goal. Even though the number of networks that can be reconstructed from the fundamental components of living systems is truly daunting, as work progresses in this area we will learn the rules that necessarily constrain the possible ranges of molecular interactions, and as a result begin to capture the more conserved network motifs that form the framework upon which all other interactions are based. The complexity revealed by a systems-biology motivated approach to elucidating complex phenotypes like disease should be embraced, given the potential to develop a better understanding of the true diversity of disease and the constellation of genes that need to be targeted to effectively treat disease.

9.11   SOFTWARE

A variety of software is available for the analysis of genetic data. See **Chapters 15**, **29**, **30**, **31**, **32** for details.
Software for Kendziorski's Mixture of Markers method: `http://www.biostat.wisc.edu/~kendzior/MOM/`

REFERENCES

Alberts, R., Terpstra, P., Bystrykh, L.V., de Haan, G. and Jansen, R.C. (2005). A statistical multiprobe model for analyzing cis and trans genes in genetical genomics experiments with short-oligonucleotide arrays. *Genetics* **171**, 1437–1439.

Anway, M.D., Cupp, A.S., Uzumcu, M. and Skinner, M.K. (2005). Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science* **308**, 1466–1469.

Barabasi, A.L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509–512.

Barabasi, A.L. and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101–113.

Brem, R.B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 1572–1577.

Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755.

Brem, R.B., Storey J.D. Whittle, J, Kruglyak, L., (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**(7051), 701–703.

Broman, K.W. (1997). Identifying quantitative trait loci in experimental crosses. Ph.D. dissertation, University of California, Berkeley, CA.

Cervino, A.C., Li, G., Edwards, S., Zhu, J., Laurie, C., Tokiwa, G., Lum, P.Y., Wang, S., Castellini, L.W., Lusis, A.J., Carlson, S., Sachs, A.B. and Schadt, E.E. (2005). Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels. *Genomics* **86**, 505–517.

Chesler, E.J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H.C., Mountz, J.D., Baldwin, N.E., Langston, M.A., Threadgill, D.W., Manly, K.F. and Williams, R.W. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics* **37**, 233–242.

Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M. and Burdick, J.T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369.

Chiellini, C., Bertacca, A., Novelli, S.E., Gorgun, C.Z., Ciccarone, A., Giordano, A., Xu, H., Soukas, A., Costa, M., Gandini, D., Dimitri, R., Bottone, P., Cecchetti, P., Pardini, E., Perego, L., Navalesi, R., Folli, F., Benzi, L., Cinti, S., Friedman, J.M., Hotamisligil, G.S. and Maffei, M. (2002). Obesity modulates the expression of haptoglobin in the white adipose tissue via TNFalpha. *Journal of Cellular Physiology* **190**, 251–258.

Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Zhang, C., Edwards, S., Sieberts, S.K., Leonardson, A., Castellini, L.W., Wang, S., Doss, S., Ghazalpour, A., Horvath, S., Drake, T.A., Lusis, A.J., Schadt, E.E. (2007). Co-expression networks reconstructed in segregating mouse populations enable a more comprehensive dissection of complex disease associated traits.

DeCook, R., Lall, S., Nettleton, D. and Howell, S.H. (2006). Genetic regulation of gene expression during shoot development in Arabidopsis. *Genetics* **172**, 1155–1164.

DePrimo, S.E., Wong, L.M., Khatry, D.B., Nicholas, S.L., Manning, W.C., Smolich, B.D., O'Farrell, A.M. and Cherrington, J.M. (2003). Expression profiling of blood samples from an SU5416 Phase III metastatic colorectal cancer clinical trial: a novel strategy for biomarker identification. *BMC Cancer* **3**, 3.

Dixon, P.M. and Pechmann, J.H.K. (2005). A statistical test to show negligible trend. *Ecology* **86**, 1751–1756.

Doss, S., Schadt, E.E., Drake, T.A. and Lusis, A.J. (2005). Cis-acting expression quantitative trait loci in mice. *Genome Research* **15**, 681–691.

Friedman N., Linial, M., Nachman I., Pe'er D. (2000). Using Bayesian networksto analyze expression data. *Journal of Computational Biology* **7**, 601–620.

Garey, M.R. and Johnson, D.S. (1979). *Computers and Intractability : A Guide to the Theory of NP-Completeness*. W. H. Freeman, San Francisco, CA.

Gargalovic, P.S., Imura, M., Zhang, B., Gharavi, N.M., Clark, M.J., Pagnon, J., Yang, W.P., He, A., Truong, A., Patel, S., Nelson, S.F., Horvath, S., Berliner, J.A., Kirchgessner, T.G. and Lusis, A.J. (2006). Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12741–12746.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E.E., Drake, T.A., Lusis, A.J. and Horvath, S. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genetics* **2**(8) e130.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P. and Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93.

Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., Musilova, A., Kren, V., Causton, H., Game, L., Born, G., Schmidt, S., Muller, A., Cook, S.A., Kurtz, T.W., Whittaker, J., Pravenec, M. and Aitman, T.J. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics* **37**, 243–253.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. and Friend, S.H. (2000). Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126.

Jansen, R.C. and Nap, J.P. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics* **17**, 388–391.

Jiang, C. and Zeng, Z.B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**, 1111–1127.

Jin, W., Riley, R.M., Wolfinger, R.D., White, K.P., Passador-Gurgel, G. and Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nature Genetics* **29**, 389–395.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R. and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**, 2141–2144.

Karp, C.L., Grupe, A., Schadt, E., Ewart, S.L., Keane-Moore, M., Cuomo, P.J., Kohl, J., Wahl, L., Kuperman, D., Germer, S., Aud, D., Peltz, G. and Wills-Karp, M. (2000). Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nature Immunology* **1**, 221–226.

Kendziorski, C.M., Chen, M., Yuan, M., Lan, H. and Attie, A.D. (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62**, 19–27.

Kim, J.K., Gabel, H.W., Kamath, R.S., Tewari, M., Pasquinelli, A., Rual, J.F., Kennedy, S., Dybbs, M., Bertin, N., Kaplan, J.M., Vidal, M. and Ruvkun, G. (2005). Functional genomic analysis of RNA interference in C. elegans. *Science* **308**, 1164–1167.

Klose, J., Nock, C., Herrmann, M., Stuhler, K., Marcus, K., Bluggel, M., Krause, E., Schalkwyk, L.C., Rastan, S., Brown, S.D., Bussow, K., Himmelbauer, H. and Lehrach, H. (2002). Genetic analysis of the mouse brain proteome. *Nature Genetics* **30**, 385–393.

Kulp, D.C. and Jagalur, M. (2006). Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics* **7**, 125.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**, 241–247.

Lee, I., Date, S.V., Adai, A.T. and Marcotte, E.M. (2004). A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558.

Lum, P.Y., Chen, Y., Zhu, J., Lamb, J., Melmed, S., Wang, S., Drake, T.A., Lusis, A.J. and Schadt, E.E. (2006). Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *Journal of Neurochemistry* **97**(s1), 50–62.

Mehrabian, M., Allayee, H., Stockton, J., Lum, P.Y., Drake, T.A., Castellani, L.W., Suh, M., Armour, C., Edwards, S., Lamb, J., Lusis, A.J. and Schadt, E.E. (2005). Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nature Genetics* **37**, 1224–1233.

Monks, S.A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J.W., Sachs, A. and Schadt, E.E. (2004). Genetic inheritance of gene expression in human cell lines. *American Journal of Human Genetics* **75**, 1094–1105.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L.C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**, 267–273.

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747.

Mural, R.J., Adams, M.D., Myers, E.W., Smith, H.O., Miklos, G.L., Wides, R., Halpern, A., *et al.* (2002). A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**, 1661–1671.

Oleksiak, M.F., Churchill, G.A. and Crawford, D.L. (2002). Variation in gene expression within and among natural populations. *Nature Genetics* **32**, 261–266.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA.

Pearl, J. (2000). *Causality*. Cambridge University Press, New York.

Petretto, E., Mangion, J., Dickens, N.J., Cook, S.A., Kumaran, M.K., Lu, H., Fischer, J., Maatz, H., Kren, V., Pravenec, M., Hubner, N. and Aitman, T.J. (2006a). Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genetics* **2**, e172.

Petretto, E., Mangion, J., Pravanec, M., Hubner, N. and Aitman, T.J. (2006b). Integrated gene expression profiling and linkage analysis in the rat. *Mammalian Genome* **17**, 480–489.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabasi, A.L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555.

Schadt, E.E. (2005). Exploiting naturally occurring DNA variation and molecular profiling data to dissect disease and drug response traits. *Current Opinion in Biotechnology* **16**, 647–654.

Schadt, E.E. (2006). Novel integrative genomics strategies to identify genes for complex traits. *Animal Genetics* **37**(Suppl. 1), 18–23.

Schadt, E.E., Edwards, S.W., GuhaThakurta, D., Holder, D., Ying, L., Svetnik, V., Leonardson, A., Hart, K.W., Russell, A., Li, G., Cavet, G., Castle, J., McDonagh, P., Kan, Z., Chen, R., Kasarskis, A., Margarint, M., Caceres, R.M., Johnson, J.M., Armour, C.D., Garrett-Engele, P.W., Tsinoremas, N.F. and Shoemaker, D.D. (2004). A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biolog* **5**, R73.

Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., Lum, P.Y., Leonardson, A., Thieringer, R., Metzger, J.M., Yang, L., Castle, J., Zhu, H., Kash, S.F., Drake, T.A., Sachs, A. and Lusis, A.J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37**, 710–717.

Schadt, E.E. and Lum, P.Y. (2006). Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *Journal of Lipid Research* **47**(12), 2601–2613.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B. and Friend, S.H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302.

Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., Wu, L.F., Altschuler, S.J., Edwards, S., King, J., Tsang, J.S., Schimmack, G., Schelter, J.M., Koch, J., Ziman, M., Marton, M.J., Li, B., Cundiff, P., Ward, T., Castle, J., Krolewski, M., Meyer, M.R., Mao, M., Burchard, J., Kidd, M.J., Dai, H., Phillips, J.W., Linsley, P.S., Stoughton, R., Scherer, S., Boguski, M.S. (2001). Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922–927.

Storey, J.D., Akey, J.M. and Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology* **3**, e267.

Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavare, S., Deloukas, P. and Dermitzakis, E.T. (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genetics* **1**, e78.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

Waring, J.F., Jolly, R.A., Ciurlionis, R., Lum, P.Y., Praestgaard, J.T., Morfitt, D.C., Buratto, B., Roberts, C., Schadt, E. and Ulrich, R.G. (2001). Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. *Toxicology and Applied Pharmacology* **175**, 28–42.

Yang, X., Schadt, E.E., Wang, S., Wang, H., Arnold, A.P., Ingram-Drake, L., Drake, T.A. and Lusis, A.J. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Research* **16**, 995–1004.

Zeng, Z.B., Liu, J., Stam, L.F., Kao, C.H., Mercer, J.M. and Laurie, C.C. (2000). Genetic architecture of a morphological shape difference between two Drosophila species. *Genetics* **154**, 299–310.

Zhu, J., Lum, P.Y., Lamb, J., GuhaThakurta, D., Edwards, S.W., Thieringer, R., Berger, J.P., Wu, M.S., Thompson, J., Sachs, A.B. and Schadt, E.E. (2004). An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenetic and Genome Research* **105**, 363–374.

Protein Structure Prediction

D.P. Klose and W.R. Taylor

Division of Mathematical Biology, National Institute of Medical Research, London, UK

In this chapter, we assume that most readers know a little about proteins and protein structure prediction, but that their knowledge might now be fuzzy and out of date. In order to refresh their knowledge and get novices up to speed, we concentrate on the most basic principles. We outline the categories of protein structure prediction before examining the underlying methods used in 'successful' prediction tools. Each method will be accompanied by examples of biological applications from recent papers and a section highlighting their advantages and disadvantages.

10.1 HISTORY

While knowledge about protein structure has a long history going back to Astbury's early experiments on hair and silk (and beyond), its current form came into being with the solution, by X-ray crystallography, of the first globular proteins at near atomic resolution. These consisted of two closely related globin structures (by Kendrew and Perutz and co-workers), followed by the first enzyme structure (lysozyme, by Phillips and co-workers). These structures were somewhat unexpected at the time: given the near crystalline regularity of Astbury's hair and silk structures (along with the structure for DNA proposed then), the globular proteins were described by Kendrew as *lacking symmetry* and *visceral*. Even worse, the structure of lysozyme did not immediately provide any clue as to how it, or any enzyme, might work!

   As more structures were solved, some of these early unexpected aspects were partially rectified: symmetry was rediscovered in the structure of triosephosphate isomerase, and plausible catalytic mechanisms for lysozyme and other enzymes were established. However, despite progress, it remains true that proteins do not yield their secrets easily. When a new structure is solved, it is often still difficult to describe its structure in a systematic way and place it in relation to other structures. Similarly, if a structure is solved for a protein of unknown function, it is not necessarily easy or possible to suggest a function from just the structure.

10.2   BASIC STRUCTURAL BIOLOGY

The proteins considered in this chapter are the globular proteins. Fibrous proteins are neglected because they are less interesting – consisting of numerous repeated units forming some trivial helical arrangement. Membrane proteins are neglected because much less is known about them structurally. What is known reveals that they are effectively confined to just two dimensions (in the membrane) and are therefore unable to explore the structural richness available in three dimensions. This leaves globular proteins, which exhibit the greatest variety of structures and functions.

Globular proteins form complex structures consisting of one or more chains of amino acids known as *polypeptides*. For over 30 years it has been generally accepted that the amino acid sequence provides sufficient information to specify the overall structure the protein adopts in three-dimensional (3D) space. This theory is largely accepted owing to work by Anfinsen and co-workers in the 1960s (Anfinsen, 1972; 1973). Since then, several schemes have been devised to exhibit the physico-chemical properties of the amino acids – naturally the authors' favourite was outlined by Taylor (1986) (see Figure 10.1).
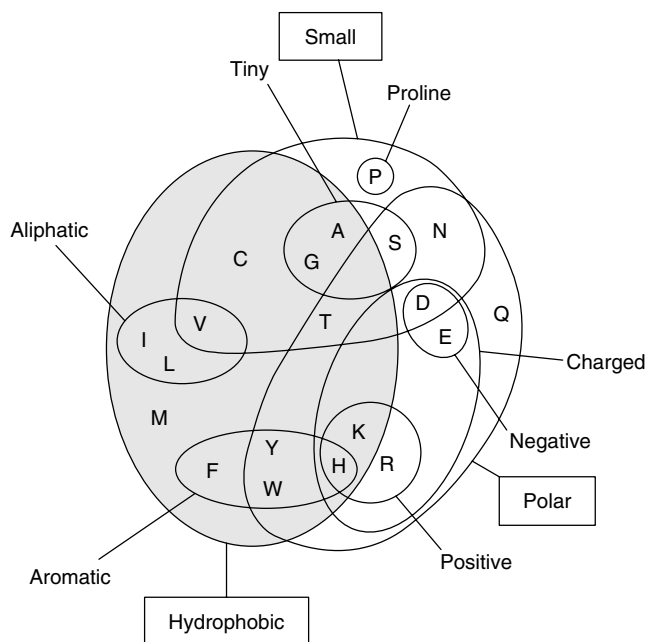
Figure 10.1   Amino acid property Venn diagram. Taylor's Venn diagram (Taylor, 1986) shows the relationships of the 20 naturally occurring amino acids and their physico-chemical properties. The two major sets are those that exhibit hydrophobic behaviour (shaded grey) and those that have a polar group (referred to as *hydrophilic*). The third largest category (labelled small) contains nine amino acids, and within this group there are the 'tiny' amino acids that contain, at most, two side-chain atoms. In this representation, cysteine is represented once; however, it is reasonable to represent it twice, in reduced form, where it behaves similarly to serine, or in its oxidised form where it behaves more like valine.

The amino acids can be roughly divided into two groups based solely on their side-chain properties. The first group, termed *hydrophilic* or *polar residues*, are soluble in water (unshaded), and the second group (shaded grey) are less soluble in water and are termed *hydrophobic* or *non-polar residues*. The properties of the amino acids result in the formation of internal structures: a hydrophobic core and secondary structure elements. These features are integral to the overall structure of the protein and from a theoretical perspective it is crucial that they can be predicted with some degree of accuracy.

10.2.1   The Hydrophobic Core

The partitioning of the two types of residues with respect to the solvent (essentially water) is the overall driving force behind protein folding and stability. While hydrogen bonding is important in maintaining specific structural features, it is crucial to remember that for every hydrogen bond formed internally two bonds with water are lost and one water–water bond formed – a simple bond count reveals no net gain. The resulting structure is a core of hydrophobic residues (from which water has been excluded), surrounded by a shell of hydrophilic residues, which interface with the solvent thus making the protein soluble. Experimental biologists should recognise that the white precipitate in their buffers is most likely the result of the collapse of the residue partition. Bearing this in mind, it should be clear how crucial it is to be able to create a properly formed hydrophobic core in predicted structures.

10.2.2   Secondary Structure

Here three broad categories of secondary structure will be introduced. The categories are alpha (α); beta (β), and coil/loop regions. The α and β categories are defined by main chain hydrogen bonding and combined phi (ϕ) and psi (ψ) angle repetitions. Helical structures are formed by local hydrogen bonding while β structures can be formed by distant parts of the backbone. The archetypal α structure is formed by hydrogen bonding between the CO of residue i and residue $i + 4$ with $\phi \approx -60^{\circ}$ and $\psi \approx -40^{\circ}$. Beta sheets have $\phi \approx -120^{\circ}$ and $\psi \approx 140^{\circ}$. Beta structures are less local and modular than helices owing to the fact that the hydrogen bonds occur between strands, and the result is not a single β strand but a pair, which can bond either in parallel or in anti-parallel arrangements. Over 30 years ago, computers were first used to predict secondary structure (Nagano, 1973). Initially, and for some time, the average accuracy of predictions ranged between 50 and 60 %, only 10 % above that of a random prediction – 40 % based on the percentage occurrence of secondary structure elements within the protein data bank (PDB) (Simossis, 2005). Today, using methods that are introduced later in the chapter, it is possible to obtain accuracies of up to 80 % while the average has increased to $\sim$75 % (Montgomerie *et al.*, 2006).

10.3   PROTEIN STRUCTURE PREDICTION

To take a protein sequence and predict the three-dimensional structure into which it folds has been a goal for many years. Yet, despite faster computers and better theories of protein

folding, no one has been able to succeed in this task for anything larger than a peptide of 30–40 residues (Duan and Kollman, 1998) – all of which have simple folding involving only local contacts. An alternative to this direct (*ab initio*) approach is to use what is known from the structures that have been seen to provide constraints on the structures that are to be predicted. These can range from local constraints – such as the preference of secondary structures to connect in a particular way – up to global constraints, which can be inferred if there is some sequence similarity to a protein of known fold. The latter end of this range can obviously lead to good predictions of structure (perfect when there is 100 % identity between the two sequences) but as the sequence similarity diminishes, then the less certain local rules must be relied on more heavily. The key to the success of this approach is to know when a similarity is significant.

10.3.1   Homology Modelling

Where there is clear sequence similarity, the problem of constructing a 3D model is largely how to substitute the existing side chains with the new side chains to which they have been matched (aligned). Since there is clear similarity, many of these will be the same and most will involve only minor substitution of groups (e.g. Asp → Asn). In addition, most of the substitutions will occur on the surface of the protein, leaving much of the tightly packed core intact. This is due to evolutionary constraints, which mean that it is easier to change a side chain when it is not buried. In this situation, the simple axiom, 'if it ain't broke, don't fix it' is the best advice to follow. Indeed, even better advice is to let it all be done automatically as there are now several programs that can construct good models provided the sequences are clearly related. Many of these are commercial but the Swiss-model (Schwede *et al.*, 2003) program can be used freely over the internet for non-commercial purposes.

In situations with less sequence similarity, the problem of indels (relative insertions and deletions) becomes important. These also imply that the protein backbone will need to be remodelled (to close the gaps after deletion, or add new chain for an insertion). Fortunately, again most of these changes will be found on the protein surface where there is usually space to make larger changes. If changes apparently do not occur on the surface, then this is a strong indication that the alignment between sequence and structure is incorrect. While such problems can be attempted to be solved using programs like Swiss model, the limiting factor becomes being able to specify the correct alignment on which to base the model – here the realm of threading must be entered.

10.3.2   Threading

The strategy of aligning the protein sequences using standard alignment methods and then building a model is no longer the standard approach for distantly related sequences. Interaction is needed between the emerging model and the alignment. As mentioned above, if an insertion is found in the core, then the answer is usually to change the alignment – not the structure. Historically, this was carried out in a series of iterations with manual realignment at each stage as, for example, in the construction of the HIV protease model (Pearl and Taylor, 1987). Eventually it became apparent that this progress could become more automated with the alignment and model being calculated simultaneously.

10.3.3 True Threading

To thread a sequence over a structure, two components are necessary: a packing measure for the substituted amino acid and an alignment method that can optimise the sum of packing scores. The former is available in the 'rough' empirical potentials of Sippl (1990) (referred to as *potentials of mean force*). These are 'rough' in that they do not directly consider side-chain interactions (to do so would be impossible until the full model was constructed) but capture the preference of an amino acid to be in a particular environment and (indirectly) secondary structure state. The second component, the alignment method, is readily available from sequence alignment methods, but could not be used directly. One solution is to apply the alignment in a series of iterations, gradually substituting new residues into the existing structure. Another solution is to take the double dynamic programming method (from structure comparison) and modify it (Jones *et al.*, 1992).

10.3.4 3D/1D Alignment

The sequence/structure matching problem was also approached from the sequence alignment side. Beginning with a pure sequence alignment, structural features are predicted (such as secondary structure state and degree of burial), which are then matched to the known corresponding features of a protein structure (along with its sequence) (Bowie *et al.*, 1991; Luthy *et al.*, 1991; Rice and Eisenberg, 1997). Unlike the 'true' threading methods described above, this approach does not take into account the 3D interactions in the final calculation of the alignment, so it can use the dynamic programming algorithm without any complications.

   In theory, the 3D/1D approach is less powerful than the 'true' threading methods. However, when applied to very distant relationships, there is little perceptible difference in the methods. This probably results from the more common incorporation of multiple sequence data into the 3D/1D methods and from accurate prediction of secondary structure. The 'true' threading methods also make the assumption that the basic core structure of the model protein will be the same as the structure on which it was built: for distant relationships this is seldom completely true.

10.3.5 New Fold (NF) Prediction (*Ab initio* and *De novo* Approaches)

The last resort – although arguably the most exciting – is *ab initio* prediction. *Ab initio* (from first principles) prediction relies on the assumption that natively folded proteins exist in a state of low free energy. To obtain the structure of a protein, one simply has to compute all possible interactions between all residues in a sequence until the lowest free energy conformation is found! In reality, this problem is far from trivial – in fact, it has only been done for short (30–40 residue) polypeptides (Duan and Kollman, 1998). Frustrated with a lack of progress, *ab initio* modellers have begun to use existing structural information, often in the form of fragment packing, allowing for proteins up to 100 residues to be predicted – although not from first principles (Rohl and Baker, 2002; Bradley *et al.*, 2005). In order to retain correct nomenclature – as well as to keep physicists happy – the name of the field has been changed to *de novo* or new fold (NF) modelling.

10.4   MODEL EVALUATION

Regardless of the method used to generate the protein structures, many thousands of models can be produced by a single prediction attempt. Some models resemble real proteins, with well-formed secondary structure and a hydrophobic core, while the majority will be poorly formed with little or no ordered secondary structure and 'unnatural' packing in the hydrophobic core (exposed hydrophobics and buried hydrophilics). To build and evaluate models that resemble real proteins, it is crucial to use reliable prediction and evaluation functions, such as YASPIN (Lin *et al.*, 2005) for secondary structure prediction and PHOBIC, an updated version of burial/hydrophobic matching used in Taylor *et al.* (2006), for hydrophobic core evaluation. These functions are constructed using a diverse mixture of methods to numerous to review here (Skolnick, 2006). A number of 'popular' techniques that have been used individually as well as in a combinatorial fashion to predict secondary structure to overall model structure are presented here.

10.4.1   Decision Trees

The classification and regression tree (CART) algorithm was popularised by Breiman *et al.* (1984). Decision trees (DTs) describe a tree structure wherein the leaves represent classifications and branches represent conjunctions of features that lead to associated classifications. DTs are frequently referred to as *classification trees* – used for prediction of class membership (i.e. is residue i buried or exposed) – or *regression trees* – used for assigning a continuous value (i.e. the number of contacts residue i has). As with other methods, the input to a CART should follow the form $(x, y) = (x_1, x_2, x_3, \ldots x_n, y)$, where y corresponds to the label or independent variable and $x_1 \ldots x_n$ correspond to the attributes that define y. The most fundamental concept is splitting the original dataset into subsets based on a particular attribute. This process is recursive and is terminated when a further splitting is either non-feasible or a classification can be made. The result is structurally similar to that shown in Figure 10.2.

There are two primary formulas employed by DTs: the first is called *Gini impurity* – based on squared probability of class membership – and the second is *information gain* – based on entropy as used in information theory (Shannon, 1997). Both deal with the diversity of the data at each node. The application of these formulas allows the current node to be split into further nodes by calculating what is called *impurity loss* or *goodness of split*.

If $y = (x_1, x_2, x_3, \ldots, x_n)$ and $f(i, j)$ is the frequency of value j in node i, then the Gini impurity is defined as shown in (10.1).

$$I_{\mathrm{G}}(i) = 1 - \sum_{j=1}^{n} f(i, j)^2, \tag{10.1}$$

while the information gain is defined as shown in (10.2), which is analogous to Shannon's entropy (10.3).

$$I_{\mathrm{E}}(i) = -\sum_{j=1}^{n} f(i, j) \log_2 f(i, j), \tag{10.2}$$

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i). \tag{10.3}$$
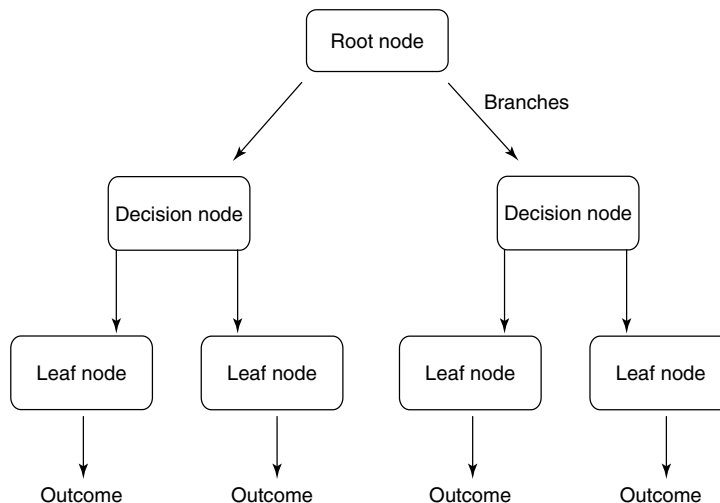
Figure 10.2 CART general outline. Once constructed, the DT can be viewed as shown here – a set of choices represented by a node that branches out from the root node (starting point). An unknown sample is fed into the system at the root node, and the sample is recursively processed until a classification is made, hopefully correctly. Although this representation shows branches of equal length, this is not a constraint. One of many internal DT methods is a branch-trimming algorithm that attempts to refine the tree during training, reducing complexity and speeding up execution.

Typically DTs are used for data mining, a task that is not directly associated with protein structure prediction. However, Selbig *et al.* (1999) used a DT to derive better consensus secondary structure predictions. Selbig combined predictions from existing programs as input for the DT. The role of the DT was to improve the final consensus prediction – something that is now very popular. At that time (1999) Selbig's program, CoDe, achieved an accuracy of 72.9 %, which outperformed the methods that existed then. Today, more advanced learning machines are used in this role, typically a support vector machine (SVM) or artificial neural network (ANN), which is considered below.

10.4.1.1   Summary of CART Methods

Advantages
- Conceptually simple.
- Data requires very little pre-processing.
- White box – it is possible to look at a DT and see how it works.
- Handle regression and classification problems.
- Handle large datasets – it is possible to complete large tasks using a desktop pc.
- Many existing architectures including Weka (Ian and Witten, 2005) and Yet Another Learning Environment (YALE) (Mierswa *et al.*, 2006).

Disadvantages
- Rubbish in, rubbish out – can be heavily influenced by small amounts of low-quality input.
- Requires qualitative input to give complete picture.

- Probabilities are estimated.
- Real-time data problems.
- Trees become complex – over trained and require trimming.

10.4.2 Genetic Algorithms

Genetic algorithms (GAs) were developed by John Holland and co-workers at the university of Michigan during the 1970s. GAs are based on the idea of natural selection and are designed to approximate solutions to optimisation and search problems. The starting point for a GA is a string, which is also known as a *chromosome*. The first step in a GA is to encode the problem. However, to restrict this chapter only to the basics, this step will be neglected as it often involves complex encoding. Examples applied to 3D structure can be found in Dandekar and Argos (1994; 1996a; 1996b; 1997) and Petersen and Taylor (2003). Each chromosome begins life with a fitness function (f) and it is this feature that the GA aims to maximise via four steps:

1. reproduction

2. 'genetic' Crossover

3. mutation

4. termination.

The reproduction method copies individual strings according to their fitness function. The fitness function determines the probability of selecting a chromosome. Therefore chromosomes with a higher fitness are more likely to be selected from the mating pool for breeding, and thus they are more likely to make the transition to new generations.

The crossover method creates diversity and drives the evolution of the chromosome. There are two steps to crossover: the selection of two parent strings at random from the breeding pool and completion of the crossover itself. Given two strings of length (l), a random position (n) that matches the criterion $1 \leq n \leq l - 1$ is chosen. For example, if two strings (A and Z) are selected at random from a breeding pool:

$$A = 01000001,$$

$$Z = 01011010,$$

and $n = 4$, then the chromosomes are split as shown below.

$$A = 0100|0001,$$

$$Z = 0101|1010.$$

To generate the next generation, the strings are pulled apart at the boundary (|) and recombined, resulting in A' and Z'.

$$A' = 01001010,$$

$$Z' = 01010001.$$

Mutation plays a secondary role in GAs (somewhat unlike biology). The mutation operation exists to compensate for overzealous eradication of useful genetic information,

or simply put, mutation prevents the loss of useful genetic information in the early stages of evolution. Mutation operations generally occur at a low frequency; however it can (and should) be optimised by the user.

All good (and bad) simulations must come to an end and this is taken care of by the termination method. Termination is typically based on the time a method has been running, the number of generations that have passed, or on completion (i.e. f does not increase over a number of generations).

GAs have a long and successful history in computational biology and recent work by Arunachalam *et al.* (2006) and Armano *et al.* (2005) has used GAs to good effect in structure prediction. In Arunachalam *et al.* (2006), the prediction of several all α proteins is described using mutually orthogonal Latin squares and GAs. The method clearly explains the basics of GAs as well as showing how simple it is to extend the simple GA to work on 'real' numbers as opposed to binary strings. Arunachalam's GA is of particular interest as a supplementary method that aims to aid convergence on local minimum is described. Armano *et al.* (2005) introduces a hybrid ANN – GA approach to secondary structure prediction. While the paper does not highlight a ground breaking discovery, it does describe a method that is comparable to MUPRED (Bondugula and Xu, 2007), a method that is discussed in the next section.

10.4.3   k and Fuzzy k-nearest Neighbour

Given a residue (x), the question is, is it part of a helix, sheet, or coil? This 'simple' classification problem lends itself to a k-nearest neighbour (kNN) algorithm (10.4). A kNN is a simple yet powerful method for assigning a class to an unknown sample. Given a number of known samples, the distances between k and these samples is calculated using a distance metric (see below). For the sake of simplicity, consider x to be defined by two coordinates. In this space, the Euclidean (EC) distance (10.5) can be calculated quickly. Initially the class of x will be known, and this allows for the number of neighbours (k) to be optimised – to produce the most accurate prediction. When a new sample x is introduced, the nearest k neighbours will be used to infer the structural class of x. The result is a probability of x belonging to the helix, sheet, or coil classes. The beauty of this method is that only one parameter has to be optimised – the value of k.

$$\mu_c(x) = \frac{\sum_{j=1}^{k} \mu_c(x_j) d_{ij}^{-2}}{\sum_{j=1}^{k} d_{ij}^{-2}}, \quad (10.4)$$

where $1 \ldots k$ are the kNNs, $\mu_c(x_j)$ is the class membership of neighbour j, and d_{ij} is the distance between features i and j. There are two commonly used distance metrics to calculate d_i, the EC distance and the city block (CB) distance (10.6). The formula for calculating the EC distance is shown in (10.5):

$$d_i = \sqrt{\sum_{i=0}^{n} (X_i - Y_i)^2}. \quad (10.5)$$

The EC distance is computationally expensive to calculate, so typically the CB distance is used if the dataset is large. The CB distance (10.5) is less accurate than the EC distance. However, the increase in execution time on large datasets is noticeably reduced.

$$d_i = \sum_{i=0}^{n} |X_i - Y_i|. \tag{10.6}$$

The most crucial and time consuming part of the kNN procedure is the optimisation of k. There are several procedures for optimisation and they come as part of existing kNN packages. The first step is to split the data (typically $60:40$) into two sets – training and hold out test (HOT). The training set is then subjected to either a leave one out or an n-fold validation. The concept of the leave one out validation is simple – remove an example, train the program, establish k, calculate accuracy, and repeat. The n-fold (where n is a number defined by the user) breaks the data into n equal sections, one section is left out and the algorithm is trained on the remaining data, and this is repeated n times. At the end of the validation routes, a plot of k against percentage accuracy should reveal the optimal value for k. When k is defined, each sample in the HOT set should be analysed and an overall accuracy obtained.

A kNN can be improved using another user-definable parameter m. This parameter acts as a 'fuzziness' parameter, weighting the class contribution for each k by its distance. The introduction of this term means that the kNN algorithm is now defined as a *fuzzy k-nearest neighbour* (*fkNN*) *algorithm*. (10.7) shows the fkNN algorithm. Note the difference between this and the previous equation is the inclusion of $-2/(m-1)$.

$$\mu_{\mathrm{c}}(x) = \frac{\sum_{j=1}^{k} \mu_{\mathrm{c}}(x_j) d_{ij}^{-2/(m-1)}}{\sum_{j=1}^{k} d_{ij}^{-2/(m-1)}}. \tag{10.7}$$

The kNN and fkNN approaches have been applied to both secondary structure (Bondugula and Xu, 2007) and solvent accessibility (Sim *et al.*, 2005), both individually and in combination with more sophisticated methods such as ANNs.

10.4.3.1   Summary of kNN/fkNN

   Advantages
- Simple to implement
- Powerful
- Existing architectures for moderate-sized datasets.

   Disadvantages
- Computationally expensive
- Time consuming – have to run over full dataset for each prediction
- Classification problems only.

10.4.4   Bayesian Approaches

Bayesian theory, like the kNN, is a simple approach – indeed it reduces to 'just counting'. In protein structure prediction, this approach has been successfully applied by Thompson and Goldstein for solvent accessibility (Thompson and Goldstein, 1996) and secondary structure (Thompson and Goldstein, 1997) prediction and later by Lathrop for sequence structure alignment (Lathrop *et al.*, 1998).

Bayes' theorem (10.5) provides a method for adjusting the degree of belief in light of new information.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \qquad (10.8)$$

The term $P(A)$ is the prior probability; it represents the probability of seeing A in a random population. $P(B|A)$, the conditional probability, represents the probability of observing B given knowledge of A. $P(B)$ is referred to as the *marginal probability* and is the probability of observing B in a random population. The left term, $P(A|B)$, is the posterior probability; it represents the probability of A given knowledge of B.

Given a residue (i) and a surrounding window of sequence ($\{A_j\}$), Bayes' theorem can be used to predict protein features (S_k) such as solvent accessibility and secondary structure using a library of known examples (k) (10.9).

$$P(\omega_i|\{A_j\}) = \sum_k \frac{P(\{A_j\}|S_k)}{P(\{A_j\})} P(S_k)\delta(s_j^k \in \omega_i), \qquad (10.9)$$

where s is the structural context (buried β sheet), and s_j^k is a window of known structure (k denotes a particular segment while j is an index in window k). The term $\delta(s_j^k \in \omega_i)$ is essentially a Boolean statement – zero unless the residue in structural context s_j^k has the structural state ω_i (otherwise one). $P(S_k)$ is the probability of witnessing the structural window k at random and is trivially calculated from a library. $P(\{A_j\})$ is the probability of witnessing a window of sequence at random. $P(\{A_j\}|S_k)$ is the probability of observing a sequence with structural context S_k, and again this must be calculated from a library.

When it comes to biological applications, some assumptions have to be made, for example, when considering the probability of seeing a window of 15 residues. The number of possible combinations in a 15 residue window is 15^{20}, which means that it is highly unlikely that all possible combinations of the amino acids will be witnessed in a sequence library – the sequence partner to the structural library. To get around this problem, an assumption is made that the product of the individual amino acid probabilities will be an accurate approximation of the probabilities of the windows (10.10).

$$\frac{P(\{A_j\}|S_k)}{P(\{A_j\})} = \prod_j \frac{P(A_j|s_j^k)}{P(A_j)}. \qquad (10.10)$$

This assumption changes the overall equation (10.9) to (10.11):

$$P(\omega_i|\{A_j\}) = \sum_k \left(\prod_j \frac{P(A_j|s_j^k)}{P(A_j)} \right) P(S_k)\delta(s_i^k \in \omega_i). \qquad (10.11)$$

This approach, despite being 'old', remains one of the most effective tools for predicting solvent accessibility and secondary structure ($\sim$75 % accuracy). For a more general introduction to Bayesian statistics, it is suggested that the reader refers to the chapter by Richard Goldstein in Jorde *et al.* (2005) and to Thompson and Goldstein (1996; 1997) for detailed biological application.

10.4.4.1   Summary of Bayesian Methods

>   *Advantages*
> - Resistant to overtraining.
> - Easy to update given new example data.
> - Several existing architectures, e.g. perl AI::Categorizer::Learner::NaiveBayes, which is fast although it is not the most accurate.
> - Remains one of the most accurate tools for solvent accessibility prediction and solvent accessibility prediction.

>   *Disadvantages*
> - Slow – has to complete a scan against a pattern library for each prediction.
> - Some of the assumptions do not apply to all situations as detailed in Thompson and Goldstein (1997).

10.4.5   Artificial Neural Networks (ANNs)

ANNs mark a significant step forward in complexity and power allowing for modelling of complex non-linear relationships between input and output. The architecture of an ANN is deceptively simple: at its most basic level is the neuron, a simple processing element that is linked to form a complex machine. The ANN is made up of layers of neurons, which are linked together by weighted connections (see Figure 10.3). It is the combination of these weighted connections and transfer functions used in the nodes that defines the overall performance of the machine. There are three categories of transfer functions: linear, where the output is proportional to the total weighted input; threshold, where the output is set to one of two states depending upon whether the input is greater or less than a threshold; sigmoid, where the output varies continually but not linearly with the input.

The first neural networks appeared in 1957 when Rosenblatt created the perceptron, a model that contained a single layer. The perceptron, or single-layer neural net, is the most

Input layer   Hidden layer   Ouput layer
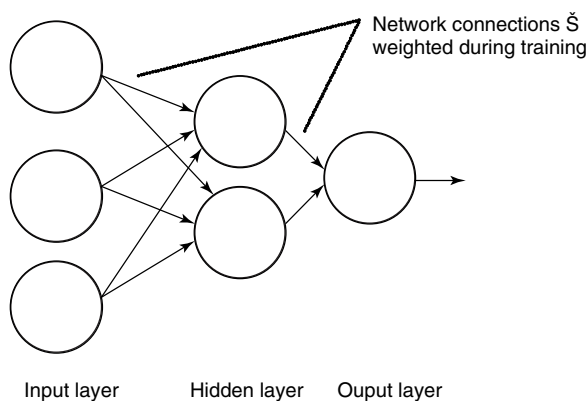
Figure 10.3   A generic ANN layout. The most basic neural networks consist of a single hidden layer that is fully connected to the input and output layers. The connections between each node (black arrows) is given a default value before training, and when training begins the weights are adjusted according to an error function. After training, the weights should be optimised to accurately map input space to output space.

basic of ANNs and is only capable of solving linearly separable problems. The concept behind the network is simple: take the sum of the products of the weights and inputs per node; if the value is above a specific threshold the neuron fires and obtains an 'activated' value or 'higher state'. The simplicity is further reflected in training, which is completed using the 'depth' walk. This function calculates the error between the sample and training set and adjusts the network weights accordingly – this procedure is also referred to as *gradient descent*, a method that is used in more complicated frameworks.

An excellent example of neural network usage in protein structure prediction is YASPIN (Lin *et al.*, 2005). Lin *et al.* presented a feed-forward perceptron network with a single hidden layer to predict seven-state secondary structure, which is then optimised using a hidden Markov model (HMM). YASPIN used a soft-max transition function with a window of 15 residues, with each residue in that window being represented by 20 units. Each unit represents a transition state in a position specific scoring matrix (PSSM) generated by PSI-BLAST. There is an additional element that is used where a terminal is spanned, and the result is a 315 (21×15) dimension input. The hidden layer comprises 15 elements while the output layer contains 7 – one for each structure classification. When the seven-state prediction is made, it is passed to the HMM. The HMM uses the viterbi algorithm (Durbin, 1998) to optimally segment the predictions, the result being transition probabilities between each of the states. The final output to the user is a three-state prediction ($\alpha, \beta, -$) with confidence values.

10.4.5.1 *Summary of ANN Methods*

Advantages
- Very powerful if trained correctly.
- Solve linear and non-linear problems.
- Existing frameworks to build ANNs are available, for example, JOONE.

Disadvantages
- Black box – once they are built and implemented you have no idea how they work.
- Can be slow to train.
- Prone to over training.
- Computationally expensive.

10.4.6   Support Vector Machines

SVMs are supervised learning machines for regression and classification problems; they were introduced by Boser, Guyon, and Vapnik in the 1980s. SVMs are capable of solving linear and non-linear problems by taking an *n*-dimensional input space and mapping it into a higher dimension, often referred to as the *kernel trick*, and once in this space a linear classifier is constructed (see Figure 10.4).

The key part of the SVM is the kernel – a mathematical function that performs the transition from a linear space to a non-linear space, or in the case of a radial basis function (RBF) to a Gaussian infinite space. The majority of existing SVM architectures implement four kernel functions (see Table 10.1), polynomial, radial basis, Gaussian radial basis, and sigmoid. In most cases, the RBF proves to be a suitable starting point as it is able to handle linear and non-linear relationships. Indeed it is the kernel of choice for most biologists.
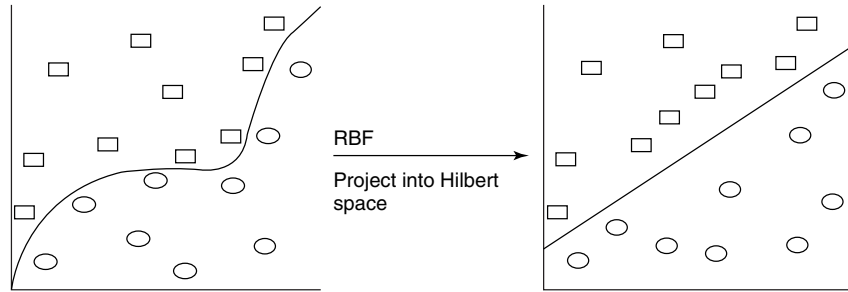
Figure 10.4 Linear classification of a non-linear problem. The initial step is to determine whether the problem is linearly separable or not. If the problem is linear, then a linear kernel should be applied, skipping the transformation into multidimensional space. If the problem is non-linear (or if checking seems like too much work), the radial basis function (RBF) provides an ideal starting point. This is because it is capable of mapping linear and non-linear problems. Once in the higher dimensional space, a linear separation of the data can be performed.

Table 10.1 Four commonly used kernel methods. Of all the kernels, the best starting point is one of the two RBF functions. This is because they are capable of solving both linear and non-linear problems.

Kernel	Mathematical definition
Polynomial kernel	$k(x, x') = (x \cdot x')^d$
Radial basis function (RBF)	$k(x, x') = \exp(-\gamma \|x - x'\|^2)$ where γ is defined by optimisation routines
Gaussian RBF	$k(x, x') = \exp - \left(\dfrac{\|x - x'\|^2}{2\sigma^2} \right)$
Sigmoid	$k(x, x) = \tanh(kx \cdot x + c)$ for $k > 0$ and $c < 0$

The simplest example of an SVM is a linear classification. Consider a problem where there are two classes. The first class describes residues that are buried, the second those that are exposed. Using a set of features that describe each class, a plot can be constructed as shown in Figure 10.5.

As the problem is clearly linearly separable, a linear kernel can be applied as shown in Figure 10.6.

The central line (dividing hyperplane) splits the two classes, while the parallel hyperplane's function is to maximise the space between the dividing hyperplane and each class. When an unknown sample is passed to the SVM, it should be possible to classify it as buried (-1) or exposed ($+1$). The classification is achieved using (10.12), implying that the problem reduces to defining the support vectors to be able to make a prediction – all other data points can be discarded! To make the machine more robust, error (slack) variables (ξ) can be introduced. Error variables allow for a solution to be found where the data is not so clearly divided.

$$c_i = \{1, \text{ if } \omega \cdot x + b \geq 0; \; -1 \text{ if } \omega \cdot x + b \leq 0\}. \tag{10.12}$$
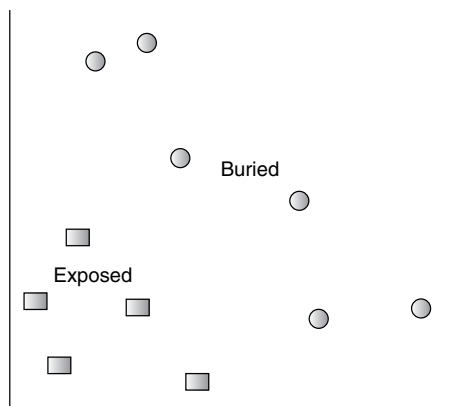
Figure 10.5 Over simplification of residue burial/exposure. Plotting the state of several residues using a number of features shows a clear separation of the two categories (it should be noted this is not the case in reality). Because of the clear separation in normal space, a simple linear kernel can be applied. However, readers with a creative mind can imagine that they used an RBF and now exist in Gaussian infinite space.
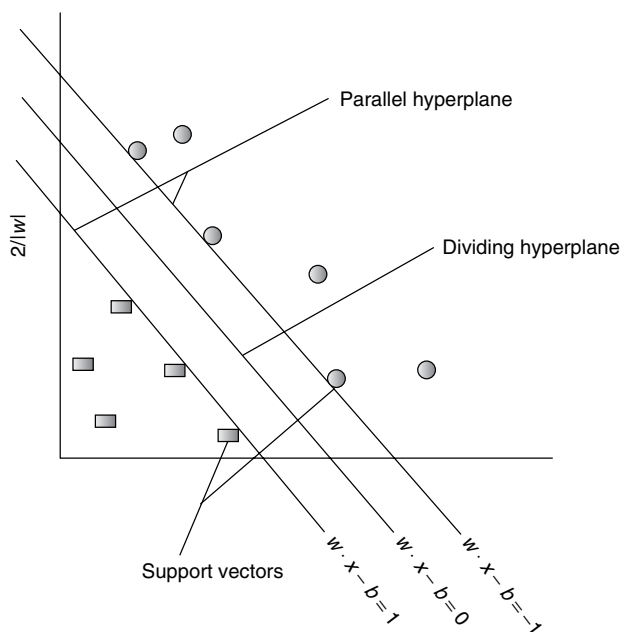
Figure 10.6 Application of a linear kernel on a linearly separable problem. The two classes can be separated by a plane (the dividing plane). To either side of the dividing plane there are parallel hyperplanes, and these define the maximum distance from the dividing hyperplane to each class. The space between the parallel hyperplanes and the dividing hyperplane is void of data points. The name support vector machine is derived from the use of support vectors to establish the position of the parallel hyperplanes. The distance between the hyperplanes is defined by geometrical techniques to be $2/|w|$ (so the aim of the SVM is to minimise $|w|$).

Regression is slightly different from classification in implementation. Generally speaking, existing tools use what is called *epsilon* (ε) support vector regression (SVR). Epsilon defines a threshold under which data is ignored; this is the exact opposite of classification. The detailed workings of this method are beyond the scope of this chapter and readers are strongly advised to read Cristianini and Shawe-Taylor (2000) in order to obtain a full understanding.

SVMs have been used successfully for both classification and for regression in biological scenarios, the easier of the two problems being classification. A good example of a classification problem is Disopred2 (Ward *et al.*, 2004). Disopred2 uses support vector classification and a neural network to predict the ordered or disordered state of a residue. Ward *et al.*, used a linear kernel to solve this problem and achieved a good degree of success as shown by their performance in the critical analysis of structure prediction (CASP) exercise.

An example of the regression problem is Yuan's work (Yuan, 2005) for the prediction of protein contact number. Previous work by Kinjo *et al.* (2005) had shown that contact number prediction was not best solved using a linear method. Yuan implemented an SVM using the RBF kernel to map the data into higher dimensional space before performing a regression to predict absolute contact number. The overall performance of the SVMs showed an improvement over existing techniques, achieving an accuracy 'greater than 77 %' (Yuan, 2005).

10.4.6.1   Summary of SVM Methods

Advantages
- Solve linear and non-linear problems.
- Kernels are hot swappable.
- Kernel addition and weighting allows for colourful combinations of SVMs.
- Simpler to implement compared to ANNs.
- Less prone to over training compared to ANNs.
- Existing frameworks – very simple to set up.

Disadvantages
- Mathematically complex.
- Kernels are hot swappable – many biologists try each kernel until they get the results they want to see. This is not wrong but can be costly.
- Black box – once trained it is not easy to understand the rules that define the system.
- On large datasets, it can be computationally very expensive to optimise all necessary variables.
- Some implementations are more robust than others.

10.5   CONCLUSIONS

Extracting the maximum amount of structural information from sequence data has had a long history of increasing diversity in the methods used and the problems to which

they have been applied. Following secondary structure methods as an example, simple frequency-based methods were used in the early 1970s (Chou and Fasman, 1974; 1978) followed by information-theory-based methods during the late 1970s (Garnier *et al.*, 1978). The 1980s saw the introduction of artificial neural net methods, which have been followed by SVMs in the last decade. Applications have also broadened from a focus on the simple three-state (α, β, other) prediction of local structure to include the solvent exposure of a residue and more recently, the predicted degree of disorder in the chain.

Despite these quantum leaps in methodology, the degree of accuracy in the predictions has improved only slowly in recent years with small, usually single figure, increases in prediction accuracy being reported. Compared to the wide range of variations in the accuracy of the predictions between proteins, these small increases are barely significant. The observed plateau in accuracy is not unexpected as most of the current methods use data derived from a multiple sequence family. Without knowing the structures of the proteins in the families, there will certainly be errors in their alignment that will introduce a base level of noise that cannot be overcome without improved alignment methods. More fundamentally, there is also variation in observed secondary structure definitions between members of the family. Taking all members of the family together means there is no unique definition of secondary structure against which accuracy can be measured. An iterative approach that would overcome this problem might be considered: starting with a wide diverse family to predict a starting state, which is gradually refined using smaller subsets of the family, finishing with one sequence. At this stage, it would then probably be necessary to predict the full three-dimensional structure.

The application of statistical methods to new problems, such as the prediction of chain disorder, is probably where the methods are making their most significant contribution. In this particular area, it is being realised that large parts of the coding sequences in a large fraction of proteins in the genome have no intrinsic structure. They serve as locations for modification, like phosphorylation, or are induced to have structure when they encounter a specific binding partner. Their role is important in dynamic behaviour of cellular systems, such as receptor signalling and gene regulation. With the growing interest in 'systems biology', the identification and understanding of these regions will be of increasing biological importance.

Finally, it is worth considering the direct application of statistical methods to three-dimensional structure prediction. Most of the applications described above involve the prediction of local structure or structural states, which can be used to constrain 3D models. Some attempts have been made to apply ANNs to direct structure prediction but with success only at the local structural level. There may be scope for methods that use a higher level of structural description but the encoding to be used for this is not obvious. A more likely route towards a practical tool is the use of statistical methods in fold-recognition (or threading) methods. As more structures become available, this task, which is essentially a classification problem, will become easier. The only impediment along this route is that large proteins usually need to be broken down into their constituent domains before they can be recognised and these domains are not always contiguous in sequence.

In summary, while structure prediction methods are becoming more powerful and helped by the vast new sources of sequence and structural data, to some extent, their traditional areas of structure prediction are becoming less relevant. The increasing volumes of data also means that there is now a good chance of finding all the necessary structural

information on a protein from another structure related by clear sequence similarity. Interesting application areas appear to lie more in novel directions associated with protein interactions and their relevance to cellular function.

REFERENCES

Anfinsen, C.B. (1972). The formation and stabilization of protein structure. *The Biochemical Journal* **128**, 737–749.

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science* **181**, 223–230.

Armano, G., Mancosu, G., Milanesi, L., Orro, A., Saba, M. and Vargiu, E. (2005). A hybrid genetic-neural system for predicting protein secondary structure. *BMC Bioinformatics* **6**(Suppl. 4), S3.

Arunachalam, J., Kanagasabai, V. and Gautham, N. (2006). Protein structure prediction using mutually orthogonal Latin squares and a genetic algorithm. *Biochemical and Biophysical Research Communications* **342**, 424–433.

Bondugula, R. and Xu, D. (2007). MUPRED: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. *Proteins* **66**(Suppl. 3), 664–670.

Bowie, J.U., Luthy, R. and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170.

Bradley, P., Misura, K.M. and Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871.

Breiman, L. Freidman, J.H., Olshen, R. A. and Stone, C. J. (1984). Classification and Regression Trees, Wadsworth International, Belmont, CA.

Chou, P.Y. and Fasman, G.D. (1974). Prediction of protein conformation. *Biochemistry* **13**, 222–245.

Chou, P.Y. and Fasman, G.D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology and Related Areas of Molecular Biology* **47**, 45–148.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods*. Cambridge University Press, Cambridge.

Dandekar, T. and Argos, P. (1994). Folding the main chain of small proteins with the genetic algorithm. *Journal of Molecular Biology* **236**, 844–861.

Dandekar, T. and Argos, P. (1996a). Ab initio tertiary-fold prediction of helical and non-helical protein chains using a genetic algorithm. *International Journal of Biological Macromolecules* **18**, 1–4.

Dandekar, T. and Argos, P. (1996b). Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *Journal of Molecular Biology* **256**, 645–660.

Dandekar, T. and Argos, P. (1997). Applying experimental data to protein fold prediction with the genetic algorithm. *Protein Engineering* **10**, 877–893.

Duan, Y. and Kollman, P.A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744.

Durbin, R. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology* **120**, 97–120.

Ian, H. and Witten, E.F. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature* **358**, 86–89.

Jorde, L.B., Little, P.F.R., Dunn, M.J. and Subramaniam, S. (2005). *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. Wiley.

Kinjo, A.R., Horimoto, K. and Nishikawa, K. (2005). Predicting absolute contact numbers of native protein structure from amino acid sequence. *Proteins* **58**, 158–165.

Lathrop, R.H., Rogers, R.G. Jr., Smith, T.F. and White, J.V. (1998). A Bayes-optimal sequence-structure theory that unifies protein sequence-structure recognition and alignment. *Bulletin of Mathematical Biology* **60**, 1039–1071.

Lin, K., Simossis, V.A., Taylor, W.R. and Heringa, J. (2005). A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* **21**, 152–159.

Luthy, R., Mclachlan, A.D. and Eisenberg, D. (1991). Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins* **10**, 229–239.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. and Euler, T. (2006). YALE: rapid prototyping for complex data mining tasks In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Philadelphia, USA *(KDD-06)*.

Montgomerie, S., Sundararaj, S., Gallin, W.J. and Wishart, D.S. (2006). Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics* **7**, 301.

Nagano, K. (1973). Logical analysis of the mechanism of protein folding. I. Predictions of helices, loops and beta-structures from primary structure. *Journal of Molecular Biology* **75**, 401–420.

Pearl, L.H. and Taylor, W.R. (1987). A structural model for the retroviral proteases. *Nature* **329**, 351–354.

Petersen, K. and Taylor, W.R. (2003). Modelling zinc-binding proteins with GADGET: genetic algorithm and distance geometry for exploring topology. *Journal of Molecular Biology* **325**, 1039–1059.

Rice, D.W. and Eisenberg, D. (1997). A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *Journal of Molecular Biology* **267**, 1026–1038.

Rohl, C.A. and Baker, D. (2002). De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *Journal of the American Chemical Society* **124**, 2723–2729.

Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research* **31**, 3381–3385.

Selbig, J., Mevissen, T. and Lengauer, T. (1999). Decision tree-based formation of consensus protein secondary structure prediction. *Bioinformatics* **15**, 1039–1046.

Shannon, C.E. (1997). The mathematical theory of communication. 1963. *MD Computing* **14**, 306–317.

Sim, J., Kim, S.Y. and Lee, J. (2005). Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. *Bioinformatics*.

Simossis, V. (2005). From sequence to structure and back again: an alignments tale. *Faculty of Sciences*. Vrije Universiteit, Amsterdam.

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology* **213**, 859–883.

Skolnick, J. (2006). In quest of an empirical potential for protein structure prediction. *Current Opinion in Structural Biology* **16**, 166–171.

Taylor, W.R. (1986). The classification of amino acid conservation. *Journal of Theoretical Biology* **119**, 205–218.

Taylor, W.R., Lin, K., Klose, D., Fraternali, F. and Jonassen, I. (2006). Dynamic domain threading. *Proteins* **64**, 601–614.

Thompson, M.J. and Goldstein, R.A. (1996). Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins* **25**, 38–47.

Thompson, M.J. and Goldstein, R.A. (1997). Predicting protein secondary structure with proba-
    bilistic schemata of evolutionarily derived information. *Protein Science* **6**, 1963–1975.
Ward, J.J., Mcguffin, L.J., Bryson, K., Buxton, B.F. and Jones, D.T. (2004). The DISOPRED server
    for the prediction of protein disorder. *Bioinformatics* **20**, 2138–2139.
Yuan, Z. (2005). Better prediction of protein contact number using a support vector regression
    analysis of amino acid sequence. *BMC Bioinformatics* **6**, 248.

11

Statistical Techniques in Metabolic Profiling

M. De Iorio

Division of Epidemiology, Public Health and Primary Care, Imperial College, London, UK

T.M.D. Ebbels

Division of Surgery, Oncology, Reproductive Biology and Anaesthetics, Imperial College, London, UK

and

D.A. Stephens

Department of Mathematics and Statistics, McGill University, Montreal, Canada

Metabolic profiling (metabonomics or metabolomics) is a field in biomedical investigation that combines the application of nuclear magnetic resonance spectroscopy and other experimental platforms with multivariate statistical analysis in studies of the composition of biofluids, cells and tissues. It is a system-level bioinformatics discipline that involves the quantitative measurement of the multivariable metabolic response of living systems to pathophysiological stimuli or genetic modification. Metabolic profiling therefore represents an important area of biological research that is likely to have an impact on a diverse range of clinical areas. In this chapter, we give a description of some basic metabolic data gathering techniques, and the most commonly used methods of statistical analysis, namely, principal components analyis and partial least squares, cluster analysis, and other modern classification approaches such as neural networks and evolutionary algorithms. Throughout, we illustrate these concepts with specific biological examples.

11.1   INTRODUCTION

Metabolic profiling, also known as *metabonomics* (Nicholson *et al.*, 1999) or *metabolomics* (Raamsdonk *et al.*, 2001), is a rapidly developing field in biomedical science that

combines the application of spectroscopic techniques with multivariate statistical analysis in studies of the molecular composition of biofluids, cells and tissues. The experimental focus is on profiling *metabolites* – the thousands of low-molecular-weight molecules which are the building blocks of the cell. These compounds interact with macromolecules such as proteins and nucleic acids in the fundamental metabolic processes that keep organisms alive. Thus, the metabolic state of an organism, as delineated by its metabolic profile, can reveal information on the genetic, physiological or functional status the system. In addition, in contrast to DNA or RNA, metabolic profiles also reflect the entirety of internal and external influences on an organism or tissue, including environment, behaviours, disease, drugs and interactions with parasitic or symbiotic organisms (Nicholson *et al.*, 2002; Nicholson and Wilson, 2003). This leads to important applications in the molecular diagnosis and prognosis of disease, drug metabolism and toxicity, functional genomics and the investigation of fundamental biological processes. From the statistical point of view, however, metabolic profiles are complex and information rich, presenting unique challenges to data analysis and modelling.

The term *metabolome* refers to the total metabolite complement of a cell, tissue or organism. While there are currently no methods capable of delivering an exhaustive measurement of the metabolome, there are several technologies that can approach this ideal. A gold standard technique would be one that is (1) non-selective, in that, it is not biased towards particular chemical classes of compounds, (2) quantitative and accurate, thus producing data that can be meaningfully modelled by statistical procedures, (3) sensitive to a wide range of metabolite concentrations and (4) able to resolve the complex mixture into separate signals from each metabolite. These requirements are satisfied, though not perfectly, by the techniques of nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS).

11.1.1 Spectroscopic Techniques

The two spectroscopic assay technologies, NMR and MS, are highly complementary and therefore often used in parallel. NMR has the advantage that it is highly non-selective and quantitative; it also requires little sample preparation and is non-destructive, thus permitting multiple measurements on each sample. MS approaches are much more sensitive and generally have higher resolution than NMR; yet they suffer from inaccuracies introduced by differential ionisation efficiencies. These can be partially alleviated by preceding the MS detection with a physical separation step using liquid chromatography (LC) or gas chromatography (GC), leading to LC–MS or GC–MS Wilson *et al.* (2005). Figure 11.1 shows examples of NMR and LC–MS spectroscopic metabolic profiles of urine.

The resulting spectra (or profiles) consist of several thousands of individual measurements at different resonances or masses. The spectra are complex and require dedicated analysis procedures. The inherent high dimensionality immediately precludes many conventional statistical analysis approaches. Similar to other spectroscopic data, the spectral variables are usually highly collinear, a characteristic that arises for both chemical analytic and biological reasons. For example, one metabolite may generate more than one spectral signal (e.g. isotopic patterns in MS) leading to several highly correlated peaks. Alternatively, signals from different metabolites may show high correlations because they are involved in the same biological process. While the above properties are typical of data from many so-called '-omic' techniques such as those from transcriptomics and proteomics, metabolic profiling poses its own challenges. The most important challenge is
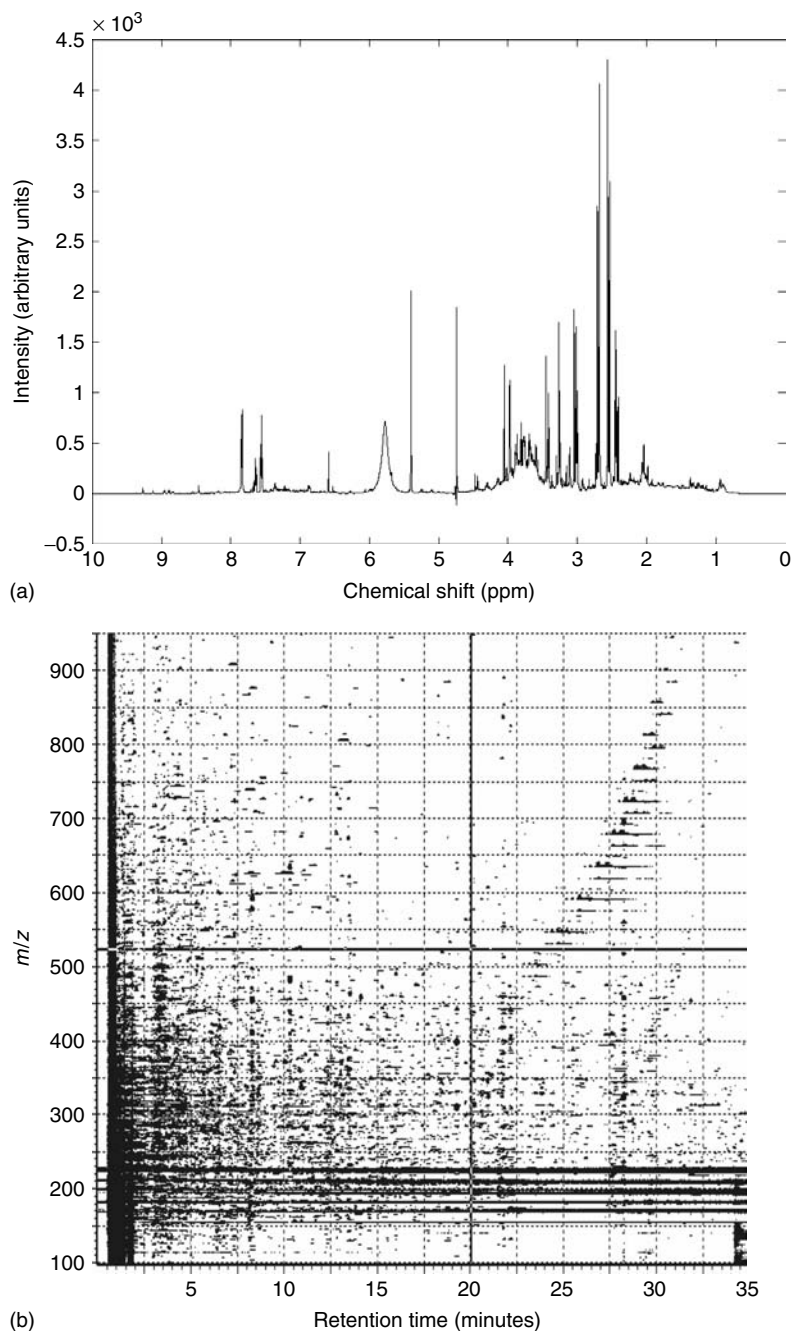
Figure 11.1 Typical metabolic profiles of rat urine. (a) a one dimensional ^{1}H NMR spectrum, (b) a two dimensional contour plot of an LC−MS profile. Both the higher number of signals and level of noise are clear in the LC−MS profile.

the problem of unidentified signals; while thousands of signals may be detected, typically only a very small fraction will be identified with a known chemical structure – so such prior chemical information is extremely helpful for correct interpretation and identification. Another difficulty is the sensitivity and resolving power of the analytical technique, as well as the complexity of the analysed mixture. A related problem is that the total number of detected analytes is in general unknown and will vary between different biological conditions. Thus, in metabolic profiling, there is a great emphasis on methods of exploratory data analysis and visualisation, which allow the analyst to probe their data for the presence of poorly detected or resolved signals and help in identifying unknown metabolites.

11.1.2   Data Pre-processing

Metabolic profiles generally undergo some pre-processing prior to more advanced statistical modelling, to account for spectral artefacts and render profiles from different samples more comparable. Pre-processing is necessary to achieve comparability between replicate samples, to achieve some form of measurement noise reduction and to more accurately distinguish the different chemical constituents by means of peak deconvolution and separation. In this chapter, we do not describe these issues in detail, but instead focus on statistical feature extraction and discrimination procedures. However, a brief discussion is given below.

One of the most serious problems affecting NMR profiles is that peak positions can be affected by the pH or ionic strength of the sample. This phenomenon seriously affects a minority of metabolites and means that the spectral intensity at a given NMR chemical shift (or wavelength) may be due to both changes in molecular concentrations as well as shifts in resonance positions, making the interpretation of statistical models difficult. While there exist techniques that account for or correct such shifts automatically (Holmes etal, 1992; Stoyanova *et al.*, 2004), these are most suited to correction of small spectral regions and have not yet been successfully applied to full-width NMR profiles. A similar difficulty occurs in LC, where obtaining highly reproducible separation of metabolites is difficult and peak alignment algorithms are often used to match LC profiles from different samples. In mass spectrometry, various processes can lead to the signal from each metabolite being broken up into multiple peaks. For example, the natural occurrence of multiple isotopes of various nuclei leads to a characteristic isotopic pattern for each primary ion, while the formation of adducts (e.g. the addition of metal ions to the primary ion) ensures each metabolite is characterised by a pattern of different peaks. In MS approaches, differential ionisation efficiency also complicates interpretation of the data. Finally, regardless of the technique employed, the profiles will be affected by gross changes in sample concentration, which can obscure important but subtle treatment-related changes. To alleviate this, profiles are often normalised to unit total intensity, so that subsequent values for any given metabolite relate to its concentration relative to the rest of the mixture, rather than to an absolute measure. In most cases, this does not detract from, and can even improve, the biological interpretation, though when there are large changes between profiles, interpretation of profiles normalised in this way is less clear.

Both NMR and MS spectra are typically data-reduced prior to statistical analysis. This may take the form of discretisation of the continuous spectra by integrating the signal in a grid of bins, or by a form of 'peak picking' and subsequent integration of

individual signals. Thus, depending on the method of preprocessing, each data variable may correspond to either a particular signal or merely an integrated region of the spectrum. The peak picking approach requires complex peak detection algorithms using many tuneable parameters and can be subject to a high degree of bias. The binning approach, on the other hand, may result in a proportion of variables containing little real signal and mainly influenced by noise. This is particularly true in the case of LC-MS, where the grid-based binning approach results in a large number of noisy variables due to the sparse nature of the two-dimensional data (see Figure 11.1b). In either case, there are often many variables whose metabolic identity is difficult to pin down (e.g. because of unidentified signals). It should be noted that, because of the issues described above (peak shifts etc.), optimal preprocessing for both NMR and MS data is an area of active current research.

Several metabolic profile analysis software packages are available. A popular and effective one is the AMDIS software (Davies, 1998; Stein, 1999), which extracts individual component spectra from LC or GC MS data files and then uses them to identify compounds by matching the spectra to a reference library of chemical compounds housed at the US National Institute for Standards and Technology (NIST). The analysis proceeds in four steps (noise analysis, component 'perception' (extraction), spectral deconvolution and compound identification) in an automatic fashions, using a variety of multivariate modelling techniques and heuristic testing procedures. The entire approach has proved very successful for analysing data from these platforms.

The aims of data analysis in metabolic profiling will depend on the scientific objectives of the study. However, the objectives of analysis typically fall into one or more of the following categories. Firstly, we wish to visualise the relationships between groups of both samples and spectral variables. For example, this could include clustering individuals or detecting significant correlations between variables. In addition, we wish to determine whether there is a significant difference between groups related to the effect of interest. The latter is a classic small n, large p inference problem (West, 2002). That is, each individual datum (metabolic profile) consists of a large vector of interrelated (dependent) observations, yet the number of samples in the study is relatively small. Finally, and perhaps most importantly, we are interested in finding out which metabolites are responsible for these changes. This chapter introduces some of the current techniques used in the statistical analysis of metabolic profiles. We give the basic principles and algorithms behind each method and illustrate their use with example datasets. We use X to denote the $n \times p$ matrix of n metabolic profiles, where p is the dimension of the spectrum (usually p is in the order of thousands). Therefore, each row of the X matrix represents a metabolic profile. We define the sample covariance matrix $\mathbf{S} = X'X/n$, assuming that each column of X is standardised to have mean 0, and take $\mathbf{y}$ to be an n-dimensional vector of responses.

11.1.3 Example Data

Several of the methods described in this chapter are illustrated with the help of data from a study profiling the metabolic consequences of hydrazine toxicity (Lindon *et al.*, 2005) (Plates 10 and 11). Thirty Sprague–Dawley rats were randomly and equally assigned to three treatment groups (control, low dose and high dose), and hydrazine was administered orally at doses of 0, 30, and 90 mg kg^{-1}, respectively. Urine samples were collected at 10 time points over 8 days, including 2 pre-treatment samples. All procedures were carried

out in accordance with relevant national legislation and were subject to appropriate local review. ^{1}H NMR spectra were measured at 600 MHz and 300 K using a robotic flow-injection system (Bruker Biospin, Karlsruhe, Germany). After phasing and baseline correction, each NMR spectrum was segmented into $M = 205$ variables by integrating the signal in regions of equal width (0.04 ppm) in the chemical shift ranges δ 0.20–4.50 and δ 5.98–10.0, excluding spectral artefacts in the region δ 4.50–5.98. All segmented spectra were then normalised to a constant integrated intensity of 100 units to take account of large variations in overall urine concentration. The principal component analysis (PCA) and partial least-squares (PLS) analysis shown were performed using SIMCA-P+ version 11 (Umetrics AB, Umea, Sweden). Other analyses were computed using in-house software written in the MATLAB programming environment (version R2006a, The MathWorks, Natick, MA).

11.2 PRINCIPAL COMPONENTS ANALYSIS AND REGRESSION

11.2.1 Principal Components Analysis

PCA is a well-known method of dimension reduction (Massy, 1965; Jolliffe, 1986), which seeks linear combinations of the columns of X with maximal variance, or equivalently, high information. It is routinely applied in chemometrics with the goal of providing the most compact representation of the data. The original p variables $X = [\mathbf{x}_1 \ldots \mathbf{x}_p]$ are transformed in a new predictor set $\mathbf{T} = [\mathbf{t}_1 \ldots \mathbf{t}_k]$, with $k \leq \min(n - 1, p)$. The new variables $\mathbf{t}_j$, called *scores*, are a weighted average of the original X variables. The *principal components* are the eigenvectors, $\mathbf{u}_j$ from the eigendecomposition of $X'X$ (and of the sample covariance matrix $\mathbf{S}$, up to a constant). PCA sequentially maximises the variance of a linear combination of the original predictor variables

$$\mathbf{u}_j = \arg \max_{\|\mathbf{u}\|=1} \mathrm{Var}(X\mathbf{u}),$$

subject to the constraint that $\mathbf{u}_i'\mathbf{S}\mathbf{u}_j = 0$ for all $1 \leq i < j$. This ensures that $\mathbf{t}_j = X\mathbf{u}_j$ is uncorrelated with all the previous linear combinations $\mathbf{t}_i = X\mathbf{u}_i$. The principal components are ordered in terms of the amount of variation of the original data they account for. The first principal component direction has the property that $\mathbf{t}_1 = X\mathbf{u}_1$ has the largest sample variance among all normalised linear combinations of the columns of X. Each subsequent component gives combinations with the largest possible variance which are uncorrelated with those that have been taken earlier.

There are various standard approaches to find the principal components, e.g. taking the singular value decomposition of X. In chemometrics it is common to estimate the principal components using the nonlinear iterative partial least-squares (NIPALS) algorithm (Wold, 1966). This is because the number of required components is usually much less than the total possible number ($k \ll p$). In fact, the NIPALS algorithm does not calculate all the principal components at once, but it first calculates $\mathbf{t}_1$ and $\mathbf{u}_1$ from the X matrix. Then the outer product $\mathbf{t}_1\mathbf{u}_1'$ is subtracted from X and the residual X_2 is calculated. In turn, this residual can be used to calculate $\mathbf{t}_2$ and $\mathbf{u}_2$.

NIPALS algorithm for PCA

---

0. Standardise each $\mathbf{x}_j$ to have mean 0.
1. Initialise $j = 1$, $X_j = X$ and $\mathbf{t}_j$ to any of the columns of X.
2. Project X_j on $\mathbf{t}_j$ to find the corresponding loadings $\mathbf{u}_j : \mathbf{u}_j = X_j'\mathbf{t}_j/\|\mathbf{t}_j\|$.
3. Normalise $\mathbf{u}_j$ to have unit length: $\mathbf{u}_j = \mathbf{u}_j/\|\mathbf{u}_j\|$.
4. Project X_j on $\mathbf{u}_j$ to find the corresponding score vector $\mathbf{t}_j : \mathbf{t}_j = X_j\mathbf{u}_j/\|\mathbf{u}_j\|$.
5. Check convergence: compare $\mathbf{t}_j$ used in step 2 and $\mathbf{t}_j$ calculated in step 4
   (check if the difference is larger than a predefined threshold, e.g. 10^{-6}).
   If they are the same, the iteration has converged; then continue to step 6.
   Otherwise return to step 2.
6. Remove the estimated PCA principal component from $X_j : X_{j+1} = X_j - \mathbf{t}_j\mathbf{u}_j'$.
7. Let $j = j + 1$ and repeat steps 1–7 until $j = k$.

If $k = \min(n - 1, p)$, then we have found all the principal components. We can now form the scores matrix $\mathbf{T}$ and the loadings matrix $\mathbf{U}$ with columns $\mathbf{t}_j$ and $\mathbf{u}_j$, respectively. These matrices are such that

$$X = \mathbf{T}\mathbf{U}',$$

where $\mathbf{U}$ is an orthogonal matrix and the $\mathbf{t}_j$ are orthogonal. Note that $\lambda_j = \|\mathbf{t}_j\|^2$ and $\mathbf{u}_j$ are eigenvalues and eigenvectors of $X'X$, respectively. The NIPALS algorithm can be modified to account for missing data (Christoffersson, 1970).

A key issue is the number of principal components necessary to describe a dataset in a parsimonious way but without loss of important information. After k runs of the NIPALS algorithm, the X matrix is decomposed as

$$X = \mathbf{t}_1\mathbf{u}_1' + \cdots + \mathbf{t}_k\mathbf{u}_k' + X_{k+1}.$$

We want to choose k such that X_{k+1} represents only noise (if $k = \min(n - 1, p)$, $X_{k+1}=\mathbf{0}$) and all the features of X are captured by the $\mathbf{t}_j\mathbf{u}_j'$, $j \leq k$.

The maximum number K of principal components is determined by the number of non-zero eigenvalues, which coincides with the rank of $\mathbf{S}$ and $K \leq \min(n - 1, p)$. However, k is usually chosen through cross-validation (CV) (Stone, 1974; Wold, 1978). The proportion of the total variance explained by a PCA model with k components is quantified by the R^2 parameter defined as

$$R_k^2 = 1 - \sum_{i=1}^{n} \frac{\|\widehat{\underline{x}_i} - \underline{x}_i\|^2}{\text{SS}},$$

where $\underline{x}_i$ denotes a row of X, $\widehat{\underline{x}_i}$ is the estimated $\underline{x}_i$ from the PCA model and SS is the total sum of squares of X, i.e. $\text{SS} = \sum_{i=1}^{n} \|\underline{x}_i\|^2$.

R^2 varies in the range 0–1, taking value one when $k = \min(n - 1, p)$. The CV procedure quantifies how robust the model is to perturbation of data and thus avoids overfitting. In each round of CV, a proportion (usually around 10 %) of the data is held out X_o, the model is computed with the remaining data X_{-o}, and the predicted values of the held out data points are computed using $\hat{X}_o = \sum_{i=1}^{k} \mathbf{t}_i\mathbf{u}_i'$. This is repeated until all data points have been held once. The parameter Q^2 is the CV equivalent of R^2 using the predicted values of the data:

$$Q^2 = 1 - \text{PRESS/SS},$$

where $\text{PRESS} = \sum_{i \in O} \|\widehat{\underline{x}}_i - \underline{x}_i\|^2$, where O is the set of individuals left out. Q^2 is bounded above by the value of R^2, and a high ratio of Q^2/R^2 indicates a robust model which is affected little by perturbations. As more components are computed, both R^2 and Q^2 generally rise as more of the variance is explained. However, a point is reached where the structure represented by the $(m + 1)$'th component is mostly noise. At this point, the predicted data deviates from the true data and Q^2 starts decreasing, indicating that $k = m$ is the correct number of components to use in the model.

Figures 11.2 and 11.3 illustrate the use of PCA on the hydrazine dataset. Figure 11.2 shows the scores plot of the first two principal components. The plot shows the characteristic L-shaped trajectory of hydrazine toxicity; high-dose profiles initially resemble controls but over the time course move to a well-separated region of the plot. Figure 11.3 shows the PCA loadings of the first two principal components. Spectral bins which have a high variance in the X space are indicated by points lying far from the origin and can be used to interpret the distribution of data on the scores plot.

11.2.2 Principal Components Regression

In most applications we are interested in predicting $\mathbf{y}$ from the sets of inputs $\mathbf{x}_j$, e.g. we are interested in studying the relation between the blood pressure and the metabolic profile of an individual. From PCA we derive a representation of the data matrix X as a score matrix $\mathbf{T}$. *Principal component regression* (PCR) uses the score vectors $\mathbf{t}_j$ as explanatory variables and regresses $\mathbf{y}$ on $\mathbf{t}_1, \ldots, \mathbf{t}_k$ for some $k \leq \min(n - 1, p)$:

$$\mathbf{y} = \alpha + \sum_{i=1}^{k} \beta_i \mathbf{t}_i + \boldsymbol{\varepsilon},$$
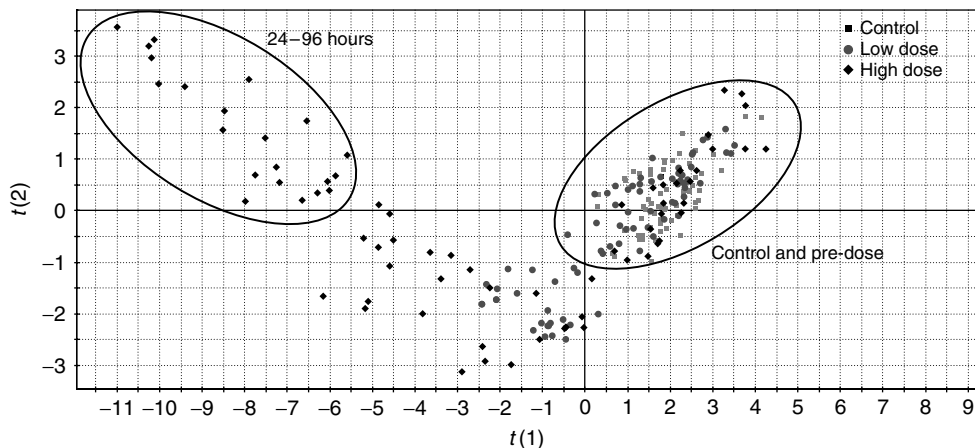
Figure 11.2 Scores plot of the first two principal components for the hydrazine dataset. Each point represents one urine sample (metabolic profile) from one animal at one time point. The different patterns indicate different treatment groups (squares, control; circles, low dose; diamonds, high dose). The plot shows the characteristic L-shaped trajectory of hydrazine toxicity; high-dose profiles initially resemble controls but over the time course move to a well-separated region of the plot. The ellipses identify regions associated with control-like profiles (on the right) and profiles showing maximum toxicity (on the left).
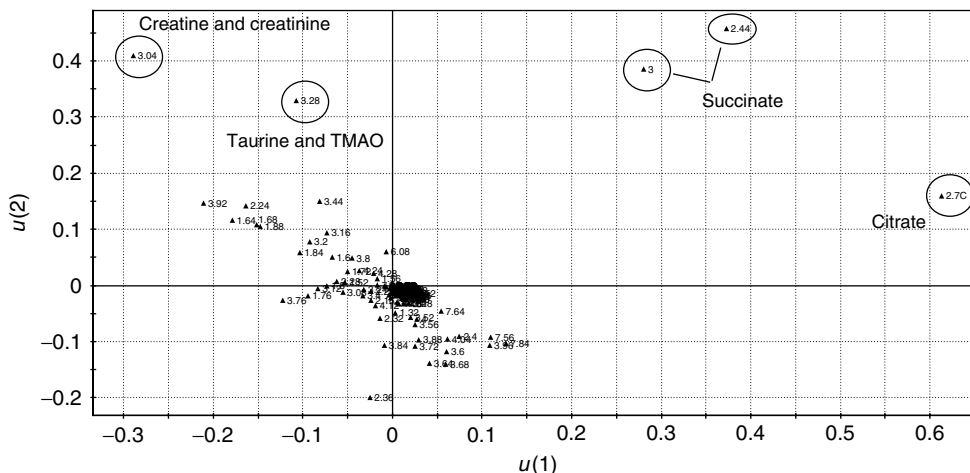
Figure 11.3 PCA loadings plot of the first two principal components for the hydrazine dataset. Each point represents one spectral bin. Points lying far from the origin indicate those bins characterised by a high degree of variation in the data and thus explaining the pattern on the scores plot. Some analytes of interest are labelled on the plot.

where α denotes the intercept, β_i are the regression coefficients and $\boldsymbol{\varepsilon}$ is a normal error term. As the $\mathbf{t}_j$ are orthogonal, PCR reduces to a sum of univariate regressions (Hastie *et al.*, 2001). Usually only the first principal components are used and the $p - k$ smallest eigenvalue components are discarded. PCR has major advantages over standard multivariate regression when X is singular and when $n < p$ due to the dimension reduction. In the latter case, the eigenvalues $\lambda_n = \cdots = \lambda_p = 0$ and so at most $n - 1$ components can be included.

11.3 PARTIAL LEAST SQUARES AND RELATED METHODS

11.3.1 Partial Least Squares

Partial least-squares regression (Wold, 1975) is extensively used in chemometrics as an alternative to ordinary least squares (OLS) in ill-conditioned problems (e.g. $n < p$). It is most often used when the explanatory variables are highly collinear and when they outnumber the observations.

PCA (and therefore PCR) finds, in some way uncritically, those latent variables (scores) that describe as much as possible of the variation in X. In PCR, the latent variables are calculated without consideration of the response and it is possible that useful predictive information for $\mathbf{y}$ is discarded as noise. If there is a lot of variation in X not correlated with the response then the latent variables found by PCR might not be adequate to describe $\mathbf{y}$. PLS tries to solve this problem by constructing latent variables that are *relevant* for describing $\mathbf{y}$. The goal of PLS is to construct linear combinations of the original variables that have simultaneously high variance and high correlation with the response (Frank and

Friedman, 1993; Hastie *et al.*, 2001):

$$\mathbf{u}_j = \arg \max_{\|\mathbf{u}\|=1} \mathrm{Corr}^2(\mathbf{y}, X\mathbf{u})\mathrm{Var}(X\mathbf{u}), \qquad (11.1)$$

subject to the constraint that $\mathbf{u}_i' S \mathbf{u}_j = 0$ for all $1 \le i < j$. Therefore the goal of PLS is to find optimal weights $\mathbf{u}_j$ to form a small number of latent variables that best predict the response $\mathbf{y}$.

PLS was introduced as a modification of the NIPALS algorithm for PCA (Wold, 1966), but it is most often presented as a latent factor regression method and produces a sequence of models $\mathbf{y}_j$. The algorithm to compute the first k latent variables is described below.

PLS algorithm

---

0. Standardise each $\mathbf{x}_j$ to have mean 0 and variance 1.
   Standardise $\mathbf{y}$ to have mean 0.
1. Initialise $j = 1$, $X_j = X$ and $\mathbf{y}_j = \mathbf{y}$.
2. Compute the univariate regression coefficient of $\mathbf{y}$ on each of the $\mathbf{x}$: $\mathbf{w}_j = X_j' \mathbf{y}_j$.
3. Normalise $\mathbf{w}_j$ to have unit length: $\mathbf{w_j} = \mathbf{w}_j / \|\mathbf{w}_j\|$.
4. Project X_j on $\mathbf{w}_j$ to find the corresponding score vector $\mathbf{t}_j : \mathbf{t}_j = X'\mathbf{w}_j$.
5. Regress $\mathbf{y}$ on $\mathbf{t}_j$ to get the ordinary least-square regression coefficient $\widehat{b}_j : \widehat{b}_j = \mathbf{t}_j' \mathbf{y} / \|\mathbf{t}_j\|$.
6. Project X_j on $\mathbf{t}_j$ to find the corresponding loadings $\mathbf{u}_j : \mathbf{u}_j = X_j' \mathbf{t}_j / \|\mathbf{t}_j\|$.
7. Orthogonalise X_j with respect to $\mathbf{t}_j : X_{j+1} = X_j - \mathbf{t}_j \mathbf{u}_j'$.
8. Let $\mathbf{y}_{j+1} = \mathbf{y}_j - \widehat{b}_j \mathbf{t}_j$.
9. Let $j = j + 1$ and repeat steps 2–9 until $j = k$.

Similar to the NIPALS algorithm for PCA, the number of components k is usually determined by CV procedures. The emphasis in PLS is not only on regression but also on uncovering latent structure in X and $\mathbf{y}$. This latent structure is made up of pairs of latent vectors, $\mathbf{t}_j$ and $\mathbf{u}_j$. The latent vectors are determined through the process of estimating the weights $\mathbf{w}_j$ for the linear combination of the X variables. The weights $\mathbf{w}_j$ correspond to directions in the space of X with highest covariance with $\mathbf{y}$ and are such that large variations in X are accompanied by large variations in $\mathbf{y}$. Steps 2 and 3 of the algorithm show that the $\mathbf{w}_j$ are obtained by normalising the covariance matrix between X and $\mathbf{y}$. In step 4 we construct the latent variable $\mathbf{t}_j$ (also called partial least-squares direction) by reweighting the $\mathbf{x}_i$ by the strength of their univariate effect on $\mathbf{y}$ (the weights are given by the covariance between $\mathbf{x}_i$ and $\mathbf{y}$). In step 5 the response $\mathbf{y}$ is regressed on $\mathbf{t}_j$ to obtain the univariate OLS estimate and then in step 6 we obtain the latent variables $\mathbf{u}_j$ by regressing the columns of X on $\mathbf{t}_j$. PLS produces a bilinear representation of the data Martens and Naes (1989):

$$X = \mathbf{t}_1 \mathbf{u}_1' + \cdots + \mathbf{t}_k \mathbf{u}_k' + X_{k+1}$$
$$\mathbf{y} = \widehat{b}_1 \mathbf{t}_1 + \cdots + \widehat{b}_k \mathbf{t}_k + \mathbf{y}_{k+1},$$

where the $\mathbf{t}_j$ are orthogonal and X_{k+1} and $\mathbf{y}_{k+1}$ are the residuals. The $\mathbf{u}_j$ and the $\widehat{b}_j$ are estimated by regression, and PLS fits a sequence of bilinear models by least squares, thus the name *partial least squares*. In steps 7 and 8 the residuals X_{j+1} and $\mathbf{y}_{j+1}$ are calculated and in each iteration only the subspace of X that is orthogonal to the earlier

linear combinations developed in the X space is used and the y space is projected on the space orthogonal to the previous X component. The basic PLS algorithm can be generalised to handle multiple responses.

The maximum number K of components is $\leq \min(n - 1, p)$. The number of components k should be chosen so that X_{k+1} contains no information on y_{k+1} (i.e. they are uncorrelated) and similar to PCA, k is usually chosen through CV. The first few PLS components are retained as they account for most of the covariance between X and y and therefore lead to a more parsimonious representation of the data. We refer to Stone and Brooks (1990), Frank and Friedman (1993), Breiman and Friedman (1997), Burnham *et al.* (1999) and Butler and Denham (2000) for a review of the statistical properties of PLS.

11.3.2   PLS and Discrimination

In many applications it is common that the response variable y is categorical, representing, say, class membership, e.g. control/dosed, affected/unaffected individuals, etc. Partial least-squares discriminant analysis (PLS-DA) consists of a classical PLS regression where the response variable is a categorical one (replaced by the set of dummy variables describing the categories). The goal of PLS-DA is to sharpen the separation between groups of observations, by rotating PCA components such that a maximum separation among classes is obtained, and to understand which variables carry the class separating information. Figure 11.4 shows the application of PLS-DA to a subset of the hydrazine data. To illustrate class discrimination, we have restricted the analysis to samples from control and low-dose animals obtained at 8–72 hours post dose. The separation between the groups is evident on the first latent variable, and metabolites responsible for such separation can be found by examining the corresponding loadings.
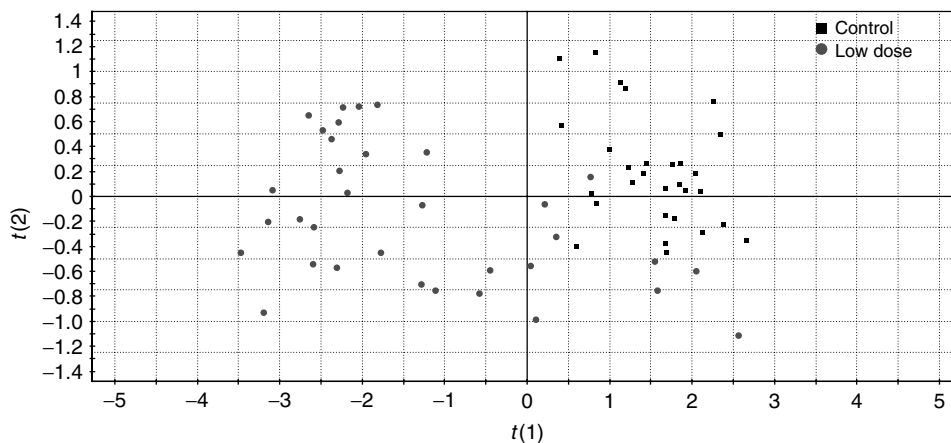
Figure 11.4   PLS-DA scores of the first two latent variables for a subset of the hydrazine dataset (control and low dose, 8–72 hours). The plot shows a clear separation between the two subgroups on the first latent variable. The metabolites responsible for the separation can be found from the corresponding loadings (results not shown).

11.3.3   Orthogonal Projections to Latent Structure

Orthogonal projections to latent structure (O-PLS) is a modification of the original PLS algorithm and was proposed with the goal of removing variation from X that is not correlated with $\mathbf{y}$ (Trygg and Wold, 2002). Spectra often contain systematic variations that are unrelated to the response $\mathbf{y}$. This systematic variation may be due to experimental effects (such as a temperature drift in the spectrometer) or systematic biological variation (e.g. differences in diets between human subjects which are not easily controlled) that often constitute the major part of the variation of the sample spectra. It is important to deal with the variation in X that is uncorrelated with $\mathbf{y}$ as it affects the construction and interpretation of statistical models, it may lead to unreliable predictions for new samples and may also affect the robustness of the model over time. In the analytical chemistry literature, differentiation and signal correction are commonly used to remove systematic variation from the sample (see, e.g. Savitsky and Golay, 1964; Geladi *et al.*, 1985).

O-PLS provides a method to remove the variation in X orthogonal to $\mathbf{y}$, therefore improving the interpretation of PLS models and reducing model complexity. It analyses the disturbing variation in each regular PLS component. Also O-PLS produces a bilinear representation of the data:

$$X = \mathbf{t}\mathbf{u}' + \mathbf{t}_{o1}\mathbf{u}'_{o1} + \cdots + \mathbf{t}_{ok}\mathbf{u}'_{ok} + X_{k+1}$$

$$\mathbf{y} = \widehat{b}\mathbf{t} + \mathbf{y}_{k+1}.$$

As in PLS, $\mathbf{t}$ represents the score vector for X and $\mathbf{y}$, $\mathbf{u}$ is the vector of orthonormal loadings, $\widehat{b}$ is a scalar and X_{k+1} and $\mathbf{y}_{k+1}$ are the respective residual matrices for X and $\mathbf{y}$. The $\mathbf{t}_{oj}$ are the scores orthogonal to $\mathbf{y}$ and the $\mathbf{u}_{oj}$ are the corresponding loadings. The O-PLS method requires only a small modification of the PLS algorithm and therefore can be embedded as an integrated part of the regular PLS modelling. It provides predictions that are similar to those provided by PLS (Cloarec *et al.*, 2005), but leads to more parsimonious models. The number of correlated O-PLS components is reduced to one in the case of a single response, thus making interpretation of the model easier. Moreover, the structured noise is modelled separately from the variation common to X and $\mathbf{y}$ and this gives the opportunity to analyse the uncorrelated variation and possibly explain its presence. An extension of the O-PLS algorithm known as *O2-PLS* has been proposed by Trygg and Wold (2003). O2-PLS improves on the original algorithm when dealing with multiple responses by allowing, e.g. the calculation of orthogonal latent vectors in the Y space.

Example. We illustrate the use of the O-PLS method with data from a study on paracetamol (acetaminophen) toxicity (Coen *et al.*, 2003). Twenty-six rats, divided into dose ($n = 17$) and control ($n = 9$) groups were exposed to 150 and $0\,\mathrm{mg\,kg}^{-1}$ paracetamol, respectively. Liver tissue was extracted and ^{1}H NMR spectra acquired to ascertain any metabolic difference due to paracetamol toxicity. Figure 11.5 shows the scores plot from an initial PCA model and shows poor separation between the groups, along with a clear drift over time. This drift was found to be the result of temperature variations in the spectrometer over the course of the experiment, introducing variation in the profiles which was unrelated to the toxicity effect. O-PLS-DA was used to remove the confounding effect of the temperature drift. This is shown in Figure 11.6, where variation orthogonal to the
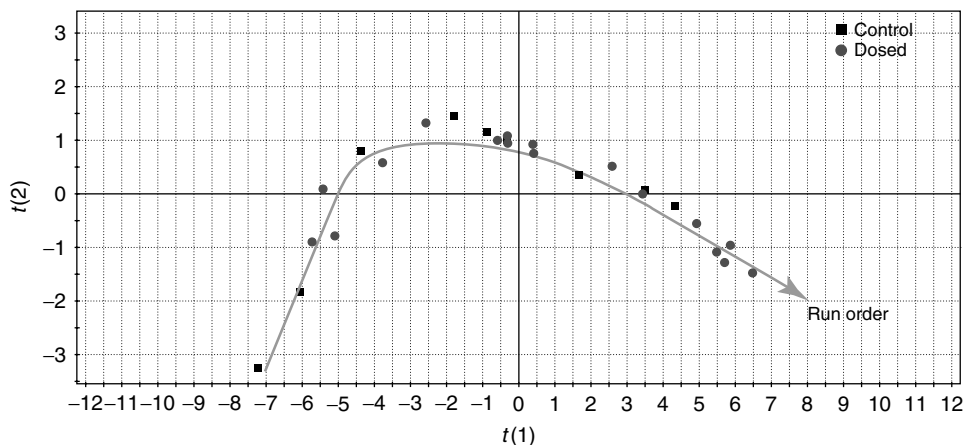
Figure 11.5 Scores plot of the first two principal components of the paracetamol dataset. The PCA score plot shows no separation and a clear trend due to run order is highlighted by the arrow.
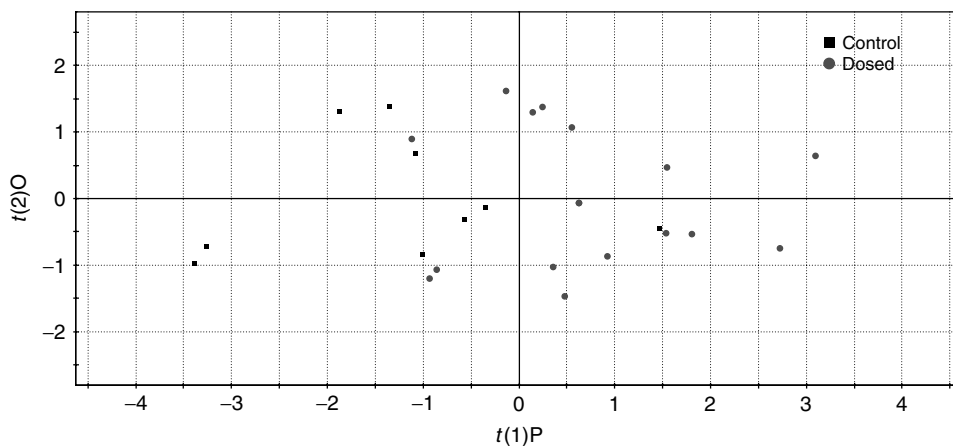
Figure 11.6 Scores plot of the first orthogonal component (y-axis) versus the first correlated component (x-axis) for the paracetamol dataset. The systematic variation visible in Figure 11.5 has been captured by the orthogonal component and it is no longer confounding the interpretation of the correlated component.

class difference has been removed from the first component and a clear separation between the groups is now observed on the scores plot. As mentioned above, a benefit of the O-PLS approach is that one may examine the orthogonal part of the data to ascertain the reasons for the confounding variation. Figure 11.7 plots the scores on the first orthogonal component and a clear trend with run order is seen, indicating that the time-ordered drift has been removed. The orthogonal loadings can also be examined (data not shown) giving an indication of the metabolites that are most affected by the time drift.
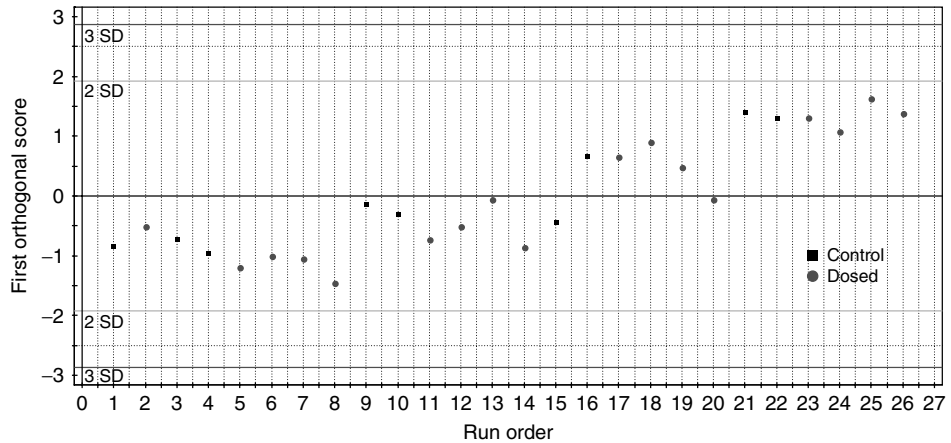
Figure 11.7   The plot shows the a strong relationship between the first orthogonal component scores (*y*-axis) and the experimental run order (*x*-axis). The orthogonal filtering captures most of the systematic variation due instrumental drift, visible in Figure 11.5.

11.4   CLUSTERING PROCEDURES

Cluster analysis identifies subgroups or clusters in multivariate data, in such a way that objects belonging to the same cluster resemble each other, whereas objects in different clusters are dissimilar. In two or three dimensions, clusters can often be visualised, but in higher dimensions, we need some kind of analytical assistance.

Datasets for clustering of N observations can have either of the following structures:

- an $N \times p$ *data* matrix, where rows contain the different observations and columns contain the different variables.
- an $N \times N$ *dissimilarity* matrix, whose (i, j) th element is d_{ij}, the *distance* or *dissimilarity* between observations i and j that has the properties

  - $d_{ii} = 0$
  - $d_{ij} \geq 0$
  - $d_{ji} = d_{ij}$.

- The most typical distance measure between two continuously measured data points i and j with measurement vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is the *Euclidean* distance

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2} = \sqrt{\left(\mathbf{x}_i - \mathbf{x}_j \right)^{\mathrm{T}} \left(\mathbf{x}_i - \mathbf{x}_j \right)}.$$

Clustering algorithms fall into two categories:

1. *Partitioning Algorithms.* A partitioning algorithm describes a method that divides the dataset into k clusters, where the integer k needs to be specified. Typically, the algorithm is run for a range of k values. For each k, the algorithm carries out the

clustering and also yields a quality index which allows *selection of* the 'best' value of k.

2. *Hierarchical Algorithms.* A hierarchical algorithm yields an entire hierarchy of clusterings for the given dataset. *Agglomerative methods* start with the situation where each object in the dataset forms its own cluster and then successively merge clusters until only one large cluster (the entire dataset) remains. *Divisive methods* start by considering the whole dataset as one cluster and then split up clusters until each object is separated.

11.4.1 Partitioning Methods

Partitioning methods specify an initial number of groups and iteratively reallocate observations between groups until some equilibrium is attained. Several different algorithms are available

1. *The **k**-means algorithm.* In the k-means algorithm the observations are classified as belonging to 1 of k groups. Group membership is determined by calculating the centroid for each group (the multidimensional version of the mean) and assigning each observation to the group with the closest centroid. The k-means algorithm iterates between calculating the centroids on the basis of the current group memberships and reassigning observations to groups on the basis of the new centroids. When a Euclidean metric is used, cluster centroids are computed (on the basis of a least squares criterion) as the arithmetic means of the observations currently assigned to each cluster, and then observations are re-assigned to the cluster whose centroid is closest. This assignment is performed in an iterative fashion, either from a starting allocation or configuration or from a set of starting centroids. The k-means algorithm is essentially model free (*though it assumes the clusters are compact and spherically symmetric*). It is heavily dependent on initial group assignments and the use of the least-squares penalty.

2. *Partitioning around medoids (PAM).* The PAM method uses medoids rather than centroids, i.e. medians rather than means in each dimension. This approach increases robustness relative to the least-squares approach given above.

11.4.2 Hierarchical Clustering

Hierarchical clustering procedures can be carried out in two ways as listed below:

- *Heuristic criteria.* The basic hierarchical agglomeration algorithm starts with each object in a group of its own. At each iteration, two groups are merged to form a new group; the merger chosen is the one that leads to the smallest increase in the sum of the *within-group sums of squares*. The number of iterations is equal to the number of objects minus 1 and in the end all the objects are together in a single group. This is known as *Ward's method*, the *sum of squares method* or the *trace method*. The hierarchical agglomeration algorithm can be used with criteria other than the sum of squares criterion, such as the *average*, *single* or *complete linkage* methods described below.

- *Model-based criteria.* Model-based clustering is based on the assumption that the data are generated by a mixture of underlying probability distributions. Specifically, it is

assumed that the population of interest consists of k different subpopulations (usually assumed to be multivariate normally distributed) and that the density of an observation from the subpopulation is *specified by* some unknown vector of parameters *which are to be determined*. We study model-based criteria in more detail below.

In conventional hierarchical clustering, the method of agglomeration or combining clusters is determined by the distance between the clusters themselves, and there are several available choices. For merging two clusters C_i and C_j, with N_1 and N_2 elements respectively, the following criteria can be used: (1) *average linkage clustering*, where the two clusters that have the smallest average distance between the points in one cluster and the points in the other are merged, (2) *connected (single linkage, nearest-neighbour) clustering*, where the two clusters that have the smallest distance between *any* point in the first cluster and *any* point in the second cluster are merged, (3) *compact (complete linkage, furthest-neighbour) clustering*, where the two clusters that have the largest distance between *any* point in the first cluster *and any point* in the second cluster are merged. Efficient algorithms to achieve hierarchical clustering exist and are readily available in standard statistical packages such as R.

11.4.3   Model-based Hierarchical Clustering

Another approach to hierarchical clustering is *model-based clustering*, which is based on the assumption that the data are generated by a mixture of K underlying probability distributions. Given that data matrix $\mathbf{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right)^{\mathrm{T}}$, let

$$\gamma = \left(\gamma_1, \ldots, \gamma_N\right)$$

denote the cluster labels, where $\gamma_i = k$ if the ith data point comes from the kth subpopulation. In the classification procedure, a maximum likelihood procedure is used to choose the parameters in the model.

Commonly, the assumption is made that the data in the different subpopulations follow multivariate normal distributions, with mean $\boldsymbol{\mu}_k$ and covariance matrix Σ_k for cluster k. If

$$\Sigma_k = \sigma^2 I_p \qquad I_p = \mathrm{diag}\left(1, \ldots, 1\right), \text{ a } p \times p \text{ matrix.}$$

then maximising the likelihood is the same as minimising the sum of the within-group sums of squares.

More general covariance models have been implemented in the statistical package R in the library `mclust`. We give brief details. The key to specifying this is the eigendecomposition of Σ_k, given by eigenvalues $\lambda_1, \ldots, \lambda_p$ and eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$, as in PCA. The eigenvectors of Σ_k, specify the orientation of the kth cluster, the largest eigenvalue λ_1 specifies its variance or size and the ratios of the other eigenvalues to the largest one specify its shape. Further, if $\Sigma_k = \sigma_k^2 I_p$, the criterion corresponds to hyperspherical clusters of different sizes; this is known as the *spherical* criterion. Another criterion results from constraining only the shape to be the same across clusters. This is achieved by fixing the eigenvalue ratios

$$\alpha_j = \frac{\lambda_j}{\lambda_1} \qquad j = 2, 3, \ldots, p$$

across clusters; different choices for the specification yield ellipsoidal, linear or spherical clusters.

11.4.4 Choosing the Number of Clusters

A hierarchical clustering procedure gives the sequence by which the clusters are merged (in agglomerative clustering) or split (in divisive clustering) according to the model or distance measure used, but does not give an indication for the number of clusters that are present in the data (under the model specification). This is obviously an important consideration. One advantage of the model-based approach to clustering is that it allows the use of statistical model assessment procedures to assist in the choice of the number of clusters. A common method is to use approximate Bayes factors to compare models with different numbers of clusters. This method gives a systematic means of selecting the parameterisation of the model, the clustering method and also the number of clusters. The *Bayes factor* is the posterior odds for one model against the other assuming neither is favoured *a priori*. A common approximation to the Bayes factor is the Bayes information criterion, which for model M is given by

$$BIC_M = -2 \log L_M + \text{const} \approx -2 \log L_M(\widehat{\theta}) + d_M \log N,$$

where L_M is the Bayesian marginal likelihood, $L_M(\widehat{\theta})$ is the maximised log likelihood of the data for the model M, N is the number of data points and d_M is the number of parameters estimated in the model. The number of clusters is not considered a parameter for the purposes of computing the BIC. The first term is a measure of fidelity to the data and the second is a penalty for the number of parameters in the model. The more negative the value of the BIC, the stronger the evidence for the model.

11.4.5 Displaying and Interpreting Clustering Results

The principal display plot for a clustering analysis is the *dendrogram*, which plots all of the individual data linked at successive levels by means of a binary 'tree'. Such a plot is displayed in Plate 10. The distance up the tree, or height, represents overall similarity between the samples and can be useful in determining the number of clusters that are present in the data.

Plate 10 displays the hierarchical clustering results for a subset of the hydrazine dataset (control and low dose, 8–72 hours). The control samples are shown in black while the dosed samples are shown in red. There are two main branches corresponding to control-like samples and dosed samples. The two groups of subjects are almost perfectly delineated by the clustering procedure if a cut is made at similarity level 0.6 units.

Hierarchical clustering has often been used to investigate metabolic profiling data. Some studies have investigated relationships between the metabolic profiles themselves (Beckonert *et al.*, 2003), while others have applied the method to elucidate dependencies between metabolite concentrations (by clustering the transposed data matrix X' (Dumas *et al.*, 2002)). In both cases, dendrograms generated intuitive visualisations of the hierarchy, allowing effects such as outliers, misclassifications and chemical structure correlations to be observed. Clustering approaches do not, of themselves, offer diagnostic information on the reasons for classification of any given object. This can be important if one wishes to know, e.g. the metabolites which are critical to determining cluster

membership. Mean cluster profiles and inter-cluster differences can be inspected; though information can still be lacking on the extent of cluster overlap and the overall relationship between clusters. Finally, note that validation of the results of a clustering exercise is recommended. This can take the form of data perturbation and reclustering to ensure that the removal of a minority of data points is not pivotal to the clustering inferences or the number of clusters selected.

11.5 NEURAL NETWORKS, KERNEL METHODS AND RELATED APPROACHES

An artificial neural network (ANN) is a multilayered statistical model that represents the variation in a potentially highly complex response or output variable to a collection of input variables via varying numbers of unobserved or latent variables linked through simple mathematical functions and a series of probabilistic dependence assumptions. ANNs are a member of a broad class of supervised learning algorithms that attempt to approximate the true data generation mechanism by mathematical models deduced from observation of cases essentially via regression arguments. These models have been perceived as mechanistic approximations to actual biological neural networks, although this interpretation is neither necessary nor uniformly helpful. We summarise the nature of such models below; see, e.g. (Ripley, 1996 Chapter 5) for a comprehensive description of statistical aspects.

11.5.1 Mathematical Formulation

The simplest mathematical formulation of an ANN involves three levels of interlinked variables;

- The *outputs*, Y (a $q \times 1$ vector), interpreted as a single/collection of continuous or categorical random variable(s).
- The *inputs*, X ($p \times 1$), interpreted as a collection of random variables believed to influence the variation in the *response* across an experimental sample of Ys.
- The *hidden variables*, Z ($d \times 1$), interpreted as a collection of unobserved random variables that form the hidden link between X and Ys.

This structure can be generalised to incorporate multiple hidden layers, but we restrict attention to the single *hidden* layer case.

In a classical biological conception of the model, the hidden variables Z have some physical interpretation as nodes in a neuronal network in the brain, but this interpretation is not necessary. Mathematically, perhaps the most useful interpretation of Z is as a *projection* of X onto a lower dimensional space ($p > d$) of *features* that facilitates modelling of the variation in Y. These features may or may not have a physical interpretation. Diagrammatically, such a network (with $q = 2$, $p = 6$ and $d = 3$) is represented in Figure 11.8.

The arrows connecting the nodes in Figure 11.8 represent functional links encompassed through mathematical functions. At the first stage, we represent the components of Z as weighted sums of functions of the input variables, typically

$$z_j = G_j \left(\alpha_j + \sum_{k=1}^{p} w_{jk} g_{jk}(x_k) \right) \qquad j = 1, \ldots, d, \qquad (11.2)$$
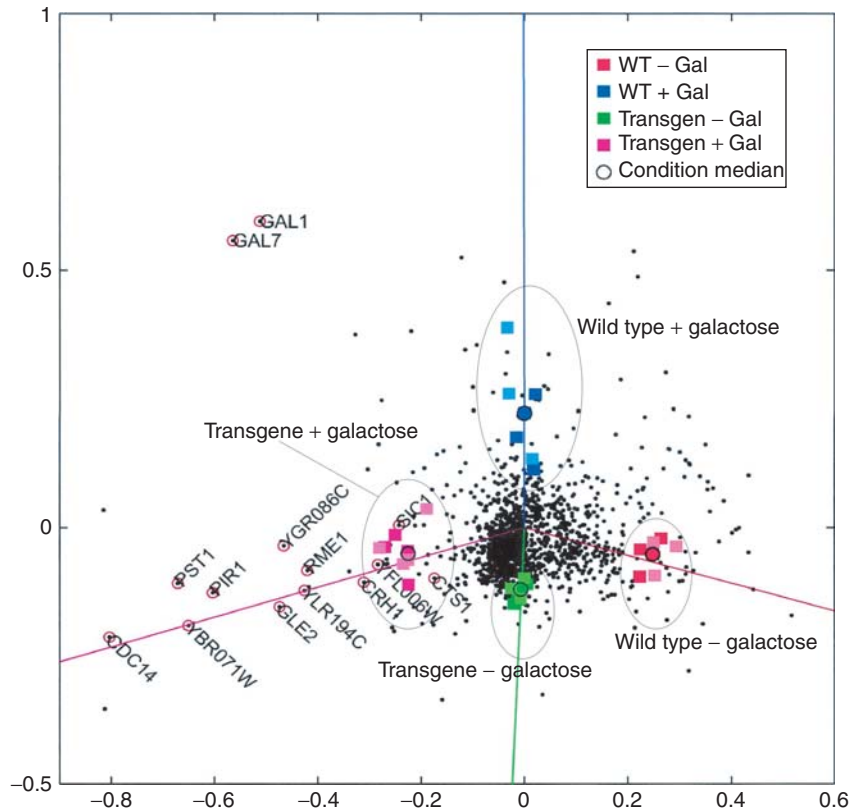
Plate 7. Correspondence analysis applied to an experiment that searched for genes expressed as a consequence of induction of the yeast cell cycle gene *CDC14* (Fellenberg *et al.*, 2001). A yeast transgene was constructed with the *CDC14* gene under a galactose-dependent promotor that allows induction of the *CDC14* gene through the addition of galactose. As a consequence, one observes upregulation of genes both due to *CDC14* induction and due to the natural reaction to galactose. Four conditions were studied: wildtype yeast with and without galactose, and the yeast transgene with and without galactose. For each condition several replicates were made. The correspondence analysis biplot shows an embedding of rows and columns of the entire data matrix with genes depicted as black dots and hybridizations depicted as small squares. Replicates for each condition cluster together, and each of the four clusters defines a direction in which the genes that are typical of the condition can be found. The bisection between the two galactose conditions points to two *GAL* genes, known to be involved in the galactose pathway. Genes in the transgene+galactose condition that are turned on in response to the addition of galactose are attracted also by the wildtype+galactose condition. Thus, the lower left direction highlights genes that are exclusively due to the *CDC14* induction. Genes are encircled which show up in a related experiment (Spellman *et al.*, 1998), too, where they are also seen to be linked to *CDC14*. (Adapted with permission from Fellenberg K *et al.* (2001) Correspondence analysis applied to microarray data. *Proc Natl Acad Sci USA*. 98:10781-6. Copyright (1998) National Academy of Sciences, U.S.A.).
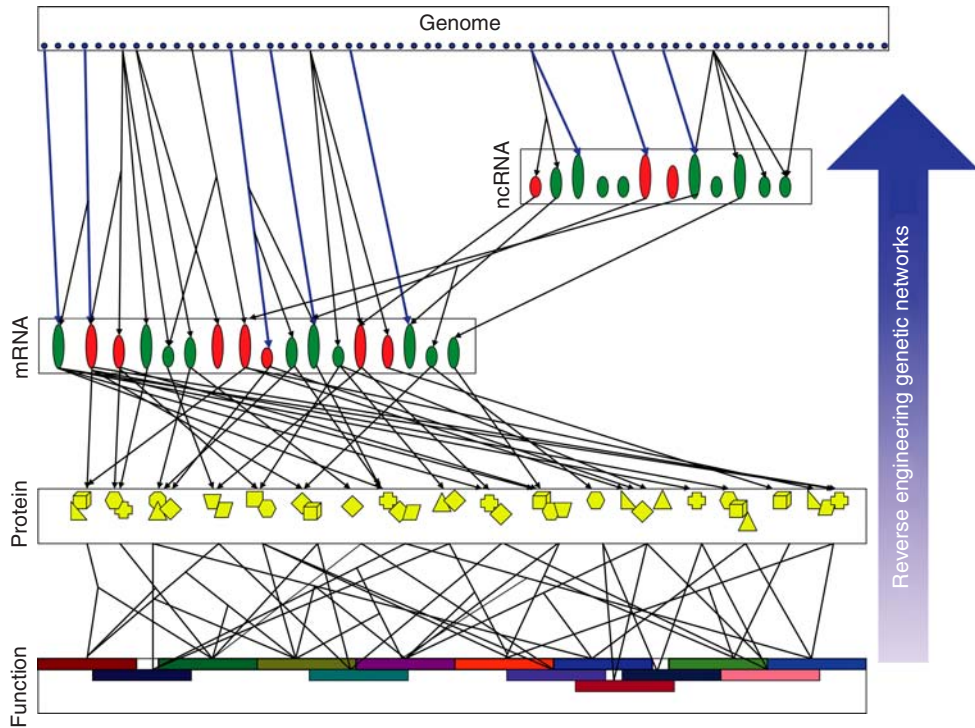
Plate 8. Simplified view of strategy for reverse engineering gene networks using genetics in a fixed environment. The top layer represents DNA in the genome, where in any given population we can associate changes in the DNA with changes in the levels of transcription of both protein-coding and ncRNA genes. DNA variations that fall within the region of the structural gene and associate with that gene's expression are referred to as proximal eQTL, as opposed to distal eQTL in which the DNA variation does not fall in the genomic region supporting the corresponding structural gene region (Doss *et al.*, 2005). The proximity of transcribed sequences to the DNA provides for increased power to detect regions of the genome affecting transcript abundance levels. Changes in RNA are then shown to induce changes in proteins, where a complex web of protein interactions can form and give rise to varied cellular functions that in turn lead to disease. The gene network reconstruction methods discussed in the text exploit the ultimate source of perturbations (changes in DNA) in a system under fixed environmental conditions to order nodes in the network.
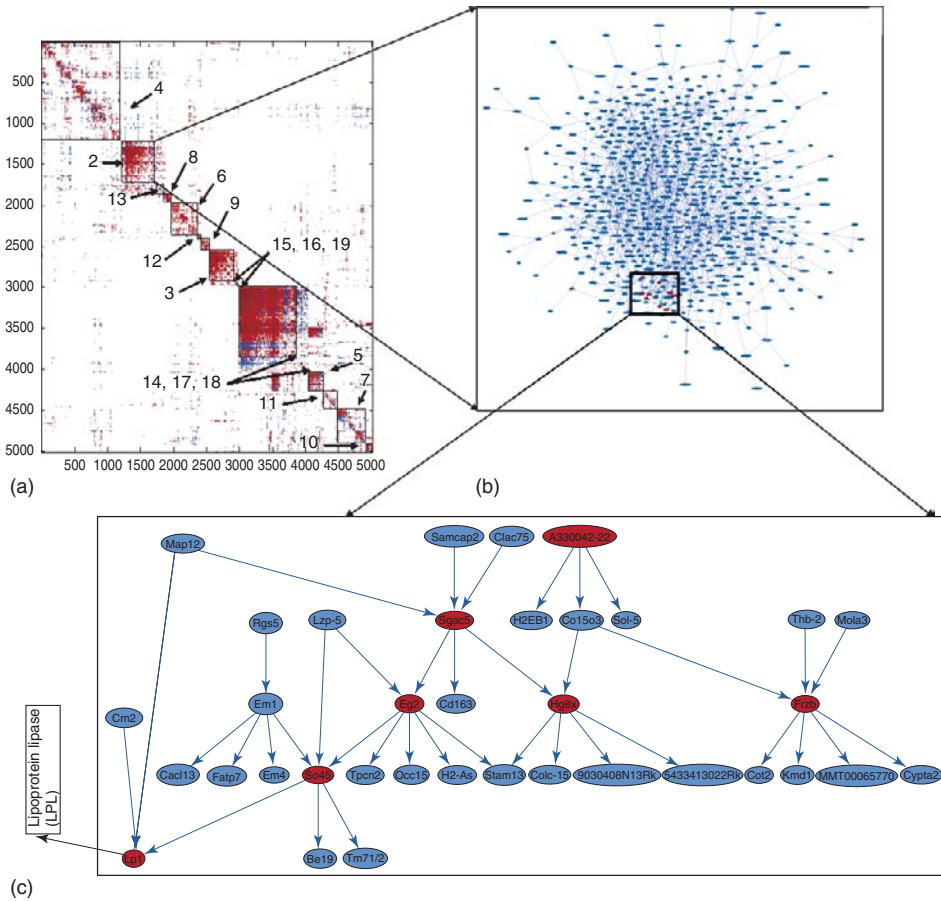
Plate 9. Coexpression and Bayesian networks from the BXH adipose expression data. (a) Topological overlap map for adipose tissue hub genes in the BXH cross. All pairs of correlations among the 5000 most highly connected genes in the BXH female adipose coexpression network are plotted in the color matrix display (red indicates positive correlation, blue indicates negative correlation, and white indicates correlation was not significant at the $p < 10^{-20}$ level). The genes are ordered along the x and y axes using an agglomerative hierarchical clustering algorithm. Tightly correlated groups of genes (modules) clearly emerge from this plot. Modules are identified as described in the text. (b) Bayesian network corresponding to genes in module 2 from (a). (c) Subnetwork consisting of 36 genes that contain the gene *Lpl*, a gene recently validated as causal for obesity (Schadt *et al.*, Submitted).
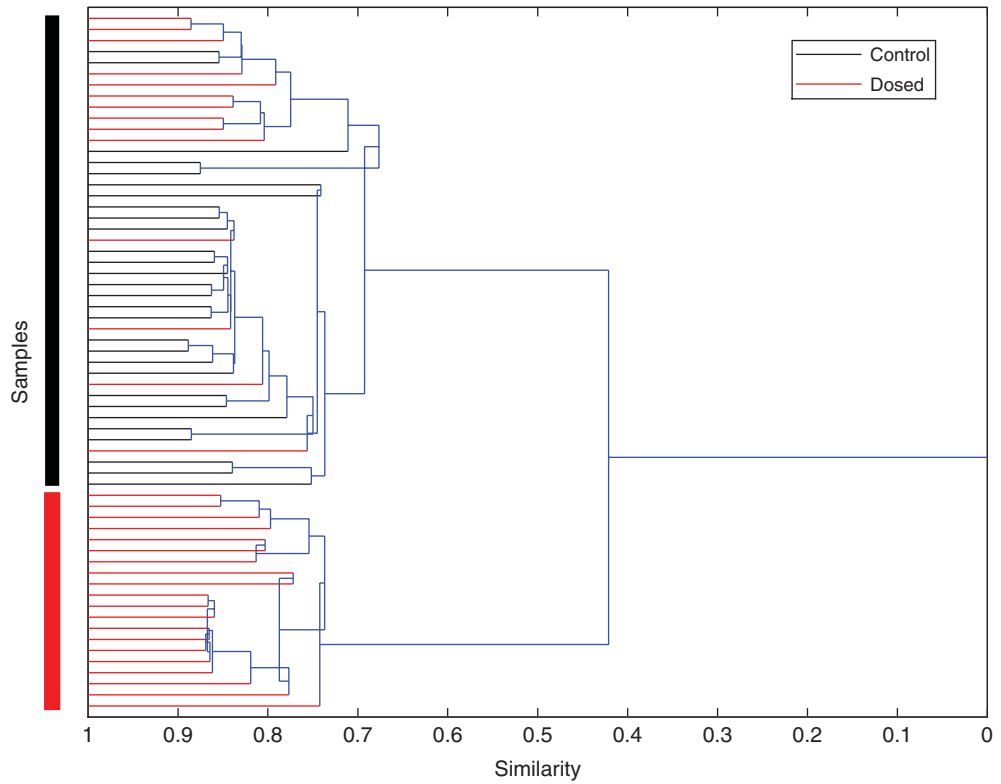
Plate 10. Hierarchical clustering for a subset of the hydrazine dataset (control and low dose 8–72 hours). The control samples are shown in black while the dosed samples are shown in red. There are two main branches corresponding to control-like samples and dosed samples. Note some dosed samples appeared to be similar to controls because the relevant animals have either recovered from or not yet responded to the toxin.
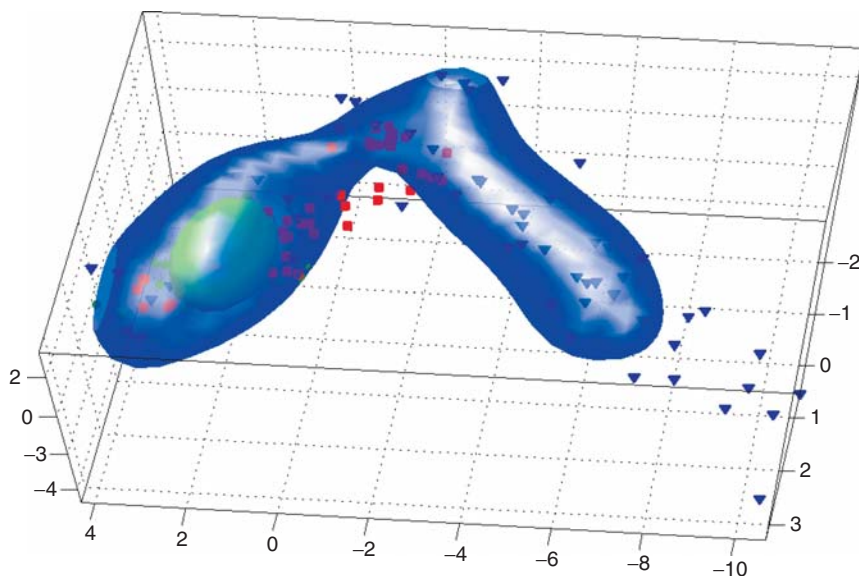
Plate 11. Visualisation using CLOUDS. The figure shows the toxicity data of Figure 11.2 (first 3 PCs) superimposed with isosurfaces of constant probability density estimated through CLOUDS. The surface for the high-dose class (blue) has been rendered transparent so as to enable viewing of the control density (green). Red squares denote samples from the low-dose class, while blue triangles denote samples from the high-dose group. One can clearly see how irregular distributions can be modelled in this way.
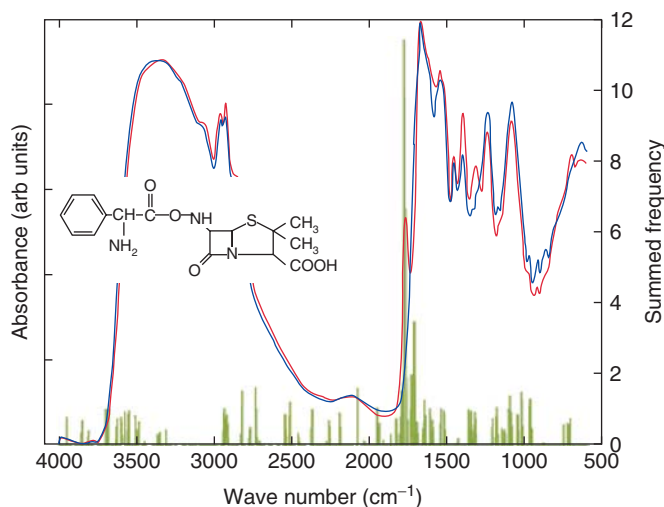
Plate 12. Summed frequency plot from GP analysis of the number of times each input (wave number) was used for the 10 evolved populations. Also shown are the normalised FT-IR spectra from *E. coli* (blue trace) and *E. coli*+5000 μg ml^{-1} ampicillin (red trace), and the structure of ampicillin. [Reprinted from Goodacre, R. (2005). Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. *Journal of Experimental Botany* **56**, 245–254, with permission from Oxford University Press.].
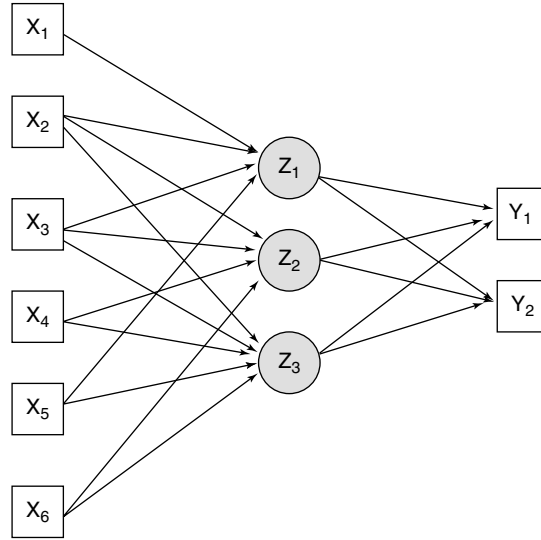
Figure 11.8 Three-level neural network with $q = 2$, $p = 6$ and $d = 3$. The hidden layer $Z = (Z_1, Z_2, Z_3)'$ links inputs X to outputs Y. Arrows indicate positive weights.

where α_j, $j = 1, \ldots, d$ are unknown constants, w_{jk}, $k = 1, \ldots, p$, $j = 1, \ldots, d$ are weights satisfying

$$0 \leq w_{jk} \leq 1 \qquad \sum_{k=1}^{p} w_{jk} = 1$$

and G_j, $j = 1, \ldots, d$ and g_{jk}, $k = 1, \ldots, p$, $j = 1, \ldots, d$ are known *link* functions, the latter often chosen to be identity functions for convenience. At the second stage we have a similar structure modelling the dependence of Y on Z,

$$y_l = H_l \left(\beta_l + \sum_{j=1}^{d} \omega_{lj} z_j \right) \qquad l = 1, \ldots, p, \tag{11.3}$$

for parameters β_l, $l = 1, \ldots, p$ and weights ω_{lj}, $j = 1, \ldots, d$, $l = 1, \ldots, p$.

In summary, therefore, we have parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{w}_j$, $j = 1, \ldots, d$ and $\boldsymbol{\omega}_l$, $l = 1, \ldots, p$ to estimate from the data. This is achieved by optimisation of some objective function characterising discrepancy of fit, i.e. for observed cases $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, we choose parameters $\widehat{\boldsymbol{\theta}}$ as

$$\widehat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} D(\boldsymbol{y}_i, \widehat{\boldsymbol{y}}_i), \tag{11.4}$$

where D measures the discrepancy between the observed $\boldsymbol{y}_i$ and the fitted value $\widehat{\boldsymbol{y}}_i$ given by using parameters $\boldsymbol{\theta}$ in (11.2) and (11.3). D is chosen to represent the continuous or categorical nature of the response variables, and the minimisation in (11.4) is achieved using numerical methods.

As described above, ANNs are flexible non-linear regression models constructed from simple mathematical functions that are learned from the observation of cases. As such, they

are ideal models for classification. However, their flexibility means that model parameters such as those determining the architecture (number of input, hidden and output nodes, etc.) are not predetermined and must be chosen on the basis of the skill and experience of the user. In addition, although good classification performance for metabolic profiles has been obtained (El-Deredy, 1997), interpretation of the rules encoded by an ANN is not straightforward. This has limited their application in the metabolic profiling arena in recent years, since model interpretation is of prime importance in applications such as biomarker screening and chemical structure elucidation. ANNs have found metabolic profiling applications in diverse areas such as classification of tumours (Howells *et al.*, 1992), preclinical toxicity prediction (Anthony *et al.*, 1995) and determination of the mode of action of herbicides in plants (Ott *et al.*, 2003). In all cases, the network weights could not be interpreted directly and other methods had to be used to determine the biochemical reasons for classification.

11.5.2   Kernel Density Estimates, PNNs and CLOUDS

Kernel density estimators (KDEs) are a well-known class of probability density estimators which have been extensively applied in many different areas of science (Parzen, 1962; Duda *et al.*, 2000). In classification problems, KDEs are usually applied to estimation of the class density (the conditional probability of observing an object, given that it is a member of the class). The estimators are especially useful when the class density deviates substantially from standard (e.g. normal) forms and allow us to estimate the density in an essentially non-parametric way. Classification of test objects is then done by calculating the class density at the coordinates of the test object and assigning the object to the class whose density is highest. Use of a Gaussian kernel leads to estimates of the form

$$p\left(\underline{x} \mid \underline{x}_{i \in A}\right) = \frac{1}{N_A (2\pi\sigma^2)^{M/2}} \sum_{i \in A} \exp\left\{-\frac{\|\underline{x} - \underline{x}_i\|^2}{2\sigma^2}\right\}, \qquad (11.5)$$

for the density of class A at point $\underline{x}$, where N_A is the number of training set objects $\underline{x}_i$ in the class, M is the number of dimensions and σ^2 is the width of the kernel. The kernel width regulates the smoothness of the density estimate and requires careful choice. It is often fixed by CV methods at the value that maximises the likelihood of class membership for objects left out of the training set in each CV round.

Probabilistic neural networks (PNNs) are a special kind of ANN whose architecture enables them to compute a kernel density estimate of the above form. Each input node corresponds to a variable in the input space and, when a test object is presented, feeds the corresponding value to each of the nodes in the second layer. Each second-layer node corresponds to one training set vector and computes the exponential of (11.5), feeding its output to a final layer of nodes, which computes the summation. Classification of unknowns by density superposition (CLOUDS) (Ebbels *et al.*, 2003) uses the KDE/PNN framework to both perform pointwise classification and detect similarity between the estimated densities. The latter function uses the overlap integral of the two densities as a measure of similarity and was developed to compare high-dimensional distributions with complex shape and topology.

The potential of PNNs to model complex high-dimensional metabolic data was first recognised by Holmes *et al.* (2001) who developed a model to classify ^{1}H NMR spectra of urine from laboratory rats treated with well-known toxins. When asked to classify

test samples into 1 of 18 toxicity-related classes, the PNN obtained higher classification accuracies than back-propagation neural networks and soft independent modelling of class analogy (SIMCA). In the area of food processing, PNNs were recently applied to classifying ^{1}H NMR spectra of fish tissue extracts according to different processing treatments (Martinez *et al.*, 2005). In order to avoid problems with the high dimensionality of the spectra, the scores from the first 20 principal components were used as inputs to the network, which successfully classified 80 % of the spectra. In metabonomic toxicology, CLOUDS was applied to several thousand ^{1}H NMR urine spectra from laboratory rats subjected to 19 different treatments known to cause toxic or other metabolic effects (Ebbels *et al.*, 2003). The approach was able to classify samples according to liver or kidney toxicity with 77 and 90 % success respectively, while experiencing a low, 2 %, rate of confusion between the classes. The class probabilities were further combined into two parameters, the confidence and uniqueness of the classification, which allowed pre-selection of the best-classified urinary profiles. Recently, the CLOUDS methodology was employed in the construction of an 'expert system' for preclinical toxicity screening using a large database of ^{1}H NMR urinary metabolic profiles from 80 different treatments (Ebbels *et al.*, 2006). A novel measure of similarity between classes was developed on the basis of the overlap integral of the density estimates. Although the system was unable to judge the likely toxicity of some of the treatments, where a decision could be made, over 92 % were classified correctly. In addition, the system correctly determined the site of toxicity for two blinded treatments.

The densities estimated by CLOUDS and similar techniques can be visualised by computing contours or isosurfaces of constant probability density and viewing them with the help of dimension reduction techniques such as PCA. Plate 11 illustrates this with the metabonomic toxicology data of Figure 11.2. The control class has a shape that might be well summarised by a multivariate normal, while the high-dose cloud extends away from the controls in a hairpinlike trajectory. The part of the high-dose density coinciding with controls is due to samples from animals that are either yet to react or have already recovered from the toxic episode.

Overall, use of KDE methods, while versatile and well suited to the characteristics of metabolic profiling data, is subject to some drawbacks. The most important of these is the large amount of data required to accurately estimate the density when the number of dimensions is high. Along with the high storage requirement and the need for the ability to interpret models in terms of spectral features, these are areas that will be the subject of further research in the near future.

11.6 EVOLUTIONARY ALGORITHMS

Evolutionary computations have attracted increasing interest as a method of solving complex problems in medicine, industry and bioinformatics. Evolutionary algorithms (EA) are ideal strategies for mining metabolite data to build useful relationships, rules and predictions. The advantages of this methodology include conceptual simplicity and broad applicability, the ability to optimise complex multimodal objective functions and to outperform standard optimisation procedures in real world applications. The algorithms are typically flexible and easily hybridised with other methods such as neural networks (Fogel and Corne, 2003). EAs are adaptive procedures, motivated by genetic processes

as they evolve a population of *chromosomal structures* (possible solutions), in order to find the *fittest* individual. Candidate solutions to the optimisation problem play the role of individuals in a population that is evolved through generations, undergoing processes inspired by biological evolution: reproduction, mutation, recombination, natural selection and survival of the fittest. In general, the evolutionary optimisation techniques use a population of possible solutions subjected to random *variation* and *selection* until some termination criterion is satisfied. The fitness of each individual in the population reflects the individual's worth in relation to some objective function.

The aim of the EA is to progressively develop better solutions to the problem under study by modifying previous solutions that exhibited good performance. Starting with an initial population of individuals, generated through some randomised process, each individual is then placed in a common environment where it competes and breeds with the other members of the population. In many applications, the environment is usually referred to as the design space and is the set of all possible solutions for a given problem. The individual's fitness shows how well it has adapted to its environment, i.e. larger fitness values correspond to better solutions. The fittest individuals have more chance to survive in the next generation and to be selected to reproduce. New individuals are generated through variation operators, the most common of which are mutation, the introduction of one or more random changes to a single parent individual, and recombination (commonly referred to as crossover), which consists in randomly taking components/characteristics from two or more parent individuals to create children. Variation operators are fundamental as they affect the algorithm's search power: when the search space has many local optima, large-scale mutations might be useful for escaping local optima, but less radical mutations might be important to proceed towards a global optimum. Following the application of variation operators, two sets of solutions exist: the parent population and the child population. Thus a selection for survival stage is required to form the new population. This step requires scoring each solution on its worth with respect to a given goal, i.e. according to its fitness. It is essential that the fitness function reflects the problem characteristics although the specification of a suitable function can, in itself, be a difficult task. Once the solutions have been scored, then the new generation is formed according to various schemes: some schemes require that the new generation is formed only by child solutions, while in other cases the new population is formed by combining individuals from both the parent and the child population. Selection for survival is often a deterministic process, in which the best solutions are selected, however stochastic schemes are also possible. Most EAs work on a generational base, applying variation and selection operators to all members of the population at a given time and iterating this process for a pre-specified number of generations or until convergence to a local or global optimum has been reached.

EAs are becoming increasingly popular in chemometrics as they can provide useful supervised learning techniques that can be utilised in many scenarios, e.g. to identify metabolites associated with a particular phenotype (Goodacre, 2005). These include genetic algorithms, evolution strategies, genetic programming (GP) and genomic computing. Among the evolutionary techniques, GP has been most commonly used in metabolic profiling applications. In the GP framework, the solution is structured as a parse tree. For example, in symbolic regression the parse tree specifies the regression relationship between the response and the predictors. One of the earliest applications of GPs to metabolic data can be found in the work of Gray *et al.* (1998). They proposed

a GP algorithm to classify tumours on the basis of ^{1}H NMR spectra of biopsy extracts. Following successful use of GP algorithms for the analysis of pyrolysis mass spectral data of fruit juice (Gilbert *et al.*, 1997), similar strategies have been developed to analyse metabolic profiling and fingerprinting data from several other spectroscopic and chromatographic techniques including Fourier Transform – Infrared (FT – IR) (Johnson *et al.*, 2000; Goodacre, 2005), high performance liquid chromatography (HPLC) (Kell *et al.*, 2001), non-linear dielectric spectroscopy Woodward *et al.* (1999) and various types of MS (Taylor *et al.*, 1998; Goodacre *et al.*, 2000; 2003). In most applications the authors highlight the ability of the GP to combine small numbers of explanatory metabolites in simple prediction rules to classify the response of interest. For example, as a test of the GP approach to modelling complex mixtures such as metabolic profiles, Goodacre (2005) developed a GP to quantify the levels of the antibiotic ampicillin in cultures of the bacterium *Escherichia coli*. Ampicillin was added to *E. coli* cultures at different concentrations, observed with FT-IR spectroscopy and the GP was used in 'symbolic regression' mode to quantify the ampicillin level using the FT-IR profiles as input. The GP was able to identify the $1767 \, \mathrm{cm}^{-1}$ vibration (corresponding to the β-lactam ring of ampicillin) as the important variable corresponding to ampicillin despite large signals from other *E. coli* metabolites, thus highlighting its promise as a method for data mining complex metabolic profiles. Plate 12 shows the frequency of the number of times each input (wavenumber) was used for 10 evolved populations, i.e. 10 runs of the GP algorithm. The area of the spectra corresponding to vibration $1767 \, \mathrm{cm}^{-1}$ is clearly dominating for *E. coli* treated with ampicillin, but absent from *E. coli* alone.

11.7 CONCLUSIONS

The techniques described here have found a wide variety of applications in metabolic profiling, many of which are mentioned in the preceding paragraphs. Two of the most important applications are biomarker screening and the structural identification of metabolites. In biomarker screening, the goal is to find molecules that differentiate between two biological states, for example, healthy and diseased populations. Almost any of the supervised pattern recognition approaches described above can be employed here. The input data are metabolic profiles (or metabolite levels derived from them) and the task is to fit a model that discriminates between samples from the healthy and diseased groups. If a discrimination is found, then the model can be examined to find out which metabolites are influential in producing the discrimination and thus might be important markers of the disease process itself. We note that some statistical techniques, for example neural networks, are much less amenable to yielding discriminatory markers than others. Additionally, such discriminators are not considered true 'biomarkers' until they have been independently validated across multiple experiments/populations and shown to be mechanistically linked with the disease or condition being studied.

As mentioned in the introduction, a large proportion of the metabolome is currently unknown and therefore the structural identification of molecules is of great importance. For this purpose, statistical analysis is directed at linking multiple signals coming from the same molecule. For example, in NMR, each spin system (group of interacting nuclei) will produce a different set of resonances. In the hands of a spectroscopist, the multiplet

structure and relationship between the resonances can yield valuable information on the likely structure of the molecule. Therefore, methods aimed at linking spectral peaks that exhibit similar statistical properties are useful. For example, signals shown to discriminate between groups in a similar way may originate from the same molecule. Alternatively, specific methods can be designed to identify statistically connected signals, which may originate from a structural relationship (Cloarec *et al.*, 2005; Crockford *et al.*, 2006).

In this chapter we have described the nature of metabolic profiling data and some of the statistical approaches that are typically used to model it. Owing to space considerations we have not undertaken a comprehensive review of all methods used in the field, but have highlighted those that have seen wide application. The reader is referred to Lindon *et al.* (2007) for a more thorough description of both the chemical and data analytic techniques used, along with in-depth reviews of many different application areas of the technology. It is clear that there are many problems still to be solved in the statistical analysis of metabolic profiles, such as peak shifts in NMR and the proper treatment of differential ionisation efficiencies in MS. The lack of structural identification of a large proportion of the signals remains a huge impediment to biological interpretation and, given the diversity of the metabolome throughout the various kingdoms of life, it seems clear that this situation is unlikely to change soon. Therefore statistical approaches which can deal with a mixture of identified and unidentified peaks are likely to have a large impact in this area. Approaches based on mixed-models and wavelet approaches have proved extremely useful (Morris *et al.*, 2006; Brown *et al.*, 2001; Clyde *et al.*, 2006). For example, sparse representation of the NMR spectrum in terms of wavelet coefficients makes wavelets an excellent tool in data compression and hence feature extraction. In addition, the wide range of chemical analytic techniques used in the generation of metabolic profiles will require statistical models that can integrate and combine information across multiple platforms. Overall, the field is a young but rapidly developing one and looks set to reap many benefits from the input of researchers in other fields, including those of machine learning and statistical analysis.

Acknowledgments

REFERENCES

Anthony, M.L., Rose, V.S., Nicholson, J.K. and Lindon, J.C. (1995). Classification of toxin-induced changes in 1H NMR spectra of urine using an artificial neural network. *Pharmaceutical and Biomedical Analysis* **13**, 205.

Beckonert, O., Bollard, E., Ebbels, T.M.D., Keun, H.C., Antti, H., Holmes, E., Lindon, J.C. and Nicholson, J.K. (2003). NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. *Analytica Chimica Acta* **490**, 3.

Breiman, L. and Friedman, J.H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society, Series B (Methodological)* **59**, 3–54.

Brown, P.J., Fearn, T. and Vannucci, M. (2001). Bayesian wavelet regression on curves with applications to a spectroscopic calibration problem. *Journal of the American Statistical Society* **96**, 398–408.

Burnham, A.J., MacGregor, J.F. and Viveros, R. (1999). A statistical framework for multivariate latent variable regression methods based on maximum likelihood. *Journal of Chemometrics* **13**, 49–65.

Butler, N.A. and Denham, M.C. (2000). The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society,Series B (Methodological)* **62**, 585–593.

Christoffersson, A. (1970). The one component model with incomplete data. Ph.D Thesis, Uppsala University.

Cloarec, O., Dumas, M.E., Craig, A., Barton, R.H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J.C., Holmes, E. and Nicholson, J., (2005). Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Analytical Chemistry* **77**, 1282–1289.

Cloarec, O., Dumas, M., Trygg, J., Craig, A., Barton, R., Lindon, J., Nicholson, J. and Holmes, E. (2005). Evaluation of the orthogonal projection to latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in 1 H NMR spectroscopic metabonomic studies. *Analytical Chemistry* **77**, 517–526.

Clyde, M.A., House, L.L. and Wolpert, R.L. (2006). Nonparametric models for proteomic peak identification and quantification. ISDS discussion papers 2006-07, Duke University.

Coen, M., Lenz, E.M., Nicholson, J.K., Wilson, I.D., Pognan, F. and Lindon, J.C. (2003). An integrated metabonomic investigation of acetaminophen toxicity in the mouse using NMR spectroscopy. *Chemical Research in Toxicology* **16**, 295–303.

Crockford, D.J., Holmes, E., Lindon, J.C., Plumb, R.S., Zirah, S., Bruce, S.J., Rainville, P., Stumpf, C.L. and Nicholson, J.K., (2006). Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: application in metabonomic toxicology studies. *Analytical Chemistry* **78**, 363–371.

Davies, T. (1998). The new automated mass spectrometry deconvolution and identification system (AMDIS). *Spectroscopy* **10**(3), 24–27.

Duda, R.O., Hart, P.E. and Stork, D.G. (2000). *Pattern Classification*. John Wiley & Sons, New York.

Dumas, M.E., Canlet, C., Andre, F., Vercauteren, J. and Paris, A. (2002). Metabonomic assessment of physiological disruptions using 1H-13C HMBC-NMR spectroscopy combined with pattern recognition procedures performed on filtered variables. *Analytic Chemistry* **74**(3), 2261–2273.

Ebbels, T.M.D., Keun, H.C., Beckonert, O., Antti, H., Bollard, M., Holmes, E., Lindon, J.C. and Nicholson, J.K. (2003). Toxicity classification from metabonomic data using a density superposition approach: 'clouds'. *Analytica Chimica Acta* **490**, 109.

Ebbels, T.M.D., Keun, H.C., Beckonert, O., Bollard, E., Lindon, J.C., Holmes, E. and Nicholson, J.K. (2006). Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data. in preparation.

El-Deredy, W. (1997). Pattern recognition approaches in biomedical and clinical magnetic resonance spectroscopy: a review. *NMR Biomedicine* **10**, 99–124.

Fogel, G.B. and Corne, D.W. (2003). An introduction to evolutionary computation for biologists. In *Evolutionary Computation in Bioinformatics*, G.B. Fogel and D.W. Corne, eds. Morgan Kaufmann Publishers, pp. 19–38.

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–148.

Geladi, P., MacDougall, D. and Martens, H. (1985). Linearization and scatter- correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy* **3**, 491–500.

Gilbert, R.J., Goodacre, R., Woodward, A.M. and Kell, D.B. (1997). Genetic programming: a novel method for the quantitative analysis of pyrolysis mass spectral data. *Analytical Chemistry* **69**, 4381–4389.

Goodacre, R. (2005). Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. *Journal of Experimental Botany* **56**, 245–254.

Goodacre, R., Shann, B., Gilbert, R.J., Timmins, E.M., McGovern, A.C., Alsberg, B.K., Kell, D.B. and Logan, N.A. (2000). Detection of the dipicolinic acid biomarker in bacillus spores using Curie-point pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. *Analytical Chemistry* **72**, 119–127.

Goodacre, R., York, E.V., Heald, J.K. and Scott, I.M. (2003). Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. *Phytochemistry* **62**, 859–863.

Gray, H.F., Maxwell, R.J., Martinez-Perez, I., Arus, C. and Cerdan, S. (1998). Genetic programming for classification and feature selection: analysis of 1H nuclear magnetic resonance spectra from human brain tumour biopsies. *NMR in Biomedicine* **11**, 217–224.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Holmes, E., Bonner, F.W., Sweatman, B.C., Lindon, J.C., Beddell, C.R., Rahr, E. and Nicholson, J.K. (1992). Nuclear-magnetic- resonance spectroscopy and pattern-recognition analysis of the biochemical processes associated with the progression of and recovery from nephrotoxic lesions in the rat induced by mercury(ii) chloride and 2-bromoethanamine. *Molecular Pharmacology* **42**, 922.

Holmes, E., Nicholson, J.K. and Tranter, G. (2001). Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks. *Chemical Research in Toxicology* **14**, 182.

Howells, S.L., Maxwell, R.J., Peet, A.C. and Griffiths, J.R. (1992). An investigation of tumor 1H nuclear magnetic resonance spectra by the application of chemometric techniques. *Magnetic Resonance Medicine* **28**, 214–236.

Johnson, H.E., Gilbert, R.J., Winson, M.K., Goodacre, R., Smith, A.R., Rowland, J.J., Hall, M.A. and Kell, D.B. (2000). Explanatory analysis of the metabolome using genetic programming of simple, interpretable rules. *Genetic Programming and Evolvable Machines* **1**, 243.

Jolliffe, I.T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

Kell, D.B., Darby, R.M. and Draper, J. (2001). Genomic computing. explanatory analysis of plant expression profiling data using machine learning. *Plant Physiology* **126**, 943–951.

Lindon, J.C., Holmes, J.L. and Tranter, G.E. (2007). *A Handbook of Metabonomics and Metabolomics*. Elsevier.

Lindon, J.C., Keun, H.C., Ebbels, T.M.D., Pearce, J.M., Holmes, E. and Nicholson, J.K. (2005). The consortium for metabonomic toxicology (COMET): aims, activities and achievements. *Pharmacogenomics* **6**, 691–699.

Martens, H. and Naes, T. (1989). *Multivariate Calibration*. John Wiley & Sons Inc, New York.

Martinez, I., Bathen, T., Standal, I.B., Halvorsen, J., Aursand, M., Gribbestad, I.S. and Axelson, D.E. (2005). Bioactive compounds in cod (*gadus morhua*) products and suitability of 1H NMR metabolite profiling for classification of the products using multivariate data analyses. *Journal of Agricultural and Food Chemistry* **53**, 6889–6895.

Massy, W.F. (1965). Principal component regression in exploratory statistical research. *Journal of the American Statistical Association* **60**, 234–246.

Morris, J., Brown, P., Baggerly, K. and Coombes, K. (2006). Chapter analysis of mass spectrometry data using Bayesian wavelet-based functional mixed models. In *Bayesian Inference for Gene Expression and Proteomics*, Cambridge University Press, pp. 269–292.

Nicholson, J.K., Connelly, J., Lindon, J.C. and Holmes, E. (2002). Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery* **1**, 153.

Nicholson, J.K., Lindon, J.C. and Holmes, E. (1999). Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **29**, 1181.

Nicholson, J.K. and Wilson, I.D. (2003). Understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nature Reviews Drug Discovery* **2**, 668.

Ott, K.H., Aranibar, N., Singh, B. and Stockton, G.W. (2003). Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts. *Phytochemistry* **62**, 971–985.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics* **33**, 1065.

Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M.C., Berden, J.A., Brindle, K.M., Kell, D.B., Rowland, J.J., Westerhoff, H.V., van Dam, K. and Oliver, S.G. (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology* **19**, 45–50.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Savitsky, A. and Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* **36**, 1627–1639.

Stein, S.E. (1999). An integrated method for spectrum extraction and compound identification from gc/ms data. *Journal of the American Society of Mass Spectrometry* **10**, 770–871.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). **36**, 111–147.

Stone, M. and Brooks, R.J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. **52**, 237–269.

Stoyanova, R., Nicholls, A.W., Nicholson, J.K., Lindon, J.C. and Brown, T.R. (2004). Automatic alignment of individual peaks in large high-resolution spectral data sets. *Journal of Magnetic Resonance* **170**, 329–335.

Taylor, J., Goodacre, R., Wade, W.G., Rowland, J.J. and Kell, D.B. (1998). The deconvolution of pyrolysis mass spectra using genetic programming: application to the identification of some eubacterium species. *FEMS Microbiology Letters* **160**, 237–246.

Trygg, J. and Wold, S. (2002). Orthogonal projections to latent structures (O- PLS). *Journal of Chemometrics* **16**, 119–128.

Trygg, J. and Wold, S. (2003). O2-pls, a two-block (x-y) latent variable regression (LVR) method with an integral osc filter. *Journal of Chemometrics* **17**, 53–64.

West, M. (2002). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics 7*, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, eds. University Press, Oxford, pp. 723–732.

Wilson, I.D., Plumb, R., Granger, J., Major, H., Williams, R. and Lenz, E.M. (2005). Hplc-ms-based methods for the study of metabonomics. *Chromatography B, Analytical Technologies in the Biomedical and Life Sciences* **817**, 67–76.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis, Proceedings of the International Symposium, June 1965*, P.R. Krishnaiah, ed. Academic Press, New York, pp. 391–420.

Wold, H. (1975). Soft modelling by latent variables; the non-linear iterative partial least squares approach. In *Perspectives in Probability ans Statistics, Papers in Honour of M.S. Bartlett*, J. Gani, ed. Academic Press, London.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **20**, 397–404.

Woodward, A.M., Gilbert, R.J. and Kell, D.B. (1999). Genetic programming as an analytical tool for non-linear dielectric spectroscopy. *Bioelectrochemistry and Bioenergetics* **48**, 389.

Part 3

Evolutionary Genetics

12

Adaptive Molecular Evolution

Z. Yang

Department of Biology, University College London, London, UK

This chapter reviews statistical methods for detecting adaptive molecular evolution by comparing synonymous and nonsynonymous substitution rates in protein-coding DNA sequences. A Markov process model of codon substitution, which forms the basis for all later discussions in this chapter, is introduced first. The case of comparing two sequences to estimate the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site is then considered. The maximum likelihood (ML) method and a number of *ad hoc* counting methods are evaluated. The rest of the chapter deals with joint analyses of multiple sequences on a phylogeny. Review is provided on Markov models of codon substitution that allow the nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$) to vary among branches in a phylogeny or among amino acid sites in a protein. Those models can be used to construct likelihood ratio tests (LRTs) to identify evolutionary lineages under episodic Darwinian selection or to infer critical amino acids in a protein under diversifying selection. Real-data examples are used to demonstrate the application of the methods. The chapter finishes with a discussion of the limitations of current methods.

12.1 INTRODUCTION

While biologists accept Darwin's theory of evolution by natural selection for morphological traits, the importance of selection in molecular evolution has been much debated. The neutral theory (Kimura, 1983) maintains that most observed molecular variation (both diversity within species and divergence between species) is due to random fixation of mutations with fitness effects so small that random drift rather than natural selection dominates their fate. Population geneticists have developed a number of tests of neutrality (for reviews, see Kreitman and Akashi, 1995 and **Chapter 22**). Those tests have been applied to identify genes under positive selection from genome-wide analysis of within-species polymorphism (Fay *et al.*, 2001; Smith and Eyre-Walker, 2002). However, they seldom provide unequivocal evidence for positive selection as it is often difficult to distinguish natural selection from demographic processes.

Another class of methods designed to detect adaptive molecular evolution relies on comparison of synonymous (silent) and nonsynonymous (amino acid−changing)

substitution rates in protein-coding genes. The synonymous and nonsynonymous rates (d_S and d_N, or K_s and K_a by some authors) are defined as the numbers of synonymous and nonsynonymous substitutions per site, respectively. The ratio of the two rates, $\omega = d_N/d_S$, then measures selective pressure at the protein level. If selection has no effect on fitness, nonsynonymous mutations will be fixed at the same rate as synonymous mutations, so that $d_N = d_S$ and $\omega = 1$. If nonsynonymous mutations are deleterious, purifying selection will reduce their fixation rate, so that $d_N < d_S$ and $\omega < 1$. If nonsynonymous mutations are favored by Darwinian selection, they will be fixed at a higher rate than synonymous mutations, resulting in $d_N > d_S$ and $\omega > 1$. A significantly higher nonsynonymous rate than the synonymous rate is thus evidence for adaptive evolution at the molecular level. This criterion has been used to identify many cases of positive selection, including the human major histocompatibility complex (MHC, Hughes and Nei, 1988), primate stomach lysozyme (Messier and Stewart, 1997), abalone sperm lysin (Lee *et al.*, 1995), vertebrate visual pigments (Miyamoto and Miyamoto, 1996), and HIV-1 *env* genes (Bonhoeffer *et al.*, 1995; Mindell, 1996; Yamaguchi and Gojobori, 1997). Development of powerful methods, such as those reviewed in this chapter, has led to identification of many more cases of molecular adaptation (Yang and Bielawski, 2000; Yang, 2006: Chapter 8) providing important insights into the mechanisms of molecular evolution.

The ω ratio has most often been calculated as an average over all codons (amino acids) in the gene and over the entire evolutionary time that separates the sequences. The criterion that such an average ω is greater than 1 is a very stringent one for detecting positive selection (e.g. Kreitman and Akashi, 1995). Many amino acids in a protein must be under strong functional constraints, with ω close to 0. Many proteins also appear to be under purifying selection during most of the evolutionary history. Adaptive evolution most likely occurs at a few time points and affects only a few amino acids (Gillespie, 1991). In such a case, the ω ratio averaged over time and over sites will not be greater than 1 even if Darwinian selection has been operating.

A remedy for this problem is to examine the ω ratio over a short evolutionary time period or in a short stretch of the gene such as functionally important domains. For example, Messier and Stewart (1997) (see also Zhang *et al.*, 1997) used inferred ancestral genes to calculate d_N and d_S for each branch in the tree and identified two lineages in the lysozyme phylogeny for primates that went through positive selection. In a similar vein, Hughes and Nei (1988) found that $d_N > d_S$ at 57 amino acids in the MHC that constitute the antigen-recognition site, although $d_N < d_S$ in the whole gene. Those ideas have also been implemented in the likelihood framework. Markov process models of codon substitution have been developed that account for different ω ratios among branches in the tree (Yang, 1998; Yang and Nielsen, 1998). They can be used to construct likelihood ratio tests (LRTs) of adaptive evolution along specific lineages, and have the advantage of not relying on inferred ancestral sequences. Models have also been developed that allow the ω ratio to vary among amino acid sites (Nielsen and Yang, 1998; Yang *et al.*, 2000). They do not require knowledge of functionally important domains and may be used to test for the presence of critical amino acids under positive selection, and, when they exist, to identify them.

This chapter reviews statistical methods for phylogenetic analysis of protein-coding DNA sequences, with a focus on comparing synonymous and nonsynonymous substitution rates to understand the mechanisms of protein sequence evolution. At first, a brief

introduction to the probability theory of Markov process of codon substitution is provided. This theory forms the basis for maximum likelihood (ML) estimation of d_N and d_S between two sequences as well as ML joint analysis of multiple sequences on a phylogeny. Different methods for comparing two sequences to estimate d_N and d_S are discussed. Besides ML (Goldman and Yang, 1994), there are about a dozen so-called counting methods for this estimation (e.g. Miyata and Yasunaga, 1980; Nei and Gojobori, 1986; Li *et al.*, 1985; Li, 1993; Ina, 1995; Yang and Nielsen, 2000). These will be evaluated. Models that account for variable ω ratios among lineages and among sites are then discussed with real-data examples to explain their use in ML analysis. This chapter uses ML as the general framework. ML is known to have nice statistical properties, and indeed offers insights into heuristic methods as well, which may not be based on an explicit probabilistic model. A brief introduction to ML estimation and LRT is provided in the Section 12.3.2. For a detailed and rigorous treatment of likelihood methods, the reader should consult a statistics textbook, such as Edwards (1992), Kalbfleisch (1985), and Stuart *et al.* (1999).

12.2 MARKOV MODEL OF CODON SUBSTITUTION

In molecular phylogenetics, we use a continuous-time Markov process to describe the change between nucleotides, amino acids, or codons over evolutionary time. See Whelan *et al.* (2001) or the chapters by Huelsenbeck and Bollback, and Thorne and Goldman (**Chapter 14** and **Chapter 15**) for use of Markov processes to model nucleotide or amino acid substitution. In this chapter, our focus is analysis of protein-coding DNA sequences, and the unit of evolution is a codon in the gene. We use a Markov process to describe substitutions between the sense codons. We exclude stop codons as they are usually not allowed in a protein. With the 'universal' genetic code, there are 61 sense codons and thus 61 states in the Markov process.

The Markov process is characterized by a rate (generator) matrix $Q = \{q_{ij}\}$, where q_{ij} is the substitution rate from sense codon i to sense codon $j (i \neq j)$. Formally, $q_{ij}\Delta t$ is the probability that the process is in state j after an infinitesimal time Δt, given that it is currently in state i. The basic model we use in this chapter is simpler than the model of Goldman and Yang (1994) but more complex than that of Muse and Gaut (1994). It accounts for the transition–transversion rate difference, unequal synonymous and nonsynonymous substitution rates, and unequal base/codon frequencies. Mutations are assumed to occur independently among the three codon positions, and so only one position is allowed to change instantaneously. Since transitions (changes between T and C, and between A and G) are known to occur more frequently than transversions (all other changes), we multiply the rate by κ if the change is a transition. Parameter κ is the transition–transversion rate ratio. Typical estimates of κ are 1.5–5 for nuclear genes and 3–30 for mitochondrial genes. To account for unequal codon frequencies, we let π_j be the equilibrium frequency of codon j and multiply substitution rates to codon j by π_j. We can either use all π_j's as parameters, with $60 (= 61 - 1)$ free parameters used, or calculate π_j from base frequencies at the three codon positions, with $9 = 3 \times (4 - 1)$ free parameters used.

To account for unequal synonymous and nonsynonymous substitution rates, we multiply the rate by ω if the change is nonsynonymous; ω is thus the nonsynonymous/synonymous

rate ratio, also termed the 'acceptance rate' by Miyata *et al.* (1979). In models considered here, the relationship holds that $\omega = d_N/d_S$. For most genes, estimates of ω are much less than 1. Parameters κ and π_j characterize processes at the DNA level, including natural selection, while selection at the protein level has the effect of modifying parameter ω. If natural selection operates on the DNA as well as on the protein, the synonymous rate will differ from the mutation rate.

Thus, the relative substitution rate from codon i to codon $j (i \neq j)$ is

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition.} \end{cases} \qquad (12.1)$$

For example, consider substitution rates to codon CTG (which encodes amino acid Leu). We have $q_{\text{CTC,CTG}} = \pi_{\text{CTG}}$ since the CTC (Leu) $\rightarrow$ CTG (Leu) change is a synonymous transversion, $q_{\text{TTG,CTG}} = \kappa\pi_{\text{CTG}}$ since the TTG (Leu) $\rightarrow$ CTG (Leu) change is a synonymous transition, $q_{\text{GTG,CTG}} = \omega\pi_{\text{CTG}}$ since the GTG (Val) $\rightarrow$ CTG (Leu) change is a nonsynonymous transversion, and $q_{\text{CCG,CTG}} = \kappa\omega\pi_{\text{CTG}}$ since the CCG (Pro) $\rightarrow$ CTG (Leu) change is a nonsynonymous transition. Also $q_{\text{TTT,CTG}} = 0$ since codons TTT and CTG differ at two positions.

The diagonal elements of the rate matrix $Q = \{q_{ij}\}$ are determined by the requirement that each row in the matrix sums to 0

$$\sum_j q_{ij} = 0, \text{for any } i, \qquad (12.2)$$

(e.g. Grimmett and Stirzaker, 1992, p. 241). Furthermore, molecular sequence data do not allow separate estimation of rate and time, and only their product can be identified. We thus multiply matrix Q by a constant so that the expected number of nucleotide substitutions per codon is 1:

$$-\sum_i \pi_i q_{ii} = \sum_i \pi_i \sum_{j \neq i} q_{ij} = 1. \qquad (12.3)$$

This scaling means that time t is measured by distance, the expected number of (nucleotide) substitutions per codon. The transition probability matrix over time t is

$$P(t) = \{p_{ij}(t)\} = e^{Qt}, \qquad (12.4)$$

where $p_{ij}(t)$ is the probability that codon i will become codon j after time t. As long as the rate matrix Q can be constructed, $P(t)$ can be calculated for any t using matrix diagonalization. Note that over any time interval, there is a nonzero probability that any codon i will change to any other codon j, even if they are separated by two or three nucleotide differences; that is, for any $t > 0$, $p_{ij}(t) > 0$ for any codons i and j.

Lastly, the model specified by (12.1) is time reversible; that is, $\pi_i q_{ij} = \pi_j q_{ji}$ for any i and j. This means that

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \text{for any } t, i, \text{ and } j. \qquad (12.5)$$

Note that $\pi_i p_{ij}(t)$ measures the amount of change from codons i to j over time t, while $\pi_j p_{ji}(t)$ measures the change in the opposite direction. Equation (12.5), known as the 'detailed balance', means that we expect to see equal numbers of changes from i to j and from j to i. Some implications of reversibility are mentioned in later sections.

12.3 ESTIMATION OF SYNONYMOUS (d_S) AND NONSYNONYMOUS (d_N) SUBSTITUTION RATES BETWEEN TWO SEQUENCES

12.3.1 Counting Methods

We want to estimate the number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) between two protein-coding DNA sequences. In the past two decades, about a dozen methods have been proposed for this estimation. They are intuitive and involve treatment of the data that cannot be justified rigorously. Important basic concepts were developed in the early 1980s (Miyata and Yasunaga, 1980; Perler *et al.*, 1980; Gojobori, 1983; Li *et al.*, 1985), which we explain here with a hypothetical example. The critical question is by how much natural selection at the protein level has increased or decreased the nonsynonymous substitution rate. Suppose the gene has 300 codons and we observe five synonymous and five nonsynonymous differences (substitutions) between the two sequences. Can we conclude that synonymous and nonsynonymous substitution rates are equal with $\omega = 1$? The answer is 'No'. An inspection of the genetic code table suggests that all changes at the second codon position and most changes at the first position are nonsynonymous, and only some changes at the third position are synonymous. As a result, we do not expect to see equal proportions of synonymous and nonsynonymous mutations even if there is no selection at the protein level. Indeed, if mutations from any one nucleotide to any other occur at the same rate, we expect 25.5 % of mutations to be synonymous and 74.5 % to be nonsynonymous (Yang and Nielsen, 1998). If we use those proportions, it is clear that selection on the protein has decreased the fixation rate of nonsynonymous mutations by about three times, since $\omega = (5/5)/(74.5/25.5) = 0.34$. There are 900 nucleotide sites in the sequence, so the numbers of synonymous and nonsynonymous sites are $S = 900 \times 25.5\% = 229.5$ and $N = 900 \times 74.5\% = 670.5$, respectively. We then have $d_S = 5/229.5 = 0.0218$ and $d_N = 5/670.5 = 0.0075$.

All counting methods roughly follow the above intuitive procedure (see Ina, 1996; Yang and Nielsen, 2000 for reviews). They involve three steps. The first step is to count the numbers of synonymous (S) and nonsynonymous (N) *sites* in the two sequences; that is, the number of nucleotide sites in the sequence is classified into the synonymous and nonsynonymous categories, measuring substitution opportunities before the operation of selection on the protein. This step is complicated by factors such as transition–transversion rate difference and unequal base/codon frequencies, both of which are ignored in our hypothetical example. The second step is to count the numbers of synonymous and nonsynonymous *differences* between the two sequences; that is, the observed differences between the two sequences are classified into the synonymous and nonsynonymous categories. This is straightforward if the two compared codons differ at one codon position only. When they differ at two or three codon positions, there exist four or

six pathways from one codon to the other. The multiple pathways may involve different numbers of synonymous and nonsynonymous differences and should ideally be weighted appropriately according to their likelihood of occurrence, although most counting methods use equal weighting. The third step is to apply a correction for multiple substitutions at the same site since an observed difference may be the result of two or more substitutions. In our hypothetical example, we ignored the possibility of multiple hits and treated the observed differences as substitutions. All counting methods have used multiple-hit correction formulas based on nucleotide-substitution models, which assume that each nucleotide can change to one of three other nucleotides. When those formulas are applied to synonymous (or nonsynonymous) sites only, this basic assumption of the Markov model is violated (Lewontin, 1989). Nevertheless, such corrections appear usable when the sequence divergence is low.

The method of Miyata and Yasunaga (1980) and its simplified version of Nei and Gojobori (1986) are based on the nucleotide-substitution model of Jukes and Cantor (1969), and ignore transition–transversion rate difference or unequal base/codon frequencies. As transitions are more likely to be synonymous at the third positions than transversions are, ignoring the transition–transversion rate difference leads to underestimation of S and overestimation of N. This effect is well known, and a number of attempts have been made to account for different transition and transversion rates in counting sites and differences (Li *et al.*, 1985; Li, 1993; Pamilo and Bianchi, 1993; Comeron, 1995; Ina, 1995). The effect of unequal base/codon frequencies was not so much appreciated (but see Moriyama and Powell, 1997). However, Yang and Nielsen (1998; 2000) found that extremely biased base/codon frequencies can have devastating effects on the estimation of d_N and d_S, often outweighing the effect of the transition–transversion rate difference. A counting method taking into account both factors was implemented by Yang and Nielsen (2000).

12.3.2 Maximum Likelihood Estimation

Likelihood is a powerful and flexible methodology for estimating parameters and testing hypotheses. Since the data are observed, we view the probability of observing the data as a function (the likelihood function) of the unknown parameters. The likelihood or log-likelihood function is our inference tool and contains all information about the parameters in the model. We estimate the unknown parameters by maximizing the likelihood function. Furthermore, the log-likelihood value under a model measures the fit of the model to data, and we compare two models by comparing their log-likelihood values. This is known as the LRT. When two models are nested, twice the log-likelihood difference between the two models can be compared with the χ^2 distribution, with the difference in the number of parameters between the two models as the degrees of freedom (df). The χ^2 approximation to the likelihood ratio statistic relies on large sample sizes (long sequences). How large the sample should be for the χ^2 approximation to be reliable depends on the specific model being tested as well as other factors such as sequence divergence. In a few cases where LRTs applied to phylogenetics were examined by computer simulation, the χ^2 approximation appears very good with as few as 100 or 200 nucleotides in the sequence (e.g. Nekrutenko *et al.*, 2001). When the sequences are too short or when the two models are not nested, the correct distribution of the test statistic can be derived by Monte Carlo simulation (Goldman, 1993).

The ML method for estimating d_N and d_S (Goldman and Yang, 1994) is described below. The data are two aligned protein-coding DNA sequences. As a numerical example, we will use the human and mouse acetylcholine receptor α genes. The first 15 codons of the gene are shown below.

```
Human GAG CCC TGG CCT CTC CTC CTG CTC TTT AGC CTT TGC TCA GCT GGC ...
Mouse GAG CTC TCG ACT GTT CTC CTG CTG CTA GGC CTC TGC TCC GCT GGC ...
```

We assume that different codons in the sequence are evolving independently according to the same Markov process. As a result, data at different sites are independently and identically distributed. Suppose there are n sites (codons) in the gene, and let the data at site h be $\mathbf{x}_h = \{x_1, x_2\}$, where x_1 and x_2 are the two codons in the two sequences at that site (see Figure 12.1a). In the above example, the data at site $h = 2$ are $x_1 = $ CCC and $x_2 = $ CTC. The probability of observing data $\mathbf{x}_h$ at site h is

$$f(\mathbf{x}_h) = \sum_{k=1}^{61} \pi_k p_{kx_1}(t_1) p_{kx_2}(t_2). \tag{12.6}$$

The term in the sum is the probability that the ancestor has codon k and the two current species have codons x_1 and x_2 at the site. This probability is equal to the prior probability that the ancestor has codon k, given by the equilibrium frequency π_k, multiplied by the two transition probabilities along the two branches of the tree (Figure 12.1a). Since the ancestral codon k is unknown, we sum over all possibilities for k. Time reversibility of the Markov process implies that

$$f(\mathbf{x}_h) = \sum_{k=1}^{61} \pi_{x_1} p_{x_1 k}(t_1) p_{kx_2}(t_2) = \pi_{x_1} \sum_{k=1}^{61} p_{x_1 k}(t_1) p_{kx_2}(t_2) = \pi_{x_1} p_{x_1 x_2}(t_1 + t_2). \tag{12.7}$$

The last step follows from the Chapman–Kolmogorov theorem (e.g. Grimmett and Stirzaker, 1992, pp. 239–246). Thus the data are probabilistically identical whether we consider the two sequences to be descendants of a common ancestor (as in Figure 12.1a) or we consider one sequence to be ancestral to the other (as in Figure 12.1b). In other words, the root of the tree cannot be identified and only $t = t_1 + t_2$ can be estimated, but not t_1 and t_2 individually. Parameters in the model are thus the sequence divergence t, the transition–transversion rate ratio κ, the nonsynonymous/synonymous rate ratio ω, and the codon frequencies π_j's. The log-likelihood function is then given by

Figure 12.1 The tree for two sequences, with the codons x_1, x_2 for one site shown. Codon-substitution models considered in this chapter are all time reversible and do not allow identification of the root. As a result, parameters t_1 and t_2 cannot be estimated separately (a), and only their sum $t = t_1 + t_2$ is estimable (b).
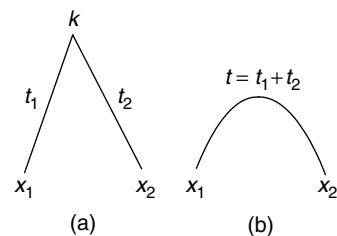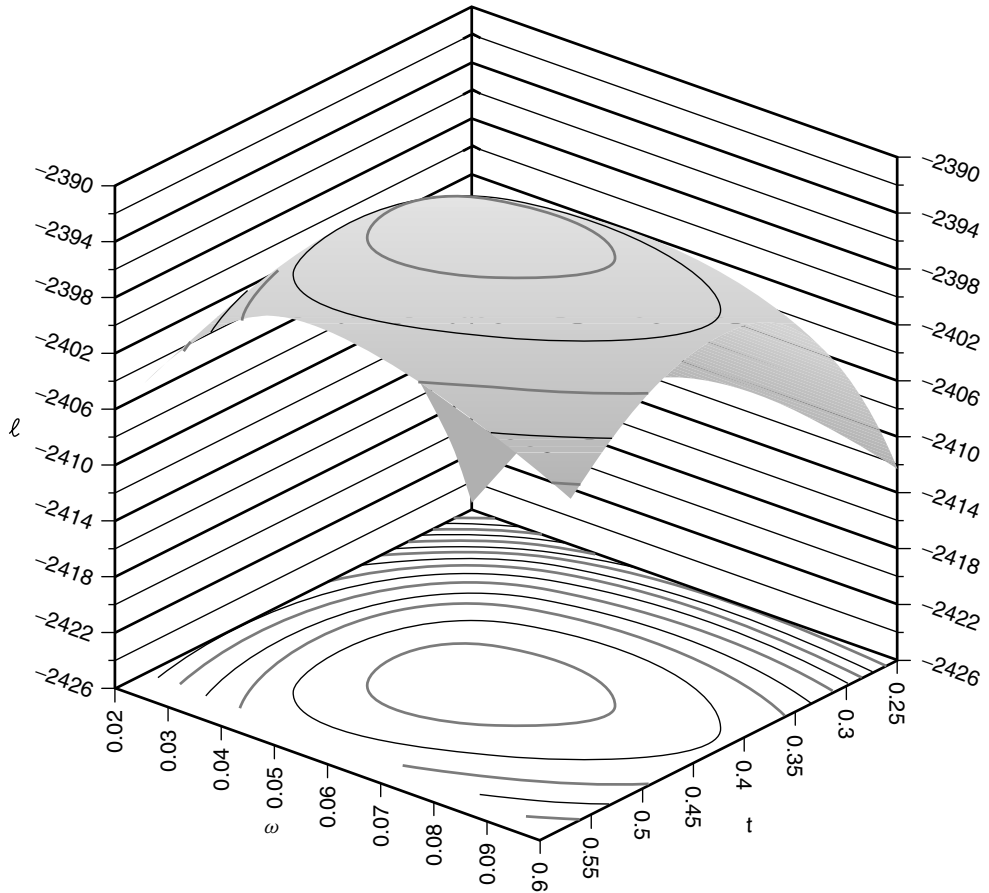
Figure 12.2 The log-likelihood surface contour as a function of parameters t and ω for the comparison of the human and mouse acetylcholine receptor α genes. The maximum likelihood method estimates parameters by maximizing the likelihood function. For these data, the estimates are $t = 0.444$, $\omega = 0.059$, with the optimum log likelihood to be $\ell = -2392.83$.

$$\ell(t, \kappa, \omega) = \sum_{h=1}^{n} \log\{f(\mathbf{x}_h)\}. \tag{12.8}$$

If some sites have the same data $\mathbf{x}$, the probability $f(\mathbf{x})$ need be calculated only once. An equivalent way of deriving the likelihood function is to note that the data follow a multinomial distribution with 61^2 categories corresponding to the 61^2 possible site patterns (configurations).

We usually estimate the codon frequencies (π_j's) by the observed base/codon frequencies. To estimate parameters t, κ, and ω, we use a numerical hill-climbing algorithm to maximize ℓ, since an analytical solution is impossible. Figure 12.2 shows a log-likelihood surface as a function of t and ω for the human and mouse acetylcholine receptor α genes. The model assumes equal transition and transversion rates and equal codon frequencies (with $\kappa = 1$ and $\pi_j = 1/61$ fixed), and involve two parameters only. This is the

model underlying the method of Miyata and Yasunaga (1980) and Nei and Gojobori (1986).

The d_N and d_S rates are defined as functions of parameters t, κ, ω, and π_j, and their ML estimates are simply functions of ML estimates of parameters t, κ, ω, and π_j. The following description thus gives both the definitions of d_N and d_S, and also the ML method for their estimation. The basic idea is the same as explained in our hypothetical example before. Here we count sites and substitutions per codon rather than for the entire sequence. First, note that the sequence divergence t is defined as the number of nucleotide substitutions per codon. We partition this number into the synonymous and nonsynonymous categories. We note that

$$\rho_S^* = \sum_{\substack{i \neq j \\ aa_i = aa_j}} \pi_i q_{ij} \tag{12.9}$$

and

$$\rho_N^* = \sum_{\substack{i \neq j \\ aa_i \neq aa_j}} \pi_i q_{ij} \tag{12.10}$$

are the proportions of synonymous and nonsynonymous substitutions, respectively, as $\rho_S^* + \rho_N^* = 1$ (12.3). The summation in ρ_S^* is taken over all codon pairs i and $j (i \neq j)$ that code for the same amino acid, while the summation in ρ_N^* is taken over all codon pairs i and j $(i \neq j)$ that code for different amino acids; aa_i is the amino acid encoded by codon i. The numbers of synonymous and nonsynonymous substitutions per codon are then $t\rho_S^*$ and $t\rho_N^*$, respectively.

Next, we calculate the proportions of synonymous and nonsynonymous *sites*. Let these be ρ_S^1 and ρ_N^1. As noted before, these measure the substitution opportunities before the operation of selection at the protein level, that is, when $\omega = 1$ (Goldman and Yang, 1994; Ina, 1995). They are calculated similar to (12.9) and (12.10), using the transition–transversion rate ratio κ and codon frequencies (π_j), except that $\omega = 1$ is fixed. We assume there are three nucleotide sites in a codon (see Yang and Nielsen, 1998 for a discussion of the effect of mutations to stop codons). The numbers of synonymous and nonsynonymous sites per codon are then $3\rho_S^1$ and $3\rho_N^1$, respectively. The numbers of synonymous and nonsynonymous substitutions per site are then $d_S = t\rho_S^*/(3\rho_S^1)$ and $d_N = t\rho_N^*/(3\rho_N^1)$, respectively. Note that $\omega = d_N/d_S = (\rho_N^*/\rho_S^*)/(\rho_N^1/\rho_S^1)$, where the numerator ρ_N^*/ρ_S^* is the ratio of the numbers of (observed) substitutions while the denominator ρ_N^1/ρ_S^1 is the ratio of the (expected) numbers of substitutions if $\omega = 1$.

Interpretation of d_N and d_S and definitions of a few other distances between two protein-coding genes were given by Yang (2006, Chapter 2). While the basic concepts discussed in the hypothetical example underlie both the ML and the counting methods for estimating d_N and d_S (and their ratio ω), differences exist between the two classes of methods. In the ML method, the probability theory (i.e. calculation of the transition probabilities by (12.4)) accomplishes several difficult tasks in one step: estimating mutational parameters such as κ, correcting for multiple hits, and weighting evolutionary pathways between codons. The Chapman–Kolmogorov theorem mentioned above ensures that the likelihood calculation accounts for all possible

pathways of change between two codons, weighting them appropriately according to their relative probabilities of occurrence. When we partition the number of substitutions (t) into synonymous and nonsynonymous categories, we only need to do it at the level of instantaneous rates (12.9 and 12.10), where there are no multiple changes.

In the counting methods, each of the three steps offers a challenge. For example, some methods ignore the transition–transversion rate difference. Others take it into account but it has been difficult to estimate κ reliably. Ina (1995) used the third codon positions and Yang and Nielsen (2000) used so-called four-fold degenerate sites and nondegenerate sites to estimate κ, assuming that substitutions at those sites are either not affected or affected equally by selection at the protein level. Both methods use nucleotide-based correction formulas to estimate κ, which seem problematic. Use of a limited class of sites also leads to large sampling errors in the estimates. The steps of counting differences, weighting pathways, and correcting for multiple hits are extremely complicated when we want to incorporate major features of DNA sequence evolution such as the transition–transversion rate difference and unequal base/codon frequencies (Yang and Nielsen, 2000). Notably, the synonymous and nonsynonymous status of a site changes over time and also with the nucleotides at other positions of the codon (Muse, 1996). As a result, nucleotide-substitution models used in counting methods are not capable of dealing with the complexity of the codon-substitution process.

After d_N and d_S are estimated, statistical tests can be used to test whether d_N is significantly higher than d_S. For the counting methods, a normal approximation is applied to the statistic ($d_N - d_S$). For ML, an LRT can be used, which compares the null model with ω fixed at 1 and the alternative model that does not place this constraint. Twice the log-likelihood difference between the two models is compared with a χ^2 distribution with 1 degree of freedom to test whether ω is different from 1. In practice, such tests rarely detect positive selection, as d_N and d_S are calculated as averages over the entire sequence. Indeed, an interesting use of this LRT is to predict protein-coding potentials of genomic regions, making use of the fact that in almost all genes, d_N is significantly lower than d_S (Nekrutenko *et al.*, 2001).

12.3.3   A Numerical Example and Comparison of Methods

To see the differences among methods for estimating d_N and d_S, we compare the human and mouse acetylcholine receptor α genes, using ML as well as several counting methods (Table 12.1). The data set is the first (alphabetically) of the 49 genes analyzed by Ohta (1995). The sequence has 456 codons (1368 nucleotides) after the start and stop codons are removed. With the ML method, we examine the effects of model assumptions. Some models ignore the transition–transversion rate ratio (with $\kappa = 1$ fixed), while others account for it (with κ estimated). Some ignore biased codon frequencies (Fequal), while others account for it to some extent (F1 × 4, F3 × 4, and F61) (see legend to Table 12.1 for definitions of these models).

Most of these models are nested, and the χ^2 approximation can be used to perform LRTs. For example, we can compare models A and B in Table 12.1 to test whether the transition and transversion rates are equal. Model A is the null hypothesis and assumes

Table 12.1 Estimation of d_N and d_S between the human and mouse acetylcholine receptor α genes.

Model	$\hat{\kappa}$	$\hat{S}$	$\hat{d}_N$	$\hat{d}_S$	$\hat{d}_N/\hat{d}_S(\hat{\omega})$	ℓ
Counting Methods						
Nei and Gojobori (1986)	1	321.2	0.030	0.523	0.058	
Li (1993)	N/A	N/A	0.029	0.419	0.069	
Ina (1995)	6.1	408.4	0.033	0.405	0.081	
YN00 (Yang and Nielsen, 2000)	2.1	311.2	0.029	0.643	0.045	
ML methods						
(A) Fequal, $\kappa = 1$	1	348.5	0.029	0.496	0.059	−2392.83
(B) Fequal, κ estimated	2.8	396.7	0.031	0.421	0.073	−2379.60
(C) F1 × 4, $\kappa = 1$, fixed	1	361.0	0.029	0.513	0.057	−2390.35
(D) F1 × 4, κ estimated	2.9	406.5	0.031	0.436	0.071	−2376.12
(E) F3 × 4, $\kappa = 1$, fixed	1	281.4	0.029	0.650	0.044	−2317.72
(F) F3 × 4, κ estimated	3.0	328.1	0.030	0.545	0.055	−2303.33
(G) F61, $\kappa = 1$, fixed	1	261.5	0.028	0.736	0.038	−2251.92
(H) F61, κ estimated	3.0	319.5	0.030	0.613	0.048	−2239.33

Note: Fequal: equal codon frequencies (=1/61) are assumed; F1 × 4: four nucleotide frequencies are used to calculate codon frequencies (three free parameters); F3 × 4: nucleotide frequencies at three codon positions are used to calculate codon frequencies (nine free parameters); F61: all codon frequencies are used as free parameters (60 free parameters). ℓ is the log-likelihood value. [Data are from Ohta (1995) and Yang and Nielsen (1998).]

that transition and transversion rates are equal ($\kappa = 1$). Model B does not impose this constraint and has one more free parameter (κ) than model A. The likelihood ratio statistic, $2\Delta\ell = 2 \times (-2379.60 - (-2392.83)) = 2 \times 13.23 = 26.46$, should be compared with the χ^2 distribution with df = 1, giving a p value of 0.27×10^{-6}. So there is significant difference between the transition and transversion rates.

For these data, both the transition–transversion rate difference and unequal codon frequencies are clearly important. ML results under the most complex model (F61 with κ estimated), which accounts for both factors, are expected to be the most reliable and will be used to evaluate other methods/models. The F3 × 4 model is commonly used as it produces similar results to, and has far fewer parameters than, the F61 model. We note that counting methods give similar results to ML under similar models; for example, Ina's method gives similar estimates to ML accounting for the transition–transversion rate difference and ignoring biased base/codon frequencies (Table 12.1: MLB, Fequal, with κ estimated).

It is well known that ignoring the transition–transversion rate difference leads to underestimation of the number of synonymous sites (S), overestimation of d_S, and underestimation of the ω ratio. This effect is obvious in Table 12.1, when ML estimates with κ estimated are compared with those when κ is fixed at 1, or when the method of Nei and Gojobori (1986) is compared with those of Li (1993) or Ina (1995). Unequal codon frequencies often have opposite effects to the transition–transversion rate difference and lead to a reduced number of synonymous sites. This is the pattern we see in Table 12.1, as estimates of S under the F3 × 4 and F61 models are much smaller than under the Fequal model. The gene is GC-rich at the third codon position, with base frequencies to be 16 % for T, 43 % for C, 14 % for A, and 27 % for G. As a result, most substitutions

at the third codon position are transversions between C and G, and there are more nonsynonymous sites than expected under equal base/codon frequencies. In this data set, the effect of unequal base frequencies is opposite to and outweighs the effect of the transition–transversion rate difference. As a result, the method of Nei and Gojobori (1986), contrary to general belief, *overestimates*, rather than *underestimates*, S and ω. The method of Ina (1995) accounts for the transition–transversion rate difference but ignores the codon frequency bias, and performs more poorly than the method of Nei and Gojobori (1986). The method of Yang and Nielsen (2000) accounts for both biases, and seems to produce estimates close to ML estimates under realistic models.

In general, different methods can produce either very similar or very different estimates of ω. With very little transition–transversion rate difference and little codon usage bias, different methods tend to produce similar results. For other data sets, estimates from different methods can be three to ten times different (Yang and Nielsen, 2000; Dunn *et al.*, 2001). Such large differences can occur even with highly similar sequences, as extreme transition–transversion rate difference or codon usage bias can drastically affect the counting of sites. One feature of the estimation is that when a method overestimates d_S, it tends to underestimate d_N at the same time, resulting in large errors in the ω ratio. This is because the total number of sites (or differences) is fixed, and if the method underestimates the number of synonymous sites (or differences), it will overestimate the number of nonsynonymous sites as well, and vice versa.

Simulation studies performed to compare different methods produced results that are consistent with real-data analysis. For example, Ina (1995) compared several counting methods and concluded that none of them performed well when base frequencies were extreme. Yang and Nielsen (2000) examined the effects of the transition–transversion rate difference and unequal base/codon frequencies, and found that estimation of the ω ratio is very sensitive to both factors. A worrying result is that the method of Nei and Gojobori (1986) can both underestimate and overestimate the ω ratio, often with large biases. In general, counting methods may be used for exploratory data analysis, and the ML method accounting for both the transition–transversion rate difference and the codon usage bias should be preferred.

12.4   LIKELIHOOD CALCULATION ON A PHYLOGENY

Likelihood calculation for multiple sequences on a phylogeny may be viewed as an extension of the calculation for two sequences. The calculation is also similar to that under a nucleotide-substitution model (Felsenstein, 1981), although we now consider a codon rather than a nucleotide as the unit of evolution. We assume in this section that the same rate matrix Q (12.1) applies to all lineages and all amino acid sites. The data are multiple aligned sequences. We assume independent substitutions among sites (codons), so that data at different codon sites are independently and identically distributed. The likelihood is given by the multinomial distribution with 61^s categories (site patterns) for s species. Let n be the number of sites (codons) in the sequence and the data at site h be $\mathbf{x}_h (h = 1, 2, \ldots, n)$; x_h is a vector of observed codons in different sequences at site h. An
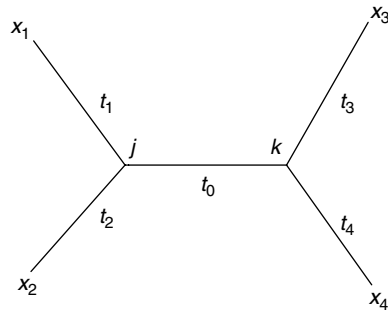
Figure 12.3 A tree of four sequences with codons at one site for nodes in the tree. Branch lengths t_0-t_4 are parameters in the model.

example tree of four species is shown in Figure 12.3. As in the case of two sequences, the root cannot be identified, and is arbitrarily fixed at the node ancestral to sequences 1 and 2. The data $\mathbf{x}_h$ can be generated by any codons j and k for the two ancestral nodes in the tree, and thus the probability of observing the data is a sum over all such possibilities.

$$f(\mathbf{x}_h) = \sum_j \sum_k [\pi_j p_{jx_1}(t_1) p_{jx_2}(t_2) p_{jk}(t_0) p_{kx_3}(t_3) p_{kx_4}(t_4)]. \tag{12.11}$$

The quantity in the square bracket is the contribution to $f(\mathbf{x}_h)$ from ancestral codons j and k, and is equal to the prior probability that the codon at the root is j, which is given by the equilibrium frequency π_j, multiplied by the five transition probabilities along the five branches of the phylogeny (Figure 12.3). For a tree of s species with $(s-2)$ ancestral nodes, the data at each site will be a sum over 61^{s-2} possible combinations of ancestral codons. In computer programs, we use the 'pruning' algorithm of Felsenstein (1981) to achieve efficient computation.

The log likelihood is a sum over all sites in the sequence.

$$\ell = \sum_{h=1}^n \log\{f(\mathbf{x}_h)\}. \tag{12.12}$$

Compared with the case of two sequences, we now have the same parameters in the substitution model (κ, ω, and the π_j's), but many more branch length parameters (e.g. $t_0, t_1, \ldots, t_4$ in Figure 12.3 instead of the single t in Figure 12.1b). Again, numerical optimization algorithms have to be used to maximize the likelihood function.

As mentioned above, the quantity in the square bracket in (12.11) is the contribution to the probability of the data $f(\mathbf{x}_h)$ by ancestral codons j and k. This contribution varies greatly depending on j and k, and the codons j and k that make the greatest contribution are the most probable codons for the two ancestral nodes at the site. This gives the empirical Bayes approach (also known as the likelihood approach) to reconstructing ancestral character states (Yang *et al.*, 1995; Koshi and Goldstein, 1996). Compared with the parsimony reconstruction (Fitch, 1971; Hartigan, 1973), the Bayes approach uses branch lengths and relative substitution rates between character states. Intuitive methods that use reconstructed ancestral sequences to detect adaptive molecular evolution will be discussed later in comparison with the ML method.

12.5   DETECTING ADAPTIVE EVOLUTION ALONG LINEAGES

12.5.1   Likelihood Calculation under Models of Variable ω Ratios among Lineages

It is easy to modify the model of the previous section to allow for different ω ratios among branches on a tree. The likelihood calculation under such a model proceeds in a similar way, except that the transition probabilities for different branches need to be calculated from different rate matrices (Qs) generated using different ωs. Suppose we want to fit a model in which the branch for species 1 of Figure 12.4 has a different ω ratio (ω_1), while all other branches have the same 'background' ratio (ω_0). Let $p_{ij}(t; \omega)$ denote the transition probability calculated using the ratio ω. Under this model, the probability of observing data, $\mathbf{x}_h$, is

$$f(\mathbf{x}_h) = \sum_j \sum_k \pi_j p_{jx_1}(t_1; \omega_1) p_{jx_2}(t_2; \omega_0) p_{jk}(t_0; \omega_0) p_{kx_3}(t_3; \omega_0) p_{kx_4}(t_4; \omega_0), \quad (12.13)$$

(compare with (12.11)).

Yang (1998) implemented models that allow for different levels of heterogeneity in the ω ratio among lineages. The simplest model (the 'one-ratio' model) assumes the same ω ratio for all branches in the phylogeny. The most general model (the 'free-ratio' model) assumes an independent ω ratio for each branch in the phylogeny. Intermediate models such as two- or three-ratio models assume two or three different ω ratios for lineages in the tree. These models can be compared using the LRT to examine interesting hypotheses. For example, the likelihood values under the one-ratio and free-ratio models can be compared to test whether the ω ratios are different among lineages. Also, we can let the lineages of interest have a different ω ratio from the background ω ratio for all other lineages in the phylogeny (as in Figure 12.4). Such a two-ratio model can be compared with the one-ratio model to examine whether the lineages of interest have a different ω ratio from other lineages. Furthermore, when the estimated ω ratio for the lineages of interest (say, ω_1 in Figure 12.4) is >1, models with and without the constraint that $\omega_1 = 1$ can be compared to test whether the ratio is different from (i.e. greater than) 1. This test directly examines the possibility of positive selection along specific lineages.
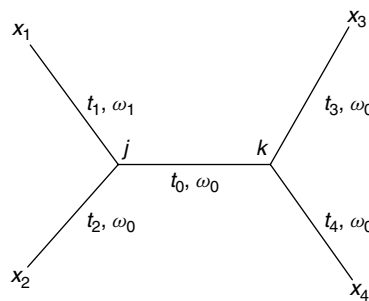
Figure 12.4   A tree of four sequences to explain a model of variable ω ratios among lineages. The ω ratio for the branch leading to species 1 (ω_1) is different from the ratio (ω_0) for all other branches.

It should be pointed out that variation in the ω ratio among lineages is a violation of the strictly neutral model, but is itself not sufficient evidence for adaptive evolution. First, relaxed selective constraints along certain lineages can generate variable ω ratios. Second, if nonsynonymous mutations are slightly deleterious but not lethal, their fixation probabilities will depend on factors such as the population size of the species. In large populations, deleterious mutations will have a smaller chance of getting fixed than in small populations. Under such a model of slightly deleterious mutations (Ohta, 1973), species with large population sizes are expected to have smaller ω ratios than species with small population sizes. At any rate, an ω ratio significantly greater than one provides convincing evidence for Darwinian selection.

12.5.2 Adaptive Evolution in the Primate Lysozyme

In the following, we use the example of the lysozyme c genes of primates (Figure 12.5) to demonstrate the use of codon-substitution models of variable ω ratios among lineages (Yang, 1998). Lysozyme is mainly found in secretions like tears and saliva as well as in white blood cells, where its function is to fight invading bacteria. Leaf-eating colobine monkeys have a complex foregut, where bacteria ferment plant material, followed by a true stomach that expresses high levels of lysozyme, where its new function is to digest these bacteria (Stewart *et al.*, 1987; Messier and Stewart, 1997). It has been suggested that the acquisition of a new function may have led to high selective pressure on the enzyme, resulting in high nonsynonymous substitution rates. In an analysis of lysozyme c genes from 24 primate species, Messier and Stewart (1997) identified two lineages with elevated ω ratios, indicating episodes of adaptive evolution in the lysozyme. One lineage, expected from previous analysis (Stewart *et al.*, 1987), is ancestral to colobine monkeys, and another unsuspected lineage is ancestral to the hominoids. The two lineages are represented by branches h and c in Figure 12.5 for a subset of the data of Messier and Stewart (1997).

As branches h and c are the lineages of interest, we test assumptions concerning three ω ratio parameters: ω_h for branch h, ω_c for branch c, and ω_0 for all other (background) branches. Table 12.2 lists log-likelihood values and ML parameter estimates under different models. The simplest model assumes one ω ratio (Table 12.2, A) while the most general model assumes three ratios (E, I, and J). The six possible two-ratio models
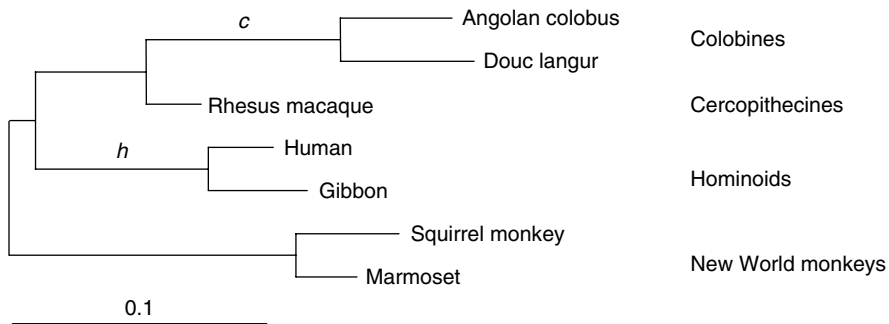
Figure 12.5 Phylogeny of seven primate species, for a subset of the lysozyme data of Messier and Stewart (1997) used to demonstrate models of variable ω ratios among branches. The analysis uses the unrooted tree, but the root is shown here for clarity. After Yang (1998).

Table 12.2 Log-likelihood values and parameter estimates under different models for the lysozyme c genes.

Model	p	ℓ	$\hat{\kappa}$	$\hat{\omega}_0$	$\hat{\omega}_h$	$\hat{\omega}_c$
A. 1 ratio: $\omega_0 = \omega_h = \omega_c$	22	-906.02	4.5	0.81	$= \hat{\omega}_0$	$= \hat{\omega}_0$
B. 2 ratios: $\omega_0 = \omega_h, \omega_c$	23	-904.64	4.6	0.69	$= \hat{\omega}_0$	3.51
C. 2 ratios: $\omega_0 = \omega_c, \omega_h$	23	-903.08	4.6	0.68	∞	$= \hat{\omega}_0$
D. 2 ratios: $\omega_0, \omega_h = \omega_c$	23	-901.63	4.6	0.54	7.26	$= \hat{\omega}_H$
E. 3 ratios: $\omega_0, \omega_h, \omega_c$	24	-901.10	4.6	0.54	∞	3.65
F. 2 ratios: $\omega_0 = \omega_h, \omega_c = 1$	22	-905.48	4.4	0.69	$= \hat{\omega}_0$	1
G. 2 ratios: $\omega_0 = \omega_c, \omega_h = 1$	22	-905.38	4.4	0.68	1	$= \hat{\omega}_0$
H. 2 ratios: $\omega_0, \omega_h = \omega_c = 1$	22	-904.36	4.3	0.54	1	1
I. 3 ratios: $\omega_0, \omega_h, \omega_c = 1$	23	-902.02	4.5	0.54	∞	1
J. 3 ratios: $\omega_0, \omega_h = 1, \omega_c$	23	-903.48	4.4	0.54	1	3.56

Note: p is the number of parameters. All models include the following 21 common parameters: 11 branch lengths in the tree (Figure 12.5), nine parameters for base frequencies at codon positions used to calculate codon frequencies, and the transition–transversion rate ratio κ.
Source: Yang (1998).

(B−D and F−H) are used as well. In models F-J, the ω ratio for the branch(es) of interest is fixed at 1.

Estimate of the ω ratio under the one-ratio model ($\omega_0 = \omega_h = \omega_c$) is 0.81, indicating that, on average, purifying selection dominates the evolution of the lysozyme. Estimates of ω_c for branch c range from 3.4 to 3.6 when ω_c is allowed free to vary (models B, E, and J in Table 12.2). Estimates of ω_h are always infinite when ω_h is assumed to be a free parameter (models C, E, and I), indicating the absence of synonymous substitutions along branch h. Estimate of the background ratio ω_0 is 0.54, when ω_h and ω_c are not constrained to be equal to ω_0 (models D, E, I, and J).

Results of LRTs are shown in Table 12.3. Tests A−E examine whether the ω ratio for the branch(es) of interest is different from (i.e. greater than) the background ratio, while tests A'−E' examine whether the ratio is greater than 1. For example, test E compares models B and E of Table 12.2 and examines the null hypothesis that $\omega_h = \omega_0$, with ω_c free to vary in both models; ω_h is significantly higher than ω_0 in this comparison. Such tests suggest that ω_h is significantly greater than the background ratio ω_0 ($P < 1\%$; Table 12.3, D and E) and also significantly greater than 1 ($P < 5\%$; Table 12.3, D' and E'). Similar tests suggest that ω_c is significantly greater than ω_0 ($P < 5\%$; Table 12.3, C), but not significantly greater than 1 (P ranges from 17–20 %; Table 12.3, B' and C'). More detailed analyses of the data set can be found in Yang (1998).

12.5.3 Comparison with Methods Based on Reconstructed Ancestral Sequences

Evolutionary biology has had a long tradition of reconstructing characters in extinct ancestral species and using them as observed data in all sorts of analyses. For molecular data, statistical methods (Yang *et al.*, 1995; Koshi and Goldstein, 1996) can be used to obtain more reliable ancestral reconstructions, taking into account branch lengths and relative substitution rates between characters (nucleotides, amino acids, or codons) (see the discussion below (12.12)). Overall, reconstructed molecular sequences appear much

Table 12.3 Likelihood ratio statistics ($2\Delta\ell$) for testing hypotheses concerning lysozyme evolution.

Hypothesis tested	Assumption made	Models compared	$2\Delta\ell$
A. $(\omega_h = \omega_c) = \omega_0$	$\omega_h = \omega_c$	A & D	8.78**
B. $\omega_c = \omega_0$	$\omega_h = \omega_0$	A & B	2.76
C. $\omega_c = \omega_0$	ω_h free	C & E	3.96*
D. $\omega_h = \omega_0$	$\omega_c = \omega_0$	A & C	5.88*
E. $\omega_h = \omega_0$	ω_c free	B & E	7.08**
A'. $(\omega_h = \omega_c) \leq 1$	$\omega_h = \omega_c$	D & H	5.46*
B'. $\omega_c \leq 1$	$\omega_h = \omega_0$	B & F	1.68
C'. $\omega_c \leq 1$	ω_h free	E & I	1.84
D'. $\omega_h \leq 1$	$\omega_c = \omega_0$	C & G	4.60*
E'. $\omega_h \leq 1$	ω_c free	E & J	4.76*

* Significant at the 5 % level.
** Significant at the 1 % level.
Source: Yang (1998).

more reliable than reconstructed morphological characters (Yang *et al.*, 1995; Cunningham *et al.*, 1998).

Messier and Stewart (1997) reconstructed ancestral sequences and used them to perform pairwise comparisons to calculate d_N and d_S along branches in the tree. Their analysis pinpointed two particular lineages in the primate phylogeny that may have gone through adaptive evolution. Crandall and Hillis (1997) took the same approach in an analysis of relaxed selective constraints in the rhodopsin genes of eyeless crayfishes living deep under the ground. Zhang *et al.* (1997) argue that the normal approximation to the statistic $d_N - d_S$ may not be reliable due to small sample sizes. These authors instead applied Fisher's exact test to the counts of differences between the two sequences, ignoring multiple hits at the same site.

A major difference between the ML method discussed in this section and the heuristic approaches using ancestral reconstruction is that ML uses all possible ancestral characters (such as codons j and k for the two ancestral nodes in the tree of Figures 12.3 and 12.4), while the approach of ancestral reconstruction uses only the most likely codons and ignores the others. Ancestral sequences reconstructed by both parsimony and likelihood involve random errors, as indicated by the calculated posterior probabilities (Yang *et al.*, 1995). Using the optimal reconstruction and ignoring the sub-optimal ones may cause a systematic bias. One kind of such bias is obvious if we use reconstructed ancestral sequences to estimate branch lengths, as both parsimony and likelihood tend to minimize the amount of evolution to select the most likely ancestral characters. Biases involved in estimation of d_N and d_S using reconstructed ancestral sequences are more complex. They can be reduced by considering sub-optimal as well as optimal reconstructions, but such modifications have not been attempted. Furthermore, pairwise comparisons along branches of the phylogeny may not be as reliable as a simultaneous comparison of all sequences by ML.

It appears advisable that ancestral reconstruction should be used for exploratory data analysis and ML be used for more rigorous tests. When the LRT suggests adaptive evolution along certain lineages, ancestral reconstruction may be very useful to pinpoint the

responsible amino acid changes. Indeed, an interesting use of ancestral reconstruction is to provide ancestral proteins to be synthesized in the lab to examine its biochemical and physiological properties. Such studies of 'paleobiochemistry' were envisaged by Pauling and Zuckerkandl (1963) several decades ago; see Golding and Dean (1998) and Chang and Donoghue (2000) for reviews.

12.6 INFERRING AMINO ACID SITES UNDER POSITIVE SELECTION

12.6.1 Likelihood Ratio Test under Models of Variable ω Ratios among Sites

Till now, we have assumed that all amino acid sites in a protein are under the same selective pressure, with the same underlying nonsynonymous/synonymous rate ratio (ω). While the synonymous rate may be homogeneous among sites, it is well known that nonsynonymous rates are highly variable. Most proteins have highly conserved amino acid positions at which the underlying ω ratio is close to 0. The requirement that the ω ratio, averaged over all sites in the protein, is greater than 1 is thus a very stringent criterion for detecting adaptive evolution. It would be much more realistic if we allow the ω ratio to vary among sites.

We can envisage two cases, which require different statistical modeling. In the first case, we may know the different structural and functional domains of the protein, and can use such information to classify amino acid sites in the protein into several classes. The different site classes are assumed to have different ω ratios, which are parameters to be estimated by ML. Suppose we have K site classes, with the corresponding ω ratios to be $\omega_1, \omega_2, \ldots, \omega_K$. The likelihood calculation under this model is rather similar to that under the model of one ω ratio for all sites (12.11 and 12.12), except that the corresponding ω ratio will be used to calculate the transition probabilities for data at each site. For example, if site h is from site class $k(k = 1, 2, \ldots, K)$ with the ratio ω_k, then $f(\mathbf{x}_h)$ of (12.11) will be calculated using ω_k. The likelihood is again given by (12.12). A few such models were implemented by Yang and Swanson (2002) and applied to MHC class I alleles in which structural information was used to identify amino acids at the antigen-recognition site. The models were termed 'fixed-sites' models.

In the second case, we assume that there are several heterogeneous site classes with different ω ratios, but we do not know which class each amino acid site belongs to. Such models were termed 'random-sites' models by Yang and Swanson (2002) and will be the focus of discussion here. They use a statistical distribution to account for the random variation of ω among sites (Nielsen and Yang, 1998). We assume that the synonymous rate is constant among sites, and allow only the nonsynonymous rate to be variable, although the same approach can be taken to deal with synonymous rate variation (Kosakovsky Pond and Muse, 2005). Branch length t is defined as the expected number of nucleotide substitutions per codon, averaged over sites. Suppose amino acid sites fall into a fixed number of K classes, with the proportions $p_0, p_1, \ldots, p_{K-1}$, and ω ratios $\omega_0, \omega_1, \ldots, \omega_{K-1}$ treated either as parameters or as functions of parameters in the ω distribution. To calculate the likelihood, we need to calculate the probability of observing data at each site, say data $\mathbf{x}_h$ at site h. The conditional probability of the data, given

ω_κ, $f(\mathbf{x}_h|\omega_k)$, can be calculated as described before (12.11). Since we do not know which class site h belongs to, we sum over all site classes (i.e. over the distribution of ω):

$$f(\mathbf{x}_h) = \sum_{k=1}^{K} p_k f(\mathbf{x}_h|\omega_k). \tag{12.14}$$

The log likelihood is a sum over all n sites in the sequence,

$$\ell = \sum_{h=1}^{n} \log\{f(\mathbf{x}_h)\}. \tag{12.15}$$

Parameters in the model include branch lengths in the tree, κ, π_j's, and parameters in the distribution of ω among sites. As before, we estimate the codon frequency parameters by the observed frequencies, and estimate the other parameters by numerical optimization of the likelihood.

A number of statistical distributions have been implemented by Nielsen and Yang (1998) and Yang *et al.* (2000). Positive selection is tested using an LRT comparing a null model that does not allow $\omega > 1$ with an alternative model that does. Computer simulations have revealed two tests to be particularly effective (Anisimova *et al.*, 2001; 2002; Wong *et al.*, 2004). The first compares the null model M1a (neutral), which assumes two site classes in proportions p_0 and $p_1 = 1 - p_0$ with $0 < \omega_0 < 1$ and $\omega_1 = 1$, and the alternative model M2a (selection), which adds a proportion p_2 of sites with $\omega_2 > 1$. M1a and M2a are slight modifications of models M1 and M2 in Nielsen and Yang (1998), which had $\omega_0 = 0$ fixed and were found to be highly unrealistic for most data sets. As M2a has two more parameters than M1a, the χ_2^2 distribution may be used for the test. However, the regularity conditions for the asymptotic χ^2 approximation are not met, as M1a is equivalent to M2a by fixing $p_2 = 0$, which is at the boundary of the parameter space, and as ω_2 is not identifiable when $p_2 = 0$. The use of χ_2^2 is expected to be conservative. The second test compares the null model M7 (beta), which assumes a beta distribution for ω, and the alternative model M8 (beta & ω), which adds an extra site class of positive selection with $\omega_s > 1$. The beta distribution, beta(p, q), can take a variety of shapes depending on its parameters p and q, such as L-, inverted L-, U-, and inverted U-shapes, but is restricted to the interval (0, 1). It is thus a flexible null model. M8 has two more parameters than M7, so that χ_2^2 may be used to conduct the LRT. As in the comparison between M1a and M2a, use of χ_2^2 is expected to make the test conservative.

Another model, called M3 (discrete), is sometimes useful as well. This assumes a general discrete model, with the frequencies and the ω ratios (p_k and ω_k in (12.14)) for K site classes estimated as free parameters. All models discussed here may be considered special cases of this general mixture model. As is typical in such models, often only a few classes can be fitted to real-data sets. Model M3 may be compared with model M0 (one-ratio) to construct an LRT to test whether the selective pressure varies among sites.

After ML estimates of model parameters are obtained, we can use the empirical Bayes approach to infer the most likely site class (and thus the ω ratio) for any site. The marginal probability of the data $f(\mathbf{x})$ (12.15) is a sum of contributions from each site class k, and the site class that makes the greatest contribution is the most likely class for the site, that

is, the posterior probability that a site with data $\mathbf{x}_h$ is from site class k (with rate ratio ω_k) is

$$f(\omega_k|\mathbf{x}_h) = \frac{p_k f(\mathbf{x}_h|\omega_k)}{f(\mathbf{x}_h)} = \frac{p_k f(\mathbf{x}_h|\omega_k)}{\sum_j p_j f(\mathbf{x}_h|\omega_j)}, \qquad (12.16)$$

(Nielsen and Yang, 1998). When the ω estimates for some site classes are greater than 1, this approach can be used to identify sites from such classes, which are potential targets of positive selection. The posterior probability provides a measure of accuracy. This is known as the *naïve empirical Bayes approach (NEB)* (Nielsen and Yang, 1998). A serious drawback with this approach is that it uses the maximum likelihood estimations (MLEs) of parameters as fixed constants in (12.16), ignoring their sampling errors. This may be a sensible approach in large or medium-sized data sets. However, in small data sets, the information content may be low and the parameter estimates may involve large sampling errors. This deficiency appears to be the major reason for the poor performance of the procedure in small data sets in several simulation studies (e.g. Anisimova *et al.*, 2002; Wong *et al.*, 2004; Massingham and Goldman, 2005; Scheffler and Seoighe, 2005). A more reliable approach, known as the *Bayes empirical Bayes* (*BEB*), was implemented by Yang *et al.* (2005). BEB accommodates uncertainties in the MLEs of parameters in the ω distribution by integrating numerically over a prior for the parameters. Other parameters such as branch lengths are fixed at their MLEs, as these are expected to have much less effect on inference concerning ω. A hierarchical (full) Bayesian approach is implemented by Huelsenbeck and Dyer (2004), using markov chain monte carlo (MCMC) to average over tree topologies, branch lengths, as well as other substitution parameters in the model. This approach involves more computation but may produce more reliable inference in small uninformative data sets, where the MLEs of branch lengths may involve large sampling errors (Scheffler and Seoighe, 2005).

12.6.2 Methods That Test One Site at a Time

An intuitive approach to examining selective pressure indicated by the ω ratio at individual sites is to reconstruct ancestral sequences and then count synonymous and nonsynonymous changes at each site. Comparison of the observed counts with a 'neutral' expectation may then allow us to decide whether the site is evolving under purifying selection or positive selection. On a large phylogeny, some sites may have accumulated many changes at a single site for this approach to be feasible. Fitch *et al.* (1997) performed such an analysis of the hemagglutinin (HA) gene of human influenza virus type A and considered a site to be under positive selection if it has more nonsynonymous substitutions over the gene than the average. Suzuki and Gojobori (1999) compared the counts with the neutral expectation that $d_S = d_N$ at the site, using the method of Nei and Gojobori (1986) to estimate d_S and d_N. A large number of sequences are needed for the test to have any power, as confirmed in computer simulations (Suzuki and Gojobori, 1999; Wong *et al.*, 2004). Both the approaches of Fitch *et al.* (1997) and Suzuki and Gojobori (1999) used the parsimony algorithm to infer ancestral sequences.

The use of reconstructed ancestral sequences may be a source of concern since the reconstructed sequences are not real observed data. In particular, positively selected sites are often the most variable sites in the alignment, at which ancestral reconstruction is the least reliable. This problem may be avoided by taking a likelihood approach, averaging over all possible ancestral states. Indeed Suzuki (2004), Massingham and Goldman (2005),

and Kosakovsky Pond and Frost (2005) implemented methods to estimate one ω parameter for each site using ML. Other parameters in the model such as branch lengths are usually estimated from the whole data set and fixed when the ω ratio is estimated for each site. Then at every site, an LRT is used to test the null hypothesis, $\omega = 1$. This is called the *site-wise likelihood ratio (SLR) test* by Massingham and Goldman (2005). A problem with those methods is that the number of ω ratios estimated in the model increases without bound with the increase of the sequence length. Likelihood methods often perform poorly under such parameter-rich models. However, in computer simulations Massingham and Goldman (2005) found that the SLR test achieved good false-positive rates and reasonably high power.

The methods discussed above test every site for positive selection, and are similar to the BEB procedure for identifying individual amino acid sites under positive selection (Yang *et al.*, 2005). To test whether the gene is under positive selection (i.e. whether the gene has any codon with $\omega > 1$), a correction for multiple testing should be applied (Wong *et al.*, 2004).

12.6.3   Positive Selection in the HIV-1 *vif* Genes

An example data set of HIV-1 *vif* genes from 29 subtype-B isolates is used here to demonstrate the likelihood models of variable ω ratios among sites. The data set was analyzed by Yang *et al.* (2000). The sequence has 192 codons. Several models are used in ML estimation, with the results shown in Table 12.4. Only parameters involved in the ω distribution are listed, as other parameters (branch lengths in the phylogeny, the transition–transversion rate ratio κ, and the base frequencies at the three codon positions) are common to all models. The model codes are those used in Yang *et al.* (2000) and in the CODEML program in the PAML package (Yang, 1997).

Table 12.4   Likelihood values and parameter estimates under models of variable ω ratios among sites for HIV-1 *vif* genes.

Model code	ℓ	d_N/d_S	Estimates of parameters
M0. one-ratio (1)	-3499.60	0.644	$\hat{\omega} = 0.644$
M1a. neutral (2)	-3393.83	0.438	$\hat{p}_0 = 0.611, \hat{\omega}_0 = 0.080,$ $(\hat{p}_1 = 0.389), (\hat{\omega}_1 = 1)$
M2a. selection (4)	-3367.86	0.689	$\hat{p}_0 = 0.573, \hat{\omega}_0 = 0.090,$ $\hat{p}_1 = 0.346, (\hat{\omega}_1 = 1)$ $(\hat{p}_2 = 0.081), \hat{\omega}_2 = 3.585$
M3. discrete (5)	-3367.16	0.742	$\hat{p}_0 = 0.605, \hat{\omega}_0 = 0.108$ $\hat{p}_1 = 0.325, \hat{\omega}_1 = 1.211$ $(\hat{p}_2 = 0.070), \hat{\omega}_2 = 4.024$
M7. beta (2)	-3400.44	0.440	$\hat{p} = 0.176, \hat{q} = 0.223$
M8. beta & ω (4)	-3370.66	0.687	$\hat{p}_0 = 0.909, \hat{p} = 0.222, \hat{q} = 0.312$ $(\hat{p}_1 = 0.091), \hat{\omega}_s = 3.385$

Note: The number of parameters in the ω distribution is given in parentheses after the model code. d_N/d_S is the average ω over all sites in the gene. Parameters in parentheses are given to ease interpretation but are not free parameters. Estimates of the transition–transversion rate ratio κ range from 3.6 to 4.1 among models.
Source: The data are from Yang *et al.* (2000).

The one-ratio model (M0) assumes one ω ratio for all sites and gives an average ω ratio of 0.644, indicating that on average, purifying selection is the dominating force during the evolution of the gene. The selection model (M2a) suggests that about $\hat{p}_1 = 8.1\%$ of sites are under positive selection with $\hat{\omega}_2 = 3.58$, and it has a significantly higher log-likelihood value than model M1a (neutral), which does not allow for sites under positive selection. The LRT statistic for comparing M1a and M2a is $2\Delta\ell = 2 \times 25.97 = 51.94$, much greater than $\chi^2_{2,1\%} = 9.21$. Similarly estimates under model M8 (beta & ω) suggest that about $\hat{p}_1 = 9.1\%$ of sites are under positive selection with $\hat{\omega}_s = 3.385$. M8 fits the data much better than M7 (β), which does not allow for sites with $\omega > 1$. Thus, the two LRTs, which compare M1a with M2a and M7 with M8, provide significant evidence for the presence of sites in the gene under positive selection.

The discrete model (M3) suggests that about $\hat{p}_2 = 7.0\%$ of sites are under strong positive selection with $\hat{\omega}_2 = 4.0$, while a large proportion ($\hat{p}_1 = 33\%$) of sites are under weak positive selection or are nearly neutral with $\hat{\omega}_1 = 1.2$.

Figure 12.6 plots the posterior probabilities for site classes at each site under model M2a (selection), calculated using the NEB approach (12.16). Parameter estimates under M2a suggest that the three site classes have the ω ratios 0.090, 1, and 3.585 and are in proportions 57.3, 34.6, and 8.1% (Table 12.4). These proportions are the prior probabilities for site classes at every site. The observed data at the site alter these prior probabilities considerably, so that the posterior probabilities are very different from the prior. For example, the posterior probabilities for site 1 are 0.986, 0.014, and 0.000, and this site is almost certainly under purifying selection. In contrast, the probabilities at site 31 are 0.000, 0.012, and 0.988, and this site is most likely to be under strong diversifying selection (Figure 12.6).

The above calculation used the NEB procedure (12.16), which treats the estimates of the proportions and ω ratios as true values. NEB is used here as it is simple to explain. In real-data analysis, the BEB procedure should be used instead, which takes into account estimation errors in those parameters. The HIV *vif* data set is relatively large, so that the two procedures produced very similar results. For example, at the 99% cutoff, both methods identified three sites to be under positive selection: 39 F, 127 Y, and 167 K (the amino acids are from the reference sequence B_SF2). At the 95% cutoff, BEB identified five additional sites: 31 I, 33 K, 101 G, 109 L, and 122 N, while the NEB list also included 63 K and 92 K.
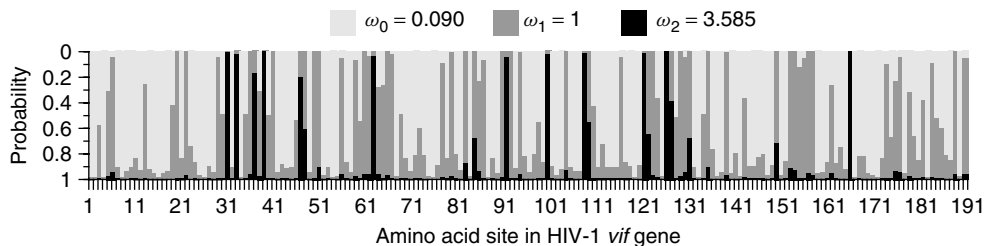
Figure 12.6 Posterior probabilities of site classes along the gene for the HIV-1 *vif* genes under the site model M2a (selection).

12.7 TESTING POSITIVE SELECTION AFFECTING PARTICULAR SITES AND LINEAGES

12.7.1 Branch-site Test of Positive Selection

A natural extension to the branch- and site-models discussed above is the *branch-site models*, which allow the ω ratio to vary both among sites and among lineages (Yang and Nielsen, 2002). Such models attempt to detect signals of local episodic natural selection (Gillespie, 1991), which affects only a few sites along particular lineages. In the models of Yang and Nielsen (2002), branches in the tree are divided *a priori* into foreground and background categories, and an LRT is constructed to compare an alternative hypothesis that allows for some sites under positive selection on the foreground branches with a null hypothesis that does not. However, computer simulations conducted by Zhang (2004) suggest that the resulting tests can produce excessive false positives when the model assumptions are violated, and in general, the tests were unable to distinguish positive selection from relaxed selective constraint along the foreground branches. Later, the branch-site models were modified, and the tests constructed using the modified models were found to be much more robust and appeared to produce reliable inference (Yang *et al.*, 2005; Zhang *et al.*, 2005).

The modified model, known as branch-site model A, is summarized in Table 12.5. Along the background lineages, there are two classes of sites: the conserved sites with $0 < \omega_0 < 1$ and the neutral sites with $\omega_1 = 1$. Along the foreground lineages, a proportion $(1 - p_0 - p_1)$ of sites come under positive selection with $\omega_2 \geq 1$. Likelihood calculation under this model is very similar to that under the site models (12.14). As we do not know *a priori* which site class each site is from, the probability of data at a site is an average over the four site classes. Let $I_h = 0, 1, 2a, 2b$ be the site class that site h is from. We have

$$f(\mathbf{x}_h) = \sum_{I_h} p_k f(\mathbf{x}_h | I_h). \tag{12.17}$$

The conditional probability $f(\mathbf{x}_h | I_h)$ of observing data $\mathbf{x}_h$ at site h, given that the site comes from site class I_h, is easy to calculate, because the site evolves under the one-ratio model if $I_h = 0$ or 1, and under the branch model if $I_h = 2a$ or 2b.

To construct an LRT, we use model A as the alternative hypothesis, while the null hypothesis is the same model A but with $\omega_2 = 1$ fixed (Table 12.5). This is known as the *branch-site test of positive selection*. The null hypothesis has one fewer parameter, but since $\omega_2 = 1$ is fixed at the boundary of the parameter space of the alternative hypothesis, the null distribution should be a 50 : 50 mixture of point mass 0 and χ_1^2 (Self and Liang,

Table 12.5 The ω ratios assumed in branch-site model A.

Site class	Proportion	Background ω	Foreground ω
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1)p_0/(p_0 + p_1)$	$0 < \omega_0 < 1$	$\omega_2 > 1$
2b	$(1 - p_0 - p_1)p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$

Note: The model involves four parameters: $p_0, p_1, \omega_0, \omega_2$.

1987). The critical values are 2.71 and 5.41 at the 5 and 1 % levels, respectively. One may also use χ_1^2 (with critical values 3.84 at 5 % and 5.99 at 1 %) to guide against violations of model assumptions.

As in the site-based analysis, the BEB approach can be used to calculate the posterior probability that a site is from site classes 2a and 2b, allowing identification of amino acid sites potentially under positive selection along the foreground lineages (Yang *et al.*, 2005).

Similar to the branch test, the branch-site test requires the foreground branches to be specified *a priori*. This may be easy if a well-formulated biological hypothesis exists, for example, if we want to test adaptive evolution driving functional divergences after gene duplication. The test may be difficult to apply if no *a priori* hypothesis is available. To apply the test to several or all branches on the tree, one has to correct for multiple testing. Several correction procedures have been suggested in the statistics literature. The simplest is the Bonferroni correction (Miller, 1981, pp. 67–70), by which one uses α/m as the significance level to test every hypotheses when m hypotheses (branches) are being tested.

12.7.2 Similar Models

Several other models of codon substitution also allow the ω ratio to vary both among lineages and among sites. Forsberg and Christiansen (2003) and Bielawski and Yang (2004) implemented the *clade* models. Branches on the phylogeny are *a priori* divided into two clades, and an LRT is used to test for divergences in selective pressure between the two clades indicated by different ω ratios. There may not be any sites under positive selection with $\omega > 1$. Clade model C, implemented by Bielawski and Yang (2004) is summarized in Table 12.6. This assumes three site classes. Class 0 includes conserved sites with $0 < \omega_0 < 1$, while class 1 includes neutral sites with $\omega_1 = 1$; both apply to all lineages. Class 2 includes sites that are under different selective pressures in the two clades, with ω_2 for clade 1 and ω_3 for clade 2. The model involves five parameters in the ω distribution: $p_0, p_1, \omega_0, \omega_2,$ and ω_3. An LRT can be constructed by comparing model C with the site model M1a (neutral), which assumes two site classes with two free parameters: p_0 and ω_0 (see Section 12.6). The χ_3^2 distribution may be used for the test. The clade models follow on the ideas of Gu (2001) and Knudsen and Miyamoto (2001), who used amino acid substitution rates to indicate functional constraint.

A *switching model* was implemented by Guindon *et al.* (2004), which allows the ω ratio at any site to switch among three different values: $\omega_1 < \omega_2 < \omega_3$. Besides the Markov process that describe substitutions between codons, a hidden Markov chain runs over time and describes the switches of any site between different selective regimes (i.e. the three ω values). The model has the same structure as the covarion model of Tuffley and Steel (1998) (see also Galtier, 2001; Huelsenbeck, 2002), which allows the substitution

Table 12.6   The ω ratios assumed in clade model C.

Site class	Proportion	Clade 1	Clade 2
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2	$p_2 = 1 - p_0 - p_1$	ω_2	ω_3

Note: The model involves five parameters: $p_0, p_1, \omega_0, \omega_2, \omega_3$.

rate for any site to switch between high and low values. The switching model is an extension of the site model M3 (discrete) discussed above, under which the ω ratio is fixed at every site. An LRT can thus be used to compare them. Guindon *et al.* (2004) found that the switching model fitted a data set of the HIV-1 *env* genes much better than the site models. An empirical Bayes procedure can be used to identify lineages and sites with high ω ratios, and it appears necessary to apply a correction for multiple testing.

12.8   LIMITATIONS OF CURRENT METHODS

Both the test of positive selection along lineages and the test of positive selection at amino acid sites discussed in this chapter are conservative. When we are testing for lineages under positive selection, we assume that the ω ratio is identical across sites. Positive selection is detected along a lineage only if the ω ratio averaged over all sites is significantly greater than 1. Since many or most sites in a protein are under purifying selection with the underlying ω ratios close to 0, this procedure constitutes a very conservative test of positive selection. Similarly, the LRT for detecting positively selected sites is based on the assumption that the ω ratio is identical among all lineages on the tree. Positive selection is detected for a site only if the underlying ω ratio averaged over all lineages is greater than 1. This assumption appears unrealistic except for genes under recurrent diversifying selection; for most genes, positive selection probably affects only a few lineages.

The branch-site and switching models (Yang and Nielsen, 2002; Guindon *et al.*, 2004; Yang *et al.*, 2005; Zhang *et al.*, 2005) allow the ω ratio to vary both among lineages and among sites and appear to have more power to detect positive selection. Nevertheless, these models are new, and more testing is needed before we know how robust they are. At some stage we will have to compromise. On one hand, we want to focus on a short time period and a few amino acid sites so that the signal of adaptive evolution will not be overwhelmed by purifying selection. On the other hand, a short time period and a few amino acid sites may not offer enough opportunities for evolutionary changes to accumulate to generate a signal that is detectable by statistical tests.

The models discussed here assume the same ω ratio for any amino acid changes; at a positively selected site, changes to any amino acid are assumed to be advantageous. This assumption is unrealistic and appears to make the test conservative. Some authors distinguish between radical and conservative amino acid replacements (i.e. between amino acids with very different and very similar chemical properties, respectively), and suggest that a higher radical than conservative rate is evidence for positive selection (Hughes *et al.*, 1990; Rand *et al.*, 2000; Zhang, 2000). However, this criterion is less convincing than the simple ω ratio and is found to be sensitive to assumptions about transition and transversion rate differences and unequal amino acid compositions (Dagan *et al.*, 2002).

The tests discussed here identify only positive selection that increases the nonsynonymous rates, and may have little power in detecting other types of selection such as balancing selection (Yang *et al.*, 2000). Furthermore, they require a moderate amount of sequence divergence to operate and tend to lack power in population data. They may be useful for species data or fast-evolving viral genes only.

12.9   COMPUTER SOFTWARE

A number of counting methods to estimate d_N and d_S between two sequences, such as those of Nei and Gojobori (1986), Li *et al.* (1985), Li (1993), and Pamilo and Bianchi (1993) have been implemented in MEGA3 (Kumar *et al.*, 2005). These and the likelihood method for estimating d_N and d_S (Goldman and Yang, 1994) have been implemented in the CODEML program in the PAML package (Yang, 1997). The latter program also implements the ML models discussed in this chapter.

REFERENCES

Anisimova, M., Bielawski, J.P. and Yang, Z. (2001). The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites. *Molecular Biology and Evolution* **18**, 1585–1592.

Anisimova, M., Bielawski, J.P. and Yang, Z. (2002). Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution* **19**, 950–958.

Bielawski, J.P. and Yang, Z. (2004). A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *Journal of Molecular Evolution* **59**, 121–132.

Bonhoeffer, S., Holmes, E.C. and Nowak, M.A. (1995). Causes of HIV diversity. *Nature* **376**, 125.

Chang, B.S. and Donoghue, M.J. (2000). Recreating ancestral proteins. *Trends in Ecology and Evolution* **15**, 109–114.

Comeron, J.M. (1995). A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of Molecular Evolution* **41**, 1152–1159.

Crandall, K.A. and Hillis, D.M. (1997). Rhodopsin evolution in the dark. *Nature* **387**, 667–668.

Cunningham, C.W., Omland, K.E. and Oakley, T.H. (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology and Evolution* **13**, 361–366.

Dagan, T., Talmor, Y. and Graur, D. (2002). Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. *Molecular Biology and Evolution* **19**, 1022–1025.

Dunn, K.A., Bielawski, J.P. and Yang, Z. (2001). Substitution rates in *Drosophila* nuclear genes: implications for translational selection. *Genetics* **157**, 295–305.

Edwards, A.W.F. (1992). *Likelihood, Expanded Edition*. John Hopkins University Press, London.

Fay, J.C., Wyckoff, G.J. and Wu, C.-I. (2001). Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Fitch, W.M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology* **20**, 406–416.

Fitch, W.M., Bush, R.M., Bender, C.A. and Cox, N.J. (1997). Long term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 7712–7718.

Forsberg, R. and Christiansen, F.B. (2003). A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. *Molecular Biology and Evolution* **20**, 1252–1259.

Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution* **18**, 866–873.

Gillespie, J.H. (1991). *The Causes of Molecular Evolution*. Oxford University Press, Oxford.

Gojobori, T. (1983). Codon substitution in evolution and the "saturation" of synonymous changes. *Genetics* **105**, 1011–1027.

Golding, G.B. and Dean, A.M. (1998). The structural basis of molecular adaptation. *Molecular Biology and Evolution* **15**, 355–369.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**, 182–198.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**, 725–736.

Grimmett, G.R. and Stirzaker, D.R. (1992). *Probability and Random Processes*. Clarendon Press, Oxford.

Gu, X. (2001). Maximum-likelihood approach for gene family evolution under functional divergence. *Molecular Biology and Evolution* **18**, 453–464.

Guindon, S., Rodrigo, A.G., Dyer, K.A. and Huelsenbeck, J.P. (2004). Modeling the site-specific variation of selection patterns along lineages. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12957–12962.

Hartigan, J.A. (1973). Minimum evolution fits to a given tree. *Biometrics* **29**, 53–65.

Huelsenbeck, J.P. (2002). Testing a covariotide model of DNA substitution. *Molecular Biology and Evolution* **19**, 698–707.

Huelsenbeck, J.P. and Dyer, K.A. (2004). Bayesian estimation of positively selected sites. *Journal of Molecular Evolution* **58**, 661–672.

Hughes, A.L. and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170.

Hughes, A.L., Ota, T. and Nei, M. (1990). Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Molecular Biology and Evolution* **7**, 515–524.

Ina, Y. (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution* **40**, 190–226.

Ina, Y. (1996). Pattern of synonymous and nonsynonymous substitutions: an indicator of mechanisms of molecular evolution. *Journal of Genetics* **75**, 91–115.

Jukes, T.H. and Cantor, C.R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, H.N. Munro, ed. Academic Press, New York, pp. 21–123.

Kalbfleisch, J.G. (1985). *Probability and Statistical Inference*, *Vol. 2: Statistical Inference*. Springer-Verlag, New York.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Knudsen, B. and Miyamoto, M.M. (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 14512–14517.

Kosakovsky Pond, S.L. and Frost, S.D.W. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution* **22**, 1208–1222.

Kosakovsky Pond, S.L. and Muse, S.V. (2005). Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution* **22**, 2375–2385.

Koshi, J.M. and Goldstein, R.A. (1996). Probabilistic reconstruction of ancestral protein sequences. *Journal of Molecular Evolution* **42**, 313–320.

Kreitman, M. and Akashi, H. (1995). Molecular evidence for natural selection. *Annual Review of Ecology and Systematics* **26**, 403–422.

Kumar, S., Tamura, K. and Nei, M. (2005). MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics* **5**, 150–163.

Lee, Y.-H., Ota, T. and Vacquier, V.D. (1995). Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Molecular Biology and Evolution* **12**, 231–238.

Lewontin, R. (1989). Inferring the number of evolutionary events from DNA coding sequence differences. *Molecular Biology and Evolution* **6**, 15–32.

Li, W.-H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* **36**, 96–99.

Li, W.-H., Wu, C.-I. and Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* **2**, 150–174.

Massingham, T. and Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**, 1753–1762.

Messier, W. and Stewart, C.-B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature* **385**, 151–154.

Miller, R.G.J. (1981). *Simultaneous Statistical Inference*. Springer-Verlag, New York.

Mindell, D.P. (1996). Positive selection and rates of evolution in immunodeficiency viruses from humans and chimpanzees. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 3284–3288.

Miyamoto, S. and Miyamoto, R. (1996). Adaptive evolution of photoreceptors and visual pigments in vertebrates. *Annual Review of Ecology and Systematics* **27**, 543–567.

Miyata, T., Miyazawa, S. and Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution* **12**, 219–236.

Miyata, T. and Yasunaga, T. (1980). Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *Journal of Molecular Evolution* **16**, 23–36.

Moriyama, E.N. and Powell, J.R. (1997). Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. *Journal of Molecular Evolution* **45**, 378–391.

Muse, S.V. (1996). Estimating synonymous and nonsynonymous substitution rates. *Molecular Biology and Evolution* **13**, 105–114.

Muse, S.V. and Gaut, B.S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**, 715–724.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**, 418–426.

Nekrutenko, A., Makova, K.D. and Li, W.-H. (2001). The K_A/K_S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Research* **12**, 198–202.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98.

Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of Molecular Evolution* **40**, 56–63.

Pamilo, P. and Bianchi, N.O. (1993). Evolution of the *Zfx* and *Zfy* genes – rates and interdependence between the genes. *Molecular Biology and Evolution* **10**, 271–281.

Pauling, L. and Zuckerkandl, E. (1963). Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. *Acta Chemica Scandinavica* **17**, S9–S16.

Perler, F., Efstratiadis, A., Lomedica, P., Gilbert, W., Kolodner, R., Dodgson, J. (1980). The evolution of genes: the chicken preproinsulin gene. *Cell* **20**, 555–566.

Rand, D.M., Weinreich, D.M. and Cezairliyan, B.O. (2000). Neutrality tests of conservative-radical amino acid changes in nuclear- and mitochondrially-encoded proteins. *Gene* **261**, 115–125.

Scheffler, K. and Seoighe, C. (2005). A Bayesian model comparison approach to inferring positive selection. *Molecular Biology and Evolution* **22**, 2531–2540.

Self, S.G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605–610.

Smith, N.G. and Eyre-Walker, A. (2002). Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024.

Stewart, C.-B., Schilling, J.W. and Wilson, A.C. (1987). Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**, 401–404.

Stuart, A., Ord, K. and Arnold, S. (1999). *Kendall's Advanced Theory of Statistics*. Arnold, London.

Suzuki, Y. (2004). New methods for detecting positive selection at single amino acid sites. *Journal of Molecular Evolution* **59**, 11–19.

Suzuki, Y. and Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution* **16**, 1315–1328.

Tuffley, C. and Steel, M. (1998). Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences* **147**, 63–91.

Whelan, S., Liò, P. and Goldman, N. (2001). Molecular phylogenetics: state of the art methods for looking into the past. *Trends in Genetics* **17**, 262–272.

Wong, W.S.W., Yang, Z., Goldman, N. and Nielsen, R. (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**, 1041–1051.

Yamaguchi, Y. and Gojobori, T. (1997). Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 1264–1269.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**, 555–556. `http://abacus.gene.ucl.ac.uk/software/paml.html`.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**, 568–573.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press, Oxford.

Yang, Z. and Bielawski, J.P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution* **15**, 496–503.

Yang, Z., Kumar, S. and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**, 1641–1650.

Yang, Z. and Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* **46**, 409–418.

Yang, Z. and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**, 32–43.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19**, 908–917.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.-M.K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449.

Yang, Z. and Swanson, W.J. (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular Biology and Evolution* **19**, 49–57.

Yang, Z., Wong, W.S.W. and Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* **22**, 1107–1118.

Zhang, J. (2000). Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *Journal of Molecular Evolution* **50**, 56–68.

Zhang, J. (2004). Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular Biology and Evolution* **21**, 1332–1339.

Zhang, J., Kumar, S. and Nei, M. (1997). Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Molecular Biology and Evolution* **14**, 1335–1338.

Zhang, J., Nielsen, R. and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* **22**, 2472–2479.

13

Genome Evolution

J.F.Y. Brookfield

Institute of Genetics, School of Biology, University of Nottingham, Nottingham, UK

Our understanding of the forces that have shaped the evolution of genome organisation is still largely rudimentary. The process of genome evolution is, in some ways, different in kind from those more typically studied by population and evolutionary geneticists. Firstly, forces of selection that operate at levels different from the individual may be important, as in the spread of selfish intragenomic parasites and in the impact of changes in genome organisation on the ability of lineages to proliferate and diversify. Secondly, alternative genotypes are often not allelic, such that the process of evolutionary change is not necessarily ended by the spread to fixation of a new genetic type. However, some progress can be made in understanding genome evolution. A number of the most controversial issues in genome evolution, such as the evolution of codon bias, the antiquity of introns and the selfishness of mobile DNAs, are reviewed. In addition, the evolutionary forces affecting gene number, linkage between genes, gene expression and the functional overlap between gene products are considered.

13.1 INTRODUCTION

Evolutionary processes involve mutation, recombination, migration, selection and genetic drift. These forces are typically thought sufficient to explain evolutionary change. Examples of the success of this microevolutionary, population genetics approach to evolutionary explanation include the many cases in which new advantageous alleles have arisen by mutation and spread in populations subjected to new environmental pressures. Indeed, it has become a byword for evolutionary explanation that, in order to explain a feature of an organism, we have to show a way in which, in a polymorphic population, a gene causing the feature would be expected to spread at the expense of its allele. The classic case of the spread of the *carbonaria* form of the peppered moth, *Biston betularia* – notwithstanding doubts about the evidential strength of the story (Majerus, 1998) – provides a case in point. Here the *carbonaria* allele spreads at the expense of the pale allele in conditions of darkened tree trunks, hypothetically as a result of its increased crypsis (Kettlewell, 1955).

This paradigm of evolutionary explanation cannot realistically be applied to some issues of genome structure, which are dealt with in this chapter. There are three reasons. The first is that the levels of comparisons are different. We cannot discuss a situation in which the presence or absence of introns in genes, for example, can be fully explained by a microevolutionary process in which two alternative types of organisms exist, one with introns in all its genes and the ability to process these from the transcripts, and another competing type in which no introns are present. In long-term evolutionary processes such as this, we still need to consider the mechanisms that may cause loss or gain of introns at the mutational level and the microevolutionary forces acting, which allow new variants to spread in the populations in which they first arise. In addition, however, we need to discuss whether the groups of organisms which possess introns will, in the longer term, have a greater ability to evolve to exploit new niches and opportunities. Selection, in other words, is partially operating at the species level, or, indeed, at higher levels. This means that we have to invoke selective forces operating over the longer timescale, involving the ability of species to proliferate in numbers and diversify into successful families, orders, classes and phyla.

The second concept that is different from our paradigm of microevolution, is that in the ordinary explanations of molecular evolutionary change we tend to imagine changes in which a new mutation arises that may then spread, by genetic drift or by selection, to fixation. Fixation signifies the end of the evolutionary process with respect to that particular mutation. Molecular evolutionary change, therefore, only requires a combination of a rate of mutation and probabilities of fixation of individual mutations.

For some aspects of genomic evolution, however, the concept of fixation, which itself follows from the concept of allelism, does not apply. Processes by which the numbers of transposable elements in the genome may be stabilised by a combination of forces involving selection, drift and replicative transposition are discussed later. In these cases, there can be an equilibrium copy number of a transposable element family, even when, as in *Drosophila*, individual transposable element insertions are rarely fixed in populations. So, the spread of a transposable element family may result in an equilibrium state in which few sites are fixed, and in which, indeed, individual transposable element loci are continually being created by transposition, and lost by a combination of deletion, selection and drift.

Similarly, in considering questions such as the determination of the extent of codon usage bias, an equilibrium is envisaged between mutation, selection and drift, in which an equilibrium level of codon bias is maintained despite changes in the nucleotides at individual codons.

This raises a related and more general question, which goes beyond molecular evolution, extending to evolution in general, which is whether organisms can be considered to be at equilibrium with respect to selectively important mutations. In other words, are typical wild populations so well adapted to their environments that virtually no new mutations would improve their fitness and only if the external environment changed would any mutations become selectively advantageous? Alternatively, might wild populations, and their interactions with their environments, typically be such that there are very many adaptive mutations that would be possible, but which have either not arisen or which have, by drift, failed to spread? While the truth must certainly fall somewhere between these two extremes, Li (1997) makes a strong case that the latter situation will be typical in evolution, and that, in this sense, the creation of new advantageous mutations is a limiting factor in evolutionary change. However, in studies of the experimental evolution

of micro-organisms in particular, the rate of morphological evolution in *Escherichia coli* (Mongold and Lenski, 1996), and evolution of patterns of new gene expression in *Saccharomyces cerevisiae* (Farea *et al.*, 1999) when in new environments, implies that very rapid change can arise through advantageous mutations despite the stability of phenotypes in wild populations. Our view on this question influences the way we see many aspects of the interaction between selection and genome variability, such as the likely impacts, through hitch-hiking, of advantageous allelic substitutions.

The third aspect of genome evolution which does not correspond closely to our paradigms of microevolutionary change is particularly relevant to repetitive DNA sequences, and could be summarised as genomic conflicts arising from selfish genetic elements (reviewed in Hurst and Schilthuizen, 1998). These include, e.g. *medea* (maternal effect dominant embryonic arrest) mutations, seen in *Tribolium*, e.g. where heterozygous mothers create eggs, among which only those that contain the *medea* gene itself are viable. Here, if the half of the offspring of a heterozygous mother that inherit the gene benefit from the death of their siblings, from reduced competition or from eating the sibs, a *medea* gene that is initially rare, will spread deterministically (Wade and Beeman, 1994).

More common are selfish transposable elements. Since these typically have a mechanism through which their copy number is increased in the act of transposition, there will be, all else being equal, a rise in the number of elements in the genome from one generation to the next, a rise that appears to be countered by natural selection acting against those individuals who have above-average numbers of copies of the elements. In other words, there is a conflict of interest between the transposable element and the host, and, in this regard, the elements can be regarded as parasites.

This concept of parasitic DNA sequences coexisting with their hosts makes it very unlikely that any stable copy number equilibrium that we see will be stable in the face of mutation. Host mutations which reduce the harmful effects of transposable elements, possibly by reducing copy number, will be favoured, yet transposable element mutations which increase copy number would, in many cases, be expected to spread, even if they reduce host fitness. Thus, any equilibrium that we see will reflect only the genetic properties of the sequences that currently coexist and will be unstable to mutational change. Of course, for parasitic DNAs, as with all parasites, there is strong ascertainment bias, in that we only see selfish DNAs that have succeeded and prevailed despite host efforts to remove them.

13.2   THE STRUCTURE AND FUNCTION OF GENOMES

13.2.1   Genome Sequencing Projects

A feature of the progress in the biological sciences over the last decade has been the availability of whole genome sequences. One very obvious conclusion from genome projects has been that the differences in gene number between organisms are very much less than the differences in DNA content. Thus, among eukaryotes, the largest genomes seen are those of some flowering plants, which exceed 120 000 Mb in size, while the smallest are less than 10 Mb (Brown, 1999). This range in genome size is not matched by gene number, which seems to show little relationship with genome size among eukaryotes. Thus, the estimated gene numbers for *Arabidopsis thaliana* and *Homo sapiens*

are each around 25 000 genes (International Human Genome Sequencing Consortium, 2004; *Arabidopsis* Genome Initiative, 2000), notwithstanding the human genome, at just over 3 Mb (in the haploid) being around 20 times larger than that of *Arabidopsis*. Other animal genome sequencing projects include mouse (*Mus musculus*), rat (*Rattus norvegicus*), chicken (*Gallus gallus*), dog (*Canis familiaris*), zebrafish (*Danio rerio*), two species of pufferfish (*Tetraodon nigroviridis* and *Takifugu rubripes*), chimpanzee (*Pan troglodytes*), rhesus macaque (*Macaca mulatta*), wallaby (*Macropus eugenii*), sea squirt (*Ciona savignyi*), many species of *Drosophila*, the mosquito *Anopheles gambiae*, the beetle *Tribolium castaneum*, the honey bee *Apis mellifera*, the nematode *Caenorhabditis elegans* and the sea urchin *Strongylocentrotus purpuratus* (Gibson and Muse, 2004), while plants also include rice (*Oryza sativa*), maize (*Zea mays*), plus *Populus, Medicago, Sorghum, Lycopersicon, Brassica* and *Solanum* (Paterson, 2006).

More than 300 genomes of unicellular organisms have now been sequenced (TIGR, 2006), and these, in particular, have given rise to the science of comparative genomics. A number of important questions have been addressed by the comparison of prokaryotic genomes, in particular. Firstly, when orthologous genes are found in different sequenced genomes, we can ask what can be said about the relative divergences, and thus rates of evolution, of different genes. Do different types of genes evolve at systematically different rates? Another issue is whether incongruencies in the estimated phylogenetic trees of different genes suggest that horizontal gene transfer has taken place in the evolution of these genomes.

Another very important question is whether the phenotypic and metabolic differences between organisms result from differences in gene contents or differences in the details of the functions of homologous genes. It is tempting to imagine that it is the differences in gene content that underlie phenotypic differences between prokaryotic species, since a catalogue of genes is available as a result of sequencing projects, and genes present in one species but without orthologues in another can easily be identified. If two organisms possess orthologues, the temptation is to assume that these genes are carrying out the same function in the two organisms. By this hypothesis, the changes in amino acid sequence are either neutral or are simply involved in adapting the protein product to the changed physiological and metabolic requirements of a different species, and also perhaps to subtle changes in the amino acid sequences of interacting proteins.

However, often a gene in one species might have two homologues in another, owing to a duplication in the lineage separating the second species from the common ancestor. The two genes in this second species may also have mutant phenotypes that are different from each other. Certainly, as we see later (Section 13.3.5), the maintenance of a gene duplication implies that the two genes have partially non-overlapping functions. One type of explanation is that the function of a gene in the first species is in some way shared between these two genes in the second, a concept referred to as *subfunctionalisation* (Force *et al.*, 1999; Lynch and Force, 2000). The prediction of this model would be that the wild-type allele of the single gene in the first species could simultaneously rescue the mutant phenotypes of both genes in the second. However, another, more radical idea is that the two functions of the genes in the second species are not only different in kind from each other, but also fundamentally different from the role of the gene in the first species. In other words, conservation of sequence should not be regarded as always implying conservation in function (e.g. conserved genes might still clearly encode serine/threonine kinases, yet their targets could differ between species).

13.2.2 The Origins and Functions of Introns

Spliceosomal introns are found in protein-coding genes in the majority of plants, animals and fungi. However, they are absent from the genomes of bacteria, archaea and the most primitive eukaryotes. Intron number is high in vertebrates and plants but lower in the genes of *Drosophila melanogaster* and other invertebrates and introns are almost completely absent from the genes of the yeast *S. cerevisiae*, although *Schizosaccharomyces pombe* has considerably more introns (Jeffares *et al.*, 2006). The debate about the origins and causes of introns is a good illustration of the way in which considerations of the evolution of genome structure depend partly on the advantageous nature of genomic change over the long timescale.

There are two related issues that are important in the study of the evolution of introns. The first is the historical question of when they were introduced into the genomes of organisms (Roy and Gilbert, 2006). The introns-early view (Gilbert, 1987) is that the intron–exon structure seen in the genomes of higher eukaryotes is a relic of the structure of the first genes in the progenote (the first living thing, ancestral to all life today), and what today are exons are the remains of what were shorter ancestral genes in the progenote. The introns-late view suggests that spliceosomal introns are derived from type II bacterial introns and have been being added to eukaryotic genomes, only, throughout their evolution. There is now no doubt that there have been extensive intron gains and losses in eukaryote genes. Rogozin *et al.* (2003), comparing conserved genes from species with fully sequenced genomes, identified many introns conserved between genes from organisms from different kingdoms, while there were very many intron gains and losses also observed. The question, however, remains whether there were any introns present in the progenote – the introns-early view does not require all introns currently seen to be early genomic components, merely that some were. This creates an unsatisfactory situation, in that an introns-early theory in which the proportion of the current introns that are ancient might be very small does not make many very specific predictions about today's intron distribution. One prediction that can be arrived at is that, if ancient introns separate exons that were the original functioning minigenes in the progenote, then ancient exons should correspond to protein domains. In ancient proteins, exons of zero phase, where the intron falls precisely between codons, indeed show the strongest correlation with protein domains (de Souza *et al.*, 1998).

The fundamental problem with the introns-early hypothesis is its requirement that the ancestors of all extant prokaryotic groups would have had to lose all their introns, while the lineage that was to give rise to eukaryotes would have had to retain them. Furthermore, Palmer and Logsdon (1991) noted that, given the then-accepted phylogeny of the early eukaryotes, genomes with introns fall into a rather restricted subgroup of the eukaryotic tree. They suggested that, in order for introns to be present in the progenote, they must have been lost independently in six basal eukaryote taxa, yet maintained in the equally acellular eukaryotes that were our ancestors. However, genome-wide intron loss is a real phenomenon. Parsimony arguments based on intron numbers in other species of fungi indicate that intron losses, in the lineage leading to *S. cerevisiae* in particular, have been extensive. Furthermore, there is evidence that the spliceosomal machinery was ancestrally found in a wider group of eukaryotes than those that now possess introns, and probably in the common ancestor of all extant eukaryotes (Collins and Penny, 2005; Lynch and Richardson, 2002). There is a potential mechanism for intron loss in organisms such as yeast, which contain retrotransposons encoding reverse transcriptase. It is the loss by

gene conversion involving cDNAs as a template. The role of this process as a general mechanism is supported by observations of the simultaneous loss of adjacent introns (Frugoli *et al.*, 1998).

A second issue, conflated with this first, is the mechanistic explanation for the gain and loss of introns, both in terms of the molecular processes creating these changes in the DNA and the population genetics and evolutionary processes determining whether mutated genomes with intron gains or losses proliferate. The mechanism of intron loss by gene conversion by reverse transcribed cDNAs, while elegant in conception, has not been established to be involved in most intron losses. Intron gains, save for the special case of reinsertion of an ancestral intron through gene conversion by a paralogue that has retained it (Hankeln *et al.*, 1997), require an insertion of DNA from a source somewhere else in the genome. The obvious sources are transposable elements and other introns. Coghlan and Wolfe (2004) examined introns gained in *C. elegans* and *C. briggsae* since their split 100 million years ago and showed that the sequences of some of these resembled those of other introns. However, at least one case is known of intron gain by the introduction of a mobile DNA (a short interspersed nuclear element (SINE) element in plants) (Iwamoto *et al.*, 1998). There remains a question whether the rate of DNA sequence evolution of introns is sufficient to explain why introns do not typically have identifiable homologues elsewhere in the genome, notwithstanding the continuing process of intron insertion, which must require a donor element. (Even if this insertion was conservative in the initial creation of the first intron-bearing allele, in that the donor DNA physically moved without replication into the gene to form the intron, recombination during the slow spread of this allele by drift would (unless the intron and its source were tightly physically linked) place it in a genetic background with the donor DNA still present elsewhere in the genome.) The genomic uniqueness of introns means that they cannot typically be derived from currently active families of mobile DNAs. We can thus imagine a model in which, at some time in the past, t million years ago, an intron was inserted into a gene, and the source of that intron and the intron itself have both since evolved in a selectively unconstrained way, with a rate of base substitution of x changes per base per million years. Under this model, the expected base identity between the intron and its source is given (from the simplest Jukes and Cantor (1969) one-parameter model) by Identity $= 0.25 + 0.75e^{-8xt/3}$. Suppose we believe that the identification of the source of a particular intron requires that it shares at least 80 % sequence identity, this predicts, if we use an approximate rate of evolution, x, of 10^{-8} changes per base per million years, that introns inserted in the last 11.65 million years would have their sources identified (provided that the sources themselves had not subsequently been lost from the genome). This would imply that a time-homogeneous process of intron insertion where an intron was inserted somewhere in the genome at least once every 5 million years or so would be likely to result in identifiable sources of recently inserted introns if whole genome sequences are available. In *Caenorhabditis* the rate of intron insertion per genome is estimated to be almost 1–2 per million years (Coghlan and Wolfe, 2002).

What is the mechanism for a new allele with an intron gain or loss to spread through the population? In principle, an allele with a newly inserted intron could be advantageous, neutral or deleterious (possibly weakly so) relative to its progenitor. While arguments for advantageous effects of introns have been suggested (see below) almost none of these have an instantaneously beneficial effect. If so, the implication is that the initial spread of introns must be by genetic drift, which can, in the case of small populations, also

bring about the spread of weakly deleterious alleles. Indeed, Lynch (2006) argues that the smaller effective population sizes of multicellular eukaryotes make introns effectively neutral. He argues that, since genes can be inactivated by mutations in introns, there will be selection against an allele bearing an intron relative to one lacking an intron, but the strength of selection will be dependent on the mutation rate per base, multiplied by the number of bases in the intron where mutational changes create inactivation. The product of these two is likely to be greater than the reciprocal of the large effective population size of unicellular organisms, causing effective selection against intron insertions in these species, while selection of the same strength would be ineffective in the smaller effective population sizes of multicellular organisms, and alleles with inserted additional introns would be able to spread to fixation by genetic drift.

Four types of adaptive functions have been ascribed to introns. The first is that introns play an important role in the process of nonsense-mediated decay. In nonsense-mediated decay, transcripts with premature nonsense (stop) codons are eliminated prior to translation, whether a stop codon is inherited by the organism, is the result of a somatic mutation or represents a transcription error (Maquat, 2004). The effect is a diminution in the amount of truncated and potentially harmful proteins produced. The obvious question is how the cell recognises the difference between these premature stop codons and the codon marking the end of the wild-type polypeptide. Proteins that attach to the junctions between exons being spliced together in the nucleus allow this recognition, with the prediction that introns may play a role in protection against the effects of these premature stop codons (Lynch and Richardson, 2002).

The second advantage for introns is that downstream sequences controlling transcription may evolve in introns, which may be more evolutionary labile than exons, since they will have no constraint arising from a role encoding amino acids, and many examples of intronic enhancers are now known. Thirdly, introns will allow alternative splicing pathways, where functionally different proteins will be produced from the same gene (Xing and Lee, 2006). This has probably been elucidated most clearly in the *D. melanogaster* somatic sex determination pathway, in which a pathway of transcripts from the genes *sex-lethal, transformer, transformer 2* and *doublesex* are alternatively spliced to yield functional female-specific proteins only in a female (XX) cellular environment (Cline, 1993). Finally, introns allow the creation of new genes by exon shuffling. This is where ectopic intronic recombination brings about the creation of hybrid genes combining exons from different sources, which then are spread to fixation by selection or drift. While genes created by exon shuffling were first identified and are most familiar in the vertebrates, they have also been seen in other metazoan groups. They are, however, apparently absent from fungi, plants and protists (Patthy, 1999). Furthermore, many of the genes where exon shuffling may be important are genes that allow animal multicellularity, suggesting that exon shuffling may have played a significant role in its evolution. Note that it is implicit in all these last three adaptive effects of introns that they are long-term advantages, which would require introns to be present in genes long before these advantages could be felt.

A model for the proliferation of introns thus has two components, one of which is the process for the insertion of introns into genes – this insertion process could be seen as a process of selfish spread, with the individual intron insertions being neutral or weakly deleterious (or weakly advantageous as a result of enhancement of nonsense-mediated decay) and with numbers of introns being built up by this mutational pressure. Subsequently, the long-term advantages would give benefit at higher levels of

selection, in which intron-bearing lineages may be able to proliferate and diversify at the expense of intronless competitor lineages. This kind of explanation is, however, still quite unsatisfactory, in that there is no good quantitative theory of selective processes operating at levels higher than the individual.

13.3   THE ORGANISATION OF GENOMES

13.3.1   The Relative Positions of Genes: Are They Adaptive?

Genome projects reveal the relative positions of genes. In prokaryotes, genes are organised into operons, such that non-homologous genes involved in the same metabolic pathway are adjacent and are transcribed into a polycistronic message, as a mechanism for integrating their activities. In eukaryotes, functional clustering of this kind is not seen, and genes involved in a shared metabolic pathway would typically be located on different chromosomes. The overall picture is one in which the positions of genes cannot, by and large, be predicted from their functions (save for a tendency for paralogues to cluster – see Section 13.3.3 below – as a result of the mechanisms for their generation). Are gene positions therefore random? The answer is 'no', and the causes of non-randomness can be seen as a combination of a small contribution from functionality and a larger contribution from the vagaries of evolutionary history.

The most straightforward way to imagine a chromosome is as a segregation device. In this context, linkage between genes is an inevitable consequence of there being more genes than chromosomes. An important issue, therefore, is whether the relative positions of genes on chromosomes have an adaptive significance that follows from the levels of recombination between them.

The fact that patterns of linkage between genes are adaptive has been a major theme in the development of ecological and population genetics. In the Malaysian swallowtail butterfly *Papilio memnon*, the females are mimetic, but are highly polymorphic, and indeed each female morph mimics a different distasteful model butterfly. Clarke and Sheppard (1968; 1969) carried out crosses to find out the pattern of inheritance of the morphs, and found that, at first sight, the complex mimetic patterns were apparently inherited as alleles at a single locus. However, exceptional butterflies were found that combined features of the different mimetic patterns, and seemed to be poor mimics of any of the model species. More detailed crosses revealed that these butterflies showing scrambled mimetic patterns were recombinants. From these experiments, Clarke and Sheppard concluded that at least five tightly linked genes were controlling the variation between mimetic morphs, and that the genes were linked for adaptive reasons, in order to maintain linkage disequilibrium between alleles at the loci, and thereby prevent poor mimics from appearing by segregation. They called such a set of adaptively linked polymorphic loci a *supergene*.

On the basis of these and earlier data, there was considerable discussion of the evolution of supergenes (see, e.g. Ford, 1964). Genes which carried out functionally related tasks, such as the colour and banding loci of *Cepaea nemoralis* (Cain *et al.*, 1960), were hypothesised to physically translocate in the chromosomes, so as to occupy tightly linked chromosomal positions. The results from genome projects, in general, strongly support the maintenance of synteny over evolutionary time periods longer than the duration of

individual functional polymorphisms, and show that genes do not, typically, move between or within chromosomes for reasons connected with the establishment of linkage disequilibrium between polymorphic, functionally distinct alleles. However, there is evidence for some level of clustering on non-homologues in eukaryotic genomes. In *S. cerevisiae*, essential genes show significant evidence of clustering (Pal and Hurst, 2003), as do genes with similar expression patterns (Hurst *et al.*, 2004). Most remarkably, *S. cerevisiae*'s DAL cluster of six genes involved in allantoin degradation has been shown to have been created by the movement of genes from previously unrelated locations (Wong and Wolfe, 2005).

Clustering of coexpressed non-homologues is probably not generated as a mechanism to ensure linkage disequilibrium, as was postulated in the original supergene concept. Rather, the clustering is related to the control of gene expression. The locus control sites of the globin gene clusters are just one example of a mechanism of gene expression control having the result that adjacent genes need to remain so because their joint expression is controlled by cis-acting sequences operating simultaneously on many linked genes. This effect apparently operates through the modulation of chromatin structure (for a review of the case of the β-globin gene cluster, see Li *et al.*, 1999). However, the clustering in the human genome of genes regulated by the aire transcription factor appears not to be the result of an effect of chromatin remodelling (Johnnidis *et al.*, 2005).

13.3.2 Functional Linkage Among Prokaryotes

In bacteria, operons form an example of phylogenetically unrelated genes brought together as a means of coordinating their expression. Genome comparisons have shown that operon structures can be evolutionarily unstable (Itoh *et al.*, 1999). Genes themselves may be added to bacterial genomes by horizontal transfer. Comparisons between the *E. coli* genome and that of *Salmonella enterica* have revealed that between 8 and 18 % of the *E. coli* sequences have been added to the genome by horizontal transfer following the split between the two species (Lawrence and Ochman, 1998). Indeed, it has been suggested that, with such high rates of gene transfer being observed between bacterial cells, the operon itself might be thought of as the unit of function (Lawrence, 1997). The clustering of genes in operons arises not solely from the need for simultaneous expression of all the genes, but also from the ability of a transferred group of functionally related genes to create a complete functional pathway when transformed into a naïve host. One example of collections of functionally related genes are the pathogenicity islands (Hacker *et al.*, 1997). These are groups of genes encoding pathways that allow bacteria to infect specific eukaryotic hosts. Their horizontal transfer is revealed through the sharing of recent ancestry for pathogenicity islands from distantly related bacterial species (Lawrence, 1999). Indeed, such 'genomic islands' of genes of related function are not solely related to pathogenicity, but to a range of ecological adaptations (Dobrindt *et al.*, 2004), and their rapid and promiscuous movement between bacterial lineages has the effect that there can be very major differences in gene content between very closely related genomes (Gogarten and Townsend, 2005).

13.3.3 Gene Clusters

Many vertebrate genes exist as linked clusters of homologous genes. The paradigm for the explanation of this type of organisation is the human globin genes. Here, there are two

clusters. The β-globin genes, consisting of those encoding the adult β and δ globins, the foetal $^G\gamma$ and $^A\gamma$ globins and the embryonic ε globin, along with a $\psi\beta$ pseudogene, are clustered on chromosome 11. The θ globin (the function of which remains unknown (Cooper *et al.*, 2005)), together with the two adult α globins, and the embryonic ζ globin, plus $\psi\alpha$ and $\psi\zeta$ pseudogenes, are in the α-globin cluster on chromosome 16. All genes and pseudogenes are in the same orientation within each cluster. In these α- and β-globin clusters it is clear that the mechanism for the generation of the clusters and thus families is local duplication by unequal crossing over. Unequal crossovers would be expected to create deletions of genes as well as duplications. However, while the deleted chromosome would be expected to have low fitness, at least in the homozygous state, and thus would be eliminated by selection, a haplotype bearing a duplicated gene might be neutral, and thus might (with a probability of $1/2N$, where N is the population size) spread to fixation by genetic drift. Under such a piecemeal process, one would expect that numerous duplications might accumulate, in which case the duplicated gene is, at least initially, truly redundant, in that silencing mutations in it would be neutral. A pair of genes created following duplication have three possible fates. If one is not lost by the reciprocal of the unequal recombination event that created the duplication, then either they could evolve to be functionally different (either through subfunctionalisation or neofunctionalisation, see Section 13.3.5), such that selection maintains each gene, or one copy could be silenced to form a pseudogene. True redundancy of the genes would not be stably maintained (see Section 13.3.6). A cycle of neutral duplication, followed by silencing, followed by decay, would thus be expected, which would lead to an increase in genome size with time. There are abundant examples of such clustered gene families, many containing pseudogenes, in the chromosomes of humans and other vertebrates. However, the extent to which pseudogenes arise and are maintained differs greatly between genomes. Unlike vertebrates, in *Drosophila* pseudogenes are almost unknown. The one case thought to be of a processed *Adh* pseudogene (i.e. a pseudogene generated by the insertion of reverse transcribed mRNA) in the species *D. teissieri* and *D. yakuba* (Jeffs and Ashburner, 1991) turned out to be a new functional gene, *jingwei*, created by the insertion of an Adh cDNA into a coding sequence from another gene (Long and Langley, 1993). The explanation for the absence of pseudogenes seems to be that *Drosophila* shows a high rate of evolutionary loss of DNA for sequences that are not actively maintained by selection (Petrov *et al.*, 1996; Petrov and Hartl, 1998). Indeed, almost the entire non-coding region of the *Drosophila* genome appears to be under selective constraint (Halligan and Keightley, 2006). This purging of the *Drosophila* genome of non-functional sequences may be the result of weak selection, which may be effective in *Drosophila*'s large population size, whereas the same selection strength would be ineffective in the smaller effective population sizes of vertebrates. Alternatively, the selection coefficients associated with DNA loss in *Drosophila* may be larger than their equivalents in other species, possibly reflecting selection for rapid genome division in syncytial early embryonic development, when nuclear division occurs every 9 minutes (Lawrence, 1992).

13.3.4 Integration of Genetic Functions

Systems biology seeks to interpret the effects of genes in terms of the complex networks connecting genes and their products. Thus, there are networks showing causal links between genes whose products influence the expression of other genes, and there are

networks of interactions between protein molecules. Such networks can be described in a Boolean way, concerned only with presence and the signs of interactions of the genes, or in a more quantitative way, in which the strengths of the interactions are also documented. The goal of such integrative approaches is ambitious – by understanding the ways in which genes and their products interact, we should be able to predict the likely phenotypes associated with mutational changes to genes, including gene losses.

Some have noted a scale-free property in a variety of networks in biology and elsewhere (Barabasi and Albert, 1999). The scale-free property is that the probability that a node in the network is connected to i other nodes is given by $P(i) = ci^{-\gamma}$, with c and γ being constants, such that a log–log plot of $P(i)$ against i will give a straight line of slope $-\gamma$. This property has been identified, e.g. in the network of protein–protein interactions in yeast. It implies that networks grow by new nodes preferentially attaching themselves to nodes that already have many connections. It has also been seen approximately in the probability that a gene family has i members in a given genome (although here it has the modified form or a Pareto function, where $P(i) = c(a + i)^{-\gamma}$, where a is a further constant). Karev *et al.* (2003; 2004) have shown that this family size distribution can be the result of a linear model involving gene births by duplication, deaths and innovations (changing the family to which a given gene belongs), although it is not clear whether the timescale of gene family evolution has been sufficient for real families to reach these models' equilibrium states. The scale-free distribution identified in networks of protein interaction also raises the question which modes of gene and thus protein duplication and divergence in protein function would be expected to result in networks with this property (Hughes and Friedman, 2005). Others have also approached gene family evolution from the standpoint of stochastic birth–death models (e.g. Hahn *et al.*, 2005; Csuros and Miklos, 2006, who also include horizontal gene transfer). It should also be said that many argue that the scale-free property is not seen, in detail, in the networks available (Khanin and Wit, 2006), also, it would not be expected to be seen in the sub-networks typically examined, even if it held in networks as a whole (Stumpf *et al.*, 2005).

13.3.5 Gene Duplications as Individual Genes or Whole Genome Duplications?

One of the great recent myths in molecular biology was the very high estimates of gene number in vertebrates, with 80 000 being typical, which were current and in undergraduate textbooks (e.g. Brown, 1999) around the start of this decade, prior to the completion of sequencing projects. These were quoted by, among others, me in the first edition of this book. These inaccurately high estimates arose from measurements of gene density in well-studied regions of the genome, but gene density is itself highly heterogeneous across genomes and well-studied regions were typically chosen because of the genes that they contained, with the result that they constituted a biased sample of genomic material. However, there is still no general agreement about the number of human genes, with recent estimates being as low as 25 000 (International Human Genome Sequencing Consortium, 2004). In fact, the identification of genes in eukaryotes is difficult, mainly as a result of introns: while an open reading frame of the length of a typical polypeptide is statistically highly unlikely, open reading frames the length of exons are not unlikely enough to constitute a strong signal.

We have seen the way in which the globin genes have been created by a series of gene duplications, which have also given rise, ultimately, to pseudogenes. Are gene families therefore typically created by piecemeal gene addition (Lynch and Conery, 2000), or,

rather, do they typically arise from genome-wide duplication events? A clear example of genome doubling is that of the *Xenopus laevis* genome, which is derived from at tetraploidisation event around 30 million years ago. Notwithstanding the presence of effectively four copies of each gene in a diploid *X. laevis*, it appears that both loci of each gene are under selective constraint, in that each shows a reduced rate of amino acid sequence change relative to synonymous changes (Hughes and Hughes, 1993). An earlier duplication has been detected in the budding yeast, *S. cerevisiae* (Wolfe and Shields, 1997) by comparing the *S. cerevisiae* genome to the related *Kluyveromyces*, which does not share the duplication. Genome sequencing of *Kluyveromyces waltii* has confirmed this interpretation (Kellis *et al.*, 2004).

Ohno (1970) suggested that it is through gene duplication that new functions can be developed. He further suggested that there could have been two rounds of tetraploidisation in the origin of the vertebrates (the '2R' hypothesis). While the apparent support for this model that was derived from the apparent fourfold increase in gene number in vertebrates relative to invertebrates proved to be illusory, there remains some evidence to support this hypothesis. One source of support has been the *Hox* genes, clusters of genes encoding transcription factors involved in antero-posterior differentiation, identified initially in *Drosophila*, but since seen in all metazoan animals. These are found as a single cluster in all invertebrates (albeit a broken one in *Drosophila* itself) but exist as four clusters, now called *A*, *B*, *C* and *D*, in tetrapods. The suggestion has been made that the first of the 2R duplications may have occurred initially prior to the divergence of the two clades of agnathans (lampreys and hagfish) from the lineage leading to the jawed vertebrates, followed by a second duplication in the gnathostomes (Holland *et al.*, 1994). The cephalochordate *Branchiostoma,* which is the closest relative to the vertebrates among invertebrates, has a single *Hox* cluster (Garcia-Fernandez and Holland, 1994). However, the situation is not the simple one of there always being four *Hox* clusters in jawed vertebrates. While tetrapods have four *Hox* clusters, the zebrafish (*D. rerio*) has seven clusters (Amores *et al.*, 1998), the result of a teleost-specific genome duplication, with one cluster (a D) having been subsequently lost. Puffer fish *Spheroides nephelus* and *T. rubripes* also have seven clusters, but have lost a C cluster rather than a D (Amores *et al.*, 2004).

The agnathans form an important test of the '2R' hypothesis, since, by this hypothesis, they should either share all four of the gnathostome Hox clusters or have only two clusters. The sea lamprey *Petromyzon marinus* (Force *et al.*, 2002) has at least four Hox clusters, but phylogenetic arguments suggest that the second duplication event was independent of that in the gnathostomes, and it has even been suggested that the common ancestor of lampreys and gnathostomes had a single Hox cluster (Fried *et al.*, 2003). The hagfish *Eptatretus stoutii* has many *Hox* genes in, apparently, at least four clusters, and possibly up to seven, and orthologies with mammalian Hox clusters suggest that at least two Hox clusters were present in the hagfish–gnathostome ancestor (Stadler *et al.*, 2004).

If the 2R hypothesis is correct, there is a strong prediction that, if there is a gene family found as a single gene in invertebrates and as four genes in tetrapods, then the four tetrapod genes should form two clades, each of two genes, in the rooted phylogeny. Bailey *et al.* (1997) looked at the *Hox* gene clusters in mammals and tried to estimate the topology of the phylogeny linking clusters A, B, C and D. While this was impossible with the *Hox* genes themselves, since too many individual genes had been lost from the clusters since their creation, the phylogeny was estimated using collagen genes linked

to each of the four clusters. An unrooted tree showed AD and BC clusters, but rooting with an outgroup placed the root on the branch connecting cluster D to the other three clusters. The implication is that, if the clusters had been derived from whole genome duplications, there would have been three such duplications, not two. The first would create the duplication distinguishing D from the ancestor of the other three clusters; the second would distinguish the ancestor of A from the ancestor of B and C (the duplicate of the D cluster created by this event must subsequently have been lost). Finally, a third genome duplication would be required to create the B and C clusters, after which the duplicates of the A and D clusters would again have to be lost. Alternatively, the clusters could have duplicated individually rather than as parts of whole genome duplications. This failure to find the topology expected from genome-wide duplications has turned out to be typical of developmentally important genes in vertebrates (Hughes, 1999). Clearly if there are multiple genome duplications followed by extensive loss of duplicates, the theory of genome duplications no longer creates any prediction of the number of gene family members expected to be seen or of the phylogeny connecting them (since that will depend on the subsequent pattern of gene loss). Indeed, genome-wide duplications followed by piecemeal and rapid loss of duplicated genes or clusters of genes will be indistinguishable from piecemeal segmental duplications. Only if piecemeal segmental losses affected different genomic regions in different lineages could the extent of the duplicated region at the time the lineages split be reconstructed.

Thus, while the genome duplication in the teleosts is well established, the status of the 2R hypothesis in early vertebrate ancestry is uncertain (Panopoulou and Poustka, 2005). There was undoubtedly a burst of gene duplication at the time suggested for the genome duplications in the 2R hypothesis (Blomme *et al.*, 2006). Furthermore, analysis of chromosomal regions bearing genes duplicated prior to the human−*Tetraodon* split using the ascidian *Ciona*'s genome as an outgroup reveals extensive regions showing fourfold paralogy (Dehal and Boore, 2005).

Whether genes are duplicated singly or as part of the regional or genome duplication, the possible fates, of subsequent loss, formation of a pseudogene or continued functionality are mentioned above (Section 13.3.4). There has been considerable recent debate (Force *et al.*, 1999; Lynch and Force, 2000) as to whether the acquisition of new functions for genes (neofunctionalisation) is preceded by, or indeed replaced by, a process of subfunctionalisation. In the subfunctionalisation model, it is imagined that a gene might have a series of sequences controlling the function of the gene, which can be partitioned into separate subfunctions. The most obvious way that this could come about is for there to be separate enhancer sequences, typically 5′ to the start of transcription, controlling the expression of the coding sequence of a gene at different times or in different tissues. In the simplest case, imagine that there are two enhancer sequences, A and B, which independently bring about the gene's expression in different tissues. Following duplication, which yields two gene copies, I and II, the possibility exists that a mutation arises which inactivates subfunction A in gene copy I. Such a mutation would be effectively neutral and could spread to fixation. A further mutation, causing the inactivation of subfunction B in gene copy I or indeed a mutation inactivating the coding sequence in gene copy I would also be able to spread neutrally. However, another neutral possibility would be a mutation inactivating subfunction B in gene copy II. Spread of such a variant would leave the genes subfunctionalised, with copy I carrying out subfunction B and copy II carrying out subfunction A. Calculations

and simulations (Lynch and Force, 2000; Force *et al.*, 2005) reveal that, provided the effective population size is considerably less than the gene's rate of inactivating mutations (whether they are in the coding sequence or subfunction enhancers), the probability of subfunctionalisation can be reasonably high (particularly when the mutation rate to subfunction loss is large relative to the mutation rate for inactivation of the whole gene). Once the subfunctions of the genes are being carried out by different duplicates, advantageous mutations in the coding sequences of the two genes may spread, adapting the genes for their now subtly different roles. The key element of this model, named *DDC* for 'duplication–degeneration–complementation', is that the molecular events that lead to the sharing of the function of the gene between the two copies are not driven to fixation by natural selection. The prediction is thus that the DDC model's effects should be sensitive to effective population sizes, particularly when these approximate the reciprocal of the rates of inactivating mutations (which might be around $10^{-6} - 10^{-5}$ per generation).

The DDC model has been of considerable interest to those studying the changes to gene functions following gene duplication events. In particular, the events following the genome duplication event in the teleosts have been examined for evidence of subfunctionalisation. Prince and Pickett (2002) showed that the *Hoxb1a* and *Hoxb1b* genes of the zebrafish (*D. rerio*) collectively have an expression pattern that duplicates that of the ancestor – inferred from the expression pattern in mouse – and that, of two sequences controlling the ancestral gene's expression, a autoregulatory sequence 5′ to the gene and a retinoic acid response element 3′ to the coding sequence, the *Hoxb1a* and *Hoxb1b* genes have retained the former and the latter, respectively. The whole genome duplication of the teleosts will supply a rich database for the identification of further cases of subfunctionalisation (Hurley *et al.*, 2005). Is this sharing of functions, expected from the DDC model, typical of the outcomes following this genome duplication? Chiu *et al.* (2002) examined controlling sequences in the HoxA cluster, and identified those conserved between humans and the horn shark (*Heterodontus francisci*). Do these controlling sequences get partitioned between the two zebrafish HoxA clusters? It was found that while there was very high conservation of cluster organisation between shark and human, the duplicated zebrafish clusters were far more diverged in structure. The human–shark-conserved putative cis-acting elements showed no clear pattern of reciprocal partitioning between the zebrafish clusters. Similarly, in duplicated genes of *Arabidopsis*, some pairs show signs of subfunctionalisation and others of neofunctionalisation only (Duarte *et al.*, 2006). Rates of amino acid evolution in gene duplicates are often, and unsurprisingly, very unequal following duplication, a result consistent with either adaptive evolution or reduced constraint in one of the pair (Conant and Wagner, 2003).

The DDC model requires not only that an ancestral gene's function is shared among the duplicates, but that the events that create this functional partitioning are themselves 'degeneration'-neutral events spreading to fixation. However, an outcome in which functions are shared could equally come about through changes that are all individually adaptive. As we have seen above, the DDC functional sharing, itself hypothesised to result in gene duplicate retention, should be more probable in organisms with small population sizes. Yet Shiu *et al.* (2006) demonstrate that the mouse genome has retained more gene duplicates than the human genome. If one makes the fairly strong assumption that the effective size of the lineage leading to the mouse has typically been larger than the effective size of populations leading from the human–mouse common ancestor to humans, then this result is more consistent with selection leading to the retention of

duplicated genes, rather than neutrally spreading degenerative mutations followed by selection. However, an effect of this kind could equally have been created by an enhanced rate of gene duplication in the mouse lineage. However, more genes already duplicated prior to the human–mouse split have been retained in the mouse lineage, an observation which could not be explained in this way.

The teleosts are the most speciose vertebrate group, and as the presence of the extra genes might be seen as creating opportunities for adaptive change, it is tempting to explain the very large number of teleost species as a consequence of their genome duplication. In particular, it has been suggested that the Hox cluster duplications may be correlated with a burst of adaptive radiation in vertebrates (Wagner *et al.*, 2003). Crow and Wagner (2005) suggest that the effect of the teleost genome duplication on the generation of the high species number seen in extant teleosts came through teleosts having a slower rate of extinction than other fish groups. However, the finding that teleosts show this (or any other) property of evolutionary persistence, diversity or adaptability does not demonstrate that the genome duplications in any way caused or facilitated this evolutionary property.

13.3.6 Apparent Genetic Redundancy

One consequence of major gene duplications is that there may be more overlap between functions of genes in duplicated genomes. The results of mouse knockout experiments have often shown that the phenotypes of homozygous null mutations, of genes thought to be important through studies of the biochemical roles of their products, are slight or undetectable (Brookfield, 1997). But is there true redundancy between duplicated genes, in which the selective advantage or disadvantage of loss of a second gene is zero, given that the first gene is present? We have seen that, immediately after a gene duplication event that has spread to fixation by genetic drift, there will be true redundancy, in this sense. We can define partial redundancy as situations in which the fitness loss associated with simultaneous mutations in genes A and B is much greater than the sum of the fitness losses associated with mutations in A and B individually. However, provided the fitness effects of mutation in genes A and B are individually greater than the reciprocal of the effective population size, selection (albeit of a weak and subtle kind) can maintain both genes. In budding yeast, *S. cerevisiae*, the selective impact of gene deletions is less for genes that possess paralogues elsewhere in the genome, but this effect gradually attenuates as the divergence in sequence between the copies increases (Gu *et al.*, 2003).

One method that could be suggested for the creation of a subtle selective advantage required to maintain partial redundancy in genes A and B, deletions in which may not be individually associated with a visible phenotype, is that the possession of two genes may create phenotypic robustness, not just to mutation but to the effects of environmental changes (Kitano, 2004). While having extra gene copies might be expected to buffer an organism against the harmful effects of recessive somatic mutations, it is not clear that somatic mutations are a major force in determining organismal phenotypes. Does a developmental process with multiple genetic pathways being involved in the production of a particular phenotypic trait make the probability of production of the trait (and concomitant Darwinian fitness) higher in an unpredictable environment than would a developmental process involving a single pathway? In my view, a positive and general answer to this question cannot be derived from first principles and is an empirical matter

only. At present, we do not know whether, in general, partial genetic redundancy increases robustness to environmental insults in the way that it does (by definition) to genetic lesions.

One major problem is the identification of the evolutionarily realistic environment in which to carry out tests for the effects of mutants. Deletion strains of *S. cerevisiae* have been created in which, in individual strains, every open reading frame in the genome has been inactivated (Winzeler *et al.*, 1999). Many of these knockouts show no effect on growth rate in a range of laboratory environments, but the conclusion that this represents true genetic redundancy cannot be drawn, since the fitness measurement is imprecise and the appropriate environment might not have been used (Giaever *et al.*, 2002). In fact, systematic characterisation of the visible phenotype of these deletion strains reveals morphological abnormalities even in deletion strains lacking a growth phenotype (Ohya *et al.*, 2005).

13.4   POPULATION GENETICS AND THE GENOME

13.4.1   The Impact of Chromosomal Position on Population Genetic Variability

Does the position of genes have any effect on their evolution? Comparison of the level of genetic polymorphism between *Drosophila* genes has shown that there is a strong correlation between the rate of recombination per unit DNA length and the level of DNA sequence polymorphism, as measured, for example by the gene diversity, π (Begun and Aquadro, 1992; Charlesworth and Wright, 2001). Since there is much smaller apparent effect of the local recombination rate on the divergence between *Drosophila* species, the correlation observed with diversity cannot be completely explained by a directly mutagenic effect of recombination. Rather, it must result from alleles sampled from low-recombination regions having recent shared descent. A number of processes have been suggested to explain how this could have come about.

One of these is a selective sweep. If a new advantageous mutation arises in a non-recombining region, it will arise initially in a particular haplotype and, in spreading, it will take with it this haplotype – extending over a length of chromosome which will be longer the lower is the effective recombination rate (Barton, 2000). This idea has been used to identify the selective sweeps in progress in the human genome, where rapidly spreading, new advantageous alleles generate 'extended haplotype homozygosity' (Sabeti *et al.*, 2002) – the sign of a haplotype which, while at high frequency, shows strong linkage disequilibrium extending into flanking markers – a pattern indicative of the recent rapid spread of the allele from a single progenitor. While sweeps undoubtedly happen, their impact is conditional upon a particular model for adaptive change, in which the rate-limiting step in adaptive allelic substitution is the generation of new mutations. However, another view of adaptive change is of weakly deleterious mutations being generated constantly and being present in an equilibrium population in a mutation–selection balance. An environmental change then changes the sign of the selection coefficient, creating an advantage for some formerly deleterious, phenotypically equivalent mutations. The advantageous alleles spread but, since they exist at the start of the sweep in a variety of different genetic backgrounds, the selective 'sweep' will create less reduction in diversity than would a more classical example of a sweep, starting from a unique mutation. Sweeps

of this kind have been called *soft sweeps* (Hermisson and Pennings, 2005; Pennings and Hermisson, 2006).

While selective sweeps are an effect on diversity resulting from an advantageous mutation, deleterious mutations will also tend to reduce diversity in low-recombination regions, in diverse ways. The first mechanism is 'background' selection (Charlesworth, 1996). Here mutations that are disadvantageous (and with selection coefficients greater than $1/N_e$) will eliminate chromosomes from the population, reducing the effective population size for the chromosomes that remain, and thus reducing neutral variability, and, as with selective sweeps, the size of the reduction generated increases as the local recombination rate drops. Even if the strength of the selection effect is weak, $s < 1/N_e$, if there are very many linked sites subject to selection, the neutral diversity will be reduced (McVean and Charlesworth, 2000) through what has been called the *Hill–Robertson effect* (Hill and Robertson, 1966). Finally, the process of Muller's ratchet (in which, in the absence of recombination, the occasional random irreplaceable loss from the population of the chromosomal class with the smallest number of deleterious mutations will steadily erode fitness) will also cause a reduction in the standing level of variability relative to a neutral population (Gordo *et al.*, 2002).

13.4.2 Codon Usage Bias

The genetic code is degenerate and thus allows different codons to be used to specify the same amino acid. However, not all codons for any given amino acid are used equally frequently, and since the initial finding of such biased codon usage, controversy has existed as to whether the bias is the result of mutation bias or of selection. There are many reasons why selection might be expected to generate a biased set of codons. Those that have been most favoured have suggested that codons are preferred by natural selection if they are complementary to anticodons on abundant tRNA species in the cell. This result was first demonstrated by Ikemura (1981; 1982) for strongly expressed genes in *E. coli* and *S. cerevisiae*. In each species, genes that were weakly expressed were not found to show such strong bias. The bias in strongly expressed genes could potentially arise from selection for increased translation rate or increased accuracy (Akashi, 1994). While the correlation between preferred codons and tRNA abundances has not been directly demonstrated in other species, in *D. melanogaster* strongly expressed genes again appear to show stronger codon bias than do weakly expressed genes (Shields *et al.*, 1988). However, the strong bias in codon usage seen in *E. coli* and yeast is not ubiquitous in bacterial genomes, being seen in only around 70 % of genomes examined (Sharp *et al.*, 2005). These genomes with codon bias seem to be those that have been selected for rapid growth in optimal conditions and are also characterised by high numbers of ribosomal RNA and transfer RNA genes. Biased codon usage also arises from deviations from a 50 % G + C ratio, as in the actinomycete *Streptomyces coelicolor*, for example, with 72 % C–G base pairs (Bentley *et al.*, 2002). Mammalian genomes contain isochores, which are blocks of DNAs of divergent C + G content (Bernardi *et al.*, 1985), the origin of which may be through mutational rather than selective causes.

For weakly expressed genes that fail to show codon bias, the explanation for their lack of bias is not selection to maintain particular non-optimal codons, since the rate of sequence evolution is high at these sites (Sharp and Li, 1989), rather, it appears that any selection operating is too weak to affect the choice of codons. For selection to be effectively neutral, the selection coefficient, *s*, has to be less than the reciprocal of the

effective population size, N_e (Bulmer, 1991). For stronger selection, the degree of codon bias depends on the product sN_e. This product can be estimated from the frequencies of preferred and unpreferred bases at polymorphic silent sites. In *Drosophila simulans*, it is estimated that sN_e is around 1, on average, implying that very weak selection is operating at these sites in this species (Akashi, 1999). If selection in *D. simulans* is only just strong enough to have a detectable effect, one would expect that in *D. melanogaster*, whose lower levels of all types of polymorphism when compared to *D. simulans* imply a smaller historical effective population size, the selection acting on silent sites might not be enough to maintain codon bias. Indeed, the codon bias in *D. melanogaster* is substantially reduced relative to *D. simulans*, and the majority of substitutions in the *D. melanogaster* lineage have been away from preferred codons. It appears that *D. melanogaster* has not reached equilibrium between mutation, selection and drift in its codon usage patterns (Akashi, 1996).

The evidence for selection on codon bias in mammals is less clear cut, but a variety of analyses have pointed to weak selection affecting a minority of synonymous codon variants, even in humans, where the effective population size has recently been particularly small (Chamary *et al.*, 2006; Comeron, 2006).

13.4.3   Effective Population Size

One aspect of population genetics that is of great relevance to the evolution of genome structure is that the impact of purifying selection depends on the strength of selection relative to the effective population size. Specifically, the probability of spread of a mutation with selective advantage s (in the heterozygous state) is given by

$$P = \frac{1 - \exp(-2sN_e/N)}{1 - \exp(-4N_e s)},$$

where N and N_e represent the census population size and the effective population size, respectively (Kimura, 1962). This relationship is important in genome evolution, as we have already seen in the determination of codon usage bias, in the likelihood of subfunctionalisation and in the likelihood that weakly deleterious intron insertions might spread to fixation. Indeed, Lynch (2006) has identified the consequences of the low effective population sizes of multicellular organisms (in particular, the low effective size relative to the reciprocal of mutation rates per generation) as being the cause of the majority of the differences in genome structure between multicellular organisms and micro-organisms. Such a view identifies the fixation process for changes to genome structure as typically being genetic drift acting on effectively neutral variants. Otto and Yong (2002), however, stress that changes to genome structure – in particular, gene duplicates, arise initially in single individuals, and their probability of spread will be greatly enhanced if they bring an instant selective benefit to their bearers (which could occur by a variety of mechanisms).

Effective population sizes may change with time. For humans, e.g. while the current census size is around 7×10^9, the diversity of the human population is more congruent with an effective size of 10 000. Here the appropriate effective size is the harmonic mean effective size over the timescale during which population genetic diversity has been generated (which is of the order of the last $4N_e$ generations). However, in explaining genome structure in terms of effective population sizes, the appropriate time window

over which N_e should be averaged is very much longer, which means that present census sizes of species with contrasting genome structures may be a very poor guide to their evolutionarily realistic $N_e s$ in terms of their genome organisation.

13.5   MOBILE DNAs

13.5.1   Repetitive Sequences

Repetitive sequences can be crudely divided into two broad classes on the basis of their pattern of repetition. Some sequences are found in tandem arrays, while others are found as single copies interspersed with unrelated sequences. The tandemly arrayed sequences can be apparently functionless, such as the microsatellite, minisatellite or longer satellite arrays, or functional, such as the ribosomal RNA genes and the histone genes. Interspersed repeats are shown, by their interspersion pattern itself, to be either currently active mobile DNAs or the descendants of sequences that have been mobile in the past. The mode of transposition is used to distinguish between what are named *Class I elements*, or *retrotransposons*, which move via an RNA intermediate, and Class II elements, which transpose as DNA. Both the RNA and DNA routes give opportunities for transposable elements to increase in the genome as a result of transposition (Brookfield, 1995). This ability to overreplicate has led to the widespread view that these elements are parasitic or selfish DNAs, maintaining their abundance in the genome through a balance between replicative transposition and the combined effects of deletion and of selection against their bearers.

In humans, the interspersed repetitive DNA sequences are dominated by Class I sequences called long interspersed nuclear element (*LINE*) and *SINE* sequences, of which there are very many copies. The long-term existence of these sequences at their current chromosomal locations means that they have almost all been inactivated, either at the moment they were inserted, or subsequently, by what has been called the *pseudogene effect* (McAllister and Werren, 1997). (Selection will not act to maintain sequences required for transposability in a sequence copy at a given genomic location, and individual copies will decay, eliminating their ability to act as donors for new insertions elsewhere in the genome.) Selection for active subfamilies operates through the ability of these families to replicate preferentially through the generation of daughter elements. Once inactivated, sequences independently (save for gene conversion) accumulate copy-specific patterns of base changes, and, from their divergence from the most closely related sequences, it is possible to identify when these sequences were inserted into their current chromosomal locations. This therefore allows the reconstruction of the history of the active elements of a mobile sequence family, as the 'extinct' subfamilies still exist in the genome.

The selfish nature of mobile DNAs does not preclude their evolving to serve roles that are functional for the host, either through 'domestication' or through their abilities as generators of diversity. Very many examples of elements serving adaptive roles are now known, and the number of examples can only increase as our knowledge grows, without threatening the paradigm that it is the capacity for selfish spread of elements through overreplication in transposition that explains the abundance of the majority of elements (Brookfield, 2005).

13.5.2 Selfish Transposable Elements and Sex

The concept of transposable elements as purely selfish DNA sequences faces major problems in the case of bacterial and other clonal populations. In clonally reproducing organisms, any individuals lacking selfish DNA sequences will inexorably outcompete those possessing the sequences, and replicative transposition of selfish elements in lineages possessing them will merely hasten this outcome. It is interesting to note that the purely asexual Bdelloid rotifers lack actively mobile retrotransposing DNAs (Arkhipova and Meselson, 2000). The transition to asexuality in this group may have been successful only because the asexual founders lacked active retroelements at the transition to clonality (Arkhipova and Meselson, 2005).

If selfish spread of mobile DNAs is impossible in clonal populations and if the mobile DNAs lack selectable trans-acting functions (as do bacterial insertion sequences), the obvious way in which these sequences could be maintained is by selection at the level of the host for their role as mutators. In clonal populations evolving adaptively, the frequency of alleles increasing the mutation rate is expected to be higher than in sexual populations, since mutator alleles are spread by hitch-hiking with the advantageous alleles that they generate (Taddei *et al.*, 1997). By a corresponding mechanism involving linkage disequilibrium, Edwards and Brookfield (2003) have produced a model for the selective maintenance of mobile DNAs through their ability to create null mutations which can create an environment-dependent selective advantage but can subsequently revert. Null mutations creating selective advantages were demonstrated in *E. coli* in a companion paper (Edwards *et al.*, 2002). Similarly, in experimental populations of *E. coli*, allowed to adapt to laboratory culture for more than 20 000 generations (Schneider and Lenski, 2004), many of the advantageous mutations were generated by insertion sequences (Schneider *et al.*, 2000).

13.5.3 Copy Number Control

The selfish spread of elements would be expected to continue indefinitely unless the elements themselves evolve a self-limitation in transposition or unless a balance with selection is generated. Examples of self-limitation are known, as in the restriction of the *D. melanogaster I* element's movements by homology-dependent gene silencing (Jensen *et al.*, 1999). However, Charlesworth and Langley (1986) showed that the area of parameter space over which harmful transposable elements would be expected to evolve self-limitation is extremely restricted. However, it has been shown (Brookfield, 1991; 1996) that, if the harmful effects of mobile DNAs arise through the process of transposition itself (rather than through the actions of elements after they have inserted in the genome), the evolution of self-limited transposition is more likely. One way in which copy number could be stabilised is through ectopic recombination between element copies from diverse locations (Langley *et al.*, 1988). Such a process is likely to be most effective in organisms such as *Drosophila*, where there is very high variation between individuals in the positions of their mobile DNAs, such that, in a typical meiosis, an element will find itself without a partner on the homologue.

13.5.4 Phylogenies of Transposable Elements

As there are many copies of a given mobile sequence family in the genome, it is possible to estimate the phylogeny of copies of a repetitive sequence family from the DNA sequence differences between copies. It is also possible to examine members of the same family

from other species and draw a tree including copies from both species. If the time to common ancestry of elements within each of two species is more recent than is the time to common ancestry of sequences from the two species, the elements will be seen to show 'concerted evolution', where species-specific substitutions will distinguish between the collections of elements from the two species (Figure 13.1a). One way in which variation between mobile DNAs from different species has been analysed is through the creation of consensus sequences for the sequences within each host species and comparison between these consensus sequences being used to examine between-species evolution of the family. However, there are problems with such an approach, illustrated in Figure 13.1. Figure 13.1(a) illustrates a tree showing three mobile DNAs sampled from each of two species, A and B, with $t1$, $t2$ and $t3$ representing times in generation between bifurcations in the ancestry of the tree. Here there is a clear split in the tree separating the sequences from species A and those from species B. This is the type of tree expected if the time to common ancestry of sequences in species A and the time to common ancestry of sequences in species B were each more recent than the split between the two species, and thus concerted evolution will be seen. The time to common ancestry of any A sequence with any B sequence is $(t1 + t2 + t3)$. The mean time to common ancestry for any two A sequences (or any two B sequences) is $t1 + 2t2/3$. In the absence of back or recurrent mutation (i.e. assuming an infinite sites model) the consensus sequence for species A would correspond to that ancestral to the most closely related pair of A sequences, and the consensus sequence of species A and the consensus of species B would be separated by $2(t2 + t3)$ generations of evolution. Figure 13.1(b) represents a different situation, where the first split in the tree pre-dated the split of species A and B. However, the representation of the two clades defined by this initial split is different in the two species (or at least in the small samples drawn from the two species). The result is that now, while the difference between the consensus sequence is still created by $2(t2 + t3)$ generations of evolution, A sequences share common ancestry, on average, $t1 + 2(t2 + t3)/3$ generations ago, while A and B sequences share ancestry, on average, $t1 + t2 + 5t3/9$ generations ago. The same difference in sequence between the consensuses masks a much-reduced ratio of the between- to within-species diversity. Figure 13.1(c) shows a case where the mean time to common ancestry within elements within species is $t1 + t2 + 2t3/3$ generations, while that between species is $t1 + (7t2 + 5t3)/9$ generations. Here there is no reason to suspect any difference in transposable element populations between the species, and the mean time to common ancestry within species in this tree is greater (for any $t1$, $t2$ and $t3$ values) than that between species, and yet there is a difference between the species' consensus sequences corresponding to $2t3$ generations of evolution.

What determines the time to common ancestry of copies of a repetitive sequence family within a given genome? In *D. melanogaster*, many studies have shown that transposable element site frequencies approximately follow the infinite alleles distribution (e.g. see Biemont *et al.*, 1994). A model that would, at least approximately, generate such a distribution would be one in which duplicative transpositions into a very large number of potential target sites could occur, with a process of genetic drift with weak selection generating the infinite alleles distribution at equilibrium. Such a model was initially explored by Langley *et al.* (1983). Suppose that we have an effective population size of N_e and that elements transpose duplicatively at a rate of v per element per generation. It is assumed that v represents the rate of insertion into neutral sites, and that at equilibrium this also represents the rate of deletion of the elements. If the number of available
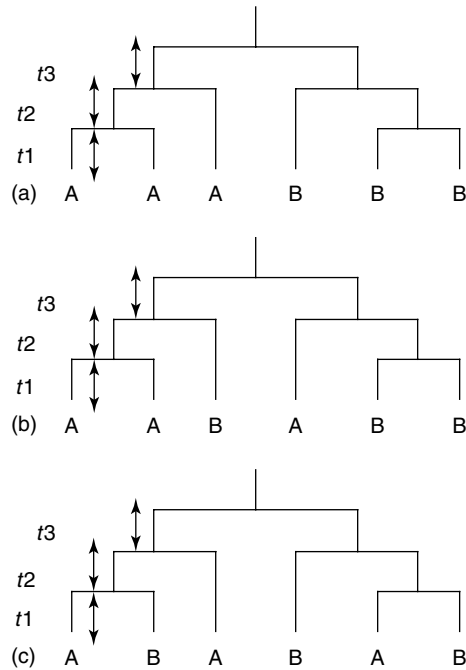
Figure 13.1   Three possible phylogenetic relationships between six transposable elements sampled from two species, A and B. $t1$, $t2$ and $t3$ represent the times to bifurcations in the tree. The extent of divergence of elements from the two species diminishes in going from (a) to (c). For each of the trees, (a), (b) and (c), it is possible to calculate the mean time to shared ancestry for sequences from A (owing to symmetry the mean time for sequences from B will be the same). It is also possible to calculate the mean time to common ancestry for pairs of sequences, one from A and one from B. The evolutionary time separating the consensus sequences can also be calculated as a function of $t2$ and $t3$ in each case.

sites for transposable element insertion is very high, the equivalence of the rates of element insertion and deletion will have the consequence that the frequency distribution of transposable elements will follow an infinite allele distribution with parameter $4N_e v$.

This model was extended by the author (Brookfield, 1986) to the expected time to common ancestry of randomly chosen transposable elements from different genomic locations, on the basis of an assumption of functional equivalence (as measured by transposition probability) of all elements at all sites. This model showed that the expected time in generations to common ancestry, t, was approximately given by

$$t \approx \frac{(n(1 + 4N_e v))}{2v}, \tag{13.1}$$

where n is the number of element copies per haploid genome. Similar results were found by others (Slatkin, 1985; Ohta, 1985).

This formula gives intuitively reasonable time estimates for the *Drosophila* transposable elements, but is clearly inappropriate for mobile DNAs from, e.g. vertebrates. One reason for this is that there will be very many element insertions which are inactive and the rate

of turnover of the active subset of elements would be greater than that predicted by this formula. Indeed, if only a proportion A of new insertions are active, the corresponding formula to (13.1) is

$$t \approx \frac{(2(1-A) + An(1+4N_e v))}{2v}. \tag{13.2}$$

The pseudogene effect leads to a different kind of model, in which the inactivation process can affect all element copies equally, rather than being restricted to elements inactivated at the moment of their insertion (so-called dead-on-arrival elements). A pseudogene effect further increases the rate of turnover of active elements and relates the rate of turnover of active elements to the proportion of the elements in a gene family that are active. Suppose now that there is a rate of inactivation of κ per element per generation and a rate of deletion of d per element per generation and, thus, a rate of replicative transposition into neutral sites of $(\kappa + d)$ per active element per generation at equilibrium. The expected time to common ancestry of two randomly chosen copies is

$$t \approx \frac{3}{(2\,d)} + \frac{n_a(1+4N_e\kappa)}{2\kappa}, \tag{13.3}$$

where n_a is the number of active elements in the genome (with $n_a >> 1$) and $\kappa >> d$.

However, these considerations of expected times to common ancestry mask another important aspect of transposable element phylogenies, which is the expected phylogeny of transposable element copies sampled from a genome. It had been noted that, for both LINEs and SINEs in mammals, the phylogeny resembles that expected under a 'master gene' model (Deininger $et\ al.$, 1992; Clough $et\ al.$, 1996) in which only a single element serves as the template for all transpositions. The result is a phylogeny in which all the bifurcations occur in a single branch.

Thus, we (Brookfield and Johnson, 2006) considered the expected phylogeny of mobile DNA sequences in an equilibrium model, where we imagine that the sample of copies is taken of members of a mobile sequence family from the genome, and we are interested in the phylogenetic tree connecting these sequences back to their common ancestor. We imagine a mobile sequence family, which was in equilibrium between three processes, replicative transposition of active elements, inactivation of elements (which may occur at the moment of transposition or result from inactivating mutations $in\ situ$ – the pseudogene effect) and deletion of elements. The relative rates of these processes define the expected numbers of copies of active and of inactive elements of this family in the genome, and this population of elements can then be seen as a population from which our sample and their ancestors can be traced following a coalescent approach. In large gene families the majority of elements will be inactive, and so we imagine the coalescent process for a sample of inactive elements. In the ancestry of our sample, tracking backwards in the standard approach of the coalescent, two types of events can occur – activations and coalescences. In the simplest model, where inactivation occurs via a pseudogene effect and never at the moment of transposition itself, coalescence in the ancestry of the sample can only occur between active lineages. Thus, in tracing the sample back to the common ancestor, there will be activation events, and coalescences. The expected shape of the phylogeny thus depends very greatly on the relative size of the two variables determining the probability of activation of a lineage, d (which is, at equilibrium, equal to the rate of deletion), and the probability of coalescence between two active lineages, which we

call T. If $d >> T$, the ancestors rapidly activate and slowly coalesce, which means that at the time of coalescences, there will be very many active lineages and the phylogeny will be complex (with coalescence events affecting many branches). Alternatively, if $T >> d$, coalescence follows rapidly for any pair of ancestral lineages that have activated, with the result that most coalescences involve a single branch. The tree is now as expected if there is a single, active master gene. The relative sizes of T and d are also related to the proportion of elements in the equilibrium population that are active. In the simplest case of pseudogene effect inactivations, having $T > d$ and, thus, a master gene–like tree, is associated with an equilibrium where the number of active elements is less than the square root of the total number.

13.5.5   Functional Variation between Element Copies

In addition to the presence of active and inactive elements and the impact of their presence on the expected element phylogeny, another important issue here is the possibility of advantageous element copies arising in a mobile sequence family. The advantage for such variants would typically be an increased rate of replicative transposition. With a large number of active elements per genome, n, and $2N$ genomes in the host population, one might expect new, advantaged variants to arise frequently in this population of $2Nn$ elements. A new variant of this kind would be expected to sweep through the sequence family in a two-speed process, initially replacing all active elements and subsequently replacing inactive elements much more slowly. However, the probability of spread of a new variant might be low (as a result of a low intrinsic rate of replicative transposition). Imagine that the rate of transposition per active copy is 10^{-3}, e.g. and that there is equilibrium between transposition and the combined effects of inactivation and deletion, such that an active element's probability of loss or inactivation is the same as its probability of copying itself. Now, imagine a more active variant, which has a probability of transposition raised by 10 % of its former value. Such a variant, instead of having, on average, one active descendant after one generation, would have, on average, 1.0001 descendants, and the probability that this variant would spread as a result of its advantage would only be approximately 0.0002. The important point is that the variant will arise initially in a single sequence copy, and its probability of spreading, rather than being lost by drift, may be very small.

13.6   CONCLUSIONS

This necessarily qualitative outline of some of the issues of genome evolution raises some questions concerning the forces that have shaped genome structures. Some of these processes, such as those concerned with mobile genetic elements and biases in codon usage, have long been subject to approaches involving quantitative modelling. Others, such as the effects of gene and genome duplications of gene families and the control of gene expression and its evolution, have mainly been dealt with in a descriptive and qualitative way and are only now being subjected to a quantitative approach. A major problem is whether, if we regard the process of genome duplication, for example, as an example of a Poisson process with an underlying expected rate, we will ever have enough data from the rare events which have actually taken place in evolving lineages to calculate what the underlying rate is likely to be. For many of these issues, the fundamental

problem is whether we can consider the state of the genome as being an equilibrium state in which a balance has been reached between dynamical evolutionary processes acting on the genome or whether the parameters of these underlying processes are themselves changing at such a high rate, relative to the rate of the individual events through which individual genomes change, that attempts to model equilibrium states on the basis of fixed parameters will be futile.

REFERENCES

Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**, 927–935.

Akashi, H. (1996). Molecular evolution between *Drosophila melanogaster* and *Drosophila simulans*: reduced codon bias, faster rates of amino acid substitution and larger proteins in *Drosophila melanogaster*. *Genetics* **144**, 1297–1307.

Akashi, H. (1999). Within- and between-species DNA sequence variation and the footprint of natural selection. *Gene* **238**, 39–51.

Amores, A., Force, A., Yan, Y.L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.L., Westerfield, M., Ekker, M. and Postlethwaite, J.H. (1998). Zebrafish hox clusters and vertebrate genome evolution. *Science* **282**, 1711–1714.

Amores, A., Suzuki, T., Yan, Y.L., Pomeroy, J., Singer, A., Amemiya, C. and Postlethwaite, J.H. (2004). Developmental roles of pufferfish Hox clusters and genome evolution in ray-finned fish. *Genome Research* **14**, 1–10.

Arabidopsis Genome Initiative (2000). Analysis of the Genome Sequence of the Flowering Plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.

Arkhipova, I. and Meselson, M. (2000). Transposable elements in sexual and ancient asexual taxa. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 14473–14477.

Arkhipova, I. and Meselson, M. (2005). Deleterious transposable elements and the extinction of asexuals. *Bioessays* **27**, 76–85.

Bailey, W.J., Kim, J., Wagner, G.P. and Ruddle, F.H. (1997). Phylogenetic reconstruction of vertebrate Hox cluster duplications. *Molecular Biology and Evolution* **14**, 843–853.

Barabasi, A.L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509–512.

Barton, N.H. (2000). Genetic hitch-hiking. *Philosophical Transactions of the Royal Society of London Series B* **355**, 1553–1562.

Begun, D.J. and Aquadro, C.F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. *Nature* **356**, 519–520.

Bentley, S.D., Chater, K.F., Cerdeño-Tárraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C.W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C.H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M.A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B.G., Parkhill, J. and Hopwood, D.A. (2002). Complete genome sequence of model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958.

Biémont, C., Lemeunier, F., Garcia Guerreiro, M.P., Brookfield, J.F., Gautier, C., Aulard, S. and Pasyukova, E.G. (1994). Population dynamics of the *copia, mdg1, mdg3, gypsy* and

P transposable elements in a natural population of *Drosophila melanogaster*. *Genetical Research* **63**, 197–212.

Blomme, T., Vanderpoele, K., De Bodt, S., Simillion, C., Maere, S. and Van der Peer, Y. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biology* **7**, R43.

Brookfield, J.F.Y. (1986). A model for DNA sequence evolution within transposable element families. *Genetics* **112**, 393–407.

Brookfield, J.F.Y. (1991). Models of repression of transposition in P-M hybrid dysgenesis by P-cytotype and by zygotically-encoded repressor proteins. *Genetics* **128**, 471–486.

Brookfield, J.F.Y. (1995). Transposable elements as selfish DNA. In *Mobile Genetic Elements-Frontiers in Molecular Biology*, D. Sherratt, ed. Oxford University Press, pp. 131–153.

Brookfield, J.F.Y. (1996). Models of the spread of non-autonomous selfish transposable elements when transposition and fitness are coupled. *Genetical Research* **67**, 199–209.

Brookfield, J.F.Y. (1997). Genetic redundancy. *Advances in Genetics* **36**, 137–155.

Brookfield, J.F.Y. (2005). The ecology of the genome-mobile DNA elements and their hosts. *Nature Reviews Genetics* **6**, 128–136.

Brookfield, J.F.Y. and Johnson, L.J. (2006). The evolution of mobile DNAs-when will transposons create phylogenies that look as if there is a master gene? *Genetics* **173**, 1115–1123.

Brown, T.A. (1999). *Genomes*. BIOS Scientific, Oxford.

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897–907.

Cain, A.J., King, J.M.B. and Sheppard, P.M. (1960). New data on the genetics of polymorphism in the snail *Cepaea nemoralis*. *Genetics* **45**, 393–411.

Chamary, J.V., Parmley, J.L. and Hurst, L.D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics* **7**, 98–108.

Charlesworth, B. (1996). Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genetical Research* **68**, 131–150.

Charlesworth, B. and Langley, C.H. (1986). The evolution of self-regulated transposition of transposable elements. *Genetics* **112**, 359–383.

Charlesworth, B. and Wright, S.I. (2001). Breeding systems and genome evolution. *Current Opinion in Genetics and Development* **11**, 685–690.

Chiu, C.-H., Amemiya, C., Dewar, K., Kim, C.B., Ruddle, F.H. and Wagner, G.P. (2002). Molecular evolution of the *HoxA* cluster in the three major gnathostome lineages. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5492–5497.

Clarke, C.A. and Sheppard, P.M. (1968). The genetics of the mimetic butterfly *Papilio memnon*. *Philosophical Transactions of the Royal Society of London Series B* **254**, 37–89.

Clarke, C.A. and Sheppard, P.M. (1969). Further studies on the mimetic butterfly *Papilio memnon*. *Philosophical Transactions of the Royal Society of London Series B* **263**, 35–70.

Cline, T.W. (1993). The *Drosophila* sex-determination signal: how do flies count to two? *Trends in Genetics* **9**, 385–390.

Clough, J.E., Foster, J.A., Barnett, M. and Wichman, H.A. (1996). Computer simulation of transposable element evolution: random template and strict master models. *Journal of Molecular Evolution* **42**, 52–58.

Coghlan, A. and Wolfe, K.H. (2002). Fourfold faster rate of genome rearrangements in nematodes than in *Drosophila*. *Genome Research* **12**, 857–867.

Coghlan, A. and Wolfe, K.H. (2004). Origins of recently gained introns in *Caenorhabditis*. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 11362–11367.

Collins, L. and Penny, D. (2005). Complex spliceosomal organization ancestral to extant eukaryotes. *Molecular Biology and Evolution* **22**, 1053–1066.

Comeron, J.M. (2006). Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 6940–6945.

Conant, G.C. and Wagner, A. (2003). Asymmetric sequence divergence of duplicate genes. *Genome Research* **13**, 2052–2058.

Cooper, S.J.B., Wheeler, D., Hope, R.M., Dolman, G., Saint, K.M., Gooley, A.A. and Holland, R.A.B. (2005). The alpha-globin gene family of an Australian marsupial, *Macropus eugenii*: the long evolutionary history of the *theta-globin* gene and its functional status in mammals. *Journal of Molecular Evolution* **60**, 653–664.

Crow, K.D. and Wagner, G. (2005). What is the role of genome duplication in the evolution of complexity and diversity? *Molecular Biology and Evolution* **23**, 887–892.

Csuros, M. and Miklos, I. (2006). A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. *Research in Computational Molecular Biology, Proceedings Lecture Notes in Computer Science* **3909**, 206–220.

Dehal, P. and Boore, J.L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* **3**, e314.

Deininger, P.L., Batzer, M.A., Hutchinson, C.A. and Edgell, M.H. (1992). Master genes in mammalian repetitive DNA amplification. *Trends in Genetics* **8**, 307–311.

Dobrindt, U., Hochhut, B., Hentshel, U. and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology* **2**, 414–424.

Duarte, J.M., Cui, L.Y., Wall, P.K., Zhang, Q., Zhang, X.H., Leebens-Mack, J., Ma, H., Altman, N. and dePamphilis, C.W. (2006). Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Molecular Biology and Evolution* **23**, 469–478.

Edwards, R.J. and Brookfield, J.F.Y. (2003). Transiently beneficial insertions could maintain mobile DNA sequences in variable environments. *Molecular Biology and Evolution* **20**, 30–37.

Edwards, R.J., Sockett, R.E. and Brookfield, J.F.Y. (2002). A simple method for genome-wide screening for advantageous insertions of mobile DNAs in *Escherichia coli*. *Current Biology* **12**, 863–867.

Farea, T.L., Botstein, D., Brown, P.O. and Rozensweig, R.F. (1999). Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 9721–9726.

Ford, E.B. (1964). *Ecological Genetics*. Methuen, London.

Force, A., Amores, A. and Postlethwaite, J.H. (2002). *Hox* cluster organization in the jawless vertebrate *Petromyzon marinus*. *Journal of Experimental Zoology* **294**, 30–46.

Force, A., Cresko, W.A., Pickett, F.B., Proulx, S.R., Amemiya, C. and Lynch, M. (2005). The origin of subfunctions and modular gene regulation. *Genetics* **170**, 433–446.

Force, A., Lynch, M., Pickett, B., Amores, A., Yan, Y.-L. and Postlethwaite, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545.

Fried, C., Prohaska, S.J. and Stadler, P.F. (2003). Independent Hox-cluster duplications in lampreys. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **229B**, 18–25.

Frugoli, J.A., McPeek, M.A., Thomas, T.L. and McClung, C.R. (1998). Intron loss and gain during evolution of the catalase gene family in angiosperms. *Genetics* **149**, 355–365.

Garcia-Fernandez, J. and Holland, P.W.H. (1994). Archetypal organization of the amphioxus *Hox* gene cluster. *Nature* **370**, 563–566.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A.P., Astromoff, A., El Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanero, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.D., Flaherty, P., Foury, F., Garfinkel, D.J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J.H., Hempel, S., Herman, Z., Jaramillo, D.F., Kelly, D.E., Kelly, S.L., Kotter, P., LaBonte, D., Lamb, D.C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C.Y., Lussier, M., Mao, R., Menard, P., Ooi, S.L., Revuelta, J.L., Roberst, C.J., Rose, M., Ross-MacDonald, P., Scherens, B., Schimmack, G.,

Shafer, B., Shoemaker, D.D., Sookhai-Mahadeo, S., Storms, R.K., Strathern, J.N., Valle, G., Voet, M., Volckaert, G., Wang, C.Y., Ward, T.R., Wilhelmy, J., Winzeler, E.A., Yang, Y.H., Yen, G., Youngman, E., Yu, K.X., Bussey, H., Boeke, J.D., Snyder, M., Phillipsen, P., Davis, R.W. and Johnston, M. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.

Gibson, G. and Muse, S.V. (2004). *A primer in genome science*, 2nd edition. Sinauer, Sunderland, MA.

Gilbert, W. (1987). The exon theory of genes. *Cold Spring Harbor Symposia on Quantitative Biology* **52**, 901–905.

Gogarten, J.P. and Townsend, J.P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology* **3**, 679–687.

Gordo, I., Navarro, A. and Charlesworth, B. (2002). Muller's ratchet and the pattern of variation at a neutral locus. *Genetics* **161**, 835–848.

Gu, Z., Steinmetz, L.M., Gu, X., Scharfe, C., Davis, R.W. and Li, W.-H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63–66.

Hacker, J., Blum-Oehler, G., Muhldorfer, I. and Tschape, H. (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on bacterial evolution. *Molecular Microbiology* **23**, 1089–1097.

Hahn, H.W., De Bie, T., Stajich, J.E., Nguyen, C. and Cristiani, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research* **15**, 1153–1160.

Halligan, D.L. and Keightley, P. T. (2006). Ubiquitous selective constraints in the Drosophila genome revealed by a genome wide-interspecies comparison. *Genome Research* **16**, 875–884.

Hankeln, T., Friedl, H., Ebersberger, I., Martin, J. and Schmidt, E.R. (1997). A variable intron distribution in globin genes of *Chironomus*: evidence for recent intron gain. *Gene* **205**, 151–160.

Hermisson, J. and Pennings, P.S. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335–2352.

Hill, W.G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294.

Holland, P.W.H., Garcia-Fernandez, J., Williams, N.A. and Sidow, A. (1994). Gene duplication and the origins of vertebrate development. In *The Evolution of Developmental Mechanisms*, M. Akam, P. Holland, P. Ingham and G. Wray, eds. Company of Biologists, Cambridge, pp. 125–133.

Hughes, A.L. (1999). Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *Journal of Molecular Evolution* **48**, 565–576.

Hughes, M.K. and Hughes, A.L. (1993). Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Molecular Biology and Evolution* **10**, 1360–1369.

Hughes, A.L. and Friedman, R. (2005). Gene duplication and the properties of biological networks. *Journal of Molecular Evolution* **61**, 758–764.

Hurley, I., Hale, M.E. and Prince, V.E. (2005). Duplication events and the origin of segmental identity. *Evolution and Development* **7**, 556–567.

Hurst, G.D.D. and Schilthuizen, M. (1998). Selfish genetic elements and speciation. *Heredity* **80**, 2–8.

Hurst, L.D., Pal, C. and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nature Reviews Genetics* **5**, 299–310.

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *Escherichia coli* translation system. *Journal of Molecular Biology* **151**, 389–409.

Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the abundance of the respective codons in protein genes: differences in synonymous codon choice patterns

of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *Journal of Molecular Biology* **158**, 573–597.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945.

Itoh, T., Takemoto, K., Mori, H. and Gojobori, T. (1999). Evolutionary comparison of operon structures disclosed by sequence comparisons of complete microbial genomes. *Molecular Biology and Evolution* **16**, 332–346.

Iwamoto, M., Maekawa, M., Saito, A., Higo, H. and Higo, K. (1998). Evolutionary relationship of plant catalase genes inferred from intron-exon structures: isozyme divergence and the separation of monocots and dicots. *Theoretical and Applied Genetics* **97**, 9–19.

Jeffares, D.C., Mourier, T. and Penny, D. (2006). The biology of intron gain and loss. *Trends in Genetics* **22**, 16–22.

Jeffs, P. and Ashburner, M. (1991). Processed pseudogenes in *Drosophila*. *Proceedings of the Royal Society of London Series B* **244**, 151–159.

Jensen, S., Gassama, M.-P. and Heidmann, T. (1999). Taming of transposable elements by homology-dependent gene silencing. *Nature Genetics* **21**, 209–212.

Johnnidis, J.B., Venanzi, E.S., Taxman, D.J., Ting, J.P.Y., Benoist, C.O. and Mathis, D.J. (2005). Chromosomal clustering of genes controlled by the aire transcription factor. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 7233–7238.

Jukes, T.H. and Cantor, C.R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, H.N. Munro, ed. Academic Press, New York, pp. 21–32.

Karev, G.P., Wolf, Y.I., Berezovskaya, F.S. and Koonin, E.V. (2004). Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evolutionary Biology* **4**, Art No 32.

Karev, G.P., Wolf, Y.I. and Koonin, E.V. (2003). Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics* **19**, 1889–1900.

Kellis, M., Birren, B.W. and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **19**, 617–624.

Kettlewell, H.B.D. (1955). Selection experiments and industrial melanism in the *Lepidoptera*. *Heredity* **9**, 323–342.

Khanin, R. and Wit, E. (2006). How scale-free are biological networks? *Journal of Computational Biology* **13**, 810–818.

Kimura, M. (1962). On the probability of fixation of mutant genes in populations. *Genetics* **47**, 713–719.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics* **5**, 826–837.

Langley, C.H., Brookfield, J.F.Y. and Kaplan, N.L. (1983). Transposable elements in Mendelian population. I A theory. *Genetics* **104**, 457–471.

Langley, C.H., Montgomery, E., Hudson, R.R., Kaplan, N. and Charlesworth, B. (1988). On the role of unequal exchange in the containment of transposable element copy number. *Genetical Research* **52**, 223–235.

Lawrence, J.G. (1997). Selfish operons and speciation by gene transfer. *Trends in Microbiology* **5**, 355–359.

Lawrence, J.G. (1999). Gene transfer, speciation, and the evolution of bacterial genomes. *Current Opinion in Microbiology* **2**, 519–523.

Lawrence, J.G. and Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 9413–9417.

Lawrence, P.A. 1992. *The Making of a Fly*. Blackwell Scientific Publications, Oxford, p. 9.

Li, Q., Harju, S. and Peterson, K.R. (1999). Locus control regions: coming of age at a decade plus. *Trends in Genetics* **15**, 403–408.

Li, W.-H. (1997). *Molecular evolution*. Sinauer, Sunderland, MA.

Long, M. and Langley, C.H. (1993). Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* **260**, 91–95.

Lynch, M. (2006). The origins of eukaryotic gene structure. *Molecular Biology and Evolution* **23**, 450–468.

Lynch, M. and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155.

Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459–473.

Lynch, M. and Richardson, A.O. (2002). The evolution of spliceosomal introns. *Current Opinion in Genetics and Development* **12**, 701–710.

Majerus, M.E.N. (1998). *Melanism: Evolution in Action*. Oxford University Press, Oxford.

Maquat, L.E. (2004). Nonsense-mediated mRNA decay: a comparative analysis of different species. *Current Genomics* **5**, 175–190.

McAllister, B.F. and Werren, J.H. (1997). Phylogenetic analysis of a retrotransposon with implications for strong evolutionary constraints on reverse transcriptase. *Molecular Biology and Evolution* **14**, 69–80.

McVean, G.A.T. and Charlesworth, B. (2000). The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* **155**, 929–944.

Mongold, J.A. and Lenski, R.E. (1996). Experimental rejection of a nonadaptive explanation for increased cell size in *Escherichia coli*. *Journal of Bacteriology* **178**, 5333–5334.

Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag, Berlin.

Ohta, T. (1985). A model of duplicative transposition and gene conversion of repetitive gene families. *Genetics* **110**, 513–524.

Ohya, Y., Sese, J., Yukawa, M., Sano, F., Nakatini, Y., Saito, T.L., Saka, A., Fukuda, T., Ishihara, S., Oka, S., Suzuki, G., Watanabe, M., Hirata, A., Ohtani, M., Sawai, H., Fraysse, N., Latge, J.P., Francois, J.M., Aebi, M., Tanaka, S., Muramatsu, S., Araki, H., Sonoike, K., Nogami, S. and Morishita, S. (2005). High-dimensional and large-scale phenotyping of yeast mutants. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 19015–19020.

Otto, S.P. and Yong, P. (2002). The evolution of gene duplicates. *Advances in Genetics Incorporating Molecular Genetic Medicine* **46**, 451–483.

Pal, C. and Hurst, L.D. (2003). Evidence for co-evolution of gene order and recombination rate. *Nature Genetics* **33**, 392–395.

Palmer, J.D. and Logsdon, J.M. Jr., (1991). The recent origins of introns. *Current Opinion in Genetics and Development* **1**, 470–477.

Panopoulou, G. and Poustka, A.J. (2005). Timing and mechanism of ancient vertebrate genome duplications- the adventure of a hypothesis. *Trends in Genetics* **21**, 559–567.

Paterson, A.H. (2006). Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nature Reviews Genetics* **7**, 174–184.

Patthy, L. (1999). Genome evolution and the evolution of exon-shuffling- a review. *Gene* **238**, 103–114.

Pennings, P.S. and Hermisson, J. (2006). Soft sweeps II- molecular population genetics of adaptation from recurrent mutation or migration. *Molecular Biology and Evolution* **23**, 1076–1084.

Petrov, D.A. and Hartl, D.L. (1998). High rate of DNA loss in *Drosophila melanogaster* and *Drosophila virilis* species groups. *Molecular Biology and Evolution* **15**, 293–302.

Petrov, D.A., Lozovskaya, E.R. and Hartl, D.L. (1996). High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**, 346–349.

Prince, V.E. and Pickett, F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. *Nature Reviews Genetics* **3**, 827–837.

Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G. and Koonin, E.V. (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology* **13**, 1512–1517.

Roy, S.W. and Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics* **7**, 211–221.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. and Lander, E.S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837.

Schneider, D., Duperchy, E., Coursange, E., Lenski, R.E. and Blot, M. (2000). Long-term experimental evolution in *Escherichia coli* IX Characterization of insertion sequence-mediated mutations and rearrangements. *Genetics* **156**, 477–488.

Schneider, D. and Lenski, R.E. (2004). Dynamics of insertion sequence elements during experimental evolution of bacteria. *Research in Microbiology* **155**, 319–327.

Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F. and Sockett, R.E. (2005). Variation in the strength of selected codon usage among bacteria. *Nucleic Acids Research* **33**, 1141–1153.

Sharp, P.M. and Li, W.-H. (1989). On the rate of DNA sequence evolution in *Drosophila*. *Journal of Molecular Evolution* **28**, 398–402.

Shields, D.C., Sharp, P.M., Higgins, D.G. and Wright, F. (1988). "Silent" sites in *Drosophila* genes are not neutral: evidence for selection among synonymous codons. *Molecular Biology and Evolution* **5**, 704–716.

Shiu, S.H., Byrnes, J.K., Pan, R., Zhang, P. and Li, W.-H. (2006). Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 2232–2236.

Slatkin, M. (1985). Genetic differentiation of transposable elements under mutation and unbiased gene conversion. *Genetics* **110**, 145–158.

de Souza, S.J., Long, M., Klein, R.J., Roy, S., Lin, S. and Gilbert, W. (1998). Towards a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 5094–5099.

Stadler, P.F., Fried, C., Prohaska, S.J., Bailey, W.J., Misof, B.Y., Ruddle, F.H. and Wagner, G.P. (2004). Evidence for Independent Hox gene duplications in the hagfish lineage: a PCR-based inventory of *Eptatretus stoutii*. *Molecular Phylogenetics and Evolution* **32**, 686–694.

Stumpf, M.P.H., Wiuf, C. and May, R.M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 4221–4224.

Taddei, F., Radman, M., Maynard Smith, J., Toupance, B., Gouyon, P.-H. and Godelle, B. (1997). Role of mutator genes in adaptive evolution. *Nature* **387**, 700–702.

The Institute of Genomic Research (TIGR) (2006). `http://www.tigr.org/`.

Wade, M.J. and Beeman, R.W. (1994). The population dynamics of maternal effect selfish genes. *Genetics* **138**, 1309–1314.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B.,Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., El Bakkoury, M., Foury, F., Friend, S.H., Gentalen, E., Giaever, G., Hegemann, J.H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D.J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J.L., Riles, L., Roberts, C.J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R.K., Veronneau, S., Voet, M., Volckaert, G., Ward, T.R., Wysocki, R., Yen, G.S., Yu, K.X., Zimmermann, K., Phillipsen, P., Johnston, M. and Davis, R.W. (1999). Functional characterization of the *Saccharomyces cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906.

Wagner, G.P., Amemiya, C. and Ruddle, F. (2003). Hox number duplications and the opportunity for evolutionary novelties. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 14603–14606.

Wolfe, K.H. and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713.

Wong, S. and Wolfe, K.H. (2005). Birth of metabolic gene cluster in yeast by adaptive gene relocation. *Nature Genetics* **37**, 777–782.

Xing, Y. and Lee, C. (2006). Alternative splicing and RNA selection pressures- evolutionary consequences for eukaryotic genomes. *Nature Reviews Genetics* **7**, 499–509.

Probabilistic Models for the Study of Protein Evolution

J.L. Thorne

Departments of Genetics and Statistics, North Carolina State University, Raleigh, NC, USA

and

N. Goldman

EMBL-European Bioinformatics Institute, Hinxton, UK

Mathematical models are indispensable tools for characterizing the process of protein evolution, many aspects of which are not easily amenable to direct experimentation. Ideally, a model of protein evolution would provide a good description of the data and would simultaneously be parameterized in a manner that facilitates biological insight. Probabilistic descriptions of protein change are especially valuable because they engender a sound basis for likelihood-based statistical inference, and can provide the foundation for local and global alignment, phylogeny reconstruction, and prediction of protein structure and function. In this chapter, models of protein evolution are reviewed. Their strengths and limitations are emphasized.

14.1   INTRODUCTION

The relationship between genotype and phenotype is central to biology. Proteins are at the heart of this relationship at its fundamental molecular level: the DNA coding for a protein is the genotype whereas a protein's structure and its pattern of expression are phenotype. Evidence that molecular biologists and biochemists have recognized the key role of proteins in this relationship can be found, for example, in the amount of money being spent to deduce all the protein-coding sequences (the proteome) for numerous species, the tremendous enthusiasm surrounding the development of microarray technologies for measuring gene expression, and the copious intellectual capital being invested in the

development of methods for determining protein tertiary structure from protein sequence data.

The relationship between genotype and phenotype is surely as important to the study of evolution as it is to the rest of biology. Largely because of the complexity of the genotype–phenotype relationship, the process of protein evolution remains poorly understood. Many probabilistic models of protein evolution have been proposed and explored, but a model that satisfactorily combines mutation, genetic drift, and natural selection is not yet on the horizon. Nevertheless, existing models provide a statistical foundation for characterizing the process of evolution and for inferring evolutionary history. While these models still have obvious flaws, there have been substantial improvements during the last two decades.

A heightened interest in the ability of probabilistic models to extract biological information has accompanied the improvements. Much recent interest is a result of the emerging field of comparative genomics. In addition to the desire to understand protein evolution on a genomic scale, there are also tasks in comparative genomics for which understanding evolution is not the goal but for which evolution should be carefully treated because common ancestry is responsible for correlations among genomes. The extent and nature of the correlation structure is determined by the phylogeny that relates species and the processes by which genomes have evolved on these phylogenies. Correlation between proteins due to common ancestry is the implicit underlying basis of established techniques of predicting protein function via sequence database searches, global sequence alignment, and homology modeling to predict protein structure. Common ancestry also plays a key role in promising techniques for predicting the phenotypic severity of new mutations and disease contributions of single nucleotide polymorphisms (Ng and Henikoff, 2001; 2002; Stone and Sidow, 2005). All of these techniques have the potential to be improved via better treatments of protein evolution.

In this chapter, we describe some of the important features of existing models of protein evolution. As well as focusing on widely applied models, we also highlight some of the most promising research directions. Molecular evolution and comparative genomics are no longer fields where the ability to collect data is the main obstacle to progress. Instead, we think that progress in these fields is mainly limited by a lack of adequate tools for extracting information from existing data and that improved evolutionary models will inevitably lead to better tools.

14.2   EMPIRICALLY DERIVED MODELS OF AMINO ACID REPLACEMENT

14.2.1   The Dayhoff and Eck Model

Dayhoff and Eck (1968; see also Dayhoff *et al.*, 1972; 1978) introduced the most influential model of protein evolution. It describes the process of amino acid replacement in terms of a continuous-time Markov process on a phylogenetic tree, and is based upon the assumption that all sites in a protein sequence evolve independently and identically. The independence assumption is convenient because it allows the likelihood of an aligned set of protein sequences to be written as a product of likelihoods for individual alignment columns (i.e. site likelihoods).

Site likelihoods can be calculated via the pruning algorithm of (Felsenstein, 1981), mentioned in **Chapter 15** Huelsenbeck and Bollback of this book. We do not repeat the algorithm here, but we emphasize that the pruning algorithm requires the calculation of transition probabilities. In molecular phylogenetics, a transition probability is the probability of a possible state at the end of a branch on a phylogenetic tree conditional upon the state at the beginning of the branch and the amount of evolution represented by the branch (i.e. the branch length), as well as upon values of other model parameters. For the Dayhoff and Eck model there are 20 possible states, each representing one of the 20 amino acids.

The parameters of the Dayhoff and Eck model are the instantaneous rates α_{ij} of replacement of amino acid type i by amino acid type j. Computation of transition probabilities from underlying instantaneous replacement rates of a Markov process is a standard procedure, described for example by Liò and Goldman (1998). The rates in Dayhoff and Eck's formulation are constrained so that the model is time-reversible. In other words,

$$\pi_i \alpha_{ij} = \pi_j \alpha_{ji} \quad \text{for all } i \text{ and } j, \tag{14.1}$$

where the equilibrium frequency of each amino acid type k is denoted π_k and is uniquely determined by the matrix of rates (α_{ij}). Informally, the time-reversibility property means that, without additional information, there is no way to determine which of two sequences is the ancestral protein and which is the descendant protein.

In fact, the Dayhoff and Eck model is the most general 20-state time-reversible homogeneous Markov model. Although four-state models of nucleotide substitution are computationally simpler, it is interesting to note that the most general four-state time-reversible model (see Tavaré, 1986; Yang, 1994a), and even the simplest four-state model (Jukes and Cantor, 1969), were not studied until after the Dayhoff and Eck model was proposed.

For models of nucleotide substitution, all parameters are typically estimated from the aligned data set of interest. The number of parameters defining a 20-state general time-reversible (GTR) rate matrix is 209, greatly exceeding the nine parameters needed to specify a four-state GTR rate matrix. As a consequence, estimates of α_{ij} for a 20-state rate matrix are unlikely to be reliable unless the amount of data being analyzed is large, and estimates of relative rates of replacement for Dayhoff–Eck-type approaches are usually obtained prior to analysis of the aligned data set of interest. These estimates are generally based upon many data sets of aligned protein sequences and are fixed when subsequent data sets of interest are studied.

Dayhoff and Eck estimated the instantaneous rates α_{ij} implicitly, via a clever technique that relied upon comparing closely related protein sequences to derive transition probability matrices. When protein sequences are closely related, they will be identical at the majority of positions. At the few positions where the sequences differ, the cause of the difference will usually be a single amino acid replacement event because the chance that two or more replacement events occurred at a position will be negligible. By comparing closely related sequences and neglecting the possibility of multiple amino acid replacements at positions, Dayhoff and Eck were able to infer implicitly the relative rates of the different kinds of amino acid replacements. Different methods for recovering the instantaneous rates α_{ij} from results such as those presented by Dayhoff and Eck (1968) were subsequently proposed, and are reviewed by Kosiol and Goldman (2005). Goldman

et al.(1998) gave an alternative procedure to estimate rates α_{ij} directly from amino acid replacement counts.

Improved techniques for estimating the α_{ij} parameters are now available (Adachi and Hasegawa, 1996; Yang *et al.*, 1998; Adachi *et al.*, 2000; Müller and Vingron, 2000; Whelan and Goldman, 2001; Holmes and Rubin, 2002; Müller *et al.*, 2002), and we refer the reader to the original Dayhoff and Eck publication (1968; see also Dayhoff *et al.*, 1972; 1978; Kosiol and Goldman, 2005) for other details about their estimation procedure. For an insightful analysis with the original Dayhoff data, see Kishino *et al.* (1990).

Subsequent to the Dayhoff and Eck (1968) publication, Dayhoff and collaborators published updated results as additional protein sequence data became available (Dayhoff *et al.*, 1972; 1978). Conventionally, references to the 'Dayhoff model' do not simply mean the 20-state GTR model of amino acid replacement but instead are usually intended to convey both the Dayhoff and Eck (1968) parameterization of the 20-state GTR model and specific estimates of the α_{ij} parameters obtained from the results of Dayhoff *et al.* (1978).

14.2.2   Descendants of the Dayhoff Model

The widely used 'Jones-Taylor-Thornton' or 'JTT' model of amino acid replacement (Jones *et al.*, 1992b) employs the same parameterization as did (Dayhoff and Eck, 1968) but refers to estimates of α_{ij} that were derived from a much larger data set. Gonnet *et al.*(1992) also published an updated set of α_{ij} estimates. More recently, it has become computationally feasible to obtain maximum likelihood estimates of all parameters in a 20-state general time-reversible model (see Adachi and Hasegawa, 1996; Yang *et al.*, 1998; Adachi *et al.*, 2000; Whelan and Goldman, 2001; Klosterman *et al.*, 2006). Whelan and Goldman (2001) used such a technique to generate the increasingly popular WAG model, based (like the models of Dayhoff and Eck, (1968), Dayhoff *et al.*(1972; 1978) and Jones *et al.*(1992b)) on data sets containing globular proteins encoded by nuclear DNA.

Adachi and Hasegawa (1996) derived amino acid replacement rate estimates from a data set consisting of all the mitochondrial proteins from each of 20 different vertebrate species. They detected some substantial differences between their estimates and those derived from nuclear-encoded proteins. Very similar findings were made by Yang *et al.*(1998), who studied mammalian mitochondrial proteins. A possible source of this variation is that mitochondrially-encoded proteins tend to be integral membrane proteins whereas nuclear-encoded proteins are often not, placing different functional constraints on the proteins and their evolution. It is reasonable to expect that the intracellular environment of a protein and its amino acid replacement process are correlated (see also Jones *et al.*, 1994), and it has been known for some time that mitochondrially encoded proteins exhibit unusual patterns of nucleotide sequence evolution (e.g. Brown *et al.*, 1982). Parameter estimates for protein evolution models derived from chloroplast genes and retroviral polymerase proteins were obtained by Adachi *et al.*(2000) and Dimmic *et al.*(2002), respectively.

14.3   AMINO ACID COMPOSITION

Amino acid replacements tend to involve chemically similar types of amino acids. There have been numerous attempts to classify amino acids on the basis of their physicochemical properties, and physicochemical distances among amino acid types have been proposed

(e.g. Grantham, 1974; Taylor and Jones, 1993). However, these physicochemical distances between amino acid types are unlikely to directly reflect the propensity for amino acid replacements to occur. Studies have been made relating amino acid properties to evolutionary patterns (e.g. Xia and Li, 1998), but these have not led to evolutionary models. The Dayhoff–Eck approach is appealing because estimates of relative amino acid replacement rates are empirically derived.

However, although the Dayhoff–Eck approach reflects the tendency for amino acid replacements to involve similar amino acid types, it describes evolution of the 'average site' in the 'average protein'. In reality, there is variability of amino acid replacement patterns among proteins. Certain proteins might be proline-rich, others might be leucine-rich, and still others might possess their own distinct patterns of amino acid composition. Dayhoff-type models can be modified to allow variation in composition of amino acids among proteins. Let us denote by $\hat{\pi}_i$ and $\hat{\alpha}_{ij}$ the amino acid frequencies and replacement rates estimated with a Dayhoff-type approach (i.e. from a database of aligned protein sequences under the assumption that all families in the database have common composition and replacement rates), and by π'_i and α'_{ij} the corresponding values we will choose to use in the analysis of a particular protein. Then, to allow variation of amino acid composition among proteins, the π'_i can be treated as parameters specific to each protein of interest. This is achieved by forcing the α'_{ij} for $i \neq j$ to obey

$$\alpha'_{ij} = \frac{\pi'_j}{\hat{\pi}_j} \hat{\alpha}_{ij},$$

(Cao *et al.*, 1994). The resulting replacement process defined by the α'_{ij} combines empirical information derived from databases (the $\hat{\alpha}_{ij}/\hat{\pi}_j$, also known as *exchangeabilities* – (Whelan and Goldman, 2001)) with the amino acid frequencies for the particular protein under study (π'_j), now estimated from that data set. Note that this approach maintains time reversibility (14.1). The benefit of this hybrid parameterization can be a sizeable improvement in model fit (Cao *et al.*, 1994). It is also attractive because it maintains the biologically realistic property of allowing amino acid replacements to occur among chemically similar amino acids, and for the pragmatic reason that estimation of the 20 amino acid frequencies π'_j from individual protein alignments is computationally tractable and seems statistically robust. Goldman and Whelan (2002) have obtained even greater improvements in model fit with a more general parameterization that makes the α'_{ij} a function of $\hat{\alpha}_{ij}$, the frequencies of the amino acid being replaced (π'_i and $\hat{\pi}_i$), and the frequencies of the replacement amino acid (π'_j and $\hat{\pi}_j$). The use of these parameterizations has thus been both to increase our understanding of the importance of particular amino acid residues in different proteins and to improve robustness of phylogenetic inferences.

14.4   HETEROGENEITY OF REPLACEMENT RATES AMONG SITES

Not only there is a variability of amino acid replacement among proteins but also there is variability of amino acid replacement rates among sites within a single protein. Yang (1994b) introduced a practical method for incorporating heterogeneity of evolutionary rates among sites into models of nucleotide or amino acid substitution (see **Chapter 15**).

Yang's innovation was to discretize a continuous gamma distribution of rates among sites into a relatively small number C of categories. Evaluation of a likelihood with Yang's discrete gamma treatment of variation of rates among sites requires an amount of computation that is approximately a factor C more than would be needed if all sites were assumed to share identical rates. Yang (1994b) demonstrated that discretization of a gamma distribution into a relatively small number of rate categories (e.g. five or six) often fits data about as well as more general (and more computationally difficult) approaches. It is now convincingly established that, for the majority of proteins, allowing heterogeneity of nucleotide substitution rates over sites is a great improvement over models that assume homogeneity of evolutionary rate, a fact partly attributable to the triplet coding nature of DNA and to the structure of the genetic code. It is also now known that amino acid replacement models often fit data much better when allowance is made for rate variation among sites (e.g. Yang *et al.*, 1998; Goldman and Whelan, 2002). This has confirmed that rate variation over protein sites is a widespread evolutionary phenomenon, that should be considered in all phylogenetic analyses.

14.5   PROTEIN STRUCTURAL ENVIRONMENTS

Although allowing for particular proteins' distinct amino acid compositions and among-site rate variation provides a large improvement in fits of models of protein evolution, the models described so far shed little light on the cause of the variation. It is clear, from considering protein structure and function, that at least some of the variability will be associated with the structural environment of a site. For example, consider the solvent accessibility of a site. A site on the surface of a globular protein will be exposed to solvent, typically water. Sites with high solvent accessibility therefore tend to be occupied by hydrophilic amino acids. In contrast, sites buried in the interior of a protein are less accessible to solvent and are apt to be occupied by hydrophobic amino acids. Further, exposed sites tend to evolve about twice as fast on average as buried sites (Goldman *et al.*, 1998). This is most likely because residues at buried sites interact with many neighboring residues; an amino acid replacement at an exposed surface site is less likely to disrupt the position of other protein residues than is a residue at a buried site.

One of the most striking findings in the study of protein structure has been that proteins' secondary and tertiary structure usually changes very slowly during evolution. In cases where homologous proteins are known to perform the same biological functions, it is often the case that their structures are very similar even when the sequences that code for them are quite diverged. In other words, protein sequence evolution tends in some sense to occur at a higher rate than the evolution of protein structure (e.g. Chothia and Lesk, 1986; Russell *et al.*, 1997). This tendency can be exploited by models of amino acid replacement. Because structure changes more slowly than does sequence, a group of homologous protein sequences is likely to share a common underlying structure. Homologous amino acid residues (i.e. those that are in the same alignment column) are likely to be in a similar structural environment. For example, if a residue from one protein is part of an α helix, then other sequences' residues in the same alignment column are also likely to be in an α helix. If the tertiary structure of one protein sequence in an alignment has been experimentally determined, then each alignment column can be assigned a structural environment and thus a separate 20-state GTR model can be estimated for

each environment (e.g. Overington *et al.*, 1990; Lüthy *et al.*, 1991; Topham *et al.*, 1993; Wako and Blundell, 1994a; 1994b; Koshi and Goldstein, 1995; 1996; Thorne *et al.*, 1996; Goldman *et al.*, 1998).

Further, the structural environment at one site in a protein is not independent of the environment at other sites. For example, if site i is in an α-helix environment, then site $i + 1$ is also probably in the α helix. It is possible to use this knowledge to further improve models of protein evolution. One approach has been the modeling of the organization of structural environments along a protein sequence as a first-order Markov chain. In protein sequence analysis, this approach was first used to predict secondary structure from single protein sequences (Asai *et al.*, 1993; Stultz *et al.*, 1993). The basis for the prediction was the tendency for certain kinds of amino acids to be found in certain secondary structure environments. Such models are referred to as *hidden Markov models* (HMMs) because the secondary structure underlying the sequence is not directly observed. Instead, it is 'hidden' but can be estimated on the basis of the amino acids that encode the protein sequence. A first-order Markov chain for the organization of protein structure along a sequence is not ideal because in reality protein structure is three dimensional rather than linear, and the structural environment of a site may be strongly influenced by other sites that are nearby in the tertiary structure but are far separated along the linear protein sequence. However, a first-order Markov model for structural organization is computationally tractable and is clearly superior to assuming independence of structural environments among sites.

The HMM for organization of structural environment can be combined with the models of amino acid replacement for each structural environment to generate an integrated model of protein sequence evolution. This has been done for the study of globular proteins (Thorne *et al.*, 1996; Goldman *et al.*, 1998), transmembrane proteins (Liò and Goldman, 1999), and mitochondrial proteins (Liò and Goldman, 2002). With this approach, each category k of structural environment has its own equilibrium amino acid frequencies (π_i^k) and its own rates of amino acid replacement (α_{ij}^k).

The algorithms used to calculate the likelihood with a HMM of protein structure were first applied to molecular sequence data by Churchill (1989) and are given in detail by Thorne *et al.*(1996). Felsenstein's (1981) pruning algorithm is used to compute likelihoods conditional on the unobserved states of the HMM, and the HMM's transition probabilities are incorporated to deal with uncertainty about which structure state each site belongs. The likelihood calculations now have a computational burden that is approximately a factor S more than without the HMM, where S is the number of distinct structure states considered, and this remains fast enough to permit maximum likelihood phylogenetic inference.

Algorithms for predicting the underlying protein secondary structure with a HMM also exist (Churchill, 1989; Asai *et al.*, 1993; Stultz *et al.*, 1993; Goldman *et al.*, 1996) but these will not be detailed here. Predictions of secondary structure that are based upon combined explicit treatment of protein structural organization and protein evolution have several potential advantages over methods for protein secondary structure prediction that are not based upon an explicit evolutionary model (Goldman *et al.*, 1996). Although aligned sequences are likely to share a common underlying secondary structure, these sequences are not independent realizations of some process. Instead, aligned sequences are related via an evolutionary tree. It is the phylogeny relating the sequences that is responsible for the correlation structure among sequences (see Felsenstein, 1985; Harvey and Pagel, 1991). Treatment of sequences as independent realizations of some process or

any other *ad hoc* weighting of sequences may be less desirable than explicit treatment of correlation structure.

Studies using these methods have shown the importance of structural environments to protein evolution. Different secondary structure elements do exhibit typical and distinct patterns of evolution, presumably related to different residues' propensities to adopt the local structures and biochemical properties required of protein functions. It is also clear that taking a phylogenetic approach could be of advantage to other sequence analyses aimed at investigating protein function, most obviously the simple goal of protein structure prediction which remains difficult even after many years of intensive research and competition (e.g. Kryshtafovych *et al.*, 2005).

Although they change slowly, secondary structure and solvent accessibility do evolve. Kawabata and Nishikawa (2000) have empirically estimated rates of change among secondary structure categories and solvent accessibility environments. This is a challenging endeavor, particularly because protein alignment becomes difficult when proteins being compared are so diverged as to have moderately different structures. Kawabata and Nishikawa base a procedure for recognizing distantly related protein homologs on their model of secondary structure and solvent accessibility change and they show that this procedure performs well. The future development of these or other ideas (e.g. see Grishin, 1997) into a complete model of protein sequence and structure evolution would be of great value.

14.6   VARIATION OF PREFERRED RESIDUES AMONG SITES

Some, but clearly not all, variation of amino acid composition and replacement rates among sites can be explained by consideration of structural environments among sites. The specific biochemical function of a protein, and even of its individual sites, mean that the evolution of every site of every protein takes place under different constraints, potentially leading to different evolutionary patterns. Bruno (1996) (also see Halpern and Bruno, 1998) made the first progress in this respect, devising a model that allowed each site in a protein to have its own equilibrium frequencies. We denote the frequency of amino acid type i at site s by $\pi_{i,s}$. This highly flexible approach suffers from potential overparameterization problems because the 20 amino acid frequencies per site add 19 degrees of freedom per site to the model. To remedy the overparameterization problems, it is not sufficient to restrict analyses to data sets consisting of a large number of sequences because, if sequences in a data set are closely related, then the sequences will be highly correlated with one another and good estimates of $\pi_{i,s}$ may still not be obtained confidently. Therefore, the approach of Bruno should work best when a data set contains a large number of highly diverged sequences.

Although this approach of allowing variation of 'preferred' residues among sites in a model of protein evolution is not computationally convenient, it is biologically attractive. The Dayhoff-type models allow the tendency of amino acid replacements to involve physicochemically similar amino acid types to be reflected in estimates of the replacement rates α_{ij}. On the other hand, the tendency for amino acid replacements to involve chemically similar amino acids may instead be largely attributable to a tendency for the values of $\pi_{i,s}$ to be positively correlated for physicochemically similar amino acids. These two explanations for the tendency of amino acid replacements to involve

physicochemically similar amino acids are not mutually exclusive, but little is known about their relative importance.

By adopting a Bayesian perspective, Lartillot and Philippe (2004) largely overcome the overparameterization concerns associated with variation of 'preferred' amino acid types among sites. They assume each site belongs to some amino acid replacement process category, but that the category is not directly observed. Different replacement categories share a common set of amino acid exchangeability parameters taken from existing standard models, but have different equilibrium amino acid frequencies. The number of categories and their frequencies are jointly determined by a Dirichlet process prior distribution, and the set of 20 different amino acid frequencies for a given category has a Dirichlet prior. This Bayesian approach, with prior distributions for the number of replacement categories, the frequencies of the categories, and the amino acid composition of each category, enables Lartillot and Philippe to greatly reduce the potential for overparameterization that occurs otherwise. They convincingly demonstrate that their Bayesian scheme is a substantial improvement over models that do not permit variation of preferred residues among sites.

Holmes and Rubin (2002) also investigate models that assume each site of a protein belongs to an amino acid replacement category that is not directly observed. Their innovative expectation-maximization (EM) algorithm yields maximum likelihood estimates of amino acid replacement rates for each replacement category. Although their frequentist procedure may be subject to potential overparameterization problems and requires the number of replacement categories to be prespecified, it has the advantage of not assuming that a particular protein site belongs to the same replacement process call for all of its evolutionary history. Instead, the EM algorithm can simultaneously estimate the rate at which sites switch among the replacement processes being modeled.

This impressive flexibility of the Holmes and Rubin procedure has the potential to assist in identifying biologically interesting cases where the evolutionary behavior of a site changes on a phylogeny. Building upon a practical adaptation by Tuffley and Steel (1998) of ideas first described by Fitch and colleagues (e.g. Fitch and Markowitz, 1970; Fitch, 1971), several groups have shown that evolutionary patterns at sites or codons change over time and that these changes in pattern should not be ignored by evolutionary models (e.g. Galtier, 2001; Huelsenbeck, 2002; Guindon *et al.*, 2004; Philippe *et al.*, 2005; Inagaki *et al.*, 2004). This characteristic of protein evolution, somewhat related to Fitch's (1971) idea of concomitantly variable codons or 'covarions' and also known as *heterotachy* (Lopez *et al.*, 2002), may be attributable to changes in protein structural constraints over long evolutionary times. Absence of heterotachy in evolutionary models has been implicated in failures of phylogenetic inference methods, particularly in cases of long-branch attraction (e.g. Inagaki *et al.*, 2004; Delsuc *et al.*, 2005).

14.7 MODELS WITH A PHYSICOCHEMICAL BASIS

An alternative approach to statistical models that implicitly reflect the tendency for amino acid replacements to involve physicochemically similar amino acid types is for this tendency to be reflected via the explicit inclusion of physicochemical properties into a model. Koshi *et al.*(1999) chose to quantify the physicochemical attributes of amino acid types according to hydrophobicity and bulk–properties that are expected to be of great importance to protein structure. Their approach assumes that each site in a protein belongs

to exactly one of several classes, with each class k occurring with probability γ_k. The probability π_i^k of amino acid i in class k is $\pi_i^k = e^{F_k(i)} / \sum_j e^{F_k(j)}$, where $F_k(i)$ is a measure of the suitability of amino acid type i in structural environment k. Koshi *et al.*(1999) investigated the use of suitability functions $F_k(i)$ that depended on the hydrophobicity H_i and the bulk B_i of the amino acids in both linear ($F_k(i) = \beta_k^H H_i + \beta_k^B B_i$) and quadratic ($F_k(i) = \beta_k^H (H_i - H_0^k)^2 + \beta_k^B (B_i - B_0^k)^2$) forms. Here, β_k^H, β_k^B, H_0^k, and B_0^k represent a relatively small number of additional free parameters (two or four per site class), making this approach statistically attractive in comparison to the 20-state GTR model, which assumes all sites experience identical evolutionary processes and, even so, requires 209 parameters to be estimated. The instantaneous rates of amino acid replacement for the Koshi *et al.*(1999) model are then defined as:

$$\alpha_{ij}^k = \begin{cases} \gamma_k & \text{if } F_k(j) > F_k(i) \\ \gamma_k e^{F_k(j) - F_k(i)} & \text{if } F_k(j) \le F_k(i). \end{cases}$$

This approach has not been widely pursued but it can yield substantially better fits to data than Dayhoff-type models (Koshi *et al.*, 1999). It will be interesting to determine the reason or reasons for these good fits. Aside from the obvious conclusion that the suitability of amino acid types at certain sites in a protein is indeed well described by the hydrophobicity and bulk preferences, another possible reason for the good performance of this model is more mundane. This parameterization scheme is very flexible, e.g. allowing for variation of preferred residue types among sites, and some of the improvement provided by this model may be attributable simply to the parametric forms for α_{ij}^k. This could be addressed by evaluating variants of the model with the formulae for the $F_k(i)$ modified by assigning to each amino acid type i the hydrophobicity and bulk of a randomly selected amino acid type j. If permutation or randomization experiments of this sort tend to result in fits comparable to those found by Koshi *et al.*(1999), then it would be the form of the α_{ij}^k rather than the actual hydrophobicity and bulk attributes themselves that are responsible for the good fit.

14.8   CODON-BASED MODELS

In reality, evolution occurs at the DNA level, with protein change being a consequence of phenomena such as nucleotide sequence mutation, genetic drift, and natural selection. It therefore makes sense to model protein evolution in terms of codons rather than amino acid types (Schöniger *et al.*, 1990; Goldman and Yang, 1994; Muse and Gaut, 1994). Codon-based models are typically framed in terms of the 61 codons that specify amino acids in common genetic codes. The three nucleotide triplets that represent stop codons are not allowed by most codon-based models.

Existing codon-based models follow largely the same approach as the Dayhoff model for amino acids. Codon replacements are modeled as a Markov process, with the instantaneous rate (α_{ij}) of change from codon i to codon j usually being defined as the product $\alpha_{ij} = \pi_j s_{ij}$ where π_j is the frequency of codon j. The exchangeabilities s_{ij} are symmetric ($s_{ij} = s_{ji}$), ensuring reversibility (14.1), and are themselves often products of parameters representing phenomena such as transition–transversion bias and the strength of selection acting on each mutation.

Codon-based models have become quite refined. For instance, Pedersen *et al.*(1998) designed a model to reflect the fact that CpG dinucleotide levels are depressed in lentiviral genes. Codon-based models are also well suited for testing hypotheses regarding natural selection. Through their parameterization of the exchangeabilities s_{ij}, they provide a basis for estimating synonymous and nonsynonymous nucleotide substitution rates (Muse, 1996; Yang and Nielsen, 2000) and have received considerable attention for their potential ability to detect both diversifying natural selection (e.g. Nielsen and Yang, 1998; Yang *et al.*, 2000; Massingham and Goldman, 2005; see **Chapter 12** (Yang), this volume) and purifying natural selection (Massingham and Goldman, 2005).

Halpern and Bruno (1998) have made some progress reconciling codon-based substitution rate matrices with molecular population genetics, explaining variation of preferred amino acid types among sites in terms of selective forces operating at the sites. While their approach may be overly rich in parameters and based upon restrictive assumptions about natural selection, they have taken an important step toward narrowing the gulf between population genetics and the models employed in phylogenetics.

Widely used codon-based models have some limitations. For instance, most assume that all mutational events result in exactly one nucleotide position changing identity. Spontaneous mutations often result in changes that are more complicated than a simple point mutation or than an insertion or deletion (see, e.g. Ripley, 1999). It seems reasonable that some of these more complicated changes are fixed in evolution. There is evidence that a moderate number of evolutionary changes are the result of mutational events that simultaneously affect two consecutive positions in a sequence (Averof *et al.*, 2000). Evolutionary models that allow for these changes now exist (Whelan and Goldman, 2004), and should be the focus of further research in the near future.

14.9 DEPENDENCE AMONG POSITIONS: SIMULATION

A limitation of widely used models of protein evolution is the assumption that the evolution of one position is independent of the sequence at other positions. Proteins are obviously three dimensional, and evolution at one site in a protein is likely to be affected by the amino acids at sites that physically neighbor it in the tertiary structure of the protein. Covariation of residues at different sites may indicate that these sites are nearby in the tertiary structure of the protein, and has been used to make inferences about protein tertiary structure and function (e.g. Pazos *et al.*, 1997; Pollock *et al.*, 1999; Tillier and Lui, 2003).

Parisi and Echave (2001; see also Bastolla *et al.*, 2003) have introduced a noteworthy technique for simulating the evolution of protein-coding genes subject to constraints imposed by protein structure. They begin their simulations with a reference protein that has known tertiary structure, and assume that this structure does not change over time. They then employ a sequence−structure distance scoring function to assess how well protein sequences fold into the known structure. Such scoring functions have been actively developed and are well established for predicting protein folds from protein sequence data. The underlying idea behind the Parisi and Echave technique is that the sequence−structure compatibility function can be employed to help parameterize rates of nonsynonymous substitutions.

As with conventional codon models (e.g. Goldman and Yang, 1994; Muse and Gaut, 1994), Parisi and Echave force the replacement rate between codons i and j to be 0 if i and j differ by more than one nucleotide or if a stop codon is involved. If the difference between i and j represents a single synonymous nucleotide change then the replacement rate is modeled as in conventional codon models, reflecting codon frequencies and different mutational processes that act on DNA. In the Parisi and Echave approach, however, nonsynonymous replacement rates from conventional codon models are multiplied by a factor of either 0 or 1, depending on the sequence–structure distance for the newly arising sequence containing codon j. If the distance is smaller than some prespecified value then the factor has a value of 1, permitting mutations that are considered to perturb the protein structure by not more than a set amount. Otherwise, this factor has a value of 0, disallowing evolutionary changes that lead to biologically implausible sequences.

Through simulations, Parisi and Echave (2001) convincingly demonstrated both that their model captured information that would be missed by other models of protein evolution and that this information capture could occur without requiring large numbers of free parameters to be added to the model. In a specific example, the simulation studies showed tendencies of certain amino acids to preferentially occupy certain sites in the left-handed β-helix domain of UDP-N-acetyl glucosamine acetyltransferases. When a group of actual acetyltransferases with this helix domain was examined, qualitatively similar tendencies were observed. To date, the Parisi and Echave model has been employed for simulation of evolution rather than for inference from a set of observed protein-coding sequences. Phylogenetic inference with models that make allowance for dependence among sites is computationally much more challenging.

14.10   DEPENDENCE AMONG POSITIONS: INFERENCE

Fornasari *et al.*(2002; see also Bastolla *et al.*, 2006) approximated the behavior of the full Parisi and Echave model with a 20-state amino acid replacement model at each position. The resulting model has independent changes among amino acid positions but does not force all of the positions of a protein to experience identical replacement processes. This is accomplished via simulating according to the Parisi and Echave codon-substitution model many times. For each position in a protein, the number of times that simulated changes occur between each pair of amino acid types can be recorded and rates of amino acid replacement for each protein position can be obtained based upon these counts of simulated changes.

A model of change permitting dependence among sequence positions poses enormous challenges for conventional evolutionary inference procedures that use the pruning algorithm of Felsenstein (1981). The difficulty is that Felsenstein's algorithm relies upon converting instantaneous evolutionary rates to transition probabilities. With independently evolving nucleotides or amino acids or codons, the number of rows and columns in the rate and transition probability matrices equals 4 or 20 or 61, respectively. But with a relatively general dependence structure, the number of rows and columns in the rate and transition probability matrices equals the number of possible sequences. Because the number of possible sequences grows exponentially with sequence length, Felsenstein's

pruning algorithm becomes computationally intractable with dependent change among sequence sites unless sequence length is extremely short.

A promising alternative to Felsenstein's inference procedure has been explored by Jensen and Pedersen (2000) and Pedersen and Jensen (2001) as part of their efforts to understand the effects of overlapping reading frames and context-dependent mutation on molecular evolution (see also Siepel and Haussler, 2004; Hwang and Green, 2004; Christensen *et al.*, 2005). Jensen and Pedersen augmented the observed sequence data at the tips of evolutionary trees with 'sequence paths'. The sequence path on a particular branch of a phylogeny specifies all of the changes that occurred on the branch. For each change, the sequence path contains information about the time at which the change occurred, the sequence position at which the change occurred, and the nature of the change. Although the sequence path is not directly observed, Markov chain Monte Carlo techniques can be used to randomly sample possible sequence paths according to the appropriate probability density. The big advantage of the sequence path approach to inference is that it is often computationally tractable in cases where Felsenstein's pruning algorithm is difficult to apply because transition probabilities cannot be calculated or well approximated.

We illustrate the sequence path approach to inference with the technique of Robinson *et al.*(2003), an inference method that combines the innovations of Parisi and Echave (2001), Jensen and Pedersen, (2000) and Pedersen and Jensen (2001). The Robinson *et al.* model of protein-coding evolution is designed for analyzing aligned protein-coding DNA sequences that, when translated, are assumed to share a known protein tertiary structure that does not change during evolution. For simplicity, we discuss evolutionary inference for data sets consisting of only two sequences. This is the case considered by Robinson *et al.*(2003). However, the inference procedure has subsequently been extended to data sets with three (Robinson, 2003) or more (Rodrigue *et al.*, 2005) sequences.

To measure the compatibility between the encoded amino acid sequence of a DNA sequence i and the known tertiary structure, an empirically derived system originally developed for protein fold recognition with globular proteins (Jones *et al.*, 1992a; Jones, 1999) is employed. This system returns two scores for any protein sequence that is folded into a particular structure. One of these scores, $E_s(i)$, is the solvent accessibility score for sequence i when translated and folded into the known structure; the other $E_p(i)$, is the pairwise interaction score for sequence i when translated and folded into the known structure. $E_s(i)$ and $E_p(i)$ loosely resemble free energies: both will be negative when sequence and structure are compatible, and sequence–structure compatibility deteriorates as the two scores rise.

The Robinson *et al.*(2003) model has the parameter κ to differentiate between transitions and transversions and the parameter ω to differentiate between nonsynonymous and synonymous substitutions. It also has parameters π_A, π_C, π_T, and π_G to represent what the frequencies of the nucleotide types would be if all natural selection was absent. Because the model permits a general dependence structure, rates of change are defined at the whole-sequence level rather than at the codon or individual position levels. We now use α_{ij} to denote the rate of change from DNA sequence i to DNA sequence j, with the value of α_{ij} forced to be zero unless sequences i and j differ at exactly one position. If sequence j has nucleotide type $h \in \{A, C, G, T\}$ at the one position where i and j differ,

the rate of change is given by:

$$\alpha_{i,j} = \begin{cases} u\pi_h & \text{for a synonymous transversion} \\ u\pi_h\kappa & \text{for a synonymous transition} \\ u\pi_h\omega e^{s(E_s(i)-E_s(j))+p(E_p(i)-E_p(j))} & \text{for a nonsynonymous transversion} \\ u\pi_h\kappa\omega e^{s(E_s(i)-E_s(j))+p(E_p(i)-E_p(j))} & \text{for a nonsynonymous transition.} \end{cases}$$

(14.2)

The parameters s and p in (14.2) represent the importance of solvent accessibility and pairwise interactions, respectively. When s and p are both zero, protein structure does not influence protein-coding DNA evolution and substitutions occur independently among codons. When these parameters are both positive, the biologically plausible situation of proteins evolving to be compatible with their structure results. While it is formally possible for s and p to be negative, this is biologically unrealistic, representing the situation where proteins evolve to be incompatible with their tertiary structure.

Although Felsenstein's pruning algorithm is not computationally feasible with this model because of the high dimensionality of the rate and transition probability matrices, the inference ideas of Jensen and Pedersen (2000) and Pedersen and Jensen (2001) can be adopted. Using a sequence path approach, what must be calculated instead is the probability of the events described by any given path (ρ). To do this, we need to know the rate of change from each successive sequence in the path, and the probability that when each change occurs, the result is the next sequence in the path. Both of these are readily derived from (14.2). The rate of change away from any sequence i (i.e. to any sequence that differs from i) is:

$$\alpha_{i,\bullet} = \sum_k \alpha_{i,k}.$$

(14.3)

Since rates of change from one sequence to another are 0 unless the two sequences differ at exactly one position, the number of sequences that differ from sequence i at exactly one position cannot exceed the length of sequence i multiplied by 3. The sum in (14.3) is therefore computationally tractable, because most terms being added are 0.

For a particular sequence path ρ from an ancestral sequence i to a descendant sequence j, let q be the total number of nucleotide substitutions on the path and let $t(z)$ be the time of the zth substitution. The time the branch begins will be $t(0) = 0$ and the time the branch ends will be $t(q + 1) = 1$. Because the scaling parameter u can be adjusted, the time at which the branch ends can always be considered to be 1. The sequence that exists immediately after the zth substitution will be represented by $i(z)$, with the initial sequence i now written as $i(0)$ and the final sequence $j = i(q + 1) = i(q)$. The sequence path ρ is then fully specified by $t(0)$, $t(1)$, ..., $t(q)$, $t(q + 1)$ and $i(0)$, $i(1)$, ..., $i(q)$, $i(q + 1)$. Let θ be a vector representing all model parameters, the probability density of the sequence path is then given by:

$$p(\rho|\theta, i) = \left(\prod_{z=1}^{q} \frac{\alpha_{i(z-1),i(z)}}{\alpha_{i(z-1),\bullet}} \alpha_{i(z-1),\bullet} e^{-\alpha_{i(z-1),\bullet}(t(z)-t(z-1))} \right) e^{-\alpha_{i(q),\bullet}(t(q+1)-t(q))}$$

$$= \left(\prod_{z=1}^{q} \alpha_{i(z-1),i(z)} e^{-\alpha_{i(z-1),\bullet}(t(z)-t(z-1))} \right) e^{-\alpha_{i(q),\bullet}(t(q+1)-t(q))}.$$

(14.4)

Here, the product over z gives the probability density for the q sequence changes from $i(z-1)$ to $i(z)$ occurring at times $t(z)$, and the final term is the probability of no further change occurring in the time interval from the final substitution at $t(q)$ until the time $t(q+1) = 1$ at which the path ends. By combining a prior distribution for θ with (14.4) and with the expression for the stationary probability for sequence i, the Metropolis–Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) can be applied to perform Bayesian inference (Robinson *et al.*, 2003).

Initial applications of this model yielded biologically plausible positive estimates of the s and p parameters and indicated that the evolutionary information leading to the estimates of these two parameters was quite distinct from the information contributing to the ω parameter (Robinson, 2003; Robinson *et al.*, 2003). Rodrigue *et al.*(2005; 2006) have carefully evaluated evolutionary models with rate parameterizations similar to that of Robinson *et al.*(2003). Employing a new model comparison procedure (Lartillot and Philippe, 2004; 2006) as well as inference based upon sequence paths, Rodrigue *et al.*(2006) demonstrate that evolutionary models with dependence among sequence positions due to tertiary structure are statistically superior to those without dependence. However, these authors also conclude that available treatments of protein tertiary structure are not sufficient to produce satisfactory evolutionary models. They find that adding rate heterogeneity among sites via the discrete gamma technique of Yang (1994b) generates model fit improvements beyond those realized solely by incorporating tertiary structure.

One reasonable interpretation of these findings is that sequence–structure compatibility measures yield models of protein evolution that are biologically meaningful and statistically valuable, extracting information that is otherwise not utilized, yet which are still incomplete summaries of natural selection. Although the Robinson *et al.*(2003) model was designed to reflect dependence due to protein tertiary structure, the same inference strategy could be applied to add evolutionary dependence due to other aspects of phenotype (e.g. protein expression or protein function). This could be done by incorporating scores for the other phenotypic aspects into the rates of (14.2).

14.11 CONCLUSIONS

After nearly 40 years, progress continues to be made with probabilistic models that give increasingly good statistical descriptions of the patterns of protein evolution. But a useful model is not simply one that fits data well. For a model of protein evolution to help us to understand processes and pressures acting on evolving genomes, it is also important to establish a solid connection between the model parameters and the biological features that they represent. Although the best models of protein evolution are now far more realistic than the earliest models, understanding of protein evolution is still at a primitive stage. Evidence for variation in evolutionary processes and evolutionary rates among sites is strong. Unfortunately, the extent to which this variation can be partitioned into components of interest (e.g. the structural environments of a site, the protein to which the site belongs, mutational tendencies at the site, interactions with other sites, protein function, etc.) remains unclear.

Ideally, protein evolution should be linked to the DNA that codes for the protein, the structure of the protein, the expression patterns of the protein, the function of the protein, and external influences on the protein that act via natural selection. Progress on modeling protein evolution will depend in part on the advances that are made in techniques for *in silico* prediction of phenotype from genotype. Protein-coding DNA that results in deleterious phenotypes is unlikely to be ancestral to extant DNA. Accurate *in silico* prediction of phenotype from genotype would lead to evolutionary models in which substitution rates are high for selectively advantageous changes and low for deleterious changes. Existing systems for *in silico* mapping of genotype to phenotype tend to be crude and this is one reason there is so much room for improvement in models of protein evolution.

Despite the primitive nature of models of protein evolution, software implementing those models for data analysis is in great demand. It must be emphasized that even the existing models of protein evolution, however unrealistic, are better than having no models at all. Explicit evolutionary models provide a basis upon which homologous sequences can be recognized, phylogenetic history can be inferred, evolutionary hypotheses can be evaluated, protein structure can be predicted, and our understanding of evolution can be quantified. For these reasons, the development of model-based approaches to study protein evolution is an attractive, active, and valuable area of research.

Acknowledgments

REFERENCES

Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution* **42**, 459–468.

Adachi, J., Waddell, P.J., Martin, W. and Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution* **50**, 348–358.

Asai, K., Hayamizu, S. and Handa, K. (1993). Prediction of protein secondary structure by hidden Markov model. *CABIOS* **9**, 141–146.

Averof, M., Rokas, A., Wolfe, K.H. and Sharp, P.M. (2000). Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287**, 1283–1286.

Bastolla, U., Porto, M., Roman, H.E. and Vendrulscolo, M. (2003). Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *Journal of Molecular Evolution* **56**, 243–254.

Bastolla, U., Porto, M., Roman, H.E. and Vendrulscolo, M. (2006). A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank. *BMC Evolutionary Biology* **6**, 43.

Brown, W.M., Prager, E.M., Wang, A. and Wilson, A.C. (1982). Mitochondrial D.N.A. sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution* **18**, 225–239.

Bruno, W.J. (1996). Modeling residue usage in aligned protein sequences via maximum likelihood. *Molecular Biology and Evolution* **13**, 1368–1374.

Cao, Y., Adachi, J., Janke, A., Pääbo, S. and Hasegawa, M. (1994). Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *Journal of Molecular Evolution* **39**, 519–527.

Chothia, C. and Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO journal* **5**, 823–826.

Christensen, O.F., Hobolth, A. and Jensen, J.L. (2005). Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates. *Journal of Computational Biology* **12**, 1166–1182.

Churchill, G.A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51**, 79–94.

Dayhoff, M.O. and Eck, R.V. (1968). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure, 1967–68*, M.O. Dayhoff and R.V. Eck, eds. National Biomedical Research Foundation, Washington, DC, pp. 33–41.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, Vol. 5(Suppl. 3), M.O. Dayhoff, ed. National Biomedical Research Foundation, Washington, DC, pp. 345–352.

Dayhoff, M.O., Eck, R.V. and Park, C.M. (1972). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, Vol. 5, M.O. Dayhoff, ed. National Biomedical Research Foundation, Washington, DC, pp. 89–99.

Delsuc, F., Brinkmann, H. and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* **6**, 361–375.

Dimmic, M.W., Rest, J.S., Mindell, D.P. and Goldstein, R.A. (2002). rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *Journal of Molecular Evolution* **55**, 65–73.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist* **125**, 1–15.

Fitch, W.M. (1971). Rate of change of concomitantly variable codons. *Journal of Molecular Evolution* **1**, 84–96.

Fitch, W.M. and Markowitz, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics* **4**, 579–593.

Fornasari, M.S., Parisi, G. and Echave, J. (2002). Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Molecular Biology and Evolution* **19**, 352–356.

Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution* **18**, 866–873.

Goldman, N., Thorne, J.L. and Jones, D.T. (1996). Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *Journal of Molecular Biology* **263**, 196–208.

Goldman, N., Thorne, J.L. and Jones, D.T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–458.

Goldman, N. and Whelan, S. (2002). A novel use of equilibrium frequencies in models of sequence evolution. *Molecular Biology and Evolution* **19**, 1821–1831.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**, 725–736.

Gonnet, G.H., Cohen, M.A. and Benner, S.A. (1992). Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443–1445.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864.

Grishin, N.V. (1997). Estimation of evolutionary distances from protein spatial structures. *Journal of Molecular Evolution* **45**, 359–369.

Guindon, S., Rodrigo, A.G., Dyer, K.A. and Huelsenbeck, J.P. (2004). Modeling the site-specific variation of selection patterns along lineages. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12957–12962.

Halpern, A. and Bruno, W.J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution* **15**, 910–917.

Harvey, P.H. and Pagel, M.D. (1991). *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, pp. 239.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Holmes, I. and Rubin, J.P. (2002). An expectation maximization algorithm for training hidden substitution models. *Journal of Molecular Biology* **317**, 753–764.

Huelsenbeck, J.P. (2002). Testing a covariotide model of DNA substitution. *Molecular Biology and Evolution* **19**, 698–707.

Hwang, D.G. and Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 13994–14001.

Inagaki, Y., Susko, E., Fast, N.M. and Roger, A.J. (2004). Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF−1α phylogenies. *Molecular Biology and Evolution* **21**, 1340–1349.

Jensen, J.L. and Pedersen, A.-M.K. (2000). Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Advances in Applied Probability* **32**, 499–517.

Jones, D.T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology* **287**, 797–815.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992a). A new approach to protein fold recognition. *Nature* **358**, 86–89.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992b). The rapid generation of mutation data matrices from protein sequences. *CABIOS* **8**, 275–282.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1994). A mutation data matrix for transmembrane proteins. *FEBS Letters* **339**, 269–275.

Jukes, T.H. and Cantor, C.R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, H.N. Munro, ed. Academic Press, New York, pp. 21–132.

Kawabata, T. and Nishikawa, K. (2000). Protein structure comparison using the Markov transition model of evolution. *Proteins* **41**, 108–122.

Kishino, H., Miyata, T. and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution* **31**, 151–160.

Klosterman, P.S., Uzilov, A.V., Bendana, Y.R., Bradley, R.K., Chao, S., Kosiol, C., Goldman, N. and Holmes, I. (2006). XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics* **7**, 428.

Koshi, J.M. and Goldstein, R.A. (1995). Context-dependent optimal substitution matrices. *Protein Engineering* **8**, 641–645.

Koshi, J.M. and Goldstein, R.A. (1996). Mutation matrices and physical-chemical properties: correlations and implications. *Proteins* **27**, 336–344.

Koshi, J.M., Mindell, D.P. and Goldstein, R.A. (1999). Using physical-chemistry based mutation models in phylogenetic analyses of HIV-1 subtypes. *Molecular Biology and Evolution* **16**, 173–179.

Kosiol, C. and Goldman, N. (2005). Different versions of the Dayhoff rate matrix. *Molecular Biology and Evolution* **22**, 193–199.

Kryshtafovych, A., Venclovas, Ç., Fidelis, K. and Moult, J. (2005). Progress over the first decade of CASP experiments. *Proteins* **61**, 225–236.

Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21**, 1095–1109.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology* **55**, 195–207.

Liò, P. and Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome Research* **8**, 1233–1244.

Liò, P. and Goldman, N. (1999). Using protein structural information in evolutionary inference: transmembrane proteins. *Molecular Biology and Evolution* **16**, 1696–1710.

Liò, P. and Goldman, N. (2002). Modeling mitochondrial protein evolution using structural information. *Journal of Molecular Evolution* **54**, 519–529.

Lopez, P., Casane, D. and Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution* **19**, 1–7.

Lüthy, R., McLachlan, A.D. and Eisenberg, D. (1991). Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins* **10**, 229–239.

Massingham, T. and Goldman, N. (2005). Detecting amino acid sites under positive and purifying selection. *Genetics* **169**, 1753–1762.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Müller, T., Spang, R. and Vingron, M. (2002). Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Molecular Biology and Evolution* **19**, 8–13.

Müller, T. and Vingron, M. (2000). Modeling amino acid replacement. *Journal of Computational Biology* **7**, 761–776.

Muse, S.V. (1996). Estimating synonymous and nonsynonymous substitution rates. *Molecular Biology and Evolution* **13**, 105–114.

Muse, S.V. and Gaut, B.S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. *Molecular Biology and Evolution* **11**, 715–724.

Ng, P.C. and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research* **11**, 863–874.

Ng, P.C. and Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Research* **12**, 436–446.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.

Overington, J., Johnson, M.S., Šali, A. and Blundell, T.L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proceedings of the Royal Society of London, Series B* **241**, 132–145.

Parisi, G. and Echave, J. (2001). Structural constrains and the emergence of sequence patterns in protein evolution. *Molecular Biology and Evolution* **18**, 750–756.

Pazos, F., Helmer-Citterich, M., Ansiello, G. and Valencia, A. (1997). Correlated mutations contain information about protein-protein interactions. *Journal of Molecular Biology* **271**, 511–523.

Pedersen, A.-M.K., Wiuf, C. and Christiansen, F.B. (1998). A codon-based model designed to describe lentiviral evolution. *Molecular Biology and Evolution* **15**, 1069–1081.

Pedersen, A.-M.K. and Jensen, J.L. (2001). A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Molecular Biology and Evolution* **18**, 763–776.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. and Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* **5**, 50.

Pollock, D.D., Taylor, W.R. and Goldman, N. (1999). Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of Molecular Biology* **287**, 187–198.

Ripley, L.S. (1999). Predictability of mutant sequences: relationships between mutational mechanisms and mutant specificity. *Annals of the New York Academy of Sciences* **180**, 159–172.

Robinson, D.M. (2003). D.R. EVOL: three dimensional realistic evolution. Ph.D. Thesis, North Carolina State University, Raleigh, NC.

Robinson, D.M., Jones, D., Kishino, H., Goldman, N. and Thorne, J.L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution* **20**, 1692–1704.

Rodrigue, N., Lartillot, N., Bryant, D. and Philippe, H. (2005). Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* **347**, 207–217.

Rodrigue, N., Philippe, H. and Lartillot, N. (2006). Assessing site-interdependent phylogenetic models of sequence evolution. *Molecular Biology and Evolution* **23**, 1762–1775.

Russell, R.B., Saqi, M.A.S., Sayle, R.A., Bates, P.A. and Sternberg, M.J.E. (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *Journal of Molecular Biology* **269**, 423–439.

Schöniger, M., Hofacker, G.L. and Borstnik, B. (1990). Stochastic traits of molecular evolution – acceptance of point mutations in native actin genes. *Journal of Theoretical Biology* **143**, 287–306.

Siepel, A. and Haussler, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution* **21**, 468–488.

Stone, E.A. and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research* **15**, 978–986.

Stultz, C.M., White, J.V. and Smith, T.F. (1993). Structural analysis based on state-space modeling. *Protein Science* **2**, 305–314.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**, 57–86.

Taylor, W.R. and Jones, D.T. (1993). Deriving an amino acid distance matrix. *Journal of Theoretical Biology* **164**, 65–83.

Thorne, J.L., Goldman, N. and Jones, D.T. (1996). Combining protein evolution and secondary structure. *Molecular Biology and Evolution* **13**, 666–673.

Tillier, E.R.M. and Lui, T.W.H. (2003). Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* **19**, 750–755.

Topham, C.M., McLeod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S. and Blundell, T.L. (1993). Fragment ranking in modelling of protein structure: conformationally constrained substitution tables. *Journal of Molecular Biology* **229**, 194–220.

Tuffley, C. and Steel, M. (1998). Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences* **147**, 63–91.

Wako, H. and Blundell, T.L. (1994a). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins *I. Solvent accessibility classes. Journal of Molecular Biology* **238**, 682–692.

Wako, H. and Blundell, T.L. (1994b). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins *II. Secondary structures. Journal of Molecular Biology* **238**, 693–708.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* **18**, 691–699.

Whelan, S. and Goldman, N. (2004). Estimating the frequency of events that cause multiple nucleotide changes. *Genetics* **167**, 2027–2043.

Xia, X. and Li, W.-H. (1998). What amino acid properties affect protein evolution? *Journal of Molecular Evolution* **47**, 557–564.

Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39**, 105–111.

Yang, Z. (1994b). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39**, 306–314.

Yang, Z. and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**, 32–43.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.-M.K. (2000). Codon-substitution models for heterogeneous selection pressure. *Genetics* **155**, 431–449.

Yang, Z., Nielsen, R. and Hasegawa, M. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution* **15**, 1600–1611.

Application of the Likelihood Function in Phylogenetic Analysis

J.P. Huelsenbeck and J.P. Bollback

Department of Biology, University of Rochester, Rochester, NY, USA

The likelihood of a phylogenetic tree is proportional to the probability of observing the comparative data (such as aligned DNA sequences) conditional on the tree. The likelihood function is important because it is the vehicle that carries the observations. The likelihood function can be used in two ways to infer phylogeny. First, the tree that maximizes the likelihood can be chosen as the best estimate of phylogeny; this is the method of maximum likelihood. Second, a prior probability distribution on trees can be specified and inferences based upon the posterior probability distribution of trees; this is the approach taken by Bayesians. Although maximum likelihood and Bayesian inference are similar in that the same models of DNA substitution can be used to calculate the likelihood function, they differ in their interpretation of probability. Markov chain Monte Carlo (MCMC) can be used to approximate the posterior probabilities of trees. MCMC also makes it possible to perform comparative analyses that accommodate phylogenetic uncertainty.

15.1 INTRODUCTION

How can the evolutionary history of life be inferred? This problem – called the phylogeny problem – has generated intense interest since publication of *Origin of Species*. The phylogeny problem has by no means been 'solved'; debate continues to center around the merits of competing methods of phylogeny inference, the applicability of concepts such as the accuracy of competing methods, how phylogenies can be used in comparative analyses, and even whether the phylogeny problem is one of statistical inference (Farris, 1983).

Sometimes the choice of phylogenetic method does not appear to matter; the same phylogenetic tree is obtained regardless of the details of the analysis. This type of situation is, of course, an ideal one. Often, however, the details of the phylogenetic analysis do matter; different phylogenies are obtained if different phylogenetic methods are used

or even if the same method of analysis is used under different assumptions. What is the biologist to do when different phylogenetic methods give different trees? Typically, biologists use one or more of the following arguments to choose among competing methods (and by extension, among the competing trees produced by these methods):

Expected accuracy. Sometimes the choice among phylogenetic methods is based upon simulation studies of the expected accuracy of competing methods under specific evolutionary scenarios. Numerous studies have used computer simulation to investigate the ability of different phylogenetic methods to accurately reconstruct phylogeny under idealized models of DNA substitution (Charleston, 1994; Gaut and Lewis, 1995; Huelsenbeck, 1995a; 1995b; Huelsenbeck and Hillis, 1993; Kuhner and Felsenstein, 1994; Nei, 1991). These types of studies have provided useful information on the consistency (ability to correctly estimate phylogeny given an unlimited number of data), efficiency (ability to correctly estimate phylogeny with a limited number of data), and robustness (ability to correctly estimate phylogeny when the assumptions of the analysis are violated) of different phylogenetic methods under a variety of evolutionary scenarios. Specifically, we know that factors such as the overall substitution rate and the relative lengths of the branches of phylogenies are among the main determinants of the performance of methods. Several disturbing properties of phylogenetic methods have also been uncovered in these studies, such as the complete failure of many phylogenetic methods under some combinations of branch length (e.g., the well-known 'Felsenstein zone' in which many phylogenetic methods can converge to the incorrect tree even with an infinite amount of data; see Felsenstein, 1978).

Philosophy. Some methods are considered to be more acceptable than others based upon philosophical concepts. For example, the parsimony method has been justified based upon the ideas of hypothetical deductive reasoning (Gaffney, 1979), falsificationism (Wiley, 1975), and, more recently, on Sir Karl Popper's concept of corroboration (Kluge, 1997; Siddall and Kluge, 1997). Other more explicitly statistical methods have justifications based upon reasoning in the face of uncertainty and use probability models to express the uncertainty of different possible outcomes (see Edwards, 1972).

Flexibility. Some methods of phylogenetic inference can be applied to a wide variety of data types and models of evolution. For example, the maximum parsimony method can easily be used to infer phylogeny from molecular or morphological data or from monomorphic or polymorphic characters. The method of maximum likelihood can be applied to estimate parameters from even complex models of evolution.

Testability. It should be possible, in principle, to test the assumptions made in the course of a phylogenetic analysis (Penny *et al.*, 1992). This is easy to do using explicitly statistical methods, such as likelihood and distance methods (Goldman, 1993; Rzhetsky and Nei, 1995), but very difficult to do using other methods. Moreover, as biologists turn their attention to comparative questions (problems that depend upon phylogeny, but for which the phylogeny is not the central question), the ability of a method to be tailored to address specific evolutionary questions becomes more important.

In this chapter, we concentrate on phylogenetic methods that are based upon the likelihood function – the methods of maximum likelihood and Bayesian inference. We

believe that of the many methods of phylogenetic inference that have been proposed, these methods have the greatest potential to simultaneously satisfy the criteria posed above. Although maximum likelihood and Bayesian inference are very similar in that the likelihood function carries the information about phylogeny contained in the data, these two methods have very different concepts of probability. In a field that has experienced vigorous debate on the relative merits of different methods of phylogenetic inference, it is perhaps safe to predict that the next debate in systematics will focus on the merits of maximum likelihood versus Bayesian inference of phylogeny (mirroring debates in the statistical literature over the past 30 years).

15.2   HISTORY

Most authors (see Edwards, 1974; Berger and Wolpert, 1984) attribute the ideas of likelihood and the method of maximum likelihood estimation to Sir Ronald A. Fisher (1912; 1921; 1922). In Fisher's 1912 work, published while a university undergraduate, he used the Bayesian term 'inverse probability' to describe his method of maximum likelihood by which he was later (Fisher, 1915) able to generate the likelihood surface for two parameters. By 1921 Fisher recognized his inappropriate use of the term 'inverse probability' in his 1912 paper and coined the term 'likelihood' (Fisher, 1921).

15.2.1   A Brief History of Maximum Likelihood in Phylogenetics

Edwards and Cavalli-Sforza made several important contributions to phylogenetics, including the introduction of the least squares, minimum evolution or parsimony, and maximum likelihood methods of phylogenetic inference (Edwards and Cavalli-Sforza, 1964; Cavalli-Sforza and Edwards, 1967). Perhaps their most important contribution, however, was the early realization that the phylogeny problem was one of statistical inference. In fact, a number of authors later emphasized this idea (Felsenstein, 1973; 1988). Both Edwards and Cavalli-Sforza had 'sat at the feet of R.A. Fisher' (Edwards, 1998) and it was natural for them to consider the method of maximum likelihood to infer phylogeny. Indeed, they discussed maximum likelihood estimation of phylogeny at the 1964 meeting of the Systematics Association and at the Cold Spring Harbor Symposium on Quantitative Biology. In 1970 Edwards published a description of a maximum likelihood method for inferring phylogeny using gene frequency data. Neyman (1971) applied the method of maximum likelihood to analyze nucleotide or amino acid sequence data for three species using a simple probabilistic model of symmetric change similar to the Jukes–Cantor formula (Jukes and Cantor, 1969). Soon after, this method was extended to more than three species by considering triplets of species at one time and combining the results in a somewhat arbitrary manner (Kashyap and Subas, 1974). The limitation on the number of species these approaches could address, and possibly why they found little application at the time, centered around the high computational cost of likelihood calculations.

Felsenstein (1973) proposed an algorithm for the application of maximum likelihood to discrete character data, but practical applications of his approach were still hindered by the computational difficulties. Computational limitations were greatly reduced in a seminal paper by Felsenstein (1981) in which he introduced the 'pruning algorithm'. The

pruning algorithm makes possible efficient calculation of the likelihood for a large number of taxa by taking advantage of the form of the tree topology; today, all computer programs for calculating likelihoods on phylogenies use Felsenstein's pruning algorithm. By 1985 maximum likelihood estimation of phylogeny began to be applied more widely. DeBry and Slade (1985) analyzed restriction site data using maximum likelihood. Since then developments in the application of maximum likelihood to phylogenetics have accelerated, centering around construction of increasingly realistic stochastic models of change – e.g. codon models (Muse and Gaut, 1994; Goldman and Yang, 1994) – for a variety of data types, and the application of maximum likelihood estimation to evolutionary questions – e.g. testing the molecular clock (Felsenstein, 1988); testing the monophyly of a group (Huelsenbeck *et al.*, 1996); and reconstructing ancestral states (Yang *et al.*, 1995; Pagel, 1999).

15.2.2 A Brief History of Bayesian Inference in Phylogenetics

Bayesian inference of phylogeny was described in the dissertations of Mau (1996) and Li (1996) and by Rannala and Yang (1996). More recently, Markov chain Monte Carlo has made it possible to efficiently approximate the posterior probability of trees (Yang and Rannala, 1997; Mau *et al.*, 1999; Larget and Simon, 1999). The Bayesian approach has been applied to numerous evolutionary questions, such as host–parasite cospeciation (Huelsenbeck *et al.*, 1997; 2000a; Newton *et al.*, 1999), simultaneous sequence alignment and phylogeny estimation (Mitchison, 1999), inferring the history of a character (Schultz and Churchill, 1999), and estimating species' divergence times under a relaxed molecular clock (Thorne *et al.*, 1998; Huelsenbeck *et al.*, 2000b).

15.3 LIKELIHOOD FUNCTION

The likelihood of a hypothesis is proportional to the probability of observing the data given the hypothesis, or

$$\ell(\text{Hypothesis}) = C \times f(\text{Observations}|\text{Hypothesis}),$$

where $f(\text{Observations}|\text{Hypothesis})$ is the probability of observing the data given a hypothesis. The constant, C, is arbitrary. (Throughout this chapter, we follow the convention of using $f(\cdot|\cdot)$ to denote a conditional probability.) As an example, take the simple case of tossing a coin n times with the object of estimating the probability of observing a head on a single toss of the coin. The number of heads observed in the course of the n tosses is x. The probability of observing the data (x heads) given a particular value for the parameter p can be calculated using the binomial distribution,

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x},$$

where $\binom{n}{x} = n!/x!(n-x)!$. The likelihood function, then, is

$$\ell(p) = C \times \binom{n}{x} p^x (1-p)^{n-x}.$$

For the phylogeny problem, the observations are taken to be the aligned character matrix. For DNA sequence data, the aligned matrix will be denoted X. As an example of aligned sequences, consider the first and last 15 aligned sites of the *replicase* gene from nine bacteriophage species of the family Leviviridae:

PP7	GACAGC- - -CGGUUC. . . CGGAUCCCUGACACG
FR	GGCAACGGU- - -GUG. . . GCAGACCCACGCCUC
MS2	GGGAACGGA- - -GUG. . . UCAGAUCCACGCCUC
GA	GGCAACGGU- - -UUG. . . UCAGAUCCGCGACUC
SP	UCA- - -AAUAAAGCA. . . UGGGAUCCUAGAGCA
NL95	UCG- - -AAUAAAGCA. . . UGGGAUCCUAGGGUA
M11	CCUUUCAAUAAAGCA. . .UGGGAUCCUAGGGUA
MX1	CCUUUCAAUAAAGCA. . .UGGGAUCCUAGGGUA
Qβ	CCUUUUAAUAAAGCA. . .UGGGAUCCUAGGGCC

The core region of the replicase gene corresponds to amino acid residue numbers 205 to 443, referenced to Qβ (GenBank accession numbers: FR, X15031; MS2, J02467; GA, X03869; SP, X07489; NL95, AF059243; M11, AF052431; MX1, AF059242; Qβ, X14764; PP7, X80191; Bollback and Huelsenbeck, 2001). The phage PP7 is treated as an outgroup.

The individual observations are the sites. For example, the first observation in the bacteriophage matrix is the site $\mathbf{x}_1 = (G, G, G, G, U, U, C, C, C)'$. The second, third, fourth, and fifth observations are $\mathbf{x}_2 = (A, G, G, G, C, C, C, C, C)'$, $\mathbf{x}_3 = (C, C, G, C, A, G, U, U, U)'$, $\mathbf{x}_4 = (A, A, A, A, -, -, U, U, U)'$, and $\mathbf{x}_5 = (G, A, A, A, -, -, U, U, U)'$, and so on. There are a total of $c = 720$ sites in this example data set.

How can the probability of observing the individual sites $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_c)$ be calculated? The probability of observing the nucleotides at a site depends upon the topology of the tree relating the species, the lengths of the branches on this tree, and a specific model of how the sequence changes along the tree. Throughout this chapter, the topology of the tree relating the species will be denoted τ and the lengths of the b branches on this tree will be denoted $\mathbf{v} = (v_1, v_2, \ldots, v_b)$, with $b = 2s - 3$ for unrooted trees and $b = 2s - 2$ for rooted trees. The lengths of the branches are in terms of expected number of changes per site.

The model of evolution gives the probability of a change from one state to another over a specified evolutionary distance, v. These transition probabilities are denoted $p_{ij}(v, \theta)$, where i is the starting state, j is the ending state, v is the length of the branch, and θ is a vector of evolutionary parameters (such as the transition/transversion rate ratio). Currently, all stochastic models of evolutionary change on a phylogeny are examples of a continuous-time Markov chain. Figure 15.1 shows an example of a Markov chain in which there are four events of substitution (change in a DNA sequence) at a site. The substitutions occurred at times t_1, t_2, t_3, and t_4. In Figure 15.1, W_i is the waiting time for the ith substitution. The waiting times for a continuous-time Markov chain follow an exponential distribution with parameter λ (the rate may change depending on the state of the chain). The waiting time until the rth substitution ($W_r = W_1 + W_2 + \cdots + W_r$) follows a gamma distribution with shape parameter r and scale parameter λ. The number of substitutions in time interval T_1 is $N(T_1) = 3$ and the number of substitutions in the time interval T_2 is $N(T_2) = 4$. For a Poisson process, the
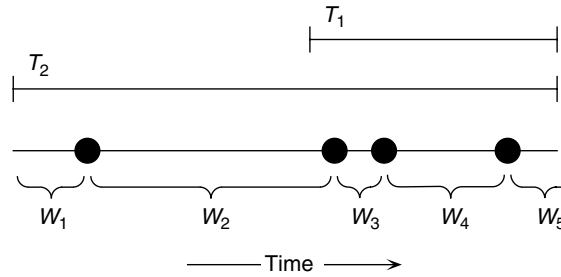
Figure 15.1 An illustration of a Poisson process. The large dots represent times at which substitutions occur.

number of substitutions in nonoverlapping time intervals are independent, the number of substitutions in a time interval of duration T follows a Poisson distribution with parameter λT, and the number of substitutions at time $T = 0$ (the beginning of the process) is $N(0) = 0$.

For models of character change, different transition types often have different rates. For example, for the simple case of two states ('0' or '1'), the rate of change from $0 \to 1$ and $1 \to 0$ might be different:

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}.$$

The rows give the rate of change from state 0 or 1 and the columns give the rate of change to state 0 or 1. This means that the rate of change from $0 \to 1$ is α and the rate of change from $1 \to 0$ is β. The diagonal entries of the matrix give the rate of change away from a particular state. The matrix $\mathbf{Q}$ is called the instantaneous rate matrix. The transition probabilities (probability of a change from 0 to 1 or 1 to 0 over a time period t) are calculated by exponentiating the matrix $\mathbf{Q}t$:

$$\mathbf{P}(t, \theta) = \{p_{ij}(t, \theta)\} = \mathrm{e}^{\mathbf{Q}t} = \begin{pmatrix} \dfrac{\beta}{\alpha + \beta} + \dfrac{\alpha \mathrm{e}^{-(\alpha+\beta)t}}{\alpha + \beta} & \dfrac{\alpha}{\alpha + \beta} - \dfrac{\alpha \mathrm{e}^{-(\alpha+\beta)t}}{\alpha + \beta} \\ \dfrac{\beta}{\alpha + \beta} - \dfrac{\alpha \mathrm{e}^{-(\alpha+\beta)t}}{\alpha + \beta} & \dfrac{\alpha}{\alpha + \beta} + \dfrac{\beta \mathrm{e}^{-(\alpha+\beta)t}}{\alpha + \beta} \end{pmatrix},$$

where $\theta = (\alpha, \beta)$ contains the parameters of the substitution model. Note that if the process is run for a very long time, that the probability that the process is in state 0 or in state 1 converges to $\beta/(\alpha + \beta)$ and $\alpha/(\alpha + \beta)$, respectively. These are the stationary or equilibrium frequencies of the process and are usually denoted π_i for the frequency of state i.

The transition probability depends upon two quantities: the rate at which a change occurs between state i and state j (q_{ij}) and the amount of time the process has been running (t). For the phylogeny problem, the observed data can be explained by either a high substitution rate and a small amount of time or by a low substitution rate and a large amount of time. For example, the transition probabilities for the two-state example given above are the same if $\alpha = 1$, $\beta = 1$, and $t = 1$ or if $\alpha = 2$, $\beta = 2$, and $t = 0.25$. For this reason, the branches of a phylogenetic tree are usually in terms of expected number of

substitutions per site, v. To make the branch lengths of a phylogenetic tree in terms of expected number of substitutions per site, the rate matrix $\mathbf{Q}$ is rescaled in such a way that the average rate of change away from some state is 1 (that is, $-\sum_{i=1}^{N} \pi_i q_{ii} = 1$, where N is the number of states).

The model of evolution can be tailored to accommodate different types of data. The 2×2 instantaneous rate matrix discussed above, for example, might be most appropriate for morphological data where only two states are typically scored (e.g. the presence or absence of a character state). For DNA or RNA sequence data, each site is assumed to take one of four states (A, C, G, or T/U) and a 4×4 instantaneous rate matrix is typically used. Details of calculating transition probabilities for models of DNA sequence evolution are discussed in Swofford *et al.* (1996). Typical models of DNA substitution include the Jukes and Cantor (1969) model where all substitution types have equal rates and the equilibrium frequencies of the four nucleotides are equal; the Kimura (1980) model, which allows transitions (A $\leftrightarrow$ G or C $\leftrightarrow$ T changes) to occur at a different rate than transversions (A $\leftrightarrow$ C, A $\leftrightarrow$ T, C $\leftrightarrow$ G, and G $\leftrightarrow$ T changes) but assumes that the frequencies of the four nucleotides are equal; and the Hasegawa *et al.* (1984; 1985) model which accommodates a transition/transversion rate bias and potentially different nucleotide frequencies. Even a model that allows there to be 11 different substitution rates has been implemented for DNA sequence evolution (Yang, 1994b). Table 15.1 summarizes the properties of models of DNA substitution commonly used in phylogenetic analysis.

One useful aspect of considering evolution as a stochastic process is that the evolutionary model can be modified to accommodate new biological processes as they are

Table 15.1 Summary of the major features of different models of DNA substitution. All of these models are time reversible.

Rate	JC69	K80	K81	F81	F84	HKY85	TN93	GTR
A $\rightarrow$ C	1	1	1	π_C	π_C	π_C	π_C	$r_{AC}\pi_C$
A $\rightarrow$ G	1	κ	κ	π_G	$\left(1 + \dfrac{K}{\pi_A + \pi_G}\right)\pi_G$	$\kappa\pi_G$	$\kappa_1\pi_G$	$r_{AG}\pi_G$
A $\rightarrow$ T	1	1	α	π_T	π_T	π_T	π_T	$r_{AT}\pi_T$
C $\rightarrow$ A	1	1	1	π_A	π_A	π_A	π_A	$r_{AC}\pi_A$
C $\rightarrow$ G	1	1	α	π_G	π_G	π_G	π_G	$r_{CG}\pi_G$
C $\rightarrow$ T	1	κ	κ	π_T	$\left(1 + \dfrac{K}{\pi_C + \pi_T}\right)\pi_T$	$\kappa\pi_T$	$\kappa_2\pi_T$	$r_{CT}\pi_T$
G $\rightarrow$ A	1	κ	κ	π_A	$\left(1 + \dfrac{K}{\pi_A + \pi_G}\right)\pi_A$	$\kappa\pi_A$	$\kappa_1\pi_A$	$r_{AG}\pi_A$
G $\rightarrow$ C	1	1	α	π_C	π_C	π_C	π_C	$r_{CG}\pi_C$
G $\rightarrow$ T	1	1	1	π_T	π_T	π_T	π_T	π_T
T $\rightarrow$ A	1	1	α	π_A	π_A	π_A	π_A	$r_{AT}\pi_A$
T $\rightarrow$ C	1	κ	κ	π_C	$\left(1 + \dfrac{K}{\pi_C + \pi_T}\right)\pi_C$	$\kappa\pi_C$	$\kappa_2\pi_C$	$r_{CT}\pi_C$
T $\rightarrow$ G	1	1	1	π_G	π_G	π_G	π_G	π_G
$\pi_A = \pi_C = \pi_G = \pi_T$	Yes	Yes	Yes	No	No	No	No	No

discovered. For example, models of DNA substitution have been modified to accommodate limited dependence among sites. Schöniger and von Haeseler (1994) accommodate nonindependent substitutions in Watson–Crick bond pairs in stem regions of rRNA genes by expanding the model of DNA substitution around doublets of nucleotides; instead of a 4×4 matrix of rates, there is a 16×16 matrix of rates from one doublet to any other doublet (e.g. from doublet AA $\rightarrow$ AC). Only one substitution in any instant of time is allowed, so many of the elements of the **Q** matrix are 0 (e.g. changes that require a change at both positions simultaneously, such as AA $\rightarrow$ GC). Similarly, models of DNA substitution have been expanded around the codon – a triplet of nucleotides (Goldman and Yang, 1994; Muse and Gaut, 1994). Hence, the instantaneous rate matrix (**Q**) is 64×64 (or, more commonly, 61×61 because stop codons are excluded from consideration). Importantly, codon models allow parameters such as the nonsynonymous/synonymous rate ratio to be directly estimated, and account for all possible histories of synonymous and nonsynonymous change.

The probability of observing the data at a particular site is a sum over all possible nucleotides that can be observed at the internal nodes of the tree. Let **y** be a generic vector of states at the internal nodes of the tree. We will assume that the tree is rooted and that the tips of the tree are labelled $1, 2, \ldots, s$ and the internal nodes of the tree are labeled $s + 1, s + 2, \ldots, 2s - 2$. Moreover, in the following, the ancestor of node k on the tree is denoted $\sigma(k)$ and the length of the kth branch is v_k. The probability of observing the data at site i, then, is

$$f(\mathbf{x}_i|\tau, \mathbf{v}, \theta) = \sum_{\mathbf{y}} \pi_{w_{2s-2}} \left(\prod_{k=1}^{s} p_{w_{\sigma(k)}x_k}(v_k, \theta) \right) \left(\prod_{k=s+1}^{2s-2} p_{w_{\sigma(k)}w_k}(v_k, \theta) \right).$$

The summation is over all 4^{s-1} possible assignments of nucleotides to the internal nodes of the tree. The reason why the probability of observing the data at a site is a sum over all possible assignments of nucleotides at the internal nodes of the tree is that we do not want our inferences to be conditioned on a particular character history; instead, the probability of observing the data is a weighted average over all possible character histories. The number of possible combinations of nucleotides becomes too large to enumerate for even moderately sized problems (e.g. for $s = 20$ species there are 2.75×10^{11} combinations of nucleotides). However, Felsenstein (1981) described a pruning algorithm that takes advantage of the tree topology to evaluate the summation over nucleotide states in a computationally efficient (but mathematically equivalent) manner.

Assuming independence of the substitutions at different sites in the sequence, the probability of observing the entire aligned sequence data set is simply the product of the probabilities of observing the data at each site:

$$f(X|\tau, \mathbf{v}, \theta) = \prod_{i=1}^{c} f(\mathbf{x}_i|\tau, \mathbf{v}, \theta).$$

Typically, this number will be very small (because many numbers between 0 and 1 are multiplied). The log of the probability is used instead so that the number can be accurately stored in computer memory.

The above calculation for the probability of observing the aligned data matrix assumes that the same set of branch lengths (**v**) apply to every site in the sequence. This is another way of saying that the rate across sites is assumed to be equal. This assumption has been

relaxed in two different ways. The first divides the observed sequences into partitions, such as first, second, and third codon positions. The rate of substitution within each partition is assumed to be homogenous but potentially different across partitions. A protein coding gene, for example, might be divided into three partitions according to codon position.

The other method for accommodating rate variation assumes that the rate at a site is unknown but drawn from some distribution. Usually, the rate at a site is assumed to follow the proportion of invariant sites model or a gamma distribution (Yang, 1993). The proportion of invariant sites model states that the rate at a site is $r = 0$ with probability p and $r = 1/(1 - p)$ with probability $1 - p$. The gamma model assumes that the rate at a site is drawn from a gamma distribution with shape and scale parameters equal to α. The gamma distribution, then, is scaled such that the mean rate is 1 and the variation in rates is $1/\alpha$. For the proportion of invariant sites model, the probability of observing the data at the ith site is

$$f(\mathbf{x}_i | \tau, \mathbf{v}, \theta, p) = \left[\sum_{\mathbf{y}} \pi_{w_{2s-2}} \left(\prod_{k=1}^{s} p_{w_{\sigma(k)} x_k}(0, \theta) \right) \left(\prod_{k=s+1}^{2s-2} p_{w_{\sigma(k)} w_k}(0, \theta) \right) \right] p$$

$$+ \left[\sum_{\mathbf{y}} \pi_{w_{2s-2}} \left(\prod_{k=1}^{s} p_{w_{\sigma(k)} x_k}(v_k/(1 - p), \theta) \right) \right.$$

$$\left. \times \left(\prod_{k=s+1}^{2s-2} p_{w_{\sigma(k)} w_k}(v_k/(1 - p), \theta) \right) \right] (1 - p),$$

which is the sum of the probability of observing the data for each category weighted by the prior probability that the site is in each category. Similarly, the probability of observing the data at a site for the gamma-distributed rate variation model is

$$f(\mathbf{x}_i | \tau, \mathbf{v}, \theta, \alpha)$$

$$= \int_0^{\infty} \left[\sum_{\mathbf{y}} \pi_{w_{2s-2}} \left(\prod_{k=1}^{s} p_{w_{\sigma(k)} x_k}(v_k r, \theta) \right) \left(\prod_{k=s+1}^{2s-2} p_{w_{\sigma(k)} w_k}(v_k r, \theta) \right) \right] f(r | \alpha, \alpha) \, \mathrm{d}r$$

where $f(r | \alpha, \alpha)$ is the gamma density with shape and scale parameters equal to α. This function cannot usually be calculated numerically. Instead, an approximation suggested by Yang (1994a) is used. The gamma distribution is divided into K categories of equal weight, with the mean rate for each category representing the rate for the entire category. The probability of observing the data at the ith site then becomes

$$f(\mathbf{x}_i | \tau, \mathbf{v}, \theta, \alpha) = \sum_{j=1}^{K} \left[\sum_{\mathbf{y}} \pi_{w_{2s-2}} \left(\prod_{k=1}^{s} p_{w_{\sigma(k)} x_k}(v_k r_j, \theta) \right) \left(\prod_{k=s+1}^{2s-2} p_{w_{\sigma(k)} w_k}(v_k r_j, \theta) \right) \right] \frac{1}{K}.$$

The logic applied to modelling rate variation across sites has also been applied to the problem of detecting sites that are under the influence of positive selection. The nonsynonymous to synonymous rate of substitution (ω) is assumed to be a random variable drawn from some distribution. The probability of observing the data at a codon position is a sum over all possible values for ω weighted by the prior probability for each ω. Nielsen and Yang (1998) used this type of model to calculate the probability that a particular site is in the positive selection category conditioned on the observed sequences at that site.

As mentioned above, many different types of data can be used in phylogenetic analyses using the likelihood function. Until recently, morphological data were a major exception. Morphological data are difficult to accommodate using stochastic models for two reasons. For one thing, morphological traits are very complex genetically; it is difficult to imagine that many important biological processes are being captured by a simple model of stochastic change. Another impediment to introducing stochastic models of evolution to morphological data has been related to sampling biases in morphological data. Two-state stochastic models have been around for decades that, in principal, could be applied to morphological data; morphological characters are typically scored as presence or absence of a trait, so a simple two-state model seems natural. However, direct application of such a model would not be valid because the calculations assume that all characters have been sampled; most morphologists do not score invariant characters in a phylogenetic analysis. A morphological study of lizards, for instance, never includes a character such as 'presence or absence of vertebral column'. Recently Lewis (1999) proposed a solution to this problem: One simply conditions on the fact that these characters are not sampled in the first place. Practically speaking, the analysis is changed only in that the probability of observing a dummy invariant character must be calculated along with the observed characters.

15.4 DEVELOPING AN INTUITION OF LIKELIHOOD

One of the strengths of likelihood methods is that the assumptions of the analysis are explicit and transparent. What is not so transparent to many is what a likelihood score is and what it means to sum over all possible ancestral states. In this section, we try to develop an understanding of likelihood with reference to the parsimony method.

Consider the simple tree of two sequences shown in Figure 15.2(a). For this tree, the nucleotide 'A' was observed for both species at a particular site. The parsimony method would reconstruct the history of this character as involving no changes. What would likelihood do? The reconstruction provided by maximum likelihood would depend upon the length of the branch separating the two species as well as upon the model of evolution assumed. In this example, we will assume that the sequences evolve under a very simple model of DNA substitution first described by Jukes and Cantor (1969) and consider several different branch lengths (v).

Table 15.2 shows how much weight likelihood and parsimony give to different reconstructions. The numbers in parentheses provide the probability of observing the site
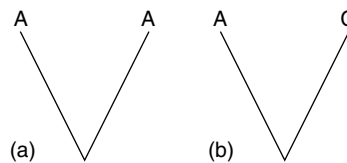
Figure 15.2 A simple phylogenetic tree of two species with two different character patterns (a) AA and (b) AC.

Table 15.2 The percentage probability of different character histories under the Jukes–Cantor model for the tree of Figure 15.2(a). The length of the branch separating the two sequences is denoted v.

Number of changes	Parsimony	$v = 0.01$ (0.2475)	$v = 0.1$ (0.2266)	$v = 0.2$ (0.20611)	$v = 1$ (0.11192)
0	100	99.99	99.83	99.31	82.17
1	0	0.00	0.00	0.00	0.00
2	0	0.0011	0.11	0.44	9.13
3	0	–	–	0.034	3.55
4	0	–	–	–	0.0027

pattern (AA) given the branch length v. The percentages are the percent of the site pattern probability that involves reconstructions of 0, 1, 2, or more changes. Parsimony places all of the weight on the reconstruction that has 0 changes across the simple two-species tree. Likelihood, on the other hand, gives some weight to *all possible reconstructions*. In a sense, likelihood hedges its bets with respect to reconstructing character history. When the branch length is small (e.g. $v = 0.01$), likelihood, like parsimony, places most of the weight on the reconstruction that has no changes. For this particular pattern of nucleotides at the branch tips, likelihood places no weight on the reconstruction that has one change because it is impossible to explain the data with only a single change on the tree. However, likelihood does place some weight on reconstructions that involve two or more changes. Note that as the branch length increases, likelihood places more weight on reconstructions involving multiple changes. In fact, when the branch length is $v = 1.0$, likelihood places 9.13 % of the weight on reconstructions involving two changes and 3.55 % weight on reconstructions involving three changes.

Parsimony and likelihood also behave similarly when there must be a character change. Consider the same tree shown in Figure 15.2(b), now with nucleotide 'A' observed for one species and 'C' for the other species. Table 15.3 gives the weights assigned to various reconstructions for this character pattern. The probability of observing a site pattern with a change increases as the branch length increases. This makes sense because as the branch becomes longer, it becomes more probable that changes will occur along that branch. Also, note that the parsimony and likelihood methods are most similar when the branch length is small; when the branch length is small, parsimony places 100 % weight on the

Table 15.3 The percentage probability of different character histories under the Jukes–Cantor model for the tree of Figure 15.2(b). The length of the branch separating the two sequences is denoted v.

Number of changes	Parsimony	$v = 0.01$ (0.00083)	$v = 0.1$ (0.00786)	$v = 0.2$ (0.01462)	$v = 1$ (0.04602)
0	0	0.00	0.00	0.00	0.00
1	100	99.66	96.64	92.36	66.54
2	0	0.33	3.22	6.22	22.19
3	0		0.12	0.48	8.61
4	0		0.003	0.023	2.05
5	0			0.0037	0.42

reconstruction involving one change whereas likelihood places 99.66 % weight on the same reconstruction.

15.5   METHOD OF MAXIMUM LIKELIHOOD

The likelihood function (probability of observing the data) depends upon several unknown parameters. For the phylogeny problem, these parameters include (minimally) the topology of the phylogenetic tree relating the species and the lengths of the branches on this tree, but may also depend upon parameters of the substitution process, such as the transition/transversion rate ratio, or parameters that describe the degree of rate variation across sites. Up to this point in the chapter, we have assumed that these parameters are fixed. However, we would like to estimate these parameters as they are generally unknown.

The method of maximum likelihood estimates parameters of a stochastic model by finding that combination of parameter values that maximizes the likelihood function. Take, for example, the coin-tossing problem, discussed above. The likelihood function was

$$\ell(p) = C \times \binom{n}{x} p^x (1 - p)^{n-x},$$

which was proportional to the probability of observing the data given a specific value for p (the unknown parameter). The method of maximum likelihood would estimate p by finding that value of p that maximized the likelihood function. This can be done by taking the first derivative of the likelihood function and finding where the first derivative equals 0. Taking the log of the likelihood function does not change where the function is maximized and makes the calculation easier:

$$\log_e \ell(p) = \log_e C + \log_e \binom{n}{x} + x \log_e p + (n - x) \log_e (1 - p).$$

Then, the first derivative is calculated:

$$\frac{d \log_e \ell(p)}{dp} = \frac{x - np}{p(1 - p)}.$$

Solving for $d \log_e \ell(p) / dp = 0$ gives $\hat{p} = x/n$. This means that the maximum likelihood estimate of the parameter p is simply the proportion of the time that heads were observed in the n tosses. This estimate does not depend upon the value of the constant C in the likelihood function.

For the phylogeny problem, the peak of the likelihood surface cannot be found analytically (as it could for the coin-tossing problem). Instead, the peak of the likelihood surface must be found numerically. Most modern computer programs take the following approach: A tree is visited and the parameters for that tree are changed such that the likelihood is maximized. The maximum likelihood value for this tree is then stored and the procedure is repeated for other trees. The tree that has the largest maximum likelihood value is the best estimate of phylogeny. The maximization of the likelihood for the phylogeny problem, then, involves two distinct problems: finding the maximum likelihood combination of parameters for distinct trees and finding the tree that has the greatest likelihood. Lewis (1998) has taken a different approach to finding the maximum likelihood estimate

of phylogeny by using a genetic algorithm. The genetic algorithm takes a population of trees and subjects the trees to mutation (changes of the parameter values on the trees) and recombination (grafting parts of one tree on to another tree). The population is allowed to evolve under the influence of selection, with those trees that have the greatest fitness (highest likelihood scores) contributing the larger number of trees in the next generation.

A practical limitation of maximum likelihood in phylogenetics is that it is very difficult to maximize the likelihood simultaneously for many parameters. Often, the biologist is interested in inferring phylogeny under a more parameter-rich (and presumably more realistic) model of DNA substitution. However, some of the parameters in a substitution model are difficult to maximize. It turns out that it is relatively easy to maximize the likelihood with respect to branch lengths because the algorithm can take advantage of information on the slope and curvature of the likelihood surface; this information tells the algorithm in which direction and by how much to change the length of a branch. Other parameters, such as the transition/transversion rate ratio or parameters specific to the model of DNA substitution, are inherently more difficult to estimate because it is difficult to obtain information on the slope and curvature of the likelihood function for those parameters. This means that it can become very time consuming to maximize the likelihood for parameter-rich models of DNA substitution.

One strategy for maximizing the likelihood under complex models of DNA substitution that can be implemented in PAUP* (Swofford, 1998) is to maximize the likelihood in a stepwise manner (Swofford *et al.*, 1996). First, a reasonable starting tree (e.g. a maximum parsimony tree) is obtained. The parameters of the substitution model are maximized on this tree. Next, the parameters of the substitution model are fixed to their maximum likelihood values and a heuristic search is performed starting from the parsimony tree. This procedure is iterated until one cannot find a better tree than the one on which the parameters were estimated. This strategy relies on the idea that, for trees near the maximum likelihood tree, parameters of the substitution model do not differ by very much. An example of a PAUP block that can implement this search strategy under the GTR $+\ \Gamma$ model of DNA substitution is as follows:

```
begin paup;
set autoclose = yes warnreset = no increase = auto criterion
    = parsimony;
hsearch;
set criterion = likelihood;
lset nst = 6 rmatrix = est basefreq = est rates = gamma shape
    = est ncat = 4 pinvar = 0;
lscores 1;
lset rmatrix = prev basefreq = prev shape = prev;
hsearch start = 1 swap = nni;
savetrees file = like.trees replace = yes;
lset rmatrix = est basefreq = est shape = est;
lscores 1;
lset rmatrix = prev basefreq = prev shape = prev;
hsearch start = 1 swap = spr;
savetrees file = like.trees append = yes;
lset rmatrix = est basefreq = est shape = est;
```

```
lscores 1;
lset rmatrix = prev basefreq = prev shape = prev;
hsearch start = 1 swap = tbr;
savetrees file = like.trees append = yes;
end;
```

This PAUP block performs three iterations of the search strategy using three different heuristic searches.

The phylogeny problem involves numerous parameters (e.g. minimally tree topology and branch lengths). Often, however, the biologist is interested in only one or a few of these parameters. The parameters that are not of interest are called nuisance parameters. (For a more extensive discussion of nuisance parameters in phylogenetic analysis, see Goldman, 1990.) Nuisance parameters can be dealt with in one of several ways. First, the likelihood can be maximized with respect to all parameters, including the nuisance parameters. This is the solution taken in most computer programs implementing maximum likelihood. Second, a prior probability distribution can be assigned to these parameters and the likelihood integrated over the range of the parameter. This is how rate variation across sites is often accommodated; a prior such as the gamma distribution is placed on the rate at a site and the overall site likelihood is integrated over all possible rate assignments. Finally, a consistent estimator (but not necessarily a maximum likelihood estimator) can be substituted for the parameter. To our knowledge, this pseudomaximum likelihood method has not been used in phylogenetics.
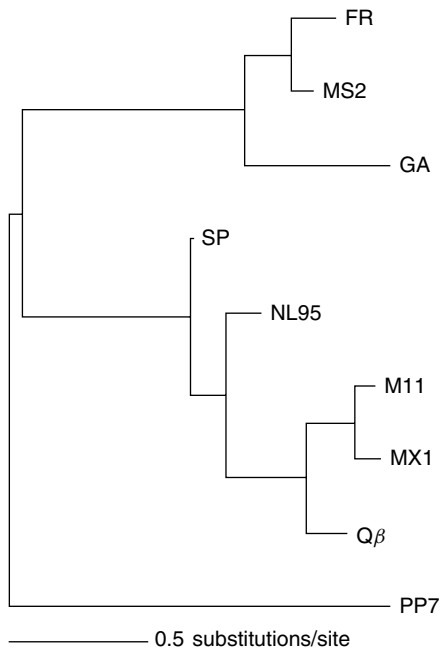
Figure 15.3 The maximum likelihood estimate of phylogeny under the HKY85 + Γ model of DNA substitution. Parameter estimates for the nuisance parameters were $\kappa = 2.33$, $\alpha = 0.69$, $\pi_A = 0.24$, $\pi_C = 0.26$, $\pi_G = 0.23$, and $\pi_T = 0.27$.

Figure 15.3 shows the maximum likelihood tree for the bacteriophage species. The Hasegawa *et al.* (1985) model of DNA substitution with gamma-distributed rate variation (Yang, 1994a) was assumed in the analysis. Nuisance parameters (e.g. the transition/transversion rate ratio, base frequencies, and branch lengths) were accommodated by maximizing the likelihood with respect to those parameters. The log-likelihood of the tree was -5695.07. The maximum likelihood estimates for the model parameters were $\kappa = 2.33$, $\alpha = 0.69$, $\pi_A = 0.24$, $\pi_C = 0.26$, $\pi_G = 0.23$, and $\pi_T = 0.27$. The analysis was performed using the program PAUP* (Swofford, 1998).

15.6   BAYESIAN INFERENCE

In a Bayesian analysis inferences are based upon what is called the posterior probability of a parameter. The posterior probability of a parameter is the probability of the parameter conditional on the observed data. This is different from maximum likelihood which maximizes the likelihood function (the probability of observing the data given the parameter). The posterior probability of a hypothesis (H_i; $i = 1, 2, \ldots, n$) can be calculated using Bayes formula,

$$f(H_i|D) = \frac{f(D|H_i)f(H_i)}{f(D)},$$

where

$$f(D) = \sum_{j=1}^{n} f(D|H_j)f(H_j)$$

for discrete H. The posterior probability is a function of the likelihood, $f(D|H_i)$, and the scientist's prior beliefs about the hypothesis, $f(H_i)$. Just as with maximum likelihood, the likelihood function is the vehicle that carries the information about the parameter contained in the data.

Figure 15.4 shows the posterior probabilities of the parameter p for the coin-tossing experiment discussed earlier. Here, the proportion of heads observed is always 0.8, for 5, 10, and 100 tosses of a coin. The likelihood function is the first column in this figure. The first (top) row of this figure shows three different priors that express the uncertainty in p before any observations have been made; the first is an uninformative prior, giving equal weight to all values of p in the interval between 0 and 1. The second and third are informative priors, placing more weight on values of p near 0.5. The other plots in Figure 15.4 show the posterior probability of p when the likelihood and prior are combined; the likelihood function for the observations of the row are combined with the prior of the column. Bayes formula is used to combine the scientist's prior beliefs with observation.

How are posterior probabilities of phylogenies calculated? The posterior probability of tree τ_i is obtained using Bayes rule,

$$f(\tau_i|X, \mathbf{v}, \theta) = \frac{f(X|\tau_i, \mathbf{v}, \theta)f(\tau_i)}{\overset{B(s)}{\underset{j=1}{\sum}} f(X|\tau_j, \mathbf{v}, \theta)f(\tau_j)},$$
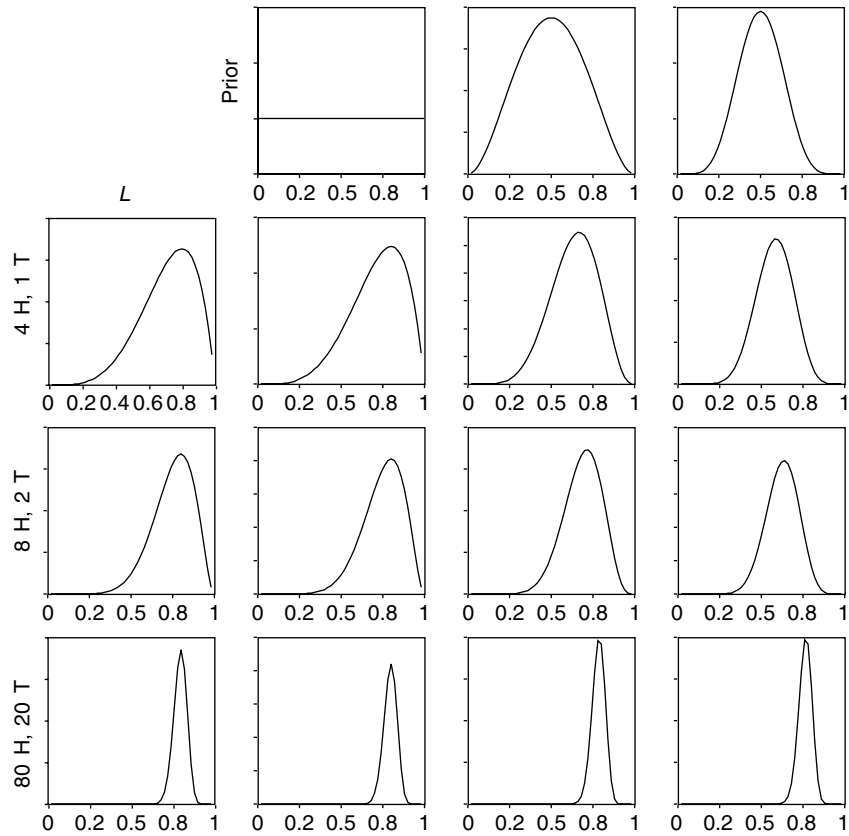
Figure 15.4   An example of Bayesian inference for the coin-tossing example with different priors.

where $B(s)$ is the number of possible trees for s species ($B(S) = (2s - 3)!/2^{s-2}(s - 2)!$ for rooted trees and $B(s) = (2s - 5)!/2^{s-3}(s - 3)!$ for unrooted trees). One possible prior for trees is to set $f(\tau_i) = 1/B(s)$, making all trees a priori equally probable.

Note that the posterior probability of a tree was conditioned on the data and several parameters, such as the branch lengths ($\mathbf{v}$) and substitution parameters (θ). One approach to accommodating these parameters is to substitute an estimate (such as the maximum likelihood estimate). This is called an empirical Bayes analysis. An alternative method integrates over uncertainty in these parameters; the inference of phylogeny, then, is not conditioned on unknown parameters taking specific values. Such an analysis is called a hierarchical Bayes analysis. The posterior probability for a hierarchical Bayes analysis is

$$f(\tau_i|X) = \frac{f(X|\tau_i)f(\tau_i)}{\sum\limits_{j=1}^{B(s)} f(X|\tau_j)f(\tau_j)},$$

where

$$f(X|\tau_i) = \int f(X|\tau_i, \mathbf{v}, \theta) \, \mathrm{d}F(\mathbf{v}, \theta)$$

and integration is over all possible combinations of branch lengths on the tree and over all possible values for the substitution parameters. An attractive feature of a hierarchical Bayes analysis is that inferences are conditioned only upon the data.

A Bayesian analysis of phylogeny requires that the scientist specify his or her beliefs about the phylogeny before observing any data. This can be viewed as either a strength or a weakness of Bayesian inference. It is a strength in that the method can potentially take advantage of any prior knowledge the scientist might have about the phylogeny. On the other hand, priors can be difficult to specify. Two different types of priors have been used in the phylogeny problem. The first was specified by Rannala and Yang (1996; also see Yang and Rannala, 1997; Thompson, 1975), who used a birth–death process of cladogenesis to specify the prior probabilities of phylogeny and branch lengths. Under a random branching model of cladogenesis (such as the birth–death process), all labeled histories have equal probability. Labeled histories are distinguished from one another not only by their topology but also by the relative speciation times (Edwards, 1970). The birth–death process has two parameters, the speciation and extinction rates. These can be fixed, or priors (called hyperpriors) can be placed on the rates of speciation and extinction. Mau (1996), Mau and Newton (1997), and Newton *et al.* (1999) place uninformative priors on topology and branch lengths. Equal weight is placed on all trees, and uniform priors (from 0 to a large number) are placed on branch lengths. The posterior probability will be mainly determined by the likelihood function when uninformative priors are used.

15.7   MARKOV CHAIN MONTE CARLO

The posterior probability of a tree involves, minimally, a summation over all possible trees and integration over branch lengths and other parameters of the substitution model. In short, the posterior probability of a tree cannot be evaluated analytically, and must be approximated. Markov chain Monte Carlo (MCMC) can be used to approximate the posterior probability of phylogenies.

MCMC is a method that makes valid, but dependent, draws from the probability distribution of interest. MCMC constructs a Markov chain that has as its state space the parameter(s) of interest. A simple example will illustrate MCMC as applied to the coin–tossing problem (for a good introduction to MCMC, see Gilks *et al.*, 1996). This example is of a variant of MCMC called the Metropolis–Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970; Green, 1995). The current state of the chain will be denoted p. If this is the first generation of the chain, then p is initialized to be some value between 0 and 1. A new state, p', is proposed. For this example, the proposal mechanism will be to pick a uniformly distributed random number in the interval $(p - \varepsilon, p + \varepsilon)$; this means that the next state of the chain will be very similar to the current state (ε might be 0.05, for example, meaning that the proposed state of the chain will be within 0.05 of the current state). The proposed state is accepted with probability

$$R = \min\left(1, \frac{\binom{n}{x} p'^x (1 - p')^{n-x} f(p')/f(x)}{\binom{n}{x} p^x (1 - p)^{n-x} f(p)/f(x)} \times \frac{f(p|p')}{f(p'|p)}\right)$$

$$= \min \left(1, \underbrace{\frac{p'^x(1-p')^{n-x}}{p^x(1-p)^{n-x}}}_{\text{likelihood ratio}} \times \underbrace{\frac{f(p')}{f(p)}}_{\text{prior ratio}} \times \underbrace{\frac{f(p|p')}{f(p'|p)}}_{\text{proposal ratio}} \right).$$

Note that the most difficult part of the posterior probability to calculate ($f(x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} f(p) \, dp$) cancels. The proposal probabilities ($f(p'|p)$ and $f(p|p')$) give the probability of proposing a move $p \to p'$ or the reverse move $p' \to p$. Many proposal mechanisms will work as long as the chain is aperiodic and irreducible. Hence, the computer programmer has much latitude in choosing proposal mechanisms, and, in fact, there is an art to constructing a Markov chain that works well. The important point, though, is that any chain will work in that it will converge on the correct distribution if the chain is run long enough.

Once the acceptance probability, R, is calculated, a uniformly distributed random number on the interval [0, 1] is generated. If this random number is less than R, then the proposed state is accepted and $p = p'$. If the new state is not accepted (i.e. the random number is greater than R), then the current state stays the same. This process of proposing a new state and then accepting or rejecting it is repeated a large number of times. The sequence of states (values of p) visited over the course of the analysis forms a Markov chain. The states visited during the analysis (i.e. different values of p) are valid, albeit dependent, draws from the posterior probability; a Markov chain that was run for
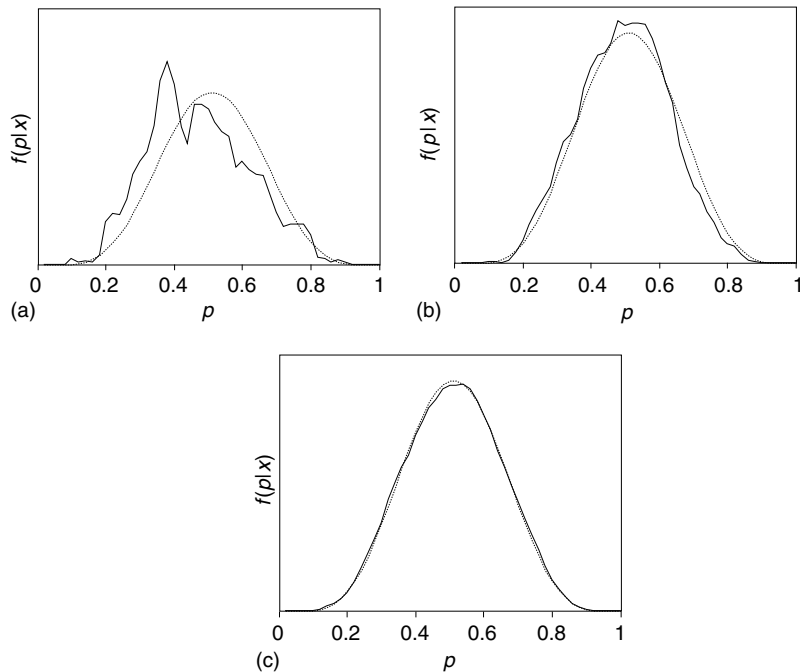
Figure 15.5 The posterior probability of p for the coin-tossing problem. The dotted line is the posterior probability calculated analytically. The solid line is the approximation of the posterior probability obtained using Markov chain Monte Carlo. The chain was run for (a) 5000 generations (b) 50 000 generations, and (c) 500 000 generations. *Source*: Huelsenbeck *et al.* (2000a).

100 steps and sampled at every step does not represent 100 independent draws from the posterior probability. However, the proportion of the time that different values of p were visited is a valid approximation of the posterior probability. Figure 15.5 shows an example of an MCMC analysis for the coin-tossing example where the data are $x = 5$ heads in $n = 10$ tosses of a coin. Note that as the length of the chain is increased, that the MCMC approximation converges to the true (analytically calculated) posterior probability.

Although MCMC was illustrated for a Bayesian analysis of coin tossing, the method can be extended to much more complex problems involving many parameters. Moreover, the method has been applied to maximum likelihood analysis where it has been used to integrate over uncertainty in coalescence histories when estimating population parameters (Kuhner *et al.*, 1994; Beerli and Felsenstein, 1999).

A limitation of MCMC is that it is not clear when the chain has been run long enough to produce a reliable approximation of the distribution of interest. The Markov chain law of large numbers guarantees that posterior probabilities can be validly estimated from long-run samples of the chain (Tierney, 1994). However, for any given analysis, it is difficult to determine if the chain has converged to the desired distribution. Several different heuristics are available to determine if the chain has converged (see Gelman, 1996). These involve running several independent chains starting from different (perhaps random) trees.

To illustrate Bayesian inference of phylogeny using MCMC, we analyzed the bacteriophage *replicase* data under the HKY85 + Γ model of DNA substitution. We used MCMC to approximate the posterior probabilities of trees using a program written by one of us (JPH). The Markov chain was run for 500 000 steps. Figure 15.6 shows the log of the likelihood function through time for the chain. Note that at first the log-likelihood was low and that it quickly reached a plateau near the maximum likelihood value (indicated by
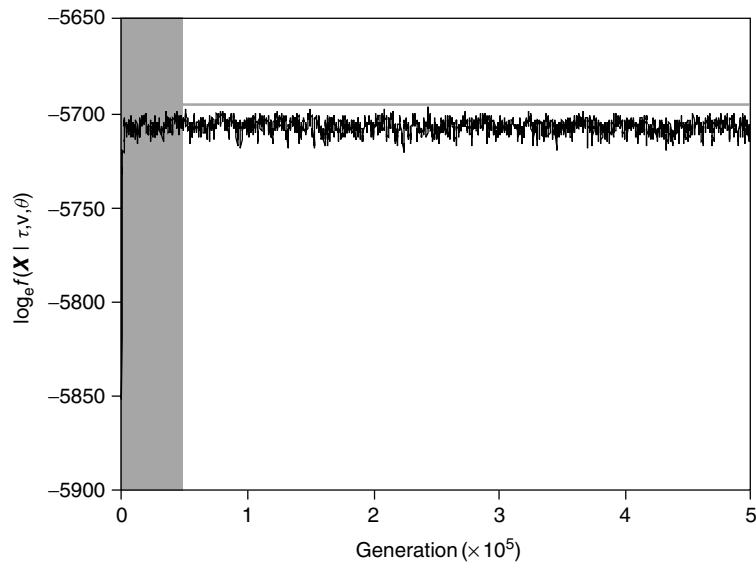
Figure 15.6 The log-likelihood function was tracked during the course of the MCMC analysis of the *replicase* sequences. The log-likelihood started at a low value, but quickly reached apparent stationarity. The shaded portion shows the part of the chain that was discarded as the 'burn-in'. The gray horizontal line shows the maximum likelihood value for the data.

a gray horizontal line). Inferences should be based on the portion of the chain at station-arity. Hence, we discarded the first 50 000 steps in the chain as the 'burn-in'. Figure 15.7 shows the 50 % majority rule tree of the sampled trees. The numbers at the interior nodes of the tree do not represent bootstrap support values but rather the posterior probability that the clade is true. Just as with maximum likelihood, the parameters of the substitution model can be estimated using Bayesian inference. Figure 15.8 shows the posterior probabilities for the transition/transversion rate ratio (κ) and the shape parameter of the gamma distribution (α) for among-site rate variation.
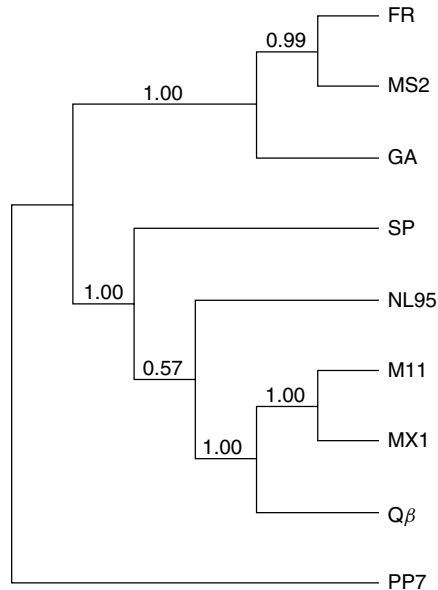
Figure 15.7 The Bayesian estimate of phylogeny. Only clades with greater than 0.5 posterior probability are shown. The numbers at the interior nodes represent the posterior probability that the clade is correct.
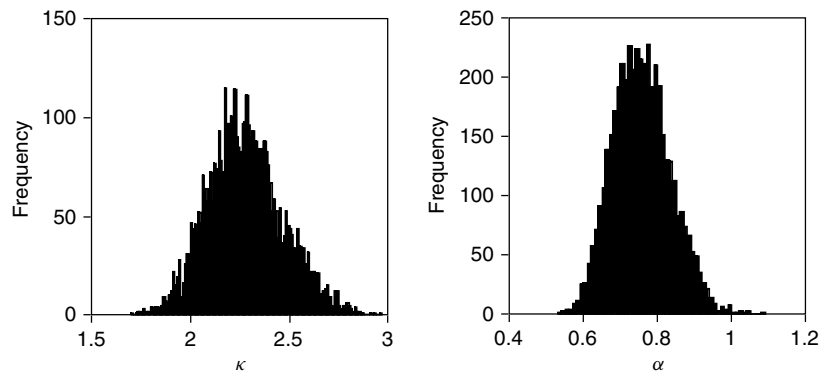
Figure 15.8 The posterior probability distribution of the transition/transversion rate ratio (κ) and the gamma shape parameter for among-site rate variation (α).

15.8   ASSESSING UNCERTAINTY OF PHYLOGENIES

The nonparametric bootstrap, introduced to the phylogeny problem by Felsenstein (1985), has become a standard method for assessing the uncertainty in phylogenetic trees. The bootstrap works by constructing many replicate data sets of the same size as the original data matrix by randomly sampling the sites with replacement. Hence, for any single bootstrap data matrix, some sites might be represented several times and others not at all. Each bootstrap data matrix is analyzed in the same manner as the original data. The proportion of the time that a particular clade is found in analysis of the bootstrapped data matrices represents the bootstrap confidence for that clade.

The appeal of the bootstrap is that it can be used to assess the accuracy of estimated phylogenetic trees ($\hat{\tau}$) without making explicit assumptions about the distribution of trees around the true tree (τ^T). However, the application of the bootstrap method to the phylogeny problem has been criticized as providing biased estimates of the accuracy of phylogenetic trees (Hillis and Bull, 1993); Hillis and Bull (1993) performed simulations in which the bootstrap distribution of trees ($\hat{\tau}^B$) was compared to the true tree that generated the simulated data (τ^T). Efron *et al.* (1996) argue that this is not the relevant comparison as the bootstrap method is intended to evaluate the (unobservable) distribution of differences between $\hat{\tau}$ and τ^T by using the (observable) distribution of differences between $\hat{\tau}^B$ and $\hat{\tau}$. The bootstrap is not severely biased in evaluating the difference between the estimated tree and the true tree, though Efron *et al.* (1996) argue that the bootstrap can be improved by performing a more elaborate bootstrap procedure. The bootstrap can also be used to construct confidence intervals for other parameters of evolutionary models. However, use of the bootstrap for this purpose is uncommon in the phylogenetics literature.

Efron *et al.* (1996) also argue that the bootstrap proportions on a tree have a Bayesian interpretation as the assessment of error in the estimated tree when assuming all phylogenetic trees have an equal prior probability. Larget and Simon (1999) point out that, if this is true, it is computationally much more efficient to perform a Bayesian analysis using MCMC than to perform a traditional nonparametric bootstrap using maximum likelihood. MCMC is very efficient in how it uses computational resources. Every calculation of the likelihood for a tree (a task that is computationally difficult) provides some information about the posterior probabilities of trees. If a move to a particular tree is often accepted then, roughly speaking, that tree has a higher posterior probability than a tree that is rarely moved to during the course of the Markov chain.

Figure 15.9 shows the maximum likelihood and Bayesian estimates of phylogeny for the bacteriophage groups with the support for the clades indicated. For the maximum likelihood tree the numbers at the interior branches represent bootstrap support for the clade, whereas for the Bayesian estimate the numbers indicate the posterior probability that the clade is correct. The confidence intervals for the parameters of the substitution model are $\kappa = 2.33$ (1.89, 2.88) for the transition/transversion rate ratio and $\alpha = 0.70$ (0.55, 0.85) for the gamma shape parameter describing the rate variation across sites. In a Bayesian analysis, the uncertainty in the parameters can be evaluated directly from the posterior probability of the parameter (see Figure 15.8). However, a 95 % credibility interval can be constructed by taking the 2.5 and 97.5 % quantiles of the posterior probability distribution. The 95 % credibility intervals for the transition/transversion rate and the gamma shape parameter are $\kappa = 2.28$ (1.93, 2.68) and $\alpha = 0.75$ (0.62, 0.92), respectively.
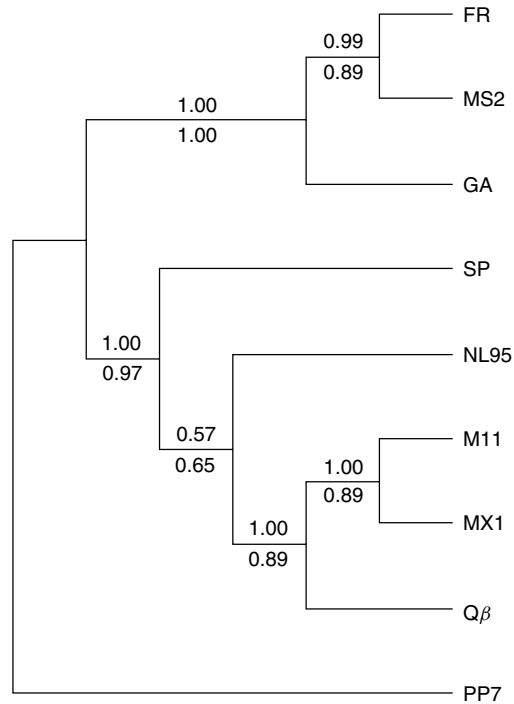
Figure 15.9   The best estimate of phylogeny for the Leviviridae for the *replicase* gene under the maximum likelihood and Bayesian criteria. The HKY85 + Γ model of DNA substitution was assumed. The numbers above the branches show the posterior probability that the clade is correct, whereas the numbers below the branches show the bootstrap support for the clade.

15.9   HYPOTHESIS TESTING AND MODEL CHOICE

There are a number of ways to construct hypothesis tests. One of the most general and important of these is based on the likelihood ratio. The likelihood ratio is the ratio of the probability of observing the same data under different models. The likelihood ratio in favor of model 1 against model 2 is

$$\Lambda_{12} = \frac{f(\text{Observations}|M_1)}{f(\text{Observations}|M_2)}.$$

When the models are nested (i.e. M_1 is a special case of M_2), twice the negative log of the ratio

$$\Lambda_{12} = \frac{\max[f(\text{Observations}|M_1)]}{\max[f(\text{Observations}|M_2)]},$$

is asymptotically χ^2 distributed, with the degrees of freedom being the difference in the number of free parameters between the two models (Wilks, 1938). Alternatively, if the models are not nested or if the assumptions of the test are otherwise violated, the null distribution of the test statistic can be simulated using the parametric bootstrap. Likelihood

ratio tests have been used extensively in phylogenetic analysis to test the molecular clock, to choose among different models of DNA substitution, and to test hypotheses on trees (see Goldman, 1993; reviewed in Huelsenbeck and Rannala, 1997).

Bayesian model choice is often guided by the Bayes factor (Kass and Raftery, 1995). The Bayes factor in favor of model 1 against model 2 is

$$B_{12} = \frac{\dfrac{f(M_1|\text{Observations})}{f(M_2|\text{Observations})}}{\dfrac{f(M_1)}{f(M_2)}},$$

which is the ratio of the posterior odds ratio to the prior odds ratio. The Bayes factor is on roughly the same scale as the likelihood ratio and should be interpreted as 'the change in the odds in favor of the hypothesis when going from the prior to the posterior' (Lavine and Schervish, 1999). The Bayes factor can be used for nested models, but is also easily extended to nonnested models. To date, Bayes factors have found only a limited application in phylogenetics. For example, Suchard *et al.* (1999) examined nested models of DNA substitution using Bayes factors, with results similar to those found using likelihood ratio tests. Also, Huelsenbeck *et al.* (2000a) examined two models of host switching using Bayes factors.

15.10   COMPARATIVE ANALYSIS

Biologists often test evolutionary hypotheses by comparing some feature of an organism across several different species (broadly referred to as the 'comparative method' here, though the term is usually applied to analyses that examine correlation in two or more characters). The comparative method removes the covariation of characters induced by a common phylogenetic history by performing the analysis in a phylogenetic context. To date, most methods that have been proposed to test specific evolutionary hypotheses assume that the phylogenetic history of a species group is known without error; this assumption is implicitly made when the phylogeny is estimated and then used in the comparative analysis without consideration of the possible errors in the phylogeny. Is it possible to avoid the assumption that the phylogeny of a species group is known without error?

Bayesian inference implemented using MCMC suggests one way of performing comparative analyses without assuming that the tree is known without error: The comparative question is addressed on all possible trees, weighted by the probability that each tree is correct. Suppose, for example, that each possible phylogenetic tree is either consistent with or inconsistent with an evolutionary hypothesis. The overall probability that the hypothesis is correct is simply the sum of the posterior probabilities of trees consistent with the hypothesis. Table 15.4 illustrates how such an analysis might be done. The sum of the posterior probabilities of trees consistent with the evolutionary hypothesis is 0.58. This might be taken as only weak evidence that the hypothesis is true.

A real example better illustrates the concept. The example concerns the evolution of the horned soldier aphids (Stern, 1998). The phylogeny of horned soldier aphids was based

Table 15.4   The posterior probabilities of different trees that are either consistent or inconsistent with some evolutionary hypothesis. The probability that the hypothesis is correct is simply the sum of the posterior probabilities of trees consistent with the hypothesis.

τ_i	$f(\tau_i \mid X)$	Tree consistent or inconsistent with hypothesis
τ_1	0.40	Consistent
τ_2	0.33	Inconsistent
τ_3	0.07	Inconsistent
τ_4	0.07	Consistent
τ_5	0.06	Consistent
τ_6	0.05	Consistent
τ_7	0.01	Inconsistent
τ_8	0.01	Inconsistent

Table 15.5   The posterior probabilities for the number of gains and losses of the horned soldier caste character.

		# Losses					
		0	1	2	3	4	5
#	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
G	1	0.0000	0.0076	0.1759	0.5819	0.0004	0.0000
a	2	0.0000	0.0316	0.1769	0.0006	0.0000	0.0000
i	3	0.0002	0.0131	0.0004	0.0000	0.0000	0.0000
n	4	0.0113	0.0000	0.0000	0.0000	0.0000	0.0000
s	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

on mitochondrial DNA sequences sampled from 34 aphid species. One question concerns the number of times that horned soldiers (a specialized caste for defense only present on the secondary hosts) evolved. Previous analysis suggests that the horned soldiers evolved once and were lost twice. We examined this hypothesis by approximating the posterior probabilities of trees using the program BAMBE (Simon and Larget, 1998; this analysis is from Huelsenbeck *et al.*, 2000c). We ran the Markov chain for 1 000 000 generations, sampling the chain every 200 generations. Inferences on the number of gains and losses of the horned soldiers were made using parsimony mapping on each of the 5000 sampled trees. Table 15.5 summarizes the results. The posterior probability of a reconstruction that has 1 gain and 2 losses is 0.1759. However, the posterior probability is spread out over a number of other reconstructions. For example, the posterior probability of reconstructions involving 1 gain and 3 losses and reconstructions involving 2 gains and 2 losses is also substantial. In fact, these three reconstructions (1 gain, 2 losses; 1 gain, 3 losses; and 2 gains, 2 losses) account for 93 % of the posterior probability. This aphid example illustrates how comparative analyses can be performed in such a way that inferences do not depend critically upon any single phylogeny.

15.11   CONCLUSIONS

The 1990s saw several exciting developments in phylogenetics. First, the models of DNA substitution assumed in statistically based methods were greatly improved to reflect basic biological observations, such as rate variation across sites and nonindependent substitutions. Second, maximum likelihood and Bayesian inference of phylogeny became more practical because of faster computers and improved computer programs implementing the methods (such as PAUP*: Swofford, 1998). Third, Bayesian inference was introduced to the phylogeny problem. Bayesian inference represents the last of the widely used methods of statistical inference to be applied to the phylogeny problem. Finally, MCMC was applied to evolutionary problems with some success. For example, MCMC was applied to maximum likelihood inference of coalescence parameters. Similarly, MCMC has been applied to approximate posterior probabilities of phylogenies in a Bayesian framework.

However, the full potential of statistical methods that are based upon the likelihood function is only beginning to be realized in phylogenetics. For example, coupled with MCMC, it should be possible to examine more complex models of DNA substitution, such as the covarion model, in a Bayesian or maximum likelihood framework. Similarly, many questions that depend upon correlation among several different trees, such as inference of horizontal gene transfer and host-parasite cospeciation, can be usefully addressed using MCMC.

Maximum likelihood and Bayesian inference are very similar in that they both use the likelihood function. However, Bayesian inference uses a subjective concept of probability whereas many of the useful concepts in maximum likelihood (such as confidence intervals and p-values) use a frequentist concept of probability. Much debate in the statistics literature has focussed on the merits of these two definitions of probability and the role that prior information should play in statistical inference (especially when it is difficult to specify the prior information as a part of the statistical model). These problems, as applied to the phylogeny problem, we leave as an exercise for the reader.

REFERENCES

Beerli, P. and Felsenstein, J. (1999). Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**, 763–773.

Berger, J. and Wolpert, R. (1984). *The Likelihood Principle: A Review, Generalizations, and Statistical Implications*, Lecture Notes – Monograph Series, Vol. 6. Institute of Mathematical Statistics, Hayward, CA.

Bollback, J.P. and Huelsenbeck, J.P. (2001). Phylogeny, genome evolution, and host specificity of single-stranded RNA bacteriophages (Family Leviviridae). *Journal of Molecular Evolution* **52**, 117–128.

Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967). Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics* **19**, 233–257.

Charleston, M.A. (1994). Factors affecting the performance of phylogenetic methods. Ph.D. thesis, Massey University, Palmerston North, New Zealand.

DeBry, R. and Slade, N. (1985). Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. *Systematic Zoology* **34**, 21–34.

Edwards, A.W.F. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society B* **32**, 155–174.

Edwards, A.W.F. (1972). *Likelihood*. Johns Hopkins University Press, London.

Edwards, A.W.F. (1974). The history of likelihood. *International Statistical Review* **42**, 9–15.

Edwards, A.W.F. (1998). History and philosophy of phylogeny methods. Paper presented to EC Summer School *Methods for Molecular Phylogenies*, Newton Institute, Cambridge, 10 August.

Edwards, A.W.F. and Cavalli-Sforza, L.L. (1964). In *Phenetic and Phylogenetic Classification*, J. McNeill, ed. Systematics Association, London, pp. 67–76.

Efron, B., Halloran, E. and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences (USA)* **93**, 7085–7090.

Farris, J. (1983). In *Advances in Cladistics*, Vol. 2, N. Platnick and V. Funk, eds. Columbia University Press, New York, pp. 7–36.

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* **22**, 240–249.

Felsenstein, J. (1978). Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology* **27**, 401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22**, 521–565.

Fisher, R.A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41**, 155–160.

Fisher, R.A. (1915). Frequency the values of the correlation coefficient in samples from indefinitely large population. *Biometrics* **5**, 507.

Fisher, R.A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron* **1**(4), 3–32.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.

Gaffney, E.S. (1979). In *Phylogenetic Analysis and Paleontology*, J. Cracraft and N. Eldredge, eds. Columbia University Press, New York.

Gaut, B.A. and Lewis, P.O. (1995). Success of maximum likelihood in the four-taxon case. *Molecular Biology and Evolution* **12**, 152–162.

Gelman, A. (1996). In *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson and D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 131–144.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Goldman, N. (1990). Maximum likelihood inference of phylogenetic trees with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Systematic Zoology* **39**, 345–361.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**, 182–198.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein coding DNA sequences. *Molecular Biology and Evolution* **11**, 725–736.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Hasegawa, M., Yano, T. and Kishino, H. (1984). A new molecular clock of mitochondrial DNA and the evolution of hominoids. *Proceedings of the Japan Academy, Series B* **60**, 95–98.

Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating the human-ape split by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160–174.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hillis, D.M. and Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Zoology* **42**, 182–192.

Huelsenbeck, J.P. (1995a). Performance of phylogenetic methods in simulation. *Systematic Zoology* **44**, 17–48.

Huelsenbeck, J.P. (1995b). The robustness of two phylogenetic methods: four taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Molecular Biology and Evolution* **12**, 843–849.

Huelsenbeck, J.P. and Hillis, D.M. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Zoology* **42**, 247–264.

Huelsenbeck, J.P. and Rannala, B. (1997). Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**, 227–232.

Huelsenbeck, J.P., Hillis, D. and Nielsen, R. (1996). A likelihood-ratio test of monophyly. *Systematic Zoology* **45**, 546–548.

Huelsenbeck, J.P., Rannala, B. and Yang, Z. (1997). Statistical tests of host-parasite cospeciation. *Evolution* **51**, 410–419.

Huelsenbeck, J.P., Rannala, B. and Larget, B. (2000a). A Bayesian framework for the analysis of cospeciation. *Evolution* **54**, 353–364.

Huelsenbeck, J.P., Larget, B. and Swofford, D.L. (2000b). A compound Poisson process for relaxing the molecular clock. *Genetics* **154**, 1879–1892.

Huelsenbeck, J.P., Rannala, B. and Masly, J.P. (2000c). Accommodating phylogenetic uncertainty in evolutionary studies. *Science* **288**, 2349–2350.

Jukes, T. and Cantor, C. (1969). In *Mammalian Protein Metabolism*, H. Munro, ed. Academic Press, New York.

Kashyap, R.L. and Subas, S. (1974). Statistical estimation of parameters in a phylogenetic tree using a dynamic model of the substitutional process. *Journal of Theoretical Biology* **47**, 75–101.

Kass, R. and Raftery, A. (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association* **90**, 773–795.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111–120.

Kluge, A.G. (1997). Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics* **13**, 81–96.

Kuhner, M.K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* **11**, 459–468.

Kuhner, M., Yamato, J. and Felsenstein, J. (1994). Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* **149**, 429–434.

Larget, B. and Simon, D. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* **16**, 750–759.

Lavine, M. and Schervish, M.J. (1999). Bayes factors: what they are and what they are not. *American Statistician* **53**, 119–122.

Lewis, P.O. (1998). A genetic algorithm for maximum likelihood phylogeny inference using nucleotide sequence data. *Molecular Biology and Evolution* **15**, 277–283.

Lewis, P.O. (1999). Paper presented to the annual meeting of the SSE, SSB, and ASN societies held in Madison, Wisconsin.

Li, S. (1996). Phylogenetic tree construction using Markov chain Monte carlo. Ph.D. dissertation, Ohio State University, Columbus, OH.

Mau, B. (1996). Bayesian phylogenetic inference via Markov chain Monte carlo methods. Ph.D. dissertation, University of Wisconsin, Madison, WI.

Mau, B. and Newton, M. (1997). Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* **6**, 122–131.

Mau, B., Newton, M. and Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* **55**, 1–12.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1091.

Mitchison, G. (1999). A probabilistic treatment of phylogeny and sequence alignment. *Journal of Molecular Evolution* **49**, 11–22.

Muse, S. and Gaut, B. (1994). A likelihood approach for comparing synonymous and non-synonymous substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**, 715–724.

Nei, M. (1991). In *Phylogenetic Analysis of DNA Sequences*, M.M. Miyamoto and J. Cracraft, eds. Oxford University Press, Oxford, pp. 90–128.

Newton, M., Mau, B. and Larget, B. (1999). In *Statistics in Molecular Biology*, F. Seillier-Moseiwitch, T.P. Speed and M. Waterman, eds. Institute of Mathematical Statistics, Hayward, CA.

Neyman, J. (1971). In *Statistical Decision Theory and Related Topics*, S.S. Gupta and J. Yackel, eds. Academic Press, New York, pp. 1–27.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.

Pagel, M. (1999). The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Zoology* **48**, 612–622.

Penny, D., Hendy, M.D. and Steel, M.A. (1992). Progress with methods for constructing evolutionary trees. *Trends in Ecology and Evolution* **7**, 73–79.

Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution* **43**, 304–311.

Rzhetsky, A. and Nei, M. (1995). Tests of applicability of several substitution models for DNA sequence data. *Molecular Biology and Evolution* **12**, 131–151.

Schöniger, M. and von Haeseler, A. (1994). A stochastic model for the evolution of autocorrelated DNA sequences. *Molecular Phylogenetics and Evolution* **3**, 240–247.

Schultz, T.R. and Churchill, G.A. (1999). The role of subjectivity in reconstructing ancestral character states: a Bayesian approach to unknown rates, states, and transformation asymmetries. *Systematic Zoology* **48**, 651–664.

Siddall, M.E. and Kluge, A.G. (1997). Probabilism and phylogenetic inference. *Cladistics* **13**, 313–336.

Simon, D. and Larget, B. (1998). *Bayesian Analysis in Molecular Biology and Evolution (BAMBE), Version 1.01 beta*. Department of Mathematics and Computer Science, Duquesne University.

Stern, D. (1998). Phylogeny of the tribe Cerataphidini (Homoptera) and the evolution of the horned soldier aphids. *Evolution* **52**, 155–165.

Swofford, D.L. (1998). *PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods*. Sinauer Associates, Sunderland, MA.

Swofford, D.L., Olsen, G., Waddell, P. and Hillis, D.M. (1996). In *Molecular Systematics*, 2nd edition, D.M. Hillis, C. Moritz and B. Mable, eds. Sinauer, Sunderland, MA, pp. 407–511.

Thompson, E. (1975). *Human Evolutionary Trees*. Cambridge University Press, Cambridge.

Thorne, J., Kishino, H. and Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**, 1647–1657.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* **22**, 1701–1762.

Wiley, E.O. (1975). Karl R. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary taxonomists. *Systematic Zoology* **24**, 233–242.

Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* **9**, 554–560.

Yang, Z. (1993). Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10**, 1396–1401.

Yang, Z. (1994a). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39**, 306–314.

Yang, Z. (1994b). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39**, 105–111.

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molecular Biology and Evolution* **14**, 717–724.

Yang, Z., Kumar, S. and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**, 1641–1650.

Phylogenetics: Parsimony, Networks, and Distance Methods

D. Penny, M.D. Hendy and B.R. Holland

Allan Wilson Center for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

The accurate recovery of evolutionary trees is a major problem in mathematical inference. The relevant theory comes from mathematics (graph theory, combinatorics, and Markov chains), statistics (likelihood, Bayesian methods, resampling, and exploratory data analysis), operations research (optimisation, and search heuristics), and computer science (complexity theory). Different ways of presenting molecular data are described and a consistent terminology is presented. The problem of the combinatorial explosion in the numbers of potential trees leads to a description of the complexity of algorithms. The use of simple Markov models for studying the evolution of DNA sequences is described. Methods for inferring trees are analysed under the three components of, an optimality criterion, searching tree space, and assumptions or knowledge about the mechanism of evolution (which can allow compensation for multiple (unobserved) changes at a site). The analysis of methods in this chapter concentrates on parsimony and distance methods for inferring trees, but also discusses the limited range of cases when maximum parsimony and maximum likelihood are equivalent. A section describes the use of networks to represent the complexities in the data or to summarise the variability of the output trees. An overview of tree-building methods includes several strategies for searching tree space, from complete searches, branch-and-bound searches, to different forms of heuristic approaches.

16.1   INTRODUCTION

The mathematics behind phylogenetic programs is interesting in that it comes from several areas of pure and applied mathematics, both discrete and continuous. These include graph theory, combinatorics, and Markov chains from mathematics, together with statistics (likelihood, Bayesian analysis, resampling), operations research (optimisation, search heuristics), and computer science (complexity theory). Fortunately, the underlying concepts are relatively simple (even if the mathematics is complex) and in this chapter we only consider the concepts. Our description of the mathematics is informal, more formal

discussions are found in the primary literature. Of course, underlying all the algorithms are (or should be) the biological models of the genetic information and biochemical processes. More detailed treatments on tree building are available in Swofford *et al.* (1996), Page and Holmes (1998), Nei and Kumar (2000), and Felsenstein (2003); and the mathematical basis is in Semple and Steel (2003).

We focus on input data that is either aligned sequences representing a set of taxa, or a matrix of differences (genetic distances) between each pair of taxa. The sequence information is usually DNA, RNA, or protein sequences, though increasingly it is information such as gene order or the position of sequence elements (short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs), respectively, Kriegs *et al.*, 2006), insertions/deletions, presence of exon boundaries, etc). The output is in a graphical form (in the sense of graph theory) and can be either a single phylogenetic tree, a set of trees, or a network (a connected graph with circuits, see later). A consequence of the use of many areas of mathematics is that non-standard terminology has been introduced into phylogeny in an *ad hoc* manner. An unfortunate consequence is that some appropriate areas of mathematics have been overlooked, important problems neglected, or erroneous assumptions made. We will keep our analysis of the problems as broad as possible while keeping to well-defined mathematical terminology that is introduced later in Figure 16.5. Basically, the terminology is both simple and powerful, and it is unclear why phylogenetics does not use a conventional terminology from mathematics.

Another theme of this chapter is the need for careful analysis of concepts in regular use. It is necessary to identify that any 'method' for inferring trees is a combination of at least three components. These are (1) an optimality criterion, (2) a search strategy, and (3) assumptions about the mechanism of evolution (including whether there is sufficient knowledge to correct for multiple changes). Similarly, a concept such as parsimony has two distinct uses: the principle of parsimony (Ockham's razor) is the main use in science generally, and an optimality criterion for inferring an evolutionary tree by minimising the number of mutations on a tree. This second meaning usually uses data without correction for multiple changes at a site, but the same algorithm works (often better) if corrections are made for multiple changes. The two usages (the principle and the optimality criterion) are quite different, and need not agree. For example, the Principle of Parsimony lead to the conclusion that in some cases maximum (average) likelihood is to be preferred over minimising the number of mutations (Steel and Penny, 2000).

It is worth reflecting on the general difficulty of inferring evolutionary trees accurately. Because the events we wish to uncover happened over time periods ranging from hundreds of years to 3000 million years ago, we usually cannot observe them directly (evolution of RNA viruses is a good exception). The subject is therefore both difficult and exciting. Exciting because it is necessary to push the current limits of quantitative science; difficult because it is not possible to check answers directly. This makes the subject ideal for mathematical and statistical reasoning. It also means that controversies will abound, and that is expected when science is stretched to its limits. However, we must never get lost in the mathematical reasoning to the exclusion of the biology.

16.2   DATA

Several ways of presenting data are illustrated in Figure 16.1: character state matrices (such as aligned sequences); distance matrices; splits (partitions, patterns, subsets of taxa);

tensors (3D matrices); gene order; and presence/absence of inserted elements and/or gaps. As used here, tensors are based on three taxa, giving a three-dimensional 'matrix'.

16.2.1 Character State Matrix

Figure 16.1(a) is a multiple alignment for short sequences from four mammals; it is a *character state matrix* or simply the *data matrix, data set*, or *data*. Mathematically a matrix or set is a singular object containing many entries – thus 'data' is singular. Each row in the matrix is a sequence and represents the information about a *taxon* (plural *taxa*). A taxon can be an individual, a group, a population, a species, or some higher taxonomic group. It is a useful general term, and the number of taxa is denoted by t. Each column of the matrix is a *site*, a *column*, a *position* (nucleotide or amino acid), or (as in the case of morphological or behavioral information) a *character*, and their number is denoted as c. An individual entry in a column may be a nucleotide, an amino acid, the presence or absence of a SINE, a code, a *colour*, or a *character state* (hence, *character state matrix*). 'Colour' is a mathematical term for discrete character states and the number of colours is usually given the symbol r. On this terminology DNA has four colours, A, G, C, and T; and proteins have 20. The number of character states is also called the *size* of the alphabet, thus forming a useful analogy with letters in an alphabet. The patterns in the full character state matrix are easily enhanced by a consensus format (see the lower part of Figure 16.1(a)). The preferred method is to make a *consensus* for each site (breaking ties arbitrarily), place it at the top, and show only deviations from the consensus. This is a natural form for introducing another way of describing the data, partitions or splits, but distances will be discussed first.

16.2.2 Genetic Distances (Including Generalised Distances)

Figure 16.1(b) is a distance matrix derived from Figure 16.1(a). In general, values can be derived from any character state matrix, or directly from a range of measurements such as DNA/DNA hybridisation, comparison of shapes of RNA or protein secondary and tertiary structures (Collins *et al.*, 2000), to the number of 'breakpoints' required to equalise gene order in two genomes (Figure 16.1e, Henz *et al.*, 2005). Distances are a subset of difference values that form a 'metric'; to be a metric, the values for any subset of three taxa i, j, and k must have the following properties:

$$d_{i,i} = 0;$$
$$d_{i,j} = d_{j,i} > 0 \text{ for } i \neq j; \text{ and}$$
$$d_{i,j} \leq d_{i,k} + d_{j,k} \text{(this being the triangle inequality).}$$

The first two are straightforward, each thing is identical to itself ($d_{i,i} = 0$); and the difference between A and B must be the same as between B and A (thus $d_{i,j} = d_{j,i}$); and the value must be positive when they differ. The third part of the definition, the triangle inequality is also a simple concept; the direct distance between London and Sydney cannot be greater than the sum of the distances from London to New York plus New York to Sydney. This simple intuitive concept of distances from everyday life is thus extended to a wide range of scientific activities, in this case, the genetic distance between organisms.

When measuring parameters such as DNA/DNA hybridisation, or distances between RNA or protein shapes, it is possible that any the three properties will be violated. There

(a) Character state matrix

Taxa names or labels	Taxon index		
		1	**2**
		| **0** | **0** |	
t_1 Monkey	1	TGAACTCAAGCACCAAAAAGGAAGACTAC	
t_2 Horse	2	TAGGCTCTAGCACCAACATGGCATACTAC	
t_3 Kangaroo	4	TAAGCCAAAGCGACAAACTAGCCGTCTAC	
t_4 Opossum	8	TAAGCCATAGCGACAAACAAGCCTATTAC	

		1	**2**
		| **0** | **0** |	
Consensus	→	**TAAGCTCAAGCACCAAAAAGGCAGACTAC**	
t_1 Monkey	1	.G.A.................A.......	
t_2 Horse	2	..G....T........C.T....T.....	
t_3 Kangaroo	4	.....CA....GA....CTA..C.T....	
t_4 Opossum	8	.....CAT...GA....C.A..CT.T...	

(b) Distances

					Divergence matrix, $\mathbf{F}_{ij}$, for monkey and horse
t_1 Monkey	0	0.28	0.41	0.45	A [9 1 1 2]
t_2 Horse	8	0	0.41	0.41	G [1 7 0 0] Monkey
t_3 Kangaroo	12	12	0	0.17	C [1 3 3 1]
t_4 Opossum	13	12	5	0	T [0 0 0 3]
	Monkey	Horse	Kanga	Oposs	A G C T
					Horse

(c) Splits   Splits (from data a)

Index	Taxa in splits	No.	Frequency	σ^2	Columns in each split
0	$\varnothing$, $\{t_1,t_2,t_3,t_4\}$	12	0.41	0.242	1, 5, 9, 10, 11, 14, 15, 16, 21, 27, 28, 29
1	$\{t_1\}$ $\{t_2,t_3,t_4\}$	3	0.10	0.090	2, 4, 22
2	$\{t_2\}$ $\{t_1,t_3,t_4\}$	2	0.07	0.065	3, 17
3	$\{t_1,t_2\}$ $\{t_3,t_4\}$	7	0.24	0.182	6, 7, 12, 13, 18, 20, 23
4	$\{t_3\}$ $\{t_1,t_2,t_4\}$	1	0.03	0.029	25
5	$\{t_1,t_3\}$ $\{t_2,t_4\}$	2	0.07	0.065	8, 24
6	$\{t_2,t_3\}$ $\{t_1,t_4\}$	1	0.03	0.029	19
7	$\{t_1,t_2,t_3\}$ $\{t_4\}$	<u>1</u>	<u>0.03</u>	<u>0.029</u>	26
		29	1	0.732	–
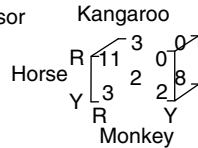
(d) Tensor   Kangaroo   (e) Gene order
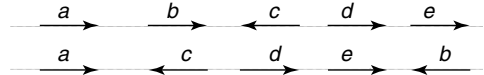
(f) Presence/absence

```
TGAACTCAAGCACCGCATTACGAAGTCCTGCCAAAAAAGGAAGACTAC
TAGGCTCTAGCACC                ...        AACATGGCATACTAC
```

may be a detectable interaction between, say, the hybridisation of DNA from the same individual ($d_{i,i} \neq 0$), the reciprocal values between two taxa may not be equal ($d_{i,j} \neq d_{j,i}$), and/or the triangle inequality may not hold. Measures based on these properties are *differences* or *dissimilarities* rather than distances; they are still useful though many tree-building algorithms are expected to work better if the data are metric distances.

Hamming distances simply count the relative number of differences between two aligned strings. The strings may be sequences, morphological characters, or even word in copies of an ancient manuscript. Many weightings are possible, e.g. transversions more than transitions. It is important to then divide by c, the length of the sequence, because corrections for multiple changes depend on proportion of sites that vary, not just on the total number of sites that differ. This gives the distance ($d_{i,j}$) as a proportion $0 \leq d_{i,j} \leq 1$. Thus dividing by c leads to more information being included in the distances.

As described, these are *observed distances* and generally, because they do not include multiple changes at a site, underestimate the total number of changes between a pair of taxa. For example the changes A $\rightarrow$ C $\rightarrow$ A between a pair of taxa will not be detected, both sequences have an 'A'. Similarly the series of changes A $\rightarrow$ C $\rightarrow$ G will appear as a single A $\rightarrow$ G change. Later we give the formulae (see Swofford *et al.*, 1996), which estimates the expected number of multiple changes and these *may* allow more accurate estimates of the real distance. Corrections are important for sequences that have diverged for longer times, but they introduce other sources of error. Small differences introduced by sampling can be exaggerated, variances increased, and a bias toward higher values introduced. Corrections become erratic when divergences are large, especially when the underlying processes vary between parts of the evolutionary tree.

These Hamming distances are not the only distances that can be obtained from sequences; the two additional distance matrices in Table 16.1 are also derived from the four sequences in Figure 16.1(a). The entries of the 4×4 matrix on the left is based solely on the nucleotide frequencies of the sequences, in this case the distance between species i and j is $d_{ij} = \sqrt{\Sigma (f_i - f_j)^2}$, where f_i and f_j are the frequencies of the nucleotides. Alternatively, entries in the matrix on the right are derived from divergence matrices, $\mathbf{F}_{ij}$, for each pair of taxa (see the right-hand side of Figure 16.1(b)) with the distance

Figure 16.1 Data for tree building; the same data is used for (a) to (d). (a) Sequence data for 29 nucleotide positions (originally derived from α-crystallin protein sequences). A consensus format is given in the lower part of (a). (b) Distances derived from (a) or (c) by: lower left, counting the number of differences between pairs of taxa; and, upper right, dividing these by the length of the sequences. A divergence matrix, $\mathbf{F}_{ij}$, for monkey and horse is shown on the right. This is used to estimate a genetic distance even when two sequences differ in nucleotide composition. (c) Splits derived from (a) by using the number of times subsets of taxa have a different code to the last taxon (t_4). This form of the data is equivalent to (a) for two-state characters, but omits the order of columns in the sequence, and which of the two codes occurs in (t_1). (d) Tensor (3-D matrix) for (a), recoded as R (purines) and Y (pyrimidines). In this example the entry 3 in front lower left counts the number of sites where monkey and kangaroo have R, while horse is Y – the RYR pattern (sites 8, 22, 24). Similarly, there are two sites (7 and 13) where monkey and horse have Y, and kangaroo R – the YYR pattern. (e) Gene order for five genes, indicating by the arrow the 5′ to 3′ direction of the gene. (f) This is an example of an insertion in one sequence, it is present in one, absent in the other.

Table 16.1 Two distance matrices derived from the data in Figure 16.1(a). The values on the left hand are derived from nucleotide frequencies; those on the right are LogDet distances that compensate for differences in nucleotide composition.

Monkey	$\begin{bmatrix} 0.000 \end{bmatrix}$	0.181	0.181	0.117	Monkey	$\begin{bmatrix} 0.000 \end{bmatrix}$	0.414	0.744	0.834
Horse	0.181	0.000	0.000	0.091	Horse	0.414	0.000	0.578	0.593
Kangaroo	0.181	0.000	0.000	0.091	Kangaroo	0.744	0.578	0.000	0.253
Opossum	0.117	0.091	0.091	0.000	Opossum	0.834	0.593	0.253	0.000
	Monkey	Horse	Kanga	Oposs		Monkey	Horse	Kanga	Oposs

$d_{ij} = \ln(\det \mathbf{F}_{ij})$, the natural logarithm of the determinant of the divergence matrix $\mathbf{F}_{ij}$ for each pair i, j of species. These LogDet (or paralinear) distances compensate partially when the sequences have different nucleotide compositions.

The important point here is that there are many ways of forming pairwise distances–distances do not just appear, they are derived from the data by a calculation that depends on a reasonable knowledge of the organisms, the data, and the mechanisms of evolution.

Three important terms are *additive distances, ultrametric distances*, and the *four-point condition*. Additive distances are values in a distance matrix **D** that can be fitted exactly on a tree. An example is shown in Figure 16.2. Check that the path lengths on the tree add up to the values in the distance matrix. Note also that in this case if the tree were built by selecting the smallest entries in **D**, then taxa 2 and 3 would be selected first – but on this tree these taxa are not joined directly. Additive distances have two interesting features; they describe a unique tree and in addition, the lengths (weights) of the edges of the tree are unique, though it is still be hard to interpret the edge lengths in any particular case. Do not get too excited about additive distances; real data is never additive, even after correcting for multiple changes. The underlying problem is the number of equations and the number of unknowns. In general, there are $(t^2 - t)/2$ distances; 10 for five taxa. Then there are $2t - 3$ edge weights (which are our unknowns, seven for five taxa). It is therefore seldom that distances are additive, although the concept is important for understanding the accuracy and reliability of tree reconstruction.

Ultrametric distances are additive distances with an additional property that can be described both biologically and mathematically. Biologically, the distances fit a rooted tree under a molecular clock so that there are (statistically) equal rates of evolution from the root to all the tips. This is equivalent to the mathematical property that for all triples

Taxon 1	0				
Taxon 2	7	0			
Taxon 3	8	5	0		
Taxon 4	13	10	9	0	
Taxon 5	11	8	7	6	0
	t_1	t_2	t_3	t_4	t_5
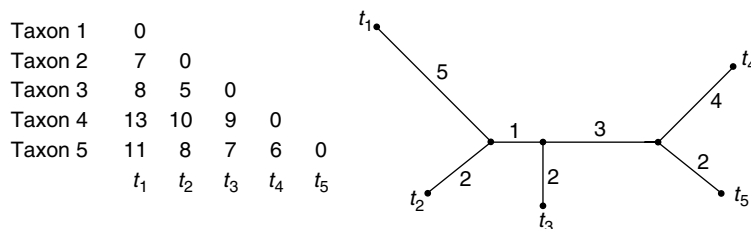
Figure 16.2 Additive distances and the corresponding tree.

i, j, and k there is some ordering so that the two largest distances are equal,

$$d_{ik} = d_{jk} \geq d_{ij},$$

('some ordering of i, j, and k' means that the indices of the three taxa can be interchanged to find an order for which this relationship holds). For example, d_{ik} may be 29 nucleotide differences, d_{jk} 31, and d_{ij} 12. In this case you might decide that, statistically speaking, you could not reject the ultrametric relationship. The ultrametric relationship is used in the relative rates test that estimates whether the same rate of change occurs on the two closest taxa. It is necessary to check that taxon k is genuinely more distantly related than the other two taxa, that it is a genuine *outgroup*. However, this test is relatively weak (Bromham *et al.*, 2000).

The third term, the 'four-point condition' is that for any taxa i, j, k, and l, the larger two of the sums

$$d_{ij} + d_{kl}, d_{ik} + d_{jl}, d_{il} + d_{jk},$$

are of equal value i.e. $(d_{ij} + d_{kl}) = (d_{ik} + d_{jl}) \geq (d_{il} + d_{jk})$ for some ordering of i, j, k, and l.

This condition is mathematically equivalent to the distances being additive. As an exercise, test any three distances from the right-hand side of Table 16.1 to see if they are ultrametric. Test also with all four whether the distances satisfy the four-point condition. You will find it helpful to look at the three unrooted trees for four taxa, as follows.
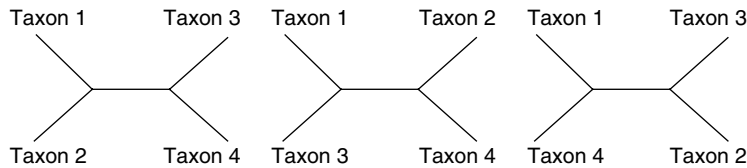
Take a set of distance values, and see how well they fit onto each of the three trees. Most data sets will not fit exactly. Programs, such as SplitsTree (Huson and Bryant, 2006), adds cycles into the graph if the distances carry a signal additional to the tree; an example is shown below on the left. This introduces an additional weight in the graph, shown by the weights, w_{12} and w_{13}, on two internal edges (the parallel edges have the same weights in this graph). This graph will more closely represent the distances in the data by summing the weights on the most direct path between any pair of taxa. This can be extended further to a *Buneman graph*, shown below on the right, where a further weight is included. Again, parallel edges are assigned the same weight.
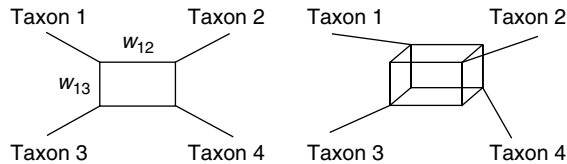
One additional concept, *generalised distances*, is described now and used later for spectral analysis. Note that the distance values above are for *pairs* of sequences and are therefore 'pairwise distances'. However, for any subset with an even number of taxa

(4, 6, 8 . . .), we can sum the minimum combination of pairwise distances to get a distance value for the subset. For example, in Figure 16.2 the generalised distance value for the subset {1,2,4,5} would be $d_{12} + d_{45} = 7 + 6 = 13$. This broader concept is useful because the total number of subsets with an even number of taxa is the same as the number of splits of the taxa (see next section). In each case there are 2^{t-1} even-order subsets (including zero taxa, which has a distance 0), and 2^{t-1} splits. This equality between the numbers of subsets and of splits is very important because it allows the Hadamard conjugation that is described later. Basically, it is possible to move between splits of the data and generalised distances without any loss of information. However, distance methods traditionally use just the pairwise distances. When distances are derived from sequence data there is loss of information because it is not possible in general to recover the frequency of splits just from the pairwise-distance values (Penny, 1982).

In this section we have shown several ways of forming distance matrices to emphasise the richness and diversity of the concept. Our purpose is to emphasise that the investigator needs to understand the alternatives and choose variants that are appropriate. Science is not a 'passive' process where you put data into a computer program and accept the answer that pops out. Rather it is an active process where the investigator is making hypotheses, testing them, looking for alternatives, challenging dogma, discussing questions with peers – altogether a much more exciting adventure. So review this section quickly with this in mind.

16.2.3   Splits (Bipartitions)

Sequence data can be represented as *splits* (or *bipartitions*), a natural extension of the character state matrix (Figure 16.1). The distribution of characters at a site is a *pattern* and the number of columns in which each pattern occurs is recorded. When only two states are present at a site the pattern can be identified by the subsets of taxa that share a common character. This pair of subsets is called a *split* or *bipartition*.

Examine the lower part of Figure 16.1(a) where the sequences are shown in consensus form. The first and the fifth columns have equivalent character states, the same pattern of taxa sharing the same character. Similarly, in the second and fourth columns the taxa have the same pattern–taxon 1 (monkey) is different from the other three, so the split is {monkey}, {horse, kangaroo, opossum}. Look at columns 2 and 4 again, the patterns are the same but one column is $[GAAA]^t$ and the other is $[AGGG]^t$ (here we *transpose* a column to show it as a row). The pattern of taxa sharing the same character state (code) is also the same as column 22 $[ACCC]^t$–thus the patterns are the same in columns 2, 4, and 22, even though the nucleotides differ. Clearly we will need a simple way of numbering these patterns so that they can be counted easily by computer.

At this point we only describe partitions for two-state characters (two colours) such as for purines and pyrimidines, or the presence or absence of a SINE at a specific site in a sequence. The concept of a *split* (or *bipartition*) comes from the observation that each column in the data set (Figure 16.1a) splits (or partitions) the taxa into two *disjoint* subsets. (Disjoint means that each taxon occurs in just one of the two subsets – no taxon is omitted and no taxon occurs in both.) With four taxa $(t_1 - t_4)$ there are eight possible splits, which we number $0 - 7$ to make it easier to automate calculations. You will find this way of recoding the data very instructive because there is both a direct relationship between splits in the data, and edges in evolutionary trees. It also allows a natural way of introducing estimates of sampling errors.
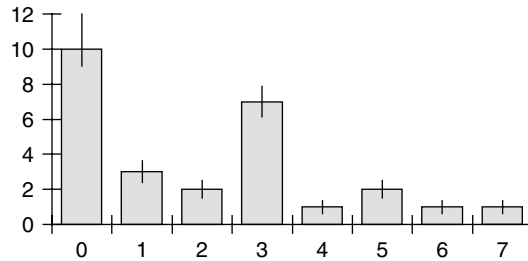
Figure 16.3   Spectrum for the splits in Figure 16.1(a).

Figure 16.3 gives the frequencies of the splits as a histogram, which is a *spectrum.* When we come to Hadamard transforms we will find that the transforms are examples of Fourier analyses and so they fit into the general framework of spectral analysis (see later).

The indexing system to identify taxa, splits (the information in columns of data), and edges in a tree is simple and beautiful. It is worth mastering because it provides a simple and unambiguous method for describing and relating edges in the tree, the taxa, and patterns in the data. Each taxon has its own *name* or *label* that is fixed for the study, but in addition we give each taxon an *index* that depends on its position in the data set. The taxon *name* (or *label*) may be *Equus caballus*, horse, *Dendroligotrichum dendroides, Drosophila*, kangaroo, Sam, t_5, ..., etc. Its *index* will be selected from successive powers of 2; i.e. 1, 2, 4, 8, 16, ..., etc. depending on whether it is the first, second, third, etc. sequence in the data set (see Figure 16.1). Thus the indices are,

$$2^0, 2^1, 2^2, 2^3, 2^4, 2^5 \cdots$$

$$1, 2, 4, 8, 16, 32 \cdots$$

This is a *binary* indexing system because it uses powers of 2. Indices for taxa are added together to index either splits in the data, or edges of a tree. The indices are shown for the taxa in Table 16.2. Using this indexing system you will find that the sum of the taxon labels of the first t taxa is one less than the label for taxon $t + 1$. For example, with four taxa the sum of the labels is $1 + 2 + 4 + 8 = 15$ and the next label added to make a five-taxon tree is $2^{5-1} = 16$.

Next we need to count all the ways in which t taxa can be *split* into two subsets–these are shown in a natural ordering in Table 16.2. When listing the two subsets, the taxon with the largest index (in this case taxon 4, t_4 with label 8) is always put in subset 2. Consequently subset 1 has all the taxa whose character state differs from the last taxon (t_4 in this case). This indexing is repeated for every site (character).

In the first split in Table 16.2 all taxa have the same code; none differ from t_4. Hence, this split has index 0 (s_0). In the second split, index 1 (s_1), t_1 is the only taxon where the character state differs from t_4, so the split has index 1. With another example, the split in column 19 of Figure 16.1(a) has index $6 = 2 + 4$ because the taxa differing from t_4 are t_2 and t_3, with indices 2 and 4. It is easy to verify that there are 2^{t-1} splits. For two-state characters, each column represents a split and summing over all columns gives the frequencies of each split. The frequencies of the splits in Figure 16.1(a) are 12, 3,

Table 16.2 Splits. A split partitions the taxa into two disjoint subsets, shown on the left as subset 1 and subset 2. All splits of four taxa, $t_1 - t_4$, are shown, together with their indices. (For completeness we also identify the *null* split, where all taxa have the same state, so subset 1 is the empty set ∅.) Check that in every case the sum of the indices is 15 $(1 + 2 + 4 + 8)$; consequently only the index of first subset need be used. The right-hand side of the table has columns representing all eight splits for two-state characters. Under a symmetric model, e.g. $[abbb]^t$ as equivalent to $[baaa]^t$. These pairs of columns are assigned the same indices, 0–7, and are shown above the columns.

					All possible patterns in columns (splits) and their indices. Both symmetric forms are given, e.g. pattern 7 = $[aaab]^t$ and $[bbba]^t$								
Subset 1	Index 1	Subset 2	Index 2	Index 1+ Index 2	Name	0	1	2	3	4	5	6	7
∅	0	$\{t_1, t_2, t_3, t_4\}$	15	15	t_1	$a\|b$	$b\|a$	$a\|b$	$a\|b$	$a\|b$	$a\|b$	$a\|b$	$a\|b$
$\{t_1\}$	1	$\{t_2, t_3, t_4\}$	14	15	t_2	$a\|b$	$a\|b$	$b\|a$	$a\|b$	$a\|b$	$b\|a$	$b\|a$	$a\|b$
$\{t_2\}$	2	$\{t_1, t_3, t_4\}$	13	15	t_3	$a\|b$	$a\|b$	$a\|b$	$b\|a$	$b\|a$	$a\|b$	$b\|a$	$a\|b$
$\{t_1, t_2\}$	3	$\{t_3, t_4\}$	12	15	t_4	$a\|b$	$a\|b$	$a\|b$	$b\|a$	$a\|b$	$b\|a$	$a\|b$	$b\|a$
$\{t_3\}$	4	$\{t_1, t_2, t_4\}$	11	15									
$\{t_1, t_3\}$	5	$\{t_2, t_4\}$	10	15									
$\{t_2, t_3\}$	6	$\{t_1, t_4\}$	9	15									
$\{t_1, t_2, t_3\}$	7	$\{t_4\}$	8	15									

2, 7, 1, 2, 1, and 1, for splits 0–7 respectively. Dividing these values by the number of columns estimates the probability of observing a particular subset in the sequence.

Extending to additional taxa is straightforward and is shown for five taxa. The eight splits in subset 1 are the same as for four taxa – then subset 2 (for four taxa, Table 16.2) is written below subset 1, in reverse order. We now have 16 indices, numbered 0–15. This completes subset 1 for five taxa; subset 2 for these is the complement of subset 1. Calculating their indices, and verifying that for each pair they sum to 31 is a check. This procedure gives a simple recursive method that can be used as t increases. There are other ways for ordering the splits, but it is *essential* to be consistent and this way of indexing is simple for both man and beast (computer). Make sure that you could determine the 32 splits for $t = 6$. There are several ways of indexing four-state characters, but this is outside the scope of this chapter (see Hendy *et al.*, 1994).

16.2.4 Sampling Error

How accurate are our data samples? Biologists are used to the concept of sampling error and any estimate for a distance value or frequency of a split should have an indication of this. For distances, the variance is $d_{i,j} \times (1 - d_{i,j})/c$ where the distances are expressed as differences per site. There is a similar formula for split i (Waddell *et al.*, 1994); its variance is $V_i = s_i(1 - s_i)/c$, where s_i is the frequency of columns with that split. The covariance for the splits i and j is $V_{ij} = s_i(1 - s_j)/c$. The variance on each distance value is expected to be relatively less than that for splits, there are $t(t-1)/2$ distance values compared with 2^{t-1} splits.

There is an additional factor for covariances with distances when a tree-like process generated the data; there is a high correlation between many pairs of distances. Consider the rooted tree in Figure 16.4. The number of differences between any pair of taxa depends on the path between them; most values will be correlated – the extent depends on the proportion of paths they share. In this example chicken–turkey and moa–kiwi distances
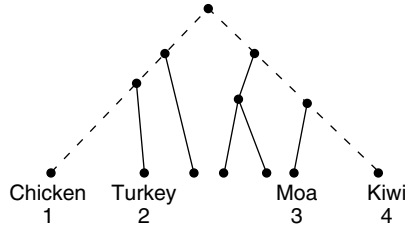
Figure 16.4 Correlation of distances. The distance between chicken and kiwi is the path length (dashed line) between them. However the distance from chicken to moa shares much of the same path and consequently its value is highly correlated. The turkey–moa distance is also highly correlated with chicken–kiwi.

Table 16.3 Covariance matrix for the splits from Figure 16.1. Variances are shown on the diagonal, and covariances off the diagonal. Check that, e.g. $V_{01} = V_{10} = -s_0 \times s_1 (-0.4138 \times 0.1034 = -0.0428)$, etc. Fill in the remaining entries.

Index	0	1	2	3	4	5	6	7
s	[0.4138	0.1034	0.0690	0.2413	0.0345	0.0690	0.0345	0.0345]
	0.2426	−0.0428	−0.0286	–	–	–	–	–
	−0.0428	0.0927	−0.0071	–	–	–	–	–
	−0.0286	−0.0071	0.0642	–	–	–	–	–
V[*s*]	−0.0998	−0.0250	−0.0166	0.1831	–	–	–	–
	−0.0143	−0.0036	–	–	0.0333	–	–	–
	−0.0286	−0.0071	–	–	–	0.0642	–	–
	−0.0143	−0.0036	–	–	–	–	0.0333	–
	−0.0143	−0.0036	–	–	–	–	–	0.0333

are not correlated, but this requires knowing the tree. Splits come from a multinomial sampling procedure. For Figure 16.1(a) the covariance matrices are given for both splits (Table 16.3) and distances (Table 16.4). These are calculated using Hadamard transforms and do not require knowledge of the tree (see later).

16.3   THEORETICAL BACKGROUND

This section gives some theoretical background to inferring evolutionary trees and networks. It covers the terminology of trees and graphs, the numbers of trees and its relation to computational complexity and to evolutionary models.

16.3.1   Terminology for Graphs and Trees

The terms we use are given in italics with alternatives given in parentheses. A graph $G = (V, E)$ is a set V of *nodes* (vertices, points) and a set E of *edges* (internodes, links, 'branches') where each edge is represented by a pair of nodes. If edge e is represented by the pair of nodes $\{v_1, v_2\}$, then we can write $e = (v_1, v_2)$, and we say the edge e is *incident*

Table 16.4 Covariance matrices for generalised distances **V[d]**. The vector of distances **d** is shown first. The largest and smallest covariances in **V[d]** are underlined. Entries 1 and 5 in **d** (the distances between taxa 1 and 4, and taxa 1 and 3, respectively) covary the most, while entries 4 and 5 are slightly negatively correlated.

	0	d_{14}	d_{24}	d_{12}	d_{34}	d_{13}	d_{23}	All 4
d	[0.0000	0.4482	0.3793	0.2759	0.1725	0.4137	0.4138	0.4584]
V[d]	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.0000	0.4947	0.2116	0.0975	0.0524	<u>0.3186</u>	0.2497	0.0594
	0.0000	0.2116	0.4709	−0.0023	0.0072	0.2378	0.3067	0.0239
	0.0000	0.0975	−0.0023	0.3996	0.1118	0.0475	0.0477	0.2116
	0.0000	0.0524	0.0072	0.1118	0.2855	<u>−0.0048</u>	0.0643	0.0547
	0.0000	<u>0.3186</u>	0.2378	0.0475	<u>−0.0048</u>	0.4851	0.2092	0.0761
	0.0000	0.2497	0.3067	0.0477	0.0643	0.2092	0.4852	0.0072
	0.0000	0.0594	0.0239	0.2116	0.0547	0.0761	0.0072	0.3663

Source: Data from Figure 16.1.

on v_1 and v_2. The graph is often drawn with a dot for each node, and a line connecting two adjacent dots for each edge. Two edges are *adjacent* if they share a common node. A *path* is a sequence of edges $e_1, e_2, \ldots, e_n$, where e_1 is adjacent to e_2, e_2 is adjacent to $e_3, \ldots, e_{n-1}$ is adjacent to e_n. If $e_1 = (v_0, v_1)$ and $e_n = (v_{n-1}, v_n)$, then we say the path connects v_0 to v_n, with v_0 and v_n being the *endpoints* of the path.

A graph is *connected* if for each pair of nodes of G there is a path connecting them. A *circuit* (cycle) is a path with two ends being the same node. Figure 16.5(a) is a graph that is not connected and contains several circuits. The *degree* (valence) of a node v is the number of edges incident on it. A node of degree 1 is a *leaf* (tip, external node, pendant vertex). Figure 16.5(a) contains vertices of degrees 0, 1, 2, 3, and 4. Sometimes we may draw two edges crossing, although the point of crossing does *not* represent a node, as in the centre of Figure 16.5(a).

A *tree* is a graph $T = (V, E)$ which is connected and has no circuits, a '*connected acyclic graph*'; Figure 16.5(b) is a tree. For phylogenetic trees a node may represent a taxon by attaching a *label* (name) to the node; Figure 16.5(c) is a tree where all nodes are labelled. Often the edges have an associated real number, the *weight* (length) of the edge. In phylogenetics this can represent the number of substitutions, time, or a probability of substitution; hence edge weights are non-negative. When the edges are weighted, T is a *weighted* tree (see Figure 16.5d). A graph or tree, or parts of the tree, can be rotated on the page and it is still the same graph or tree. Examples include (t_1, t_2) in Figure 16.5(f) and similarly (t_5, t_6) with t_4. What is essential is the relationship between the edges and vertices, not the way they are represented on the computer screen or on the page.

We often wish to identify the *root* of the tree (the most recent common ancestor). Such a tree is a *rooted* (directed) tree; the root is unique and the tree can be drawn at the top, bottom, or to the left, but with all edges directed away from the root. A rooted tree can be derived from an unrooted tree by identifying a root, which may be an existing node, or a new node (of degree 2) can be inserted into an edge. Figure 16.5(e) is a rooted tree obtained from Figure 16.5(b) where the root has been added to the central edge in Figure 16.5(b). Removing the root from Figure 16.5(e) returns to the unrooted tree Figure 16.5(b).
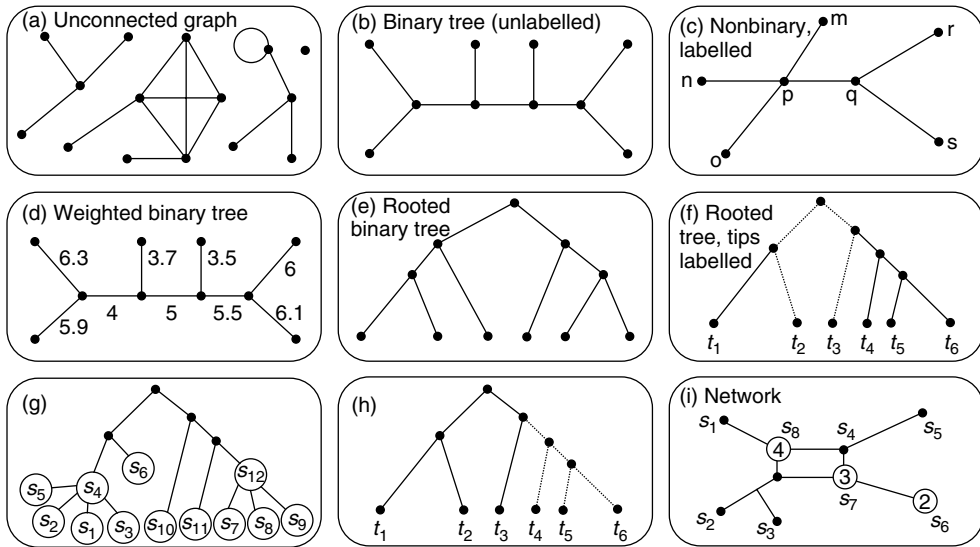
Figure 16.5 Terminology for trees. Examples of graphs (trees and networks), see text for details.

A *phylogenetic tree* (phylogeny, X-tree, cladogram) for a set of taxa S is a tree with labels on all leaves, and possibly on some internal nodes. More than one taxon can label a node. Figure 16.5(f) is a phylogenetic tree on the taxon set $S = \{t_1, t_2, \ldots, t_6\}$. A phylogenetic tree may be rooted or unrooted, weighted or unweighted, binary or nonbinary. A path connecting two labelled vertices (t_2 and t_3), is shown in Figure 16.5(f) as a dotted line. In a rooted tree we can identify a *lineage* as a path from an internal node to a labelled node; the path is directed away from the root. In addition, a *subtree* (branch, in mathematical and botanical terminology) is the set of lineages from a node v with a common first edge e. Figure 16.5(h) shows both a lineage and a subtree. In a *binary* rooted tree, any path from the root has two choices (a bifurcation) at every non-leaf node.

A *phylogenetic network* is a connected graph, again with some of the nodes labelled (Figure 16.5i). In this example the number of times a sequence has been found in a population is indicated by the size of the circle. In a tree the removal of any edge partitions the taxa into disjoint subsets, in a network a set of (parallel) edges may be required to partition the graph into two connected subgraphs (so the graph appears 'box-like' as in Figure 16.5(i) which has two sets of parallel edges).

There are two non-standard usages found in phylogenetics. The first is that a 'branch' refers to an edge (not a subtree; in both mathematical and botanical terminology a branch is a subtree). The second is that 'topology' often refers to an unweighted tree (rather than to an unlabelled tree). It is important to be aware of both alternative usages, only occasionally are they ambiguous. However, we will avoid both non-standard usages.

16.3.2 Computational Complexity, Numbers of Trees

16.3.2.1 Numbers of Trees

A fundamental problem with inferring evolutionary trees is the very large numbers of trees, even in searches limited to unrooted binary trees. The numbers of trees (and either
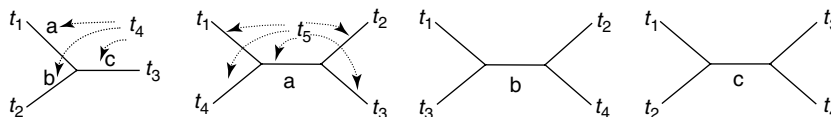
Figure 16.6 Counting unrooted binary trees. There is a single unrooted binary tree for three taxa. The fourth taxon (t_4) can be connected to any of the three edges a, b, or c. This leads to the three distinct unrooted trees for four taxa (shown as a, b, and c), each with five edges. A fifth taxon can be added to any of these five edges, thus leading to 15 trees (3 trees, times five edges), etc.

the time to calculate them, or the amount of computer storage required) increases faster than exponentially with respect to the number of taxa t, i.e. $\propto t^t$. There are simple formulae for counting tip-labelled binary trees, either unrooted or rooted. The number of unrooted binary trees is $(2t - 5)!!$ and the number of rooted binary trees $= (2t - 3)!!$, where t is the number of taxa. The 'double factorial' notation (!!) is the product of every second number, i.e. $(2t - 5)!! = 1 \times 3 \times 5 \times 7 \times \ldots \times (2t - 5)$ where the '...' indicates repeating the intermediate steps until you reach the last value $(2t - 5)$. For example, for 10 species $2t - 5 = 15$ so there are $1 \times 3 \times 5 \times \ldots \times 15 = 2\,027\,025$ unrooted binary trees. An informal way of finding the above formulae is given in Figure 16.6.

16.3.2.2 Computational Complexity

The fundamental measure of the usefulness of an algorithm (a numerical recipe) is the time and/or computer storage space it requires – its *computational complexity*. This can be the number of computational steps as a function of the size of the input data. It is usually expressed as a function of n; here it is the number of sequences (t) and so we use n and t interchangeably. If this function is bounded (limited) by a polynomial of n (such as n^2, or n^4, etc.), then the algorithm is a polynomial time algorithm and is *efficient*. The classification uses the worst case, and so an order n^2 algorithm guarantees to complete in that number of steps. The number of steps to determine a tree using the neighbor-joining (NJ) algorithm is a polynomial of degree 3 and we write $O(n^3)$ (pronounced as *order n cubed*) for the computational complexity of NJ.

In contrast, any algorithm which requires a separate calculation of each of the $(2n - 5)!!$ tip-labelled binary trees (such as maximum parsimony (MP) or maximum likelihood (ML)) cannot in general be computed in polynomial time. We write $O(n^n)$ for the number of binary trees, because $(2n - 5)!!$ is asymptotic to n^n. We consider complexity under just a single variable, the number of taxa. In practice it is necessary to also consider the complexity of an algorithm under sequence length, c. Most tree-building methods are linear with respect to c, but alignment algorithms may be of order c^2 or c^3. When an algorithm is $O(n.\log n)$ then, because $\log n$ is less than n, $n.\log n$ is less than n^2, so the algorithm is still efficient because its complexity is bounded by a polynomial in n. Some examples of efficient algorithms include:

1. Locating an item in an ordered list of n items is $O(\log n)$. For example, to find the telephone number of a friend in a telephone directory of n names, first compare your friend's name with a name in the middle of the book. Then compare it with the middle of the first, or second half, depending on whether her name is before

or after that central name. Repeating this subdivision process until finding the name requires at most $\log_2 n$ comparisons. The number might be located in fewer steps but the complexity always refers to the *worst* case; we can guarantee the algorithm will terminate in at most $\log_2 n$ steps.

2. If we needed to find the person with a specific number in this telephone directory, then potentially we need to look at each number until it is found. It could be the very first number, but also it could be nth, the very last, so this is an O(n) algorithm.

3. To sort the telephone directory by numbers (rather than by names) requires O($n.\log_2 n$) steps. This is achieved by locating the position of each successive number among those already sorted. The search algorithm of example 1 above orders each of the n numbers in $\log_2 n$ steps.

4. Finding the smallest distance for n taxa is O(n^2). The numbers of entries (distances) is n^2, and because the matrix is symmetric we need only compare the entries above the diagonal. Thus the number of comparisons is bounded by a polynomial in n^2, so the search is O(n^2).

5. In the unweighted pair-group mean average (UPGMA) algorithm (see later), the closest pair of taxa (requiring O(n^2) steps to find them, as in example 4) are amalgamated, giving a new matrix of distances between the $n - 1$ remaining 'taxa'. The closest pair in this matrix is again found, amalgamated, and the process repeated $n - 2$ times. The complete algorithm is thus O(n^3) and so is an efficient algorithm.

6. The tree-building algorithms of MP and ML, if fully implemented, require an evaluation of each of the $(2n - 5)!!$ binary trees. Even if each tree can be evaluated efficiently (as for MP) the number of trees cannot be bounded by a polynomial in n. Consequently, these algorithms are not polynomial and hence *not* efficient. In Table 16.5 we compare the growth of some of these functions for some small values of t.

7. The Hadamard conjugation used in the spectral analysis involves calculating values for each split, and there are 2^{n-1} splits. The fast Hadamard transform involved requires $2n$ steps for each split, and hence Hadamard conjugation is an O($n2^n$) algorithm. Again, this is not an efficient algorithm (even if it is the best possible).

It is easy to demonstrate the general problem resulting from the large numbers of trees. Estimate how long it would take to do all the calculations for algorithms of different efficiencies. Assume that the time for calculating each step is the same for each algorithm. If one million trees could be calculated per second, how long would it take to calculate all trees for 20 species? There are $\approx 2.2 \times 10^{20}$ trees for 20 taxa and $\approx 3.16 \times 10^7$ seconds year^{-1} so we can calculate $\approx 3 \times 10^{13}$ trees per year. Thus we require $2.2 \times 10^{20}/3 \times 10^{13} \approx 7$ million years calculating the trees for 20 taxa!

Thus for $t = 20$, a million-fold increase in computing speed is not going to make the calculation practical. In contrast, at 10^6 calculations per second, the Hadamard conjugation would require almost 21 seconds, and the NJ algorithm would require less than one-hundredth of a second.

This means that for $t > 20$ it is not practical to simply evaluate a function for each tree to find the optimal tree. Finding the optimal tree for t taxa is an example of a

Table 16.5   Values for some polynomial and exponential functions for small values of t.

t	$t \log_2 t$	t^2	t^3	$t2^t$	$(2t-5)!!$	$(2t-3)!!$
3	4.75	9	27	24	1	3
4	8.00	16	64	64	3	15
5	11.61	25	125	160	15	105
6	15.51	36	216	384	105	945
7	19.65	49	343	896	945	10 395
8	24.00	64	512	2048	10 395	135 135
9	28.53	81	729	4608	135 135	2 027 025
10	33.22	100	1000	10 240	2 027 025	34 459 425
11	38.05	121	1331	22 528	34 459 425	654 729 075
12	43.02	144	1728	49 152	654 729 075	1 374 931 575

non-deterministic polynomial (NP)-complete problem. The travelling salesman problem is another well-known example. That problem is, given the distances between all pairs of n cities, find the shortest path that allows a salesman to visit all cities and return to the start. This general class of NP-complete problems has the following properties: if an efficient solution is found for one, it will work for all; no general solution has been found for any; but there is no proof that an efficient solution cannot exist.

That is the bad news from complexity theory, which deals with the worst case scenarios. There are special cases where algorithms can run in polynomial time for the number of sequences (Fernandez-Baca and Lagergren, 2003). One example is for population data with many similar sequences; there may be only a single nucleotide difference between them. In this case, parsimony is a ML estimator, and a MinMax Squeeze program (Holland *et al.*, 2005a) has been able to guarantee the shortest possible tree for over 50 human mitochondrial genomes. However, having considered some of the theoretical computational questions, it is time to look at the models of evolution used when inferring evolutionary trees from biological data.

16.3.3   Three Parts of an Evolutionary Model

Scientific models are usefully classified into three parts: the structure of the system; the forces (mechanisms) operating on it and; some initial conditions that specify the system at a particular point in time. For an evolutionary model, the equivalencies are as follows,

1. A *tree* (or more generally a graph), is either rooted or unrooted. This is usually a tree, but cycles can arise from endosymbiosis, lateral transfer, gene conversion, allopolyploidy, hybridisation, etc. This part of the model gives the relationships between the sequences and is assumed even if there is insufficient information to specify a common mechanism (see later).

2. A *mechanism* of change (typically, changes to the sequences are stochastic and are independent and identically distributed (i.i.d.) – again, see later).

3. *Probabilities* **q** (weights) of observing a change along each edge of the tree, these are the *initial conditions* including edge lengths, transition/transversion ratios, and nucleotide composition at the starting point.

Given these three features, we calculate or predict the expected frequencies of patterns (splits) in the sequence data (Figure 16.7). For simpler mechanisms this is best done by spectral analysis using Hadamard transforms, a simple discrete-Fourier process. For small numbers of taxa, the calculations can be carried out in spreadsheet programs such as Excel. The transforms convert the vector of the observed frequencies of splits (**s**, the sequence spectrum) to the probabilities of change on each edge of a tree T (in the vector **q**). Each entry in these vectors is indexed by one of the 2^{t-1} possible splits (Section 16.2.3). The intermediate steps are generalised distances (Section 16.2.2), which are equivalent to paths through the tree (or graph). The series of vectors is thus **s** (observed frequencies of splits), **r** (observed generalised distances), **ρ** (rho, predicted (corrected) generalised distances), **γ** (γ, predicted (corrected) frequencies of splits), and finally **q** (γ values fitted onto edges of a tree). For simple models, these transforms are invertible with the use of Hadamard matrices (see later), so that **γ** can be derived from **s**, and vice versa.

The Hadamard transforms are a series of simple invertible steps involving a Hadamard matrix, **H**. This m by m matrix has entries 1 and -1 and satisfies the equation $\mathbf{H^t H} = m\mathbf{I}$, where **I** is the m by m identity matrix and $\mathbf{H^t}$ is the *transpose* of **H** (rows and columns interchanged). The transform allows a simple set of equations that add and subtract the lengths of edges on a tree to give path lengths; or conversely (the inverse) add and subtract path lengths to get lengths of individual edges. Figure 16.7 shows the relationships between data as sequences (i), and the evolutionary model (vi).

In following Figure 16.7 from (i) to (vi) you will see that the first step is translate the sequence data as a *sequence spectrum* **s**; the observed frequencies of each of the 2^{t-1} possible splits from the sequences for t taxa (see Figure 16.1c and Table 16.2). The Hadamard product $\mathbf{r = Hs}$ (a linear transform) calculates the observed *generalised*
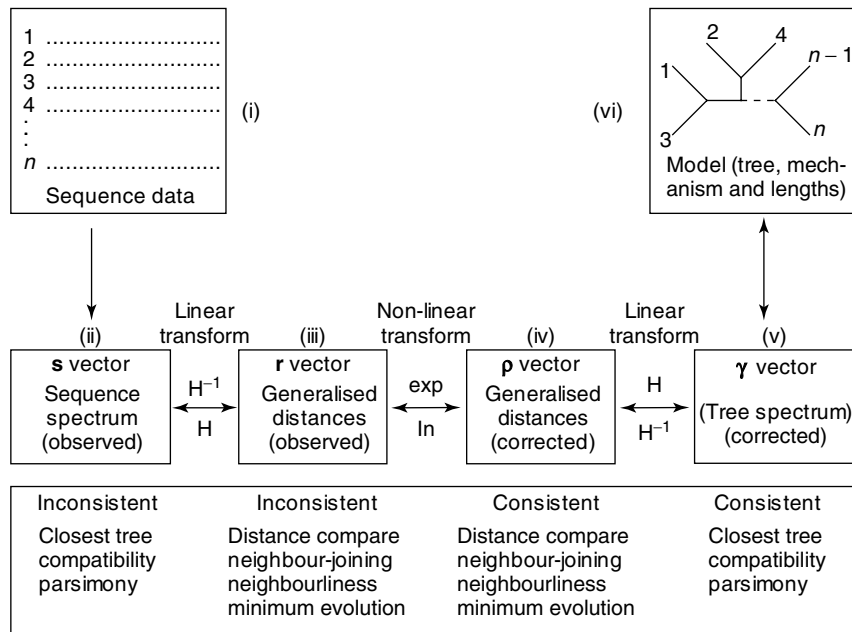
Figure 16.7 Invertible transformations between data (i) and model (vi).

distances **r** (iii). Many methods use just pairwise distances, i.e. distances that are calculated for each pair of taxa. Pairwise distances can be used to infer generalised distances, and they reduce the variance when corrections are made for multiple changes (Waddell *et al.*, 1994).

The step from **r** to **ρ** corrects, depending on the mechanism of evolution, for inferred multiple changes at a site. The values in **ρ** are *predicted generalised distances* (iv). The logarithmic function from (iv) to (iii) corrects for unobserved changes and depends on the distribution of substitution rates in the model. When all sites are evolving at the same rate (the uniform rate model), the natural logarithm is used in going from (iii) to (iv), with its inverse being the exponential function when going from (iv) to (iii). If not all sites evolve at the same rate then a Gamma distribution can be used, though other distributions are possible (Waddell *et al.*, 1997). The inverse Hadamard matrix (a linear transformation) is used to get from the generalised distances to corrected edge lengths $\gamma = \mathbf{H}\rho$, see (v). This inverts path lengths (generalised distances) to individual edges of a tree (observed splits, or splits in the underlying model). The three-step procedure from (ii) to (v) uses a Hadamard transformation, then a logarithmic transformation, and finally the inverse Hadamard transformation. Applying three functions sequentially, with the third being the inverse of the first, is (in mathematics) a conjugation. Thus the three-step procedure is called a *Hadamard conjugation*, and is invertible.

From this spectrum in (v), a tree can be selected by a number of selection criteria (parsimony, closest tree, minimum evolution, etc.) Thus, trees can also be selected anywhere along the transformation from **s** to **p**. Notice that in Figure 16.7 the Hadamard conjugation is NOT a tree selection criterion, it is an invertible method for moving between paths and edges without loss of information. The transformations are invertible, so the method can be used to calculate properties of sequences from simple models of evolution, or recover a model of evolution given observed sequences. Different tree selection procedures can be applied to components within the transformation. Distance based methods are applied to the **r** and the **ρ** vectors, while sequence based methods can be applied to the **s** and the **γ** vectors. The selection procedures are statistically consistent after the correction for multiple changes (iii) to (iv) (provided that the correct mechanism was used), but can be inconsistent otherwise.

For analysis of a set of t taxa, the Hadamard conjugation is a function of a spectrum (a vector of 2^{t-1} components) where each entry is indexed by a split. The simplest substitution process is the symmetric two character state model. The two character states could be purines (R) and pyrimidines (Y) in DNA or RNA. The model has a probability p_e for each edge e of the tree T, where p_e is the probability of a substitution (either R to Y or Y to R) between the vertices of the edge. It can also be applied (Hendy *et al.*, 1994) to the symmetric four-state models of Jukes and Cantor, and Kimura's 2ST and 3ST models (see Swofford *et al.*, 1996).

A Hadamard matrix is a square matrix with entries of 1 and −1 and where all the columns (or rows) are orthogonal. (Two vectors are orthogonal if the result is zero when their corresponding entries are multiplied and summed.) An 8×8 Hadamard matrix in an appropriate form is shown below. The order of rows and columns is critical, it must correspond to the order of the entries in **p**, **q**, **γ**, and the other vectors. That is, changing the order of rows and/or columns still gives a Hadamard matrix, but no other order would be suitable unless we also change the ordering of splits and subsets. Below we show H in two forms, the usual way on the left and on the right-hand matrix + stands for +1

and – for -1. The right-hand matrix is broken by dotted lines into four 4×4 submatrices to illustrate the recursive process for generating the matrix. The two top and the lower left submatrices are identical, the lower right has the sign reversed for every entry.

$$
\begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1
\end{bmatrix}
\quad
\begin{bmatrix}
+ & + & + & + & \cdot & + & + & + & + \\
+ & - & + & - & \cdot & + & - & + & - \\
+ & + & - & - & \cdot & + & + & - & - \\
+ & - & - & + & \cdot & + & - & - & + \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
+ & + & + & + & \cdot & - & - & - & - \\
+ & - & + & - & \cdot & - & + & - & + \\
+ & + & - & - & \cdot & - & - & + & + \\
+ & - & - & + & \cdot & - & + & + & -
\end{bmatrix}
$$
$$\mathbf{H}^{(3)} \qquad\qquad\qquad\qquad \mathbf{H}$$

For t taxa, $\mathbf{H}$ is generated recursively starting from the matrix with a single entry $\mathbf{H}^{(1)} = [1]$, and then forming the next larger Hadamard matrix as follows.

$$\text{Let } \mathbf{H}^{(1)} = [1], \text{ and } \mathbf{H}^{(m+1)} = \begin{bmatrix} \mathbf{H}^{(m)} & \mathbf{H}^{(m)} \\ \mathbf{H}^{(m)} & -\mathbf{H}^{(m)} \end{bmatrix}, \text{ for } 1 \leq m \leq t - 2, \mathbf{H} = \mathbf{H}^{(t-1)}.$$

Note that the signs of the upper left 2×2 submatrix are of the form,

$$\begin{bmatrix} + & + \\ + & - \end{bmatrix},$$

and the upper left 4×4 submatrix can be derived from this upper left 2×2 submatrix with the same pattern of sign changes. Thus, the four upper left 2×2 submatrices have the same pattern of sign changes. For example, row 6, column 5 will be the negative of row 2, column 1; row 8, column 7 will be the negative of row 4, column 3. Similarly, you can repeat the generating process once more to derive $\mathbf{H}^{(4)}$, the 16×16 matrix for five taxa, and in general $\mathbf{H}^{(t-1)}$, for the $2^{t-1} \times 2^{t-1}$ matrix for t taxa.

Hadamard matrices have several important properties. Firstly, it allows an invertible transform between splits and generalised distances. This requires the same numbers of entries in each vector, here 2^{t-1}. In mathematical terms the transform is one to one and 'onto', therefore it is invertible. A second important property is that multiplication of a vector by $\mathbf{H}^{(t)}$ can be reduced to $t2^t$ additions and subtractions (the fast Hadamard transform). Another particularly useful property is that the inverse of any Hadamard matrix ($\mathbf{H}^{-1}$) is the transpose of the matrix, $\mathbf{H}^{(t)}$ divided by the number of rows (2^{t-1} in our case), i.e. $\mathbf{H}^{-1} = \mathbf{H}^{(t)}/2^{t-1}$. Together these make the method simple but powerful. However, it is now time to consider the final aspect of theoretical background; mechanisms of evolution for sequence data.

16.3.4 Stochastic Mechanisms of Evolution

Markov models (named after a Russian mathematician) are widely used in science, including describing the evolution of sequences through time. They are very general and the simplest versions just require that the probability of substitution depends only on the current state. For sequences, this means that the probability of a substitution depends only on the present nucleotide at a site; it is independent of what the sequence might have been earlier in evolution. This is realistic for molecular evolution where, e.g. no

information is retained in a human sequence as to which nucleotide was present at a given site in early primates.

Fortunately, Markov models are well studied and relatively easy to work with. Consider a sequence that evolves from time A until time B, the total length of time is t. We are interested in both the rate of change at any point x, as well as the total amount of change from A to B. Evolutionary models are effectively reversible, we could equally consider the amount of change from B to A (though a different Markov process might be required).
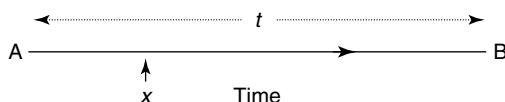
At any point x, the instantaneous rate matrix, $\mathbf{K}$, describes the rate (per site) that mutations are fixed into the sequence. The transition matrix, $\mathbf{M}$, describes the probability of states at a site being different at A and B, it depends on both $\mathbf{K}$ and the time t. (In Markov models, 'transition' is used in a general sense for any type of substitution, not as a contrast to a transversion.) The 3ST Kimura model (shown below on the left) is frequently used with nucleotides and has three classes of substitution, rates α, β, and γ (identified in the middle graph below). This model can be simplified to the Kimura 2ST model by setting $\beta = \gamma$, giving the model with separate rates for transitions and transversions. Furthermore, setting $\alpha = \beta = \gamma$ gives the simple Poisson model, the Jukes–Cantor model.

3ST rate matrix   graph of rate classes   a specific transition matrix

where $* = -(\alpha + \beta + \gamma)$.

Consider the rate matrix $\mathbf{K}$. At any point x along a lineage from A to B there is a probability (per unit time) that an adenine will be mutated to guanine (α), or to cytosine (β), or to thymine (γ). Consequently, the probability of replacing adenine by another nucleotide is $= (\alpha + \beta + \gamma)$. $\mathbf{K}$ is normalised with diagonal entries $-(\alpha + \beta + \gamma)$ so that each of the rows sums to zero; thus there is conservation of the nucleotides. The same process holds for the other nucleotides; they also have similar probabilities of changing (α, β, and γ). Under this model (which is simplistic, but useful and popular), the nucleotide composition will converge to equal proportions (25 %) of A, G, C, and T, even if the proportions of nucleotides was different when the process started.

The next step is to obtain the transition matrix, the expected probability of a difference between A and B over the time t. The expected number of substitutions of each class is αt, βt, and γt. The transition matrix, $\mathbf{M}$, is calculated from the instantaneous rate matrix $\mathbf{K}$ and t, as, $\mathbf{M} = \exp(\mathbf{K}t)$; with $\exp(\mathbf{K}) = \mathbf{I} + \mathbf{K} + \mathbf{K}^2/2! + \mathbf{K}^3/3! + \mathbf{K}^4/4! + \ldots$, which converges. This transition matrix gives probabilities for each class of substitution during the time interval t, and entries in $\mathbf{M}$ predict the number and classes of nucleotide changes between the times A and B. In the right-hand example above, there is a 3 % chance that any adenine in the sequence will have mutated to G, and 7 % to C and 7 % to thymine.

Consequently, there is an 83 % chance it is still an adenine (even though it may have mutated to another nucleotide and then mutated back again).

These matrices are simple and symmetric (the rate A → G is equal G → A). It is straightforward to compensate for unequal frequencies of nucleotides (or amino acids) if all sequences are similar, say 60 % G + C. This is modelled with a symmetric matrix plus a vector of nucleotide frequencies. However, when sequences have different nucleotide (or amino acid) compositions, an asymmetric rate matrix with its resulting asymmetric transition matrix is required. This asymmetry is not a problem (apart from requiring more parameters) if the transition matrix is being used to predict sequence data, such as in a simulation study, or for ML. However, when compensating for multiple changes it is necessary to make a 4 × 4 divergence matrix for DNA (or 20 × 20 for amino acids) for each pair of sequences (see Figure 16.1b, right). Taking the logarithm of the determinant of this divergence matrix corrects for multiple changes. This is the LogDet (or paralinear) distance correction (Lockhart *et al.*, 1994).

A frequent assumption is that changes occurring in the sequence are 'i.i.d'. Independent means that a change at any site in the sequence, and on any lineage on the tree, is independent of all other changes. Identically distributed is that all sites (nucleotides or amino acids) are selected from the same distribution of rates. Sites may have different rates, but are still drawn from the same distribution. The mathematics requires (Tuffley and Steel, 1997) that sites are always in the same rate class over the whole tree, whether fast, medium, or slower sites. This is problematic, and contradicts the biochemical knowledge that the 3D structure of proteins evolves through time. The covarion model of Walter Fitch is an early attempt to allow, over time, sites to vary in their rates. His model assumed a limited number of sites were free to change, but that this set could slowly drift as the 3D structure evolved. A hidden Markov Model (Tuffley and Steel, 1997) can approximate this covarion process, using just two additional parameters. These are the proportion of nucleotides that are (for functional reasons) free to change, and how quickly variable and fixed sites interchange. A relaxed covarion model would allow this proportion of fixed sites to vary. Future developments will probably incorporate more biochemical realism. Some mechanisms also assume the same rate of change along each edge of the tree – the *molecular clock* – and this imposes additional constraints on edge lengths. The chapter of Swofford *et al.* (1996) has a good discussion of the standard mechanisms, and how they are interrelated.

16.4   METHODS FOR INFERRING EVOLUTIONARY TREES

Most new methods for inferring trees have been developed for sequence data (DNA and protein), but most work equally well on discrete (i.e. not continuous) data from morphology, anatomy, biogeography, behaviour, etc. Methods for inferring tree are considered in three parts:

1.   an *optimality criterion*, to measure how the data 'fits' a particular tree (or graph);

2.   a *search strategy* for finding the optimal tree(s); and

3.   assumptions about the *mechanisms of evolution* for the data.

It is incorrect to speak of a 'method' of inferring trees without specifying the optimality criterion, the search strategy, and how multiple changes are handled (by the mechanism assumed). Examples of these three combinations are given with the optimality criterion identified as (1), search strategy as (2), and the mechanism as (3). A method could be, a parsimony (1) branch-and-bound search (2) on data corrected for multiple changes by the Kimura 2ST model (3). Or, a NJ search (2) for the minimum evolution tree (1) using observed (uncorrected) distances (3). Or, ML with hill climbing to find the optimal parameters on a tree (1), a general reversible model of change (3), and a heuristic search through the tree space (2).

Comparing two methods that differ in all three of above options has generated a great deal of confusion. To take just one example, several studies have compared 'MP' and 'NJ' (details are given later). These 'methods' differ in their optimality criterion, their search strategy, and whether or not they correct for multiple changes. But which of the three differences leads to a particular set of results? Such sloppy work is poor science, and unfortunately many simulation studies are not worth the silicon they are computed on. We will examine some criteria for evaluating the different approaches.

16.4.1   Five Desirable Properties for a Method

There is still no general agreement how to evaluate tree reconstruction methods. We have suggested that being *efficient, consistent, powerful, robust*, and *falsifiable* is a good starting point for analysis (Penny *et al.*, 1992). *Efficiency* is discussed earlier under computational complexity – to have any real chance of finding the global optimum for medium to large data sets, the algorithm must run in polynomial time. The concept is used in multiple alignment, in evaluating an optimality criterion on a single tree, and searching over all trees. Alignment is not discussed here, but there are efficient algorithms available for a weighting scheme of matches, mismatches, and gaps for pairs of sequences. However, in general it is necessary to optimise an alignment on a tree and there are no known efficient solutions for the combined problem. Optimality criteria such as parsimony, minimum evolution, and closest tree can be calculated efficiently for a single tree. For ML it is not yet known whether this is true.

The next two desirable features come from considering the effects of increasing the length of the sequences, c. Biologists accept that, with short sequences, there will be sampling error and are not too surprised if details of the optimal tree changed when longer sequences became available. Eventually, with long sequences, we expect convergence to a single tree. But is it the correct tree, the tree on which the sequences were generated? Two questions that arise are: does the method converge to the correct tree, and how long do sequences need to be to get *convergence*? A method (optimality criterion, search strategy, plus assumed mechanism of evolution) is said to be *consistent* if it guarantees to converge to the correct tree; conversely, it is *inconsistent* if it can converge to an incorrect tree.

There is both good news and bad news here. Any reasonable optimality criterion is consistent if the correct assumptions are made about the mechanism of evolution. But each is also inconsistent if invalid assumptions are made. Parsimony is consistent if the correct assumptions are made (and compensated for), inconsistent otherwise. ML is inconsistent if, e.g. it assumes that all sites are free to vary when most are fixed for functional reasons. Differences in nucleotide composition between sequences lead to similar problems for all optimality criteria. There has been a lot of confusion with statements that optimality

criteria are, or are not, consistent, when the real issue is the consistency of the complete method – optimality criterion, search strategy, and assumed mechanism of evolution.

In an important development J. Felsenstein showed in 1978 (see Swofford *et al.*, 1996) that even with only four taxa, but with unequal rates of evolution, uncorrected parsimony, and some distance methods using observed (uncorrected) data were not consistent even with two-state sequences under a symmetric model of substitution. With five or more taxa, uncorrected parsimony is also inconsistent even with equal rates of evolution, and even with equal edge lengths for larger numbers of taxa. We (Hendy and Penny, 1989) suggested that there was a general problem when there was a juxtaposition of long-short-long edges on a tree – the two long edges (branches) tended to join together. We called this the *long-edges-attract* problem. Although it may appear bad news, it does give biologists the chance to use their biological knowledge to select taxa that intersect long edges in the tree. Uncorrected parsimony can be improved with additional taxa that join into what would otherwise be long edges on the tree, and/or by correcting for multiple changes. A method is inconsistent under a given model if there are any possible cases where it converges to the wrong tree. It is a separate, but important, question as to whether such cases are common. This leads us into the next criterion, how long should the sequences be.

Do we need to analyse sequences of 10^3, 10^5, or 10^7 nucleotides in length? Convergence with relatively short sequences we call *powerful* and is another desirable property. Not surprisingly, considerable effort has gone into measuring the rates of convergence and the techniques available include: computer simulation; repetitive sampling (bootstrap and jackknife); congruence of trees from different sets of data; fit to a model; prior knowledge (occasionally possible); and analytical solutions.

Different rates of convergence can result from methods using different amounts of the information in sequences. As the number t of taxa increases, methods using genetic distances use only a vanishingly small fraction of the information in the data. The number of distance values increases in proportion to t^2 while the number of patterns in the data (bipartitions or splits) increases exponentially (2^{t-1}). Uncorrected parsimony omits sites that are constant or where only one character state occurs more than once (singletons). But both types of position give information about the appropriate model of evolution. In addition, such sites occur frequently, so the estimates of their frequencies are more accurate.

Many optimality criteria, especially non-parsimony methods, ignore information from insertions and/or deletions (which are highly informative), and also from other types of biochemical information (such as SINEs). Using only subsets of taxa (such as quartets) disregards most of the information that can come from incompatibilities. Variability within taxa is again ignored. Information loss is not limited to a particular method. All omit some information and consequently the power of each method is reduced. As yet, we can make only rather weak predictions about the relative power of methods. For large data sets, ML and closest tree methods use more sequence information than uncorrected parsimony, and these together use more than distance methods.

To move onto the fourth desired feature – *robustness*. A method may be consistent under one model, but become inconsistent with only small deviations from it. Everyone would like a method that is efficient, powerful, and consistent, even with sizeable deviations from the model – i.e. a method that is *robust*. Robustness is an area where there is almost a conspiracy of silence from those developing techniques. In general, we only compare data from simulations where the data is generated on a tree, and by a known mechanism.

With simulation, data can be generated under one mechanism of evolution and different assumptions tested when inferring a tree. For example, data may be generated with a γ distribution of rates at different sites, and the effect studied of recovering the tree with very different γ values (including all sites evolving at the same rate). This is an excellent start, but we still have little idea of the consistency of different approaches as real data start deviating from the simplest models.

Computer simulations can study several aspects of the problem of robustness. A method that was consistent with equal nucleotide frequencies can become inconsistent when the frequencies of character states (such as GC (guanine plus cytosine) content) vary between taxa. Simulation studies are only of limited help unless the results are integrated into a theoretical analysis of the methods. The real problem is that, considering all models, there are billions of combinations of parameters to be tested. By itself, testing four or five combinations of parameters is not that useful. Simulations are more powerful when testing predictions from theoretical studies of algorithms.

The final desired feature is the standard scientific requirement that the data, in principle, must be able to *falsify* the model. This is perhaps the least studied, and the most difficult, of our five desirable features. Perhaps the worst methods in this respect are clustering (or local search) methods that do not evaluate an optimality criterion over the whole tree. They output just a single tree, and give no ranking of the other $(2t - 5)!!$ trees. Would a network be better for a given data set? We distinguish cases when the model is very wrong, from those when the model is incomplete – missing some significant aspect. Examples are when recombination, or gene conversion, has occurred and a tree (lacking cycles) is then incomplete. Another case is when the mechanism assumes similar nucleotide compositions, when the sequences differ significantly. In such cases, features of the model such as, the tree, stochastic mechanism, similar rates, may all be reasonable; but the full model requires at least one additional parameter.

In practice it is difficult to test the full model completely. If a tree is assumed, it may be easy to test quantitatively two versions of the mechanism of evolution. Conversely, for a given mechanism, we can test between different trees. This aspect of the falsifiability of models requires more consideration. At present we are left with the status that no method (combination of optimality criterion, search strategy, and initial conditions) can meet all five of our desired features. This is a problem, but has not stopped a lot of progress in understanding the evolution of both macromolecules and organisms. It is perhaps a call for less special pleading about our own favourite method, more caution in interpreting results, and a challenge to keep improving.

16.4.2 Optimality Criteria

It is time to consider the first aspect of a method for recovering trees – the optimality criterion (objective function). Common optimality criteria include parsimony, minimal length, ML, least-squares fit, minimum evolution, etc. An optimality criterion usually estimates how well the data fits a particular tree though the splits-graph criterion is a fit of distance data to a phylogenetic network.

16.4.2.1 *General Comments on Parsimony*

Parsimony is one of the most frequently used optimality criteria, but there are ambiguities in the use of the concept. In everyday life the term is used as an adjective, e.g. when

someone is said to be parsimonious; not sharing, thrifty, or stingy. However, parsimony has a long history in science as looking for the 'simplest' explanation. Thus there are two major usages.

1. The Principle of Parsimony (Ockham's razor). This is its most common usage in science – find the simplest hypothesis that explains the data; do not add complexities to a model that works.

2. Minimising the number of mutations when fitting sequences onto a tree.

The two usages can be in contradiction. In other words, it is not always true that the simplest explanation (the Principle of Parsimony) is to minimise the observed number of mutations on a tree (Steel and Penny, 2000). It is traditional when minimising the number of mutations on a tree not to make any adjustments for multiple changes at a site. This is best called *maximum parsimony* to distinguish it from other variants such as weighted parsimony, or corrected parsimony. Weighted parsimony uses additional knowledge of the mechanism of evolution such as the transition/transversion ratio, or the relative rates of change between nucleotides or between amino acids. Similarly, corrected parsimony (Penny *et al.*, 1996) minimises the number of mutations on a tree after correcting for inferred multiple changes.

MP thus consists of two independent concepts: minimising the number of mutations on a tree (the optimality criterion), and, using the observed data (not correcting for multiple changes). Normally the only part of the model of evolution it assumes is the tree itself. However, parsimony can be modified to minimise the number of mutations of data onto a network (i.e. a connected graph with cycles). The failure to distinguish between the minimal length (number of mutations) criterion, and using observed data (uncorrected for multiple changes) has led to acrimonious debates because the problem was not analysed adequately.

16.4.2.2 Relationship between Parsimony and Maximum Likelihood

Just as with parsimony, there are many usages of the term *maximum likelihood* (Steel and Penny, 2000) and we use their terminology. All usages have the common theme of maximising the likelihood of observing the data, given some background information about the model of evolution (in this case, a tree, a mechanism of change, and initial conditions). J Huelsenbeck discusses ML itself separately. Here we only consider the relationship between parsimony and ML.

The usual form of ML in phylogeny is called *maximum average likelihood* (ML_{av}). This uses a *weighted* tree (weights associated with edges), and then sums the likelihoods of observing the data over all codings at the internal nodes. We illustrate ML_{av} with two character states 'a' and 'b' over an unrooted tree connecting four taxa (Figure 16.8). There are four ways the two internal nodes (ancestors) could be coded, both internal nodes could be 'a' [a, a, Case 1], or one 'a' and one 'b' [a, b, Case 2 or b, a, Case 3] or both 'b' [b, b, Case 4]. For simplicity we assume the same rates (probabilities) of change on each edge of the tree, namely 0.05. It is straightforward to calculate the likelihood of the pattern [a, a, b, b] by:

1. calculating the probability of a specified coding at internal nodes for Cases 1–4; and

2. summing the probabilities over all possible internal codings (four in this example).
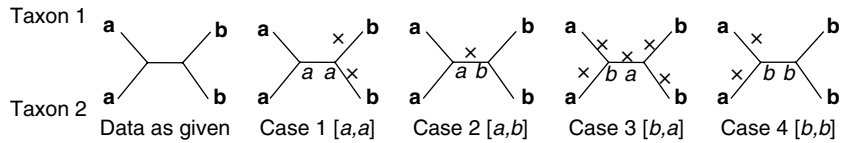
Figure 16.8   Calculating maximum average likelihood, given the pattern **a**, **a**, **b**, **b** on the external points, and with the internal points indicated as being either '*a*' or '*b*'. The symbol '×' marks each edge where a change (mutation) is required for a particular coding at internal points. There are two such changes in Case 1, one for Case 2, etc. This simple model gives the same results wherever the root is placed; the model is essentially an unrooted tree.

For 0.05 as the probability of a change, the probability of no change is 0.95 $(1 - 0.05)$. Thus, for the tree in Case 1 of Figure 16.8, the probability (to five decimal digits) is

Case 1[a, a] $0.00214 \ (= 0.95 \times 0.95 \times 0.95 \times 0.05 \times 0.05)$;

Case 2[a, b] $0.04073 \ (= 0.95 \times 0.95 \times 0.05 \times 0.95 \times 0.95)$;

Case 3[b, a] $0.00000 \ (= 0.05 \times 0.05 \times 0.05 \times 0.05 \times 0.05)$; and again

Case 4[b, b] $\underline{0.00214}$ (the same as the first). Summing these values gives

0.04501 as the probability for observing this data under this model.

This is for just a single site in the data, probabilities for each site are multiplied together (in practice, the logarithms are taken and added in order to keep the numbers manageable).

Another form of ML is to find the most-likely codings at the internal nodes of the tree, and this is maximum most-parsimonious likelihood (ML_{mp}, Steel and Penny, 2000). For the example above, Case 2 is the most likely and so the internal codes [a, b] would be selected. (In this example, the same choice would be made by MP, but this is not always true.) This form of ML would be chosen if a researcher wishes, e.g. to synthesise the most-likely ancestral form of a gene and test whether it functions as expected.

However, even more detailed information might be required; perhaps the investigator wishes to know the most-likely coding at every point along the edges of the tree, not just for the internal nodes as for ML_{mp}. This more highly specified form of ML is called *evolutionary pathway likelihood* (ML_{ep}). This form of ML is equivalent to MP (Steel and Penny, 2000), this is the first of the equivalencies between ML and MP.

These three (ML_{av}, ML_{mp}, ML_{ep}) are examples of maximum relative likelihood. In each case they are selecting a weighted tree. Another form of ML in phylogeny is maximum integrated likelihood (Steel and Penny, 2000). This form of likelihood is probably closer to what researchers really want and may, e.g. take into account the *a priori* distribution of trees. However, this is outside the scope of studying the relationship between parsimony and likelihood. To sum up, there are several forms of both parsimony and likelihood. This in itself is not a problem at all, as long as it is made explicit about which version is being used.

Having considered some forms of ML, we turn to a more detailed consideration of parsimony and ML, an area that has been controversial. The relationship between the two concepts appears complex but is being resolved. There are occasions when the MP solution is equivalent to a ML solution.
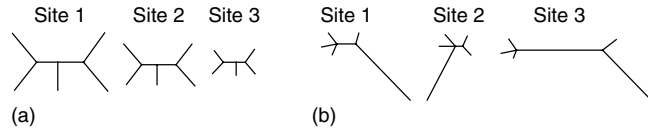
Figure 16.9 Common mechanism and no common mechanism. (a) Common mechanism – each site evolves on the same tree and proportionally has the same expected rate of change on every edge (branch). (b) No common mechanism – each character or site evolves on the same tree, but each character has different rates of change for every edge.

If there is only a single column of data, then the ML and MP solutions are the same. That is for this particular case, the ML estimator is MP. With only a single site, there is no additional information from combining additional sites. This may sound trivial but it leads to a much more interesting case, when there is no common mechanism of evolution shared by all the characters. For example, each character may have evolved on the same tree, but the potential for a character to change on a particular edge of the tree varies between characters. The concepts of common mechanism, and no common mechanism, are illustrated in Figure 16.9. An example of no common mechanism could occur with morphological data – the chance of an insect evolving eight legs may be zero, but somewhere deep in the invertebrate tree a $6 \leftrightarrow 8$ change (for example) must have been possible. Some major rearrangements in genomes and/or gene structure have similar properties. Although with no common mechanism the MP and ML solutions are the same, it is an area where ML itself does not guarantee consistency. Basically, with a common mechanism each site is an estimate of the same process. This means that additional sites give an improved estimate of the common underlying process. Conversely, with no common mechanism, each site is basically a new and unique process.

Before considering additional examples where parsimony is the ML estimator it is worth noting that with the present example it is a good point to consider whether the 'Principle of Parsimony' supports using observed data (rather than considering multiple changes at a site). Unfortunately, it does not. Advocates of Ockham's razor (the Principle of Parsimony) would, as illustrated by the following example, invoke the principle at this point. They would point out that, for some data sets, assuming a common mechanism for all sites is 'more parsimonious' than assuming 'no common mechanism' (each site evolving differently).

Table 16.6 has a small segment of homologous sequences from a pseudogene (a duplicate copy of a gene that no longer codes for a gene product). The full sequences are over 10 000 nucleotides long, and as a first approximation it is expected that there is little selection at any of the sites – most mutations would be neutral. Therefore, according to the Principle of Parsimony, it is simplest to assume one common mechanism for all sites, not 10 000 different mechanisms, one for each site. Each site is thus an estimate of the same underlying model – the same tree, the same mechanism, and same rates of change on each edge of the tree. In this case the Principle of Parsimony supports the standard (average) ML approach, rather than assuming a separate mechanism for each site (as MP does). Thus it is 'simpler' to correct for multiple changes. In this case, the two uses of parsimony (Ockham's razor, and minimising the number of mutations on observed data) are not in agreement.

Another example of the equivalency between parsimony and likelihood is with an extremely large number of character states, two changes may never occur at same site in

Table 16.6 A small segment of sequences from a pseudogene. If there is little selection at such sites, then it is more parsimonious to assume one common mechanism for all 10 000 sites.

... A C A C T G A G G G A A G G A T G A G A A A T A A A T G T G A A A G C A ...			Human
	G		Chimpanzee
	T		Gorilla
	T	C	Orang-utan
G	T		Rhesus monkey

different lineages. Here again the MP and ML solutions coincide (Steel and Penny, 2004; 2005). Examples include SINEs or LINEs inserting at a new site. Such unique events may be the most reliable evidence for oldest divergences, but this requires large amounts of sequence data to find the relevant examples. For example, each insertion of a transposable element would be at a different location in the genome. Some gene and/or chromosomal rearrangements may also be unique events and MP is then the ML solution.

A major difficulty with ML is that it is computer intensive. There is no efficient algorithm known for calculating the ML for just a single tree, even before considering the large numbers of trees. In Figure 16.8, the calculation was done for just one set of edge weights (rates). It is necessary to repeat the calculation many times, adjusting the rates slightly until the highest likelihood value is found. It has been shown that data sets can have many optima on a single tree (i.e. combinations of edge weights that cannot be improved by slight iterations, see Chor *et al.*, 2000). Again, this is for a single tree and, for large data sets, likelihood methods are limited to heuristic searches of tree space (see later).

To summarise this section, there are interesting relationships between maximum (uncorrected) parsimony and likelihood. They are equivalent on a single column of data and thus when there is no common mechanism shared by the characters. They are also equivalent under ML_{ep}, or there are infinite character states (the same mutation never occurs twice). Another example is when there are many sequences just one mutation apart (i.e. the sequences are 'abundant', as in a population study). However, sequences from different species are usually quite distinct ('sparse') and the two criteria differ, and need not arrive at the same tree. It is now time to look at the parsimony criterion itself.

16.4.2.3   The Fitch Algorithm for Parsimony

Parsimony is both easy and fast to calculate on a single tree, it is $O(t)$ – though the problem of the large number of trees still remains. For a set of t aligned sequences, and a labelled tree T, parsimony counts the minimum number of changes on the edges of T. Consider an example of five species (taxa), where at one site, the first two have A (adenine), and the other three have G (guanine). Find a tree that requires the smallest number of changes (see Figure 16.10). With only five taxa a minimal coding can be found by eye. The trees are unrooted because the parsimony score is independent of the position of the root; other information is required to identify the root.

The data A_1 and A_2, and G_3 to G_5 are the leaves of the tree. In the first tree, only a single A $\leftrightarrow$ G change is required, indicated by the arrow (unrooted trees do not specify which direction the change is going), but two changes are necessary for tree II. On the parsimony criterion, tree I is preferred over tree II for this single column of data. Tree II requires at least two changes, no matter how you code the internal nodes (all three
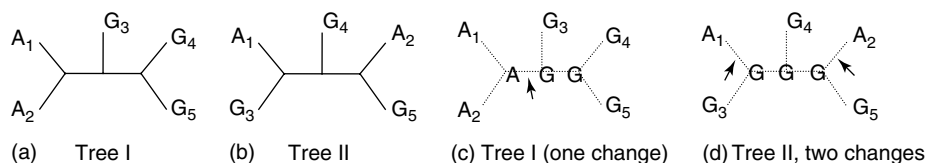
Figure 16.10 The parsimony criterion. The data is a site with character states $[A,A,G,G,G]^t$, these are shown as A_1, G_3, etc. Two trees (I and II) are shown in (a) and (b) and then minimal coding at the internal points are shown in (c) and (d); the changes (mutations) required are shown as arrows. Parsimony selects the tree with fewest changes, tree I in this example, but this is just for one site in the sequence.

internal nodes are be labelled 'G' in this case). This visual process works only for a few columns on a small number of taxa.

The Fitch algorithm quickly finds the parsimony score. It moves through the internal nodes (starting from neighbouring pairs of taxa) and gives both the minimal number of changes for the tree and a first estimate of the character states at internal nodes. There are many alternatives for moving through the tree, but the result is independent of the choice. For each internal node the algorithm checks two adjacent nodes where the character states are already defined; either because they are external nodes labelled by the sequences, or they are internal nodes for which a first estimate has already been made. The first estimate takes the intersection ($\cap$) of the character states at the two adjacent nodes. If the intersection is not empty, keep the value and move to the next internal node. If the intersection is empty, then take the union ($\cup$) of the character states and add a 'cost' to the total length of the tree. The cost (or weight) is usually equal for each column, but if the same pattern occurs in several columns then the cost is the number of times the pattern appears (or the expected number of times for corrected parsimony). For weighted parsimony the cost varies between character states, transitions are usually weighted less than transversions, and the weights need no longer be integers. The costs are summed over all columns in the data.

The tree (or trees) that requires the smallest number of changes (mutations) over the whole data set is selected. There is a second pass for the Fitch algorithm but it is only used to finalise the alternatives that may occur at an internal node on a minimal coding. The Fitch algorithm is efficient (on a single tree) because it is linear with respect to t, the number of taxa.

Many other things could be said about parsimony. As an optimality criterion it is not consistent on observed sequences, but neither are other criteria. However, we have pointed out that most criteria are consistent if appropriate corrections are made for multiple changes, and this includes parsimony. There is an interesting case under a molecular clock where MP can be inconsistent if applied to the tree as a whole, but is consistent if the tree is built up from quartets of taxa. There is still much to be learned about the properties of optimality criteria!

16.4.2.4 Minimum Evolution and Related Criteria

This section includes several optimality criteria that involve distances directly, or else edge lengths on trees (often calculated from distances). The details are well covered in

Swofford *et al.* (1996) and consequently we only give an overview here. However, the section again illustrates the necessity of distinguishing between optimality criteria and search strategies.

NJ, e.g. is one of the most popular 'methods' for distances but it is basically a tree search strategy that can be implemented with other optimality criteria, including parsimony and ML. In the standard NJ programs a pair of taxa x and y is selected to combine based on their *net divergence* – estimates of the length of the internal edges that results from pairing two taxa. Initially x and y are separate taxa, but later in the search they can belong to a cluster of taxa that has already been combined. To calculate the net divergence for any pair x and y, ($\mathbf{E}_{xy}$), the distances from each taxon in the pair to every other taxon is used, as follows,

$$\mathbf{E}_{xy} = \frac{1}{n'-2} \left(\sum_{i=1}^{n'} \left(d_{xi} + d_{yi} \right) \right) - d_{xy},$$

where n' is the current number of clusters (n' decreases during tree building; i.e. part of the tree searching section). Suppose we choose clusters x and y to amalgamate to form the new cluster, z. A new distance matrix $\mathbf{D}$ is recalculated with z replacing x and y, as follows

$$d_{zj} = \frac{d_{xj} + d_{yj} - d_{xy}}{2} = d_{jz} \text{ (the values are symmetric).}$$

The procedure is consistent if the distances are additive. Essentially, the optimality criterion is maximising the lengths of the internal edges of the tree but the search strategy only finds the maximum value for the next edge. It is relatively easy to evaluate the optimality criterion over the entire tree (not just a single edge at a time). In a general sense then, NJ is a local (greedy) search strategy using a minimum evolution optimality criterion.

Another optimality criterion just selects the smallest entry in the distance matrix, this is UPGMA, or average linkage. After each clustering the distance matrix, $\mathbf{D}$, is modified to

$$(d_{zj}) = \frac{a_x d_{xj} + a_y d_{yj}}{a_x + a_y} = (d_{jz}), \text{ where } a_x \text{ is the size of cluster } x, a_y \text{ that of cluster } y.$$

Additive distances are not sufficient to guarantee that UPGMA is consistent; it also requires that they fit a molecular clock, i.e., the distances between each taxon and the root are the same. On the other hand, when a molecular clock is appropriate UPGMA can out perform NJ (see Holland *et al.*, 2003). There are also a group of Fitch–Margoliash (FM) methods where each possible tree is tested, and edge lengths assigned to minimise

$$\sum_{1 \le i < j \le n} w_{ij} \left| d_{ij} - p_{ij} \right|^{\alpha},$$

where w_{ij} is a weighting factor, d_{ij} is the distance (which may be adjusted for multiple changes), p_{ij} is the sum of the edge lengths on the path on the tree between i and j (the patristic distance). α is a constant, usually 1 or 2 depending on whether absolute or squared deviations are to be minimised. While reasonably fast algorithms exist to find the best-fit edge lengths when $\alpha = 1$ or 2, they are not available for other values of α.

An important criterion, Closest Tree, minimises the Euclidean distance between either the inferred (or observed) bipartition spectrum γ (or s), and that expected for a particular tree T. When the split spectrum is known, as from character sequences, the square of this distance is

$$\sum_{i>0, e_i \notin T} \gamma_i^2 + (2n-2)\gamma_T^2,$$

where

$$\gamma_T = \frac{1}{2n-2}\left(\gamma_0 + \sum_{e_i \in T}\gamma_i\right) \text{ and } \gamma_0 = -\sum_{i=1}^{m-1}\gamma_i.$$

This optimality criterion, because it is implemented over the whole tree, can be implemented with branch-and-bound searches of tree space, as described in Section 16.6.1.1. Thus there is a range of optimality criteria that work either directly on distances, or on how well edge lengths on a tree predict the data. More research is still required with different combinations of optimality criteria and search strategies; there tends to be unnecessary associations of criteria, corrections, and search strategies.

Distances (pairwise or generalised) are readily corrected for multiple changes at a site for a wide range of possible mechanisms. Again the details are well handled in standard texts such as Swofford *et al.* (1996), Page and Holmes (1998), and Nei and Kumar (2000). The simplest is a standard Poisson correction, which is called *Neyman* or *Cavender/Farris* for two character states, Jukes/Cantor for four. In general, the correction is

$$d'_{xy} = -((r-1)/r) \cdot \ln(1 - (r/(r-1) \cdot d_{xy}),$$

where d' is the inferred distances after correcting for multiple changes at a site and r the number of character states ($r = 2$ for two-state characters, 4 for nucleotides and 20 for amino acids).

The main problem with the corrections is that they both increase the variance on the distances, and introduce a bias that overestimates the distances. This bias results from taking the logarithm of the observed distances and is particularly marked for relatively short and/or quite distant sequences. The formulae under-correct slightly if a particular distance is smaller than expected – but even more overcorrect if the distance is larger than expected, even by the same amount ε. This asymmetry in the correction leads to the bias. Formulae are available to reduce the bias, and in our experience should always be used. The Tajima (1993) formula is;

$$d'_{ab} = \Sigma\frac{((k_{ab}^{(i)} \times 2^{i-1})}{(i \times c^{(i)}))}, \text{ where } i \text{ is summed over 1 to } k_{ab}$$

and where

$$z^{(i)} = \frac{z!}{(z-i)!}$$

The parsimony criterion can also be used after correcting for multiple changes (Penny *et al.*, 1996) and although such a method does not use 'distances' in its final step, it does use distances (pairwise or generalised) in intermediate steps during the correction for multiple changes. This emphasises that there are many interrelationships between methods and each 'method' is a combination of an optimality criterion, a search strategy, and an assumed mechanism of evolution (included the option of insufficient knowledge

to be able to correct for multiple changes). The interrelationship of methods is still an active area of research.

16.5 PHYLOGENETIC NETWORKS

Phylogenetic networks can usefully illustrate evolutionary relationships when the evolutionary history of sequences or species may be poorly represented by a tree. They have been used to study viruses and plants species where the underlying assumption of tree-like evolution is inadequate possibly due to recombination, lateral gene transfer, or hybridisation. They display the extent of conflicting signals in the data and thus provide a visual measure of confidence in a model. Across a range of taxonomic groups, species trees and gene trees may differ due to gene duplication and loss, or lineage sorting. This gives rise to the gene tree/species tree reconciliation problem, which prompted Maddison (1997) to describe species phylogeny as *a cloud of gene histories*. In studies within interbreeding populations, networks can represent recombination. In all these cases it is of interest to reconstruct a reticulate evolutionary history and represent it with a *directed acyclic graph* (DAG). Even when the underlying historical signal fits a tree there may be conflicting non-historical signals caused by sampling error, long-branch attraction, nucleotide composition bias, or changes in the substitution rate at individual sites across the tree, as well as alignment or misreading errors. Here we require networks that can visualise conflicting signals in data, but we do not require an explicit representation of evolutionary history.

16.5.1 Reconstructing Reticulate Evolutionary Histories

Each site (character) has a tree-like history, even when sequences have a reticulate history. We expect that contiguous segments of DNA (for instance most genes) evolve along a tree, but that the complete sequence alignment may be a mosaic of different tree-like parts (Maddison, 1997). This means that the problem of representing the evolutionary history of the complete sequences often reduces to finding a DAG that displays a set of trees. There will not generally be a unique solution so we need to find minimal DAGs that minimise the number of reticulation nodes (such as hybridisation events or recombination nodes).

Bordewich and Semple (2007) have shown that in general it is NP hard to find a minimal hybrid phylogeny (a DAG with some restrictions) that displays two gene trees. Baroni *et al.* (2006) gave a non-polynomial time algorithm for computing a minimal hybrid phylogeny for two gene trees, but the problem of combining more than two gene trees is unresolved. Special cases can be solved in polynomial time. For instance, if the hybridisation events are all independent (cycles do not overlap) the minimal hybrid phylogeny can be found efficiently (Nakhleh *et al.*, 2005). Huson *et al.* (2005) have made further progress; their method solves cases where the reticulations are non-independent but overlap in a special way. These approaches all make the assumption that the gene trees are error free. Nakhleh *et al.* (2005) provide a polynomial time algorithm for constructing a hybrid phylogeny from two gene trees where error is allowed but the network can have only a single reticulation. Another special case is when one of the input trees is known to be the species tree for the gene tree/species tree reconciliation problem; the aim is to find a series of events (gene duplication, loss, or lateral transfer) that explain how a gene tree

has arisen in the species tree. Hallett and Lagergren (2001) give an efficient algorithm for this where only lateral gene transfer is considered.

Huber *et al.* (2006) infer DAGs that represent the evolutionary history of polyploids. Unlike a lateral gene transfer event, where a species acquires a novel gene or a current gene is displaced, a polyploidy hybridisation event is additive, all versions of the gene are retained. This scenario gives rise to multilabelled trees that are like standard phylogenetic trees except that the same taxon name may appear at more than one tip. Their algorithm finds a DAG that exhibits a multilabelled tree using a minimal number of recombination nodes.

Recombination is an important process within a sexual population and in this case we need to recover the ancestral recombination graph – a DAG where at recombination nodes a breakpoint indicates which sites inherit their state from the left parent and which inherit from the right parent. Work in this area usually assumes the infinite sites model of sequence evolution, which implies that each site undergoes at most one mutation, so any incompatibilities between sites must be explained by recombination events rather than parallel mutations or reversals (Hein, 1990). The problem of finding the minimal ancestral recombination graph (where minimal refers to the number of recombination events) is NP hard (Wang *et al.*, 2001). Song and Hein (2005) give a non-polynomial time algorithm for it. Gusfield *et al.* (2003) give a polynomial time algorithm for the special case where cycles introduced into the DAG by recombination nodes are all independent.

16.5.2 Displaying Conflicting Phylogenetic Signals

Rather than attempting to explicitly model reticulate evolution it is often of interest to simply display conflicting signals in the input data. This is partly because the problem of estimating reticulate evolution is computationally difficult, and partly because although we think a tree is useful we are nevertheless interested in the signals in the data that do not fit on a tree.

Splits graphs are a versatile tool for displaying conflict in many types of phylogenetic data. A split is a bipartition of the taxa set (see Section 16.2.3), so every edge in a tree corresponds to a split, since its removal separates the labels into disjoint subsets. A split system is any collection of splits, and a splits graph is a network that represents the split system. Trees are the simplest splits graphs; split systems that correspond to trees are called *compatible*, and those that do not are *incompatible*. Incompatible split systems can be represented by splits graphs but each split corresponds to a set of parallel edges (Figure 16.11). Although every splits graph corresponds to a unique split system, the same split system can be displayed by many splits graphs.

A k-clique in a split system is a collection of k pairwise incompatible splits, a k-*compatible split system* is defined by requiring that there is no $(k + 1)$ clique. Any k-compatible split system can be displayed by a k-dimensional graph. However, it is hard to interpret splits graphs in more than three dimensions, so it has been an active area of research to characterise split systems that are less complex than general split systems. *Circular split systems* arise from defining a circular ordering of the taxa and then only allowing splits that correspond to a contiguous segment of this ordering, or in graphical terms, splits that correspond to cuts though the circle. Circular split systems can be displayed by planar graphs even though they can contain k-cliques for arbitrarily large k (see Figure 16.12). *Weakly compatible split systems* are defined by requiring that there is no set of four taxa a, b, c, d where all three possible quartet trees $ab|cd, ac|bd, ad|bc$
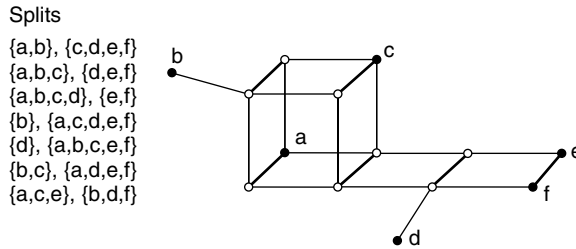
Figure 16.11 Incompatible split systems can be represented by splits graphs. Unlike a tree where each edge corresponds to a split here each set of parallel edges corresponds to a split. The edges in the splits graph that correspond to the split {a,c,e}, {b,d,f} are shown in bold.
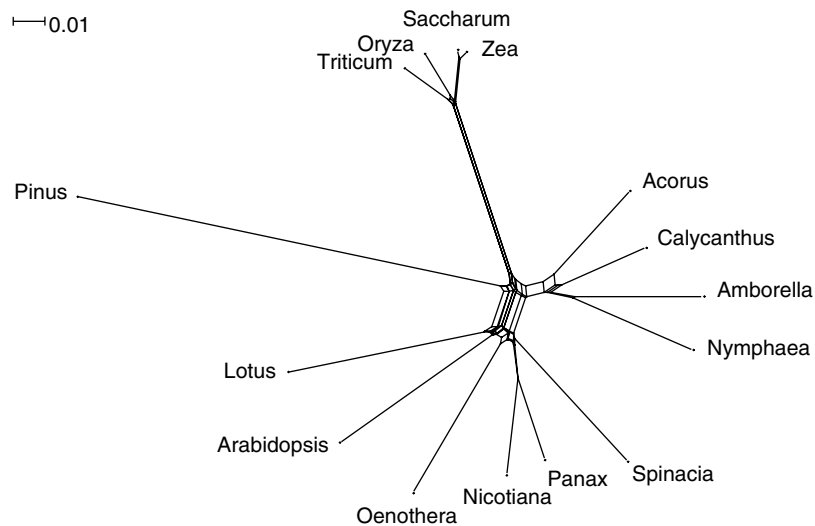
Figure 16.12 A NeighborNet of uncorrected distances from an alignment of 89 436 sites of the chloroplast genomes of 14 angiosperms and 1 gymnosperm (Pinus). The split system found by NeighborNet is circular and the resulting splits graph can be drawn in the plane.

are displayed in the split system. In other words, any two quartets can be shown on the graph, but not all three. Weakly compatible split systems can be displayed by 'almost' planar graphs.

We now give a brief description of splits-based methods. All the methods mentioned below (with the exception of quartet imputation supernetworks) are implemented in the freely available software package *SplitsTree4* (Huson and Bryant, 2006). Spectral analysis and median metworks are implemented in spectronet (Huber *et al.*, 2002).

Split decomposition (Bandelt and Dress, 1992) works from a distance matrix and identifies a set of weakly compatible splits. Split weights are assigned by a function (the *isolation index*) which is a minimum of a quantity evaluated over all quartets displayed by the split; this means that split-decomposition networks can be poorly resolved (star-like) especially for large data sets. The split residue or percentage of the distances that is not captured by the split system is summarised by the fit percentage.

NeighborNet (Bryant and Moulton, 2004) works from a distance matrix, it begins by finding a circular ordering on the taxa by an extension of the NJ algorithm. This circular ordering selects the splits and it then uses a least-squares criterion to find non-negative weights for each split. The circular ordering ensures that the splits can be displayed by a planar splits graph. NeighborNets are almost always more resolved than split-decomposition graphs.

Median networks (Bandelt, 1994) are one-connected networks where edges represent a single mutation and nodes represent sequences (extant or inferred). New sequences are inferred using the median hull operation which, given three sequences, can form a potentially new sequence by taking the majority-rule consensus, this is repeated until no new sequences can be derived. These networks are guaranteed to contain all the most-parsimonious trees. Unlike split decomposition or NeighborNet, the resulting split system need not be weakly compatible or circular, and the associated splits graph can be complex. Median networks have proven useful for analysing population data where sequences are densely sampled (Bandelt *et al.*, 1995).

Spectral analysis exploits the relationship between sites patterns and splits (Hendy and Penny, 1993). Any site where only two different states occur corresponds directly to a split of the taxon set. If more than two states are observed at a site then some recoding is required, e.g. RY coding or other projections to two states. The spectral graph displays all splits that correspond to site patterns in the data. Spectral analysis can be used directly on a sequence alignment or combined with Hadamard conjugation and applied to the γ vector (see Figure 16.7), although only with simple models (K3ST and subclasses). As with median networks, the splits graph can be complex, so usually a threshold is specified and splits with fewer than this number of supporting sites are excluded.

Consensus networks and *super-network* methods take trees as input, rather than the initial data (distances or sequences). The trees could arise from bootstrap analysis, Monte Carlo Markov Chain (MCMC) analysis, equally parsimonious trees or trees estimated from different genes. Consensus networks (Holland *et al.*, 2004; 2005b) display all splits that appear in more than some threshold t of the input trees. This threshold controls the complexity of the resulting graph which is k compatible, where k is the first integer less than or equal to $1/t$. A drawback of the consensus network method for gene trees is that all trees must include the same taxa. When the taxa set is overlapping, but not identical, supernetwork methods are required. The first supernetwork method was the Z-closure network of Huson *et al.* (2004), which is based on repeated use of a closure operation that takes two partial splits (i.e. splits on a subset of the complete taxa set) and extends them; the closure rule is applied iteratively until no further splits can be extended, and then all the full splits are displayed in a splits graph. The method is dependent on the order of the partial splits, so typically many orders are used and the union of the full splits found is displayed. A new super-network method was recently presented by Holland *et al.* (2007); it adds all missing taxa into each partial tree according to a scoring function that maximises the number of quartets in agreement with other trees in the input set. If there is enough information to complete the partial trees they are combined in a consensus network.

Other approaches also show conflict within phylogenetic data but are not within the splits-graph framework. Legendre and Makarenkov (2002) developed Reticulograms that use a least-squares criterion with distance data. They begin by constructing a tree and then augment it with additional edges that improve the fit to the data. It adds edges

one at a time by minimising a least-squares fit of the distances on the network to the observed distances. A scoring function that considers both the least-squares fit and the total number of edges decides when to stop adding edges. The method is implemented in the software package *T-Rex*. Statistical parsimony (Templeton *et al.*, 1992) creates a minimum spanning network up to some connection limit. A graph is constructed starting with the taxa as vertices and no edges; edges are added between taxa that differ at only one site, then between taxa that differ at two sites, etc until either the network is connected or the connection limit is reached. The connection limit is set by calculating how likely multiple substitutions are at a site between taxa, a certain distance apart. If the probability of multiple substitutions is greater than 5 % then connections at that distance are disallowed.

Further information on network methods can be found in recent review articles, including Linder and Rieseberg (2004) who discuss the application of network approaches to reconstructing reticulate evolution in plants; Morrison (2005) who discusses the application of networks to problems in population biology; Posada and Crandall (2001) who discuss the gene tree/species tree problem; Huson and Bryant (2006) who describe splits-based methods and their implementation in the software package SplitsTree4; and Huber and Moulton (2005) who give a detailed description of median networks, consensus networks, split decomposition, and NeighborNet.

16.6   SEARCH STRATEGIES

Optimality criterion that are evaluated on a tree have the problem, discussed earlier, of the large number of trees on which the criterion must be evaluated. This is what makes finding the optimal tree NP complete; there is no efficient algorithm for searching all trees. We consider strategies of searching for an optimal tree under three classes, exact methods that (in principle) consider all trees, heuristic searches with limited (or local) searches, and heuristic searches that move through the space of all trees.

16.6.1   Complete or Exact Searches

For small numbers of taxa (for most optimality criteria up to about 10 taxa, 2 027 025 unrooted binary trees) we can search all trees. However, even in this case, it is better to use the first branch-and-bound search strategy outlined below.

16.6.1.1   *Branch-and-bound Methods*

These are standard methods in Operations Research. The 'branch' is a structured search through the space of trees; this is the standard use of 'branch', not the non-standard use in phylogeny for an edge of a tree. The 'bound' is the cost of the optimality criterion on the best tree found (for parsimony, the tree with the fewest changes). If the 'bound' is exceeded, there is no need to search deeper from that point, thus eliminating large numbers of trees from the search.

The two most popular branch-and-bound approaches are adding taxa sequentially (thus forming subtrees with a subset of taxa), and adding sites sequentially ('TurboTree').

The first starts with a subtree of three widely separated taxa; adds a fourth to every position on that subtree, etc. (see Figure 16.6). If a subtree with, say, 10 out of 20 taxa is already longer than the best tree already found then there is no need to add the remaining 10 taxa. For example, if the best tree has 769 changes, and a subtree of 10 taxa has 770, then $17 \times 19 \times 21 \times \ldots \times 35 (> 10^{14})$ trees are excluded because each *must* be worse. It is important to eliminate early unproductive regions of the search space. A second approach, TurboTree, ranks the most frequent patterns first. It then forms a tree by adding these most frequent sites sequentially, but includes the cost of excluding the other sites. Patterns in the data that fit the optimal tree are expected to be more frequent and are added quickly. By having a gain from including an edge in a tree, and the costs of excluding other edges, it finds optimal trees quickly (especially for closest tree).

Branch-and-bound methods are readily implemented with parsimony, minimum evolution, and closest tree optimality criteria. They can be used with either observed data, or data corrected for multiple changes. Branch-and-bound methods are easy to implement, even though optimising them in order to decrease average run time still has scope for improvement. It is unfortunate that branch-and-bound methods have not yet been so useful with maximum (average) likelihood. They can be implemented quite readily but do not give the many orders of magnitude speedup that occurs with the simpler optimality criteria.

If branch-and-bound searches run to completion, they *guarantee* to have found the best trees for that optimality criterion. For MP, the searches should work for up to 20 taxa and improvements will work for up to 30 taxa in some cases, especially for corrected parsimony. But many data sets have 100+ sequences! Approximations are possible to cut running time, but this will sacrifice the certainty of finding the shortest possible tree, and so are really a heuristic searches. By using an artificial bound, a tight limit on the length of the minimal tree can be determined, even if the unrestricted search takes too long to complete. For example, a heuristic search may have found a tree with 679 changes, but the branch-and-bound search does not run to completion. You can repeat the search with a bound of 677 or of 678. The search with bound 677 may run to completion and simply report that there are no trees of that length or shorter. Even if the search trees of length 678 keeps running forever, you now know that the shortest possible tree is of length 678 or 679. Consequently, the known tree of length 679 cannot be more than one mutation longer than the global minimum.

16.6.1.2   *Converging Upper and Lower Bounds (the Minmax Squeeze)*

This approach combines two search methods simultaneously, neither of which can guarantee optimality by itself. They are: a heuristic search for shorter and shorter trees (thus reducing the 'upper bound' – the shortest tree cannot be longer than one you have already found); and partitioning sites into subsets for which a minimal tree can be found (thus increasing the lower bound – the minimum length of the shortest tree). The lower bound comes from partitioning sites into small subsets for which a guaranteed minimal tree can be found. For example, if there are 12 distinct combinations of codes for a subset of sites, and linked with steps no more than a single mutation (see Figure 16.5g), then the minimal length subtree has 11 steps. Every tree on the full data must have at least 11 parsimony steps for this subset of sites. The lower bound is the sum of the lengths of all disjoint subsets of columns. The aim is to find shorter and shorter trees (reducing

the upper bound) and larger and larger lower bounds, until they meet, hence the name MinMax Squeeze.

If (emphasise 'if') the length of the shortest tree (the upper bound) reaches the same value as the lower bound, then no tree with fewer mutations can exist. This combined approach has been used to establish a minimum length tree for 127 complete human mitochondrial genomes (Pierson *et al.*, 2006). In general, this approach is better for population samples because it works most efficiently when the sequences are close together (abundant). However, the general approach will be of growing interest with the increasing use of molecular information such as gene order, presence or absence of inserted elements, major insertions and/or deletions.

16.6.2   Heuristic Searches I, Limited (Local) Searches

Heuristic methods do not guarantee to find a global optimum; but based on previous experience, they will quickly find a 'reasonable' solution. They are sometimes called '*quick and dirty*', and a huge effort has been spent in improving them because they are used in thousands of applications worldwide. We consider them under the two categories, limited or greedy searches, and hill-climbing searches. We discuss the first group under cluster methods, and subtree searches.

16.6.2.1   *Cluster Methods*

There are a many cluster search strategies, including NJ and UPGMA, which build trees sequentially by joining pairs of taxa. They sequentially select internal edges of the tree until the tree is fully resolved. There are several optimality criteria, as well as methods of adjusting the distance values after joining two groups. The searches are essentially 'greedy' in that the programs do not evaluate an optimality criterion over the entire tree; they just accept the tree from the stepwise search. In practice, virtually all cluster methods use distance data, either observed or corrected for inferred multiple changes. In principle, the methods could function on the original sequence data, and with either parsimony or ML.

NJ is the most popular cluster method. It selects two taxa (including previously clustered taxa) that give the biggest increase in 'net divergence' (Section 16.4.2). The pair with the largest internal edge is joined together and this new cluster is treated as a single taxon thereafter (hence the name, neighbor joining). For example, if building a tree of mammals, then humans and chimpanzees are joined on the tree and information about them averaged (more or less). Later in the search, this group may be joined to the gorilla and then this threesome forms a new cluster, etc. The search strategy for t taxa is fast, $O(t^3)$, because it searches just one pathway in building a tree. As mentioned in the previous paragraph, NJ could be implemented with many optimality criteria; it is basically a search strategy. One variation of the method takes in to account variances and covariances of distances, and another allows several trees to be stored during the search, thus considering a larger range of trees.

UPGMA is a search strategy that clusters the most similar pair of taxa, and unites them as a single taxon for subsequent rounds. When a molecular clock is a good description of the data then this strategy is effective, especially if the distances are *not* corrected for multiple changes. UPGMA is no longer consistent when there are differences in the rates

of evolution. An example of data where UPGMA will fail is shown Figure 16.2, which illustrates additive distances and their corresponding tree.

16.6.2.2 Addition Trees

Other limited search algorithms use the original data (not converted to distances). A simple method is to add taxa successively into an enlarging subtree, each time the new sequence is added into its (locally) optimal position. The 'random addition tree' selects taxa randomly, and adds them into the best place on the growing subtree. This can be repeated many times and gives a broader search of tree space. In both versions the fourth taxon is added to all three edges of the single tree for three taxa (see Figure 16.6). The fifth taxon is added to just the five edges of one best tree for four taxa, etc. Overall, $1 + 3 + 5 + 7 + \ldots + (2t - 5)$ trees and subtrees are evaluated, and this is proportional to t^2. But this is a vanishingly small fraction of the $(2t - 5)!!$ unrooted binary trees. A more general but slower version is to add all remaining taxa at each point, and select the taxon that has the most stable position in the subtree. In other words, a taxon that had a unique position two steps longer than the second best tree would be preferred to a taxon whose best position was only one step longer than the second best. The latter would be preferred over a taxon that had two equally good optimal positions.

Both cluster analysis and addition trees are technically 'greedy' algorithms; select the best choice at each step, but do not go back and re-evaluate earlier decisions. These tree(s) can be used as a starting point for more extensive explorations of tree space, as in the next section.

16.6.3 Heuristic Searches II–Hill-climbing and Related Methods

There is a major literature in Operations Research on searching effectively through a complex 'landscape'; there are thousands of major real-world problems that require effective solutions, even if the global optimum is unknown. For phylogenetics the simplest approach selects a tree randomly, then starts making slight changes to it. Every time a better tree is found the new tree is taken and the procedure repeated with changes to this new tree. The process is repeated until no further changes are found that lead to a better tree.

The two aspects of the algorithm are: moving through the space of all trees; and accepting (or not) the new tree. There are many ways of changing a tree slightly; randomly moving a taxon to a new position; collapsing an internal edge and re-expanding it (a commonly used technique for moving through tree space – a crossover or nearest neighbour interchange); and a cut-and-paste move (or tree bisection and reconnection) that breaks one internal edge (branch) of a tree and reconnects the two resulting subtrees in a different way. In the last two methods, the edges can be selected randomly, and if only approach is used to move through the space of all trees then it is a MCMC.

When it comes to searching for improved trees, the simplest strategies are basically 'hill-climbing' methods; either accepting any better tree, or accepting the tree with the largest improvement (steepest ascent). This continues until no further improvement is possible. Such searches are widely used, but are prone to being caught in a *local optimum* – a reasonable solution, but not the *global optimum*. Improved methods can initially accept slightly worse trees for a while – then resume hill climbing. They may eventually end up with a better tree. We will mention just three variants.

1. Simulated annealing (the Metropolis algorithm) has a probability of accepting a worse solution at any point in the search, but gradually reduces this probability as the search continues. Thus it will eventually stop when it has hill-climbed to an optimum.

2. The next variant is the Great Deluge search strategy that accepts any solution (better or worse) above a bound. However, the bound keeps increasing, faster in a good region of the search, more slowly in a poorer region (thus giving more time to escape).

3. The third is a tabu search, which hill climbs until it reaches an optimum, backs away from it by several steps, then resumes hill climbing. By recording how it backed away from the optimum (the tabu list) it avoids immediately re-climbing the last optimum.

The latter two strategies appear especially useful for tree searches. All can be coupled with a MCMC search, with the Metropolis algorithm it is called *Metropolis-coupled*, or *MCMCMC*.

Other approaches use a population of trees at any one time, and can use the best features of each. For example, hitchhiking can be combined with the options above. Several trees are being optimised simultaneously, some are drivers that are hill climbing, other trees are hitch-hikers that accept a better solution their driver finds – if it is also an increase for them. The hitchhikers periodically change drivers, effectively giving recombination possibilities. Charleston (2001) describes some of these search strategies, as well as giving properties of the search space that can be readily measured. Understanding the landscape of the search space with real data is of critical importance, but the problem is in need of much more study.

16.6.4   Quartets and Supertrees

Finally, there are two strategies that both involve building trees from subtrees, one from quartets of taxa, the other with any sized subsets. With just three taxa, there is only one unrooted tree (though a molecular clock version with three taxa is possible). Four taxa (a quartet) is the smallest subset that leads to a decision between unrooted trees; there are three choices to consider. It might be expected (hoped?) that only having three choices leads to a greater chance of getting the correct subtree. If all quartets were chosen correctly, then they could readily be joined into one large tree with all taxa included. If all quartets of t taxa are considered the algorithm is of order t^4. In practice, not all subsets will agree and it is necessary to form the best overall tree. The best-known approach is *quartet puzzling* that evaluates quartets using ML, and then searches for the best tree that agrees with as many subsets as possible. Equal weight needs not be given to all quartets; the relative advantage of the best tree over the second best is one subsidiary criterion that could weight the reliability of quartets.

A second approach in this category of searching tree space does not limit the quartets to four taxa, they can be of any size (Bininda-Emonds *et al.*, 2003). In many studies the sequences available for the taxa may differ. In such a case, trees can be inferred for each data set, and if there is sufficient overlap of taxa between subsets (and the trees from the subsets are in reasonable agreement), then 'supertrees' can be built from the subtrees. The algorithm works efficiently only if there is at least one species common to all the subsets; this taxon effectively acts as an artificial root for building the subtree.

In conclusion, searching through the space of all trees is the limiting feature in inferring trees. It is this aspect that stops any algorithm from being efficient. This is not really a

surprise; in Operations Research there are large numbers of problems for which no efficient algorithm is known, and others for which it is proven that no efficient algorithm can exist. But because there are so many such real-world problems, the range of techniques that are useful has been studied extensively. In this section we have considered branch-and-bound, local searches, greedy algorithms, hill-climbing methods (including those that can escape local maxima), and finally trees assembled from subtrees. The searching of tree space is a major study in its own right.

16.7   OVERVIEW AND CONCLUSIONS

Although this chapter is limited to parsimony and distance methods, it still encompasses a wide field. Our central theme is the need for careful analysis of the problems involved. Many areas of mathematics are involved, not just classical statistics. Trees and graphs are a well-understood part of graph theory. It really matters little if there are two uses of 'parsimony' in phylogenetics, as long as the particular use is made clear, and as long as the assumptions are understood. Real confusion has arisen when it is forgotten that a 'method' for inferring trees consists of three procedures, (1) an optimality criterion, (2) a search strategy, and (3) a procedure for handling multiple changes. This, by itself, has caused more confusion than any other aspect of the theory of tree reconstruction. Similarly, claiming an optimality criterion was, or was not, consistent, misses the point that it is the method of handling multiple changes that leads to consistency, or otherwise. Inferring divergences from millions of years ago is at the cutting edge of modern science and as long as we all learn, the subject will continue to progress.

There is still a range of topics in the quantitative study of trees that need to be understood. We require more information on objective 'tree comparison' methods and their distributions. These help in the evaluation of results, especially in the testing of hypotheses. They do not replace standard resampling (parametric and non-parametric bootstrap and jackknife) methods. Locating the root of the tree is still particularly susceptible to error. This is largely because some well-established groups (such as mammals, birds, and flowering plants) have been separated for many millions of years from their closest living relatives. Outgroups are very distant from the in-group, a classic case for a long-edge attraction problem. If a molecular clock were known to be a realistic description of the data, then 'centering' on the average midpoint of the unrooted tree is another approach.

It is necessary at all times to keep potential sources of error in mind. Perhaps there has been too much emphasis on convenient mathematical models, rather then checking them against biochemical processes. A classic example is assuming that sites are always in the same rate class over the whole tree. This is not expected based on 3D structure of proteins, which keeps evolving. Better estimates of dates of divergence, calibrated against the fossil record, are urgently needed to test a wider range of hypotheses.

Molecular data is now the most important data for phylogeny as it allows phylogenies for all organisms and is providing a large amount of data at relatively low cost. There is still a need for improved methods and in ensuring there is a thorough testing of the results. Phylogeny is now a part of normal science – it does not depend on subjective ideas of the research worker. Darwin's 'theory of descent' can now be established quantitatively, though still with exceptions such as lateral gene transfer. Phylogenetics will continue to

fascinate both theoretical and experimental biologists. This chapter illustrates the wide range of mathematics involved, and the need for careful analysis of the basis of the methods in use.

REFERENCES

Bandelt, H.-J. (1994). Phylogenetic networks. *Verhandl. Naturwiss. Vereins Hamburg* **34**, 51–71.

Bandelt, H.-J. and Dress, A.W.M. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* **1**, 242–252.

Bandelt, H.J., Forster, P., Sykes, B.C. and Richards, M.B. (1995). Mitochondrial portraits of human population using median networks. *Genetics* **141**, 743–753.

Baroni, M., Semple, C. and Steel, M. (2006). Hybrids in real time. *Systematic Biology* **55**, 46–56.

Bininda-Emonds, O.R.P., Gittleman, J.L. and Steel, M.A. (2003). The (super) tree of life: procedures, problems and prospects. *Annual Review of Ecology and Systematics* **33**, 265–289.

Bordewich, M. and Semple, C. (2007). Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Applied Mathematics* **155**, 914–928.

Bromham, L.D., Hendy, M.D., Penny, D. and Rambaut, A. (2000). The power of relative rates tests depends on the data. *Journal of Molecular Evolution* **50**, 296–301.

Bryant, D. and Moulton, V. (2004). NeighborNet: an agglomerative algorithm for the construction of planar phylogenetic networks. *Molecular Biology and Evolution* **21**, 255–265.

Charleston, M.A. (2001). Hitch-hiking: a parallel heuristic search strategy, applied to the phylogeny problem. *Journal of Computational Biology* **8**, 79–91.

Chor, M.B., Holland, B.R., Penny, D. and Hendy, M.D. (2000). Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Molecular Biology and Evolution* **17**, 1529–1541.

Collins, L.J., Moulton, V. and Penny, D. (2000). Use of RNA secondary structure for evolutionary relationships: the case of RNase P and RNase MRP. *Journal of Molecular Evolution* **51**, 194–204.

Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Fernandez-Baca, D. and Lagergren, J. (2003). A polynomial-time algorithm for near-perfect phylogeny. *SIAM Journal on Computing* **32**, 1115–1127.

Gusfield, D., Eddhu, S., Langley, C. (2003). Efficient reconstruction of phylogenetic networks with constrained recombination. Bioinformatics conference. In *CSB 2003. Proceedings of the 2003 IEEE*, Stanford University, Stanford CA, pp. 363–374.

Hallett, M.T. and Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. In *Proceedings of the Fifth Annual International Conference on Computational Biology (RECOMB01)*, Montreal, pp. 149–156.

Hein, J. (1990). Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences* **98**, 185–200.

Hendy, M.D. and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Systematic Zoology* **38**, 297–309.

Hendy, M.D. and Penny, D. (1993). Spectral analysis of phylogenetic data. *Journal of Classification* **10**, 5–24.

Hendy, M.D., Penny, D. and Steel, M.A. (1994). Discrete Fourier spectral analysis of evolution. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 3339–3343.

Henz, S.R., Huson, D.H., Auch, A.F., Nieselt-Struwe, K. and Schuster, S.C. (2005). Whole-genome prokaryote phylogeny. *Bioinformatics* **21**, 2329–2335.

Holland, B.R., Conner, G., Huber, K. and Moulton, V. (2007). Imputing supertrees and supernetworks using quartets. *Systematic Biology*, **56**, 57–67.

Holland, B.R., Huber, K.T., Moulton, D. and Lockhart, P.J. (2004). Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular Biology and Evolution* **21**, 1459–1461.

Holland, B.R., Huber, K.T., Penny, D. and Moulton, V. (2005a). The MinMax Squeeze: guaranteeing a minimal tree for population data. *Molecular Biology and Evolution* **22**, 235–242.

Holland, B.R., Delsuc, F. and Moulton, V. (2005b). Visualizing conflicting evolutionary hypotheses in large collections of trees: using consensus networks to study the origins of placentals and hexapods. *Molecular Biology and Evolution* **22**, 66–76.

Holland, B.R., Penny, D. and Hendy, M.D. (2003). Outgroup misplacement and phylogenetic inaccuracy under a molecular clock – a simulation study. *Systematic Biology* **52**, 229–238.

Huber, K.T., Langton, M., Penny, D., Moulton, V. and Hendy, M.D. (2002). Spectronet: a package for computing spectra and median networks. *Applied Bioinformatics* **1**, 2041–2059.

Huber, K.T. and Moulton, V. (2005). Phylogenetic networks. In *Mathematics of Evolution and Phylogeny*, O. Gascuel, ed. Oxford University Press, pp 178–204.

Huber, K.T., Oxelman, B., Lott, M. and Moulton, V. (2006). Reconstructing the evolutionary history of polyploids from multilabeled trees. *Molecular Biology and Evolution* **23**, 1784–1791.

Huson, D.H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**, 254–267.

Huson, D.H., Dezulian, T., Klöpper, T. and Steel, M.A. (2004). Phylogenetic super-networks from partial trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **1**, 151–158.

Huson, D.H., Klöpper, T., Lockhart, P.J. and Steel, M.A. (2005). Reconstruction of reticulate networks from gene trees. In *Research in Computational Biology*, *Lecture Notes in Computer Science, Vol. 3500*, S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner and M. Waterman, eds. Springer-Verlag, Berlin, pp. 233–249.

Kriegs, J.O., Churakov, G., Kiefmann, M., Jordan, U., Brosius, J. and Schmitz, J. (2006). Retroposed elements as archives for the evolutionary history of placental mammals. *PLoS Biology* **4**, e91.

Legendre, P. and Makarenkov, V. (2002). Reconstruction of biogeographic and evolutionary networks using reticulograms. *Systematic Biology* **51**, 199–216.

Linder, C.R. and Rieseberg, L.H. (2004). Reconstructing patterns of reticulate evolution in plants. *American Journal of Botany* **91**, 1700–1708.

Lockhart, P.J., Steel, M.A., Hendy, M.D. and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution* **11**, 605–612.

Maddison, W.P. (1997). Gene trees in species trees. *Systematic Biology* **46**, 523–536.

Morrison, D. (2005). Networks in phylogenetic analysis: new tools for population biology. *International Journal for Parasitology* **35**, 567–582.

Nakhleh, L., Warnow, T., Linder, C.R. and St. John, K. (2005). Reconstructing reticulate evolution in species – theory and practice. *Journal of Computational Biology* **12**, 796–811.

Nei, M. and Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford.

Page, R.D.M. and Holmes, E.C. (1998). *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, Oxford.

Penny, D. (1982). Towards a basis for classification: the incompleteness of distance measures, incompatibility analysis and phenetic classification. *Journal of Theoretical Biology* **96**, 129–142.

Penny, D., Hendy, M.D., Lockhart, P.J. and Steel, M.A. (1996). Corrected parsimony, minimum evolution and Hadamard conjugations. *Systematic Biology* **45**, 593–603.

Penny, D., Hendy, M.D. and Steel, M.A. (1992). Progress with evolutionary trees. *Trends in Ecology and Evolution* **7**, 73–79.

Pierson, M.J., Martinez-Arias, R., Holland, B.R., Gemmell, N.J., Hurles, M.E. and Penny, D. (2006). Deciphering past human population movements in Oceania: provably optimal trees of 127 mtDNA genomes. *Molecular Biology and Evolution* **23**, 1966–1975.

Posada, D. and Crandall, K. (2001). Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology and Evolution* **16**, 37–45.

Semple, C. and Steel, M.A. (2003). *Phylogenetics*. Oxford University Press, Oxford.

Song, Y.S. and Hein, J. (2005). Constructing minimal ancestral recombination graphs. *Journal of Computational Biology* **12**, 147–169.

Steel, M.A. and Penny, D. (2000). Parsimony, likelihood and the role of models in molecular phylogenetics. *Molecular Biology and Evolution* **17**, 839–850.

Steel, M.A. and Penny, D. (2004). Two further links between MP and ML under the Poisson model. *Applied Mathematics Letters* **17**, 785–790.

Steel, M.A. and Penny, D. (2005). Maximum parsimony and the phylogenetic information in multistate characters. In *Parsimony, Phylogeny and Genomics*, V. Albert, ed. Oxford University Press, pp 163–178.

Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. (1996). Phylogenetic inference. In *Molecular Systematics*, 2nd edition, D.M. Hillis, C. Moritz and B.K. Mable, eds. Sinauer Associates, pp. 407–514.

Tajima, F. (1993). Unbiased estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution* **10**, 677–688.

Templeton, A.R., Crandall, K.A. and Sing, C.F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**, 619–633.

Tuffley, C. and Steel, M.A. (1997). Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences* **147**, 63–91.

Waddell, P.J., Penny, D., Hendy, M.D. and Arnold, G.C. (1994). The sampling distributions and covariance matrix of phylogenetic spectra. *Molecular Biology and Evolution* **11**, 630–642.

Waddell, P.J., Penny, D. and Moore, T. (1997). Hadamard conjugations and modeling sequence evolution with unequal rates across sites. *Molecular Phylogenetics and Evolution* **8**, 33–50.

Wang, L., Zhang, K. and Zhang, L. (2001). Perfect phylogenetic networks with recombination. *Journal of Computational Biology* **8**, 69–78.

<div style="text-align: right">

17

</div>

Evolutionary Quantitative Genetics

B. Walsh

Department of Ecology and Evolutionary Biology, Department of Plant Sciences, Department of Molecular and Cellular Biology, University of Arizona, Tuscon, AZ, USA

Evolutionary quantitative genetics is the study of how complex traits evolve over time. While this field builds on traditional concepts from quantitative genetics widely used by applied breeders and human geneticists (in particular, the inheritance of complex traits), its unique feature is in examining the role of natural selection in changing the population distribution of a complex trait over time. Our review focuses on this role of selection, starting with response under the standard infinitesimal model, in which trait variation is determined by a very large number of loci, each of small effect. We then turn to issues of measuring fitness (and hence natural selection) on both univariate and multivariate traits. We conclude by examining models that treat fitness itself as a complex trait.

17.1   INTRODUCTION

Evolutionary quantitative genetics is a vast field, ranging from population genetics at one extreme to development and functional ecology at the other. The goal of this field is a deeper understanding of not only the genetics and inheritance of complex traits in nature (traits whose variation is due to both genetic and environmental factors) but also the nature of the evolutionary forces that shape character variation and change in natural populations. Given the limitations of space, we have chosen to focus this review on evolutionary change, in particular estimation of the nature of selection and aspects of the evolutionary response to selection. A detailed treatment of the inheritance of complex traits in nature can be found in Lynch and Walsh (1998). Roff (1997) provides a good overview of the entire field, while the reader seeking detailed treatments of specific issues should consult Bulmer (1980) and Walsh and Lynch (2003). Bürger (2000) provides an excellent (but, in some places, highly technical) review of the interface between population and quantitative genetics.

17.1.1 Resemblances, Variances, and Breeding Values

A few summary remarks on quantitative genetics are in order to give the reader the appropriate background. A good introduction is Falconer and Mackay (1996), while a very detailed treatment is given by Lynch and Walsh (1998). **Chapter 18**, **Chapter 19**, **Chapter 20**, and **Chapter 21** in this Handbook also examine some of these issues in greater detail.

17.1.1.1 Fisher's Genetic Decomposition

Although Yule (1902) can be considered the first paper in quantitative genetics, the genesis of modern quantitative genetics is the classic paper by Fisher (1918). Fisher made a number of key insights, notably that with sexual reproduction only a specific fraction of an individual's genotypic value (the mean value of that genotype when averaged over the distribution of environments) is passed on to its offspring, and that we can estimate the variances associated with these various components by looking at the phenotypic covariances between appropriate relatives.

Fisher decomposed the observed phenotypic value z of an individual into a genotypic G and environment E value,

$$z = G + E. \tag{17.1a}$$

The genotypic value can be thought of as the average phenotypic value if the individual was cloned and replicated over the universe of environments it is likely to experience. As mentioned, Fisher noted that a parent does not pass along its entire G value to its offspring, as for any given locus a parent passes along only one of its two alleles to a particular offspring. Thus, the genotypic value can be further decomposed into a component passed on to the offspring A (the *additive genetic value*) and a nonadditive component, which includes the dominance D and any epistatic effects. We ignore the effects of epistasis here, whose complex features are extensively examined by Lynch and Walsh (1998) and Walsh and Lynch (2003). Thus,

$$z = \mu + A + D + E, \tag{17.1b}$$

where μ is the population mean (by construction, the mean values of A, D, and E equal zero). The additive values A are often referred to as *breeding values*, as the average value of an offspring is just the average breeding value of its parents. Hence, the expected value of an offspring is $\mu + (A_f + A_m)/2$, where A_f is the paternal (or sire) and A_m the maternal (or dam) breeding values.

17.1.1.2 Additive Genetic Variances and Covariances

When the phenotype can be decomposed as in (17.1b), Fisher showed that the phenotypic covariance between parent and offspring is half the population variance of breeding values, σ_A^2:

$$\sigma^2(z_p, z_0) = \frac{\sigma_A^2}{2}. \tag{17.2a}$$

Thus, twice the parent–offspring phenotypic covariance provides an estimate of σ_A^2, which is called the *additive genetic variance*. There are a number of potential complications in blindly applying (17.2a), such as shared environmental effects, and maternal effects when considering the mother–offspring covariance. See Falconer and Mackay (1996) and Lynch and Walsh (1998) for details on handling these complications. **Chapter 19** considers how to use not just parent–offspring information, but also at the same time all the information in general pedigrees in estimating the additive genetic variances.

The covariance in breeding values between two traits (say x and y), $\sigma_A(x, y)$, is needed for considering evolution of multiple traits. These *additive genetic covariances* can also be estimated from parent–offspring regressions, as the phenotypic covariance between one trait in the parent (p) and the other trait in the offspring (o) estimates half the additive genetic covariance of the traits, e.g.,

$$\sigma^2(x_p, y_o) = \sigma^2(y_p, x_o) = \frac{\sigma_A(x, y)}{2}. \tag{17.2b}$$

Two different mechanisms can generate a (within-individual) covariance in the breeding values of two traits: linkage and *pleiotropy*. Linked loci show an excess of parental gametes, creating a correlation between alleles at these loci. (If a parent has alleles A and B on one chromosome and a and b on the other homologous chromosome, this parent will produce more AB and ab parental gametes than Ab and aB recombinant gametes.) In this case, even if each linked locus affects only a single trait, the correlation between alleles (linkage disequilibrium) generates a short-term correlation between the breeding values of the traits. Over time these associations decay through recombination, randomizing allelic associations across loci. Alternatively, with pleiotropy a single locus can affect multiple traits. Covariances due to pleiotropic loci are stable over time.

17.1.2 Single Trait Parent–Offspring Regressions

17.1.2.1 Phenotypic and Genetic Regressions

Most of the theory of selection response in quantitative genetics assumes linear parent–offspring regressions that have homoscedastic residuals (i.e., the variance about the expected value is independent of the values of the parents). The simplest version considers the phenotypic value of the midparent, $z_{mp} = (z_m + z_f)/2$, with the offspring value z_o predicted as

$$z_o = \mu + b(z_{mp} - \mu), \tag{17.3a}$$

where μ is the population mean. Recalling that the slope b of the best linear regression of y on x, $y = a + bx$, is given by

$$b = \frac{\sigma(x, y)}{\sigma_x^2},$$

the slope of the midparent–offspring regression becomes

$$b = \frac{\sigma(z_o, z_{mp})}{\sigma^2(z_{mp})} = \frac{\sigma(z_o, z_m)/2 + \sigma(z_o, z_m)/2}{\sigma^2(x_m/2 + x_f/2)} = \frac{2\sigma_A^2/4}{2\sigma_z^2/4} = \frac{\sigma_A^2}{\sigma_z^2}. \tag{17.3b}$$

This ratio of the additive genetic variance to the phenotypic variance is usually denoted h^2 and is called the (narrow-sense) *heritability*. Thus, the midparent–offspring regression is given by

$$z_o = \mu + h^2(z_{mp} - \mu) = \mu + h^2 \left(\frac{z_m + z_f}{2} - \mu \right);$$ (17.3c)

this is the expected value for an offspring. The actual value for any particular offspring is given by

$$z_o = \mu + h^2(z_{mp} - \mu) + e,$$ (17.4a)

where the residual e has mean value zero and variance

$$\sigma_e^2 = \left(1 - \frac{h^4}{2} \right) \sigma_z^2.$$ (17.4b)

17.1.2.2 *Selection Differentials and the Breeders' Equation*

The parent–offspring regression allows us to predict the response to selection. Suppose the mean of parents that reproduce (μ_*) is different from the population mean before selection μ. Define the *directional selection differential* as $S = \mu_* - \mu$ (S is often simply called the selection differential). From (17.3c), the difference R between the offspring mean of these parents and the original mean of the population before selection is

$$R = h^2 S.$$ (17.5)

This is the *breeders' equation* which transforms the between-generation change in the mean (the selection response R) into the within-generation change in the mean (S). Notice that strong selection does not necessarily imply a large response. If the heritability of a trait is very low (as occurs with many life-history traits), then even very strong selection results in very little (if any) response. Hence, selection (nonzero S) does not necessarily imply evolution (nonzero R).

17.1.3 Multiple Trait Parent–Offspring Regressions

The (single-trait) parent–offspring regression can be generalized to a (column) vector of n trait values, $\mathbf{z} = (z_1, \ldots, z_n)^{\mathrm{T}}$. Letting $\mathbf{z}_o$ be the vector of trait values in the offspring, $\mathbf{z}_{mp}$ the vector of midparent-trait values (i.e., the ith element is just $(z_{f,i} + z_{m,i})/2$), and μ be the vector of population means, then the parent–offspring regression for multiple traits can be written as having expected value

$$\mathbf{z}_o = \mu + \mathbf{H}(\mathbf{z}_{mp} - \mu)$$ (17.6)

where the (i,j)th element in the matrix $\mathbf{H}$ is the weight associated with the value of trait i in the offspring and trait j in the midparent.

17.1.3.1 *The Genetic and Phenotypic Covariance Matrices*

In order to further decompose $\mathbf{H}$ into workable components, we need to define the phenotypic covariance matrix $\mathbf{P}$ whose (i,j)th element is the phenotypic covariance

between traits i and j. Note that $\mathbf{P}$ is symmetric, with the diagonal elements corresponding to the phenotypic variances and off-diagonal elements corresponding to the phenotypic covariances. In a similar manner, we can define the symmetric matrix $\mathbf{G}$, whose (i, j)th element is the additive genetic covariance (covariance in breeding values) between traits i and j. Using similar logic to that leading to (17.3b) (the slope of the single-trait regression), it can be shown that

$$\mathbf{H} = \mathbf{GP}^{-1}, \tag{17.7a}$$

giving the multitrait parent–offspring regression as

$$\mathbf{z}_0 = \boldsymbol{\mu} + \mathbf{GP}^{-1}(\mathbf{z}_{mp} - \boldsymbol{\mu}). \tag{17.7b}$$

17.1.3.2 The Multivariate Breeders' Equation

Letting $\mathbf{R}$ denote the column vector of responses (so that the ith element is the between-generation change in the mean of trait i), and $\mathbf{S}$ the vector of selection differentials, then (17.7b) allows us to generalize the breeders' equation to multiple traits,

$$\mathbf{R} = \mathbf{HS} = \mathbf{GP}^{-1}\mathbf{S}; \tag{17.8}$$

this equation forms the basis for discussions about selection on multiple characters (Section 17.6).

17.2 SELECTION RESPONSE UNDER THE INFINITESIMAL MODEL

17.2.1 The Infinitesimal Model

The breeders' equation predicts the change in mean following a single generation of selection from an unselected base population. The only assumption required is that the parent–offspring regression is linear, but the question remains as to when this linearity holds. Further, the breeders' equation focuses only on the change in mean, leaving open the question of what happens to the variance. The latter concern is of special relevance in evolutionary quantitative genetics, as stabilizing selection (which reduces the variance without necessarily any change in the mean) is thought to be common in natural populations. The infinitesimal model, introduced by Fisher (1918), provides a framework to address these issues.

17.2.1.1 Allele Frequency Changes under the Infinitesimal Model

Under the infinitesimal model, a character is determined by an infinite number of unlinked and nonepistatic loci, each with an infinitesimal effect. A key feature of the infinitesimal model is that while allele frequencies are essentially unchanged by selection, large changes in the mean can still occur by summing infinitesimal allele frequency changes over a large number of loci. To see this feature, consider a character determined by n completely additive and interchangeable loci, each with two alleles, Q and q (at frequencies p and $1 - p$), where the genotypes QQ, Qq, and qq contribute $2a, a$, and 0 (respectively) to

the genotypic value. The resulting mean is $2nap$ and the additive variance (ignoring the contribution from gametic-phase disequilibrium) is $\sigma_A^2 = 2na^2 p(1 - p)$. For σ_A^2 to remain bounded as the number of loci increase, a must be of order $n^{-1/2}$. The change in mean due to a single generation of selection is easily found to be $\Delta\mu = 2na\Delta p$. Assuming the frequency of Q changes by the same amount at each locus, $\Delta p = \Delta\mu/(2na)$. Since a is of order $n^{-1/2}$, Δp is of order $1/(n \cdot n^{-1/2}) = n^{-1/2}$, approaching zero as the number of loci becomes infinite. Thus the infinitesimal model allows for arbitrary changes in the mean with (essentially) no change in the allele frequencies at underlying loci.

What effect does this very small amount of allele frequency change have on the variance? Letting $p' = p + \Delta p$ denote the frequency after selection, the change in the additive genic variance is

$$\Delta\sigma_A^2 = 2na^2 p'(1 - p') - 2na^2 p(1 - p)$$
$$= 2na^2 \Delta p(1 - 2p - \Delta p)$$
$$\approx a(1 - 2p)\Delta\mu.$$

Since a is of order $n^{-1/2}$, the change in variance due to changes in allele frequencies is roughly $1/\sqrt{n}$ the change in mean. Thus, with a large number of loci, very large changes in the mean can occur without any significant change in the variance. In the limit of an infinite number of loci, there is no change in the genic variance ($\Delta\sigma_A^2 = 0$), while arbitrary changes in the mean can occur.

17.2.1.2 *Linearity of Parent–Offspring Regressions under the Infinitesimal Model*

Under the infinitesimal model, genotypic values (G) are normally (or multivariate normally if G is a vector) distributed before selection (Bulmer, 1971; 1976). Assuming environmental values (E) are also normal, then so is the phenotype z (as $z = G + E$) and the joint distribution of phenotypic and genotypic values is multivariate normal. In this case, the regression of offspring phenotypic value z_o on parental phenotypes (z_f and z_m for the father and mother's values) is linear and homoscedastic.

17.2.2 Changes in Variances

As mentioned, under the infinitesimal model (in an infinite population) there is essentially no change in the genetic variances caused by allele frequency change. Changes in allele frequencies, however, are not the only route by which selection can change the variance (and other moments) of the genotypic distribution. Selection also creates associations (covariances) between alleles at different loci through the generation of gametic-phase (or linkage) disequilibrium, and such covariances can have a significant effect on the genetic variance. Disequilibrium can also change higher-order moments of the genotypic distribution, driving it away from normality and hence potentially causing parent–offspring regressions to deviate from linearity.

17.2.2.1 *The Additive Variance under Disequilibrium*

The additive genetic variance σ_A^2 in the presence of linkage disequilibrium can be written as

$$\sigma_A^2 = \sigma_a^2 + d \tag{17.9}$$

where σ_a^2 is the additive genetic variance in the absence of disequilibrium and d the disequilibrium contribution. To formally define σ_a^2 and d, let $a_1^{(k)}$ and $a_2^{(k)}$ denote average effects of the two alleles at locus k from a random individual. Since σ_A^2 is the variance of the sum of average effects over all loci,

$$\sigma^2\left(\sum_{k=1}^{n}\left(a_1^{(k)} + a_2^{(k)}\right)\right) = 2\sum_{k=1}^{n}\sigma^2(a^{(k)}) + 4\sum_{k<j}^{n}\sigma(a^{(j)}, a^{(k)}) \qquad (17.10a)$$

$$= 2\sum_{k=1}^{n} C_{kk} + 4\sum_{k<j}^{n} C_{jk}, \qquad (17.10b)$$

where n is the number of loci and $C_{jk} = \sigma(a^{(j)}, a^{(k)})$ is the covariance between allelic effects at locus j and k. Thus $\sigma_a^2 = 2\sum C_{kk}$ is the additive variance in the absence of gametic-phase disequilibrium and the disequilibrium contribution $d = 4\sum_{j<k} C_{kj}$ is the covariance between allelic effects at different loci. The component of the additive genetic variance that is unaltered by changes in gametic-phase disequilibrium, σ_a^2, is often referred to as the *additive genic variance* (or simply the *genic variance*) to distinguish it from the additive *genetic* variance σ_A^2.

Under the infinitesimal model, selection does not change the C_{kk} (as this requires changes in the allele frequencies), and hence does not alter σ_a^2. However, selection does generate correlations between loci ($C_{ik} \neq 0$), and this can result in significant changes in the overall additive variance σ_A^2. Changes in the covariances C_{ij} between loci i and j (for $i \neq j$) are roughly of order n^{-2} (Bulmer, 1980; Turelli and Barton, 1990). Since there are n^2 terms contributing to d, the total disequilibrium is of order one ($n^2 \cdot n^{-2}$) and does not necessarily approach zero as the number of loci becomes infinite. The same reasoning holds for changes in the higher-order moments, which are caused by higher-order associations between groups of loci, and can also be nontrivial (Turelli and Barton, 1990).

17.2.2.2 Dynamics of Disequilibrium

Recall that the total genetic variance is the sum of the additive and dominance variances (plus any epistatic variances), so that $\sigma_G^2 = \sigma_A^2 + \sigma_D^2$. Under the infinitesimal model, disequilibrium changes the additive genetic, but not the dominance, variance (Walsh and Lynch, 2003). Hence, the phenotypic variance in generation t of selection is

$$\sigma_z^2(t) = \sigma_E^2 + \sigma_D^2 + \sigma_A^2(t) = \sigma_z^2 + d(t), \qquad (17.11a)$$

where $\sigma_z^2 = \sigma_z^2(0)$ is the phenotypic variance before selection in the initial (unselected) base population. The resulting heritability in generation t becomes

$$h^2(t) = \frac{\sigma_A^2(t)}{\sigma_z^2(t)} = \frac{\sigma_a^2 + d(t)}{\sigma_z^2 + d(t)}. \qquad (17.11b)$$

Assuming that the parent–offspring regression remains linear, the selection response in generation t becomes $R(t) = h^2(t)S(t)$. Assuming unlinked loci, Bulmer (1971) showed

that the change in disequilibrium is given by

$$d(t+1) = \frac{d(t)}{2} + \frac{h^4(t)}{2}\left(\sigma_{z^*(t)}^2 - \sigma_{z(t)}^2\right), \tag{17.12}$$

where $\sigma_{z^*(t)}^2 - \sigma_{z(t)}^2$ is the within-generation change in the phenotypic variance. The first term represents the removal of disequilibrium by recombination, while the second is the generation of disequilibrium by selection. Note that (17.12) is the variance analog of the breeders' equation, relating the between- $(d(t+1) - d(t))$ and within-generation $(\sigma_{z^*(t)}^2 - \sigma_{z(t)}^2)$ changes in the variance.

Starting with an unselected base population (where $d(0) = 0$), iterating (17.12) gives the disequilibrium (and hence the heritability, phenotypic variance, and response) in any desired generation. Under directional selection (selection only on the mean), most of the disequilibrium is generated in the first three to five generations, after which d is very close to its equilibrium value $\tilde{d}$. Once selection stops, the current disequilibrium is halved each generation, with the additive variation rapidly returning to its value before selection. This occurs because under the infinitesimal model, all changes in variances are due to disequilibrium, which decays via recombination in the absence of selection.

An important point regarding (17.12) is that there can be no change in the mean ($S = 0$) and yet selection can still act on the variance. Stabilizing selection, which removes extreme individuals and reduces the variance, is widely thought to be widespread in nature. The within-generation reduction in variance generates negative disequilibrium, which in turn reduces the additive variance. Under the infinitesimal model, once stabilizing selection stops, the variance returns (increases) to its initial value after a few generations. Likewise, under disruptive selection, individuals of intermediate value are selected against, increasing the variance. This generates positive d, increasing the additive variance. Again, when selection is stopped, the variance decays back to its initial (pre-selection) value.

Iteration of (17.12) allows any form of selection to be analyzed. For modeling purposes, it is often assumed that the within-generation change in the phenotypic variance can be written as

$$\sigma_{z^*(t)}^2 = (1 - \kappa)\sigma_{z(t)}^2, \tag{17.13}$$

where κ is a constant that does not change over time. Noting that $\sigma_{z(t)}^2 = \sigma_A^2(t)/h^2(t)$ and substituting (17.13) into (17.12) recovers the result of Bulmer (1974),

$$d(t+1) = \frac{d(t)}{2} - \frac{\kappa}{2}h^2(t)\sigma_A^2(t) = \frac{d(t)}{2} - \frac{\kappa}{2}\frac{[\sigma_a^2 + d(t)]^2}{\sigma_z^2 + d(t)}. \tag{17.14}$$

Again, simple iteration allows one to compute the variance in any generation.

17.2.2.3 Equilibrium Variances

The equilibrium variances, and hence the asymptotic rate of response under directional selection or the equilibrium variance under stabilizing or disruptive selection, are easily obtained. At equilibrium, (17.14) implies

$$\tilde{d} = -\kappa\tilde{h}^2\tilde{\sigma}_A^2 = -\kappa\frac{(\sigma_a^2 + \tilde{d})^2}{\sigma_z^2 + \tilde{d}}. \tag{17.15a}$$

Solving for $\tilde{d}$ gives the equilibrium additive genetic variance as

$$\tilde{\sigma}_A^2 = \sigma_z^2 \theta, \quad \text{where } \theta = \frac{2h^2 - 1 + \sqrt{1 + 4h^2(1 - h^2)\kappa}}{2(1 + \kappa)}, \tag{17.15b}$$

and the resulting heritability at equilibrium is

$$\tilde{h}^2 = \frac{\tilde{\sigma}_A^2}{\tilde{\sigma}_z^2} = \frac{\tilde{\sigma}_A^2}{\sigma_z^2 + (\tilde{\sigma}_A^2 - \sigma_A^2)} = \frac{\theta}{1 + \theta - h^2}. \tag{17.15c}$$

The simple picture that emerges from directional selection under the infinitesimal model is that while the heritability may decrease slightly from its initial value (due to generation of linkage disequilibrium), the response to selection continues without limit. Biological reality, of course, places limits on the character values that can be obtained by selection response. For example, selection may move the mean phenotype to a region on the fitness surface where stabilizing selection dominates (see Section 17.4). Likewise, selection on one character may be opposed by selective constraints imposed by other characters also under selection, and a limit can be reached despite significant additive variance and a nonzero S in the character being followed (Section 17.6).

17.2.3   The Roles of Drift and Mutation under the Infinitesimal Model

While selection does not change allele frequencies under the infinitesimal model, genetic drift (due to finite population size) and mutation can play very important roles in shaping the genetic variance. Inclusion of these two forces allows the infinitesimal model to be modified into a much more biologically realistic description of selection response.

17.2.3.1   Drift

Assuming no dominance or epistasis, with drift (assuming an effective population size of N_e) the expected genic variation σ_a^2 declines each generation from its initial value $\sigma_a^2(0)$, with the genic variance in generation t being

$$\sigma_a^2(t) = \sigma_a^2(0) \left(1 - \frac{1}{2N_e}\right)^t. \tag{17.16}$$

If dominance (or epistasis) is present, the additive variation can actually increase (at least while the level of inbreeding is moderate) before it ultimately declines to zero (Walsh and Lynch, 2003). Under the infinitesimal model, selection is so weak on any given underlying locus that the allele frequency dynamics are entirely governed by drift, and the change in the genic variance is given by (17.16).

Keightley and Hill (1987) show that, under drift, the change in the amount of disequilibrium is given by

$$\Delta d(t) = -\frac{d(t)}{2}\left(1 + \frac{1}{N_e}\right) - \frac{1}{2}\left(1 - \frac{1}{N_e}\right)\kappa h^2(t)\sigma_A^2(t). \tag{17.17}$$

Joint iteration of (17.16) and (17.17) allows for modeling of response with a finite population size. During the course of selection, genetic drift removes the initial variation,

driving the genic (and hence additive genetic) variance to zero. With directional selection a limit is reached, as the response stops as the heritability approaches zero. Robertson (1960) showed that the total change in the mean in this case is approximately $2N_e R(1)$, with $R(1)$ the response in the first generation. It takes approximately $1.4N_e$ generations for half this total response to occur.

17.2.3.2   Mutation

Balancing the removal of genetic variation by genetic drift is the introduction of new variation by mutation. The mutation variance σ_m^2 is defined as the per-generation contribution by mutation to the additive variance, with the average value of σ_m^2/σ_E^2 typically around 0.005 (Lynch and Walsh, 1998). When both mutation and drift occur, the additive variance eventually attains an equilibrium value of

$$\tilde{\sigma}_A^2 = 2N_e\sigma_m^2. \tag{17.18}$$

Assuming the infinitesimal model, completely additive loci, a base-population additive variance of $\sigma_A^2(0)$ and ignoring any effects of gametic-phase disequilibrium, the expected additive genetic variance at generation t is given by

$$\sigma_A^2(t) \simeq 2N_e\sigma_m^2 + [\sigma_A^2(0) - 2N_e\sigma_m^2]\exp(-t/2N_e). \tag{17.19a}$$

Setting $\sigma_A^2(0) = 0$ gives the additive variance contributed entirely from mutation as

$$\sigma_{A,m}^2(t) \simeq 2N_e\sigma_m^2[1 - \exp(-t/2N_e)]. \tag{17.19b}$$

Hence, the rate of response at generation t from mutational input is

$$R_m(t) = i\frac{\sigma_{A,m}^2(t)}{\sigma_z} \simeq 2N_e i\frac{\sigma_m^2}{\sigma_z}[1 - \exp(-t/2N_e)], \tag{17.19c}$$

with $i = S/\sigma_z$ (the selection intensity, which is discussed below). For $t \gg 2N_e$, the per-generation response approaches an asymptotic limit of

$$\tilde{R}_m = 2N_e i\frac{\sigma_m^2}{\sigma_z} = i\frac{\tilde{\sigma}_A^2}{\sigma_z}. \tag{17.19d}$$

A full treatment allowing for mutation, drift, and disequilibrium under the infinitesimal model is given by simply adding a σ_m^2 term to (17.16) and then jointly iterating (17.16) and (17.17).

17.3   FITNESS

Predicting the selection response under the infinitesimal model requires knowledge of the change in mean and variance after an episode of selection. In an artificial selection program, the breeder or experimentalist can not only measure these components, but also largely set their values. In nature, the currency of selection is fitness, and the change

in the phenotypic distribution is computed by first weighting individuals by their fitness values. Discussion of selection response in nature thus starts by trying to assign fitness values to particular character states. This is the linking step that allows use of the above machinery for prediction of selection response.

17.3.1　Individual Fitness

Loosely stated, the *lifetime* (or *total*) *fitness* of an individual is the number of descendants it leaves at the start of the next generation. When measuring the total fitness of an individual, care must be taken not to cross generations or to overlook any stage of the life cycle in which selection acts. To accommodate these concerns, lifetime fitness is defined as the total number of zygotes (newly fertilized gametes) that an individual produces. Measuring total fitness from any other starting point in the life cycle (e.g., from adults in one generation to adults in the subsequent generation) can result in a very distorted picture of true fitness of particular phenotypes (Prout, 1965; 1969). If generations are crossed, measures of selection on a particular parental phenotype in reality are averages over both parental and offspring phenotypes, which may differ considerably.

Systems for measuring lifetime fitness have been especially well developed for laboratory populations of *Drosophila* (reviewed by Sved, (1989)). Measurements of lifetime fitness in field situations are more difficult and (not surprisingly) are rarely made. Attention instead is usually focused on particular episodes of selection or particular phases of the life cycle. Fitness components for each episode of selection are defined to be multiplicative. For example, lifetime fitness can be partitioned as (probability of surviving to reproductive age)·(number of mates)·(number of zygotes per mating). Number of mates is a measure of *sexual selection*, while the viability and fertility components measure natural selection. A commonly measured fitness component is *reproductive success*, the number of offspring per adult, which confounds natural (fertility) and sexual selection (in males, the number of matings per adult). Clutton-Brock (1988) reviews estimates of reproductive success from natural populations.

Fitness components can themselves be further decomposed. For example, fertility in plants might be decomposed as (seeds per plant) = (number of stems per plant)·(number of inflorescences per stem)·(average number of seed capsules per inflorescence)·(average number of seeds per capsule). Such a decomposition allows the investigator to ask questions of the form: do plants differ in number of seeds mainly because some plants have more stems, or more flowers per stem, or are there tradeoffs between these?

Estimates of fitness can be obtained from either *longitudinal* or *cross-sectional* studies. A longitudinal study follows a cohort of individuals over time, while a cross-sectional study examines individuals at a single point in time. Cross-sectional studies typically generate only two fitness classes (e.g., dead versus living, mating versus unmated), and their analysis involves a considerable number of assumptions (Lande and Arnold, 1983; Arnold and Wade, 1984b). While longitudinal studies are preferred, they usually require far more work and may be impossible to carry out in many field situations. Age-structured populations pose further complications in that proper fitness measures require knowledge of the population's demography; see Charlesworth (1994), Lande (1982), Lenski and Service (1982), and Travis and Henrich (1986) for details.

17.3.2 Episodes of Selection

As mentioned, individuals are often measured over more than one episode of selection. Imagine that a cohort of n individuals (indexed by $1 \leq r \leq n$) is followed through several episodes. Let $W_j(r)$ be the fitness measure for the jth episode of selection for the rth individual. For example, for viability W_j is either zero (dead) or one (alive) at the census period. Let $\overline{W}_j$ denote the mean fitness of a random individual following the jth episode of selection. Relative fitness components $w_j(r) = W_j(r)/\overline{W}_j$ will turn out to be especially useful. At the start of the study, the frequency of each individual is $1/n$, giving for the first (observed) episode of selection

$$\overline{W}_1 = \frac{1}{n} \sum_{r=1}^{n} W_1(r). \tag{17.20a}$$

Caution is in order at this point as considerable selection may have already occurred prior to the life cycle stages being examined. Following the first episode of selection, the new fitness-weighted frequency of the rth individual is $w_1(r)/n$, implying

$$\overline{W}_2 = \sum_{r=1}^{n} W_2(r) \cdot w_1(r) \cdot \left(\frac{1}{n} \right). \tag{17.20b}$$

In general, for the jth episode of selection,

$$\overline{W}_j = \sum_{r=1}^{n} W_j(r) \cdot w_{j-1}(r) \cdot w_{j-2}(r) \ldots w_1(r) \cdot \left(\frac{1}{n} \right). \tag{17.20c}$$

Note that if $W_j(r) = 0$, further fitness components for r are unmeasured.

Letting $p_j(r)$ be the fitness-weighted frequency of individual r after j episodes of selection, it follows that $p_0(r) = 1/n$ and

$$p_j(r) = w_j(r) \cdot p_{j-1}(r) = \frac{1}{n} \prod_{i=1}^{j} w_i(r). \tag{17.21a}$$

Thus, (17.20c) can also be expressed as $\overline{W}_j = \sum W_j(r) \cdot p_{j-1}(r)$. Using these weights allows fitness-weighted moments to be calculated, e.g., the mean of a particular character following the jth episode is computed as

$$\mu_{z(j)} = \sum z(r) \cdot p_j(r), \tag{17.21b}$$

where $z(r)$ is the value of the character of individual r.

The directional selection differential S is computed as the difference between fitness-weighted means before and after an episode of selection, with the differential S_j for the jth episode given by

$$S_j = \mu_{z(j)} - \mu_{z(j-1)}. \tag{17.22a}$$

Selection differentials are additive over episodes, so that the total differential S following k episodes of selection is

$$S = \mu_{z(k)} - \mu_{z(0)} = (\mu_{z(k)} - \mu_{z(k-1)}) + \ldots + (\mu_{z(1)} - \mu_{z(0)}) = S_k + \ldots + S_1. \quad (17.22b)$$

17.3.3 The Robertson–Price Identity

As first noted by Robertson (1966), and greatly elaborated on by Price (1970; 1972a), the directional selection differential equals the covariance of phenotype and relative fitness,

$$S = \sigma(w, z). \quad (17.23)$$

This identity is quite useful for obtaining the selection differential in complex selection schemes and (as detailed below) forms the basis for a number of useful expressions relating selection and fitness.

To obtain the Robertson–Price identity, let μ_s be the fitness-weighted mean after selection and μ the mean before selection,

$$
\begin{aligned}
S = \mu_s - \mu &= \sum_{r=1}^{n} z(r) w(r) \cdot p(r) - \mu \\
&= E[zw] - 1 \cdot E[z] \\
&= E[zw] - E[w] \cdot E[z] \\
&= \sigma(w, z),
\end{aligned}
$$

where we have used the fact that (by construction) $E[w] = 1$.

17.3.4 The Opportunity for Selection

How does one compare the amount of selection acting on different episodes and/or different populations? At first thought, one might consider using the standardized selection differential,

$$i = \frac{S}{\sigma_z}, \quad (17.24)$$

which is just the directional selection differential scaled in terms of character standard deviations (i is often called the *selection intensity* and is widely used in breeding). The drawback with i as a measure of *overall* selection on individuals is that it is *character-specific*. Hence, i is appropriate if we are interested in comparing the strength of selection on a particular *character*, but inappropriate if we wish to compare the overall strength of selection on *individuals*. Two populations may have the same i value for a given character, but if that character is tightly correlated with fitness in one population and only weakly correlated in the other, selection is much stronger in the latter population. Further, considerable selection can occur without changing the mean (e.g., stabilizing selection). Standardized (quadratic) differentials also exist for the variance (Section 17.4.3), but the problem of character specificity still remains.

A much cleaner measure (independent of the characters under selection) is I, the *opportunity for selection*, defined as the variance in *relative* fitness:

$$I = \sigma_w^2 = \frac{\sigma_W^2}{\overline{W}^2}. \tag{17.25}$$

This measure was introduced by Crow (1958; reviewed in Crow (1989)), who referred to it as the *index of total selection*. Crow noted that if fitness is perfectly heritable (e.g., $h^2(\text{fitness}) = 1$), then $I = \Delta\overline{w}$, the scaled change in fitness. Following Arnold and Wade (1984a; 1984b), I is referred to as the opportunity for selection, as any change in the distribution of fitness caused by selection represents an opportunity for within-generation change. A key feature of I is that it bounds the maximal selection intensity i for *any* character. From the Robertson–Price identity (17.23), the correlation between any character z and relative fitness (which is bounded in absolute value by one) is

$$|\rho_{z,w}| = \frac{|\sigma_{z,w}|}{\sigma_z \sigma_w} = \frac{|S|}{\sigma_z \sqrt{I}} = \frac{|i|}{\sqrt{I}} \leq 1,$$

implying

$$|i| \leq \sqrt{I} \tag{17.26}$$

Thus, the most that the mean of any character can be shifted within a generation is $\sqrt{I}$ phenotypic standard deviations.

The usefulness of I as a bound of i depends on the correlation between relative fitness and the character being considered. Figure 17.1 shows scatterplots of relative fitness against two characters (z_1 and z_2) measured in the same set of individuals. The marginal distributions of fitness are identical for both characters (since the same set of individuals was measured for both traits), and hence both traits have the same opportunity
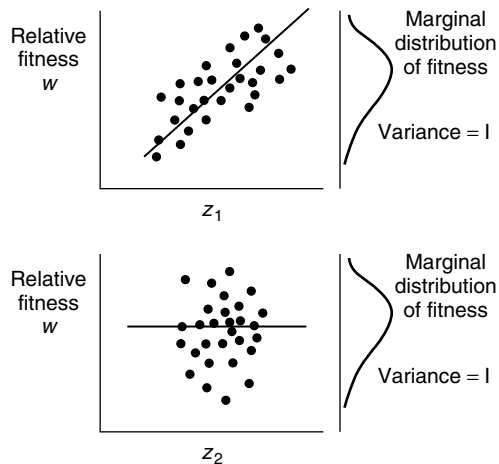
Figure 17.1 The fraction of the opportunity for selection I that is translated into a change in the mean depends on the correlation between relative fitness and the character. Characters z_1 and z_2 have the same marginal distribution of fitness, but only the regression of w on z_1 is significant.

for selection. The association between relative fitness and trait z_1 is fairly strong, while there is no relationship between relative fitness and z_2, so that z_1 realizes much, while z_2 realizes none, of the opportunity for change.

17.3.5   Some Caveats in Using the Opportunity for Selection

Despite its usefulness, there are some subtle issues in the interpretation of I. To begin with, even though I appears to remove scaling effects due to different types of fitnesses, for estimates of I to be truly comparable, they must be based on consistent measures of fitness (Trail, 1985). Second, if the variance in fitness is not independent of $\overline{W}$, comparisons of I values between populations are compromised. Such a lack of independence occurs in cross-sectional studies that measure sexual selection by simply counting the number of mating pairs. If the time scale is such that only single matings are observed, the fitness of any individual is either 1 (mating) or 0 (not mating). The resulting fitness of randomly drawn individuals is binomially distributed with mean p (the mean copulatory success for the sex being considered) and variance $p(1-p)$, hence

$$I = \frac{p(1-p)}{p^2} \simeq \frac{1}{p} \text{ if } p \ll 1. \tag{17.27}$$

The mean and variance in individual fitness are not independent, and the opportunity for selection depends entirely on mean population fitness. As the time window for observing mating pairs decreases, fewer matings are seen and p decreases, increasing I. Thus the opportunity for selection is often inflated if the observation period is short relative to the total mating period. Likewise, if one is only comparing viability selection (alive or dead after an episode of selection), (17.27) also holds.

Another example of the lack of independence between $\overline{W}$ and σ_W^2 was given by Downhower *et al.* (1987), who assumed that the number of mates for any given male follows a Poisson distribution. In this case, the variance in number of mates equals the mean number of mates, giving

$$I = \frac{\overline{W}}{\overline{W}^2} = \overline{W}^{-1}, \tag{17.28}$$

where $\overline{W}$ is the mean number of mates per male. Thus, differences in I between populations do not necessarily indicate *biological* differences in male mating ability. For example, in a population of 100 males, if only 5 females mate, average male mating success is $\overline{W} = 0.05$, while if 50 females mate, $\overline{W} = 0.5$. For this example, differences in I come solely from variation in the number of mating females, not biological differences between males in their ability to acquire mates. Downhower *et al.* conclude from this example that comparing I values with the Poisson prediction ($I = 1/\overline{W}$), or some other appropriate random distribution, may help clarify the interpretation of I. For this case, values of I less than the Poisson prediction indicate a more uniform distribution of fitness than expected if mate choice is random, while values in excess of this expectation indicate disproportionately high fitness among a limited set of individuals.

The Poisson mating example further points out that random variation (differences in individual fitness not attributable to intrinsic differences between individuals) reduces the correlation between phenotypic value and relative fitness. For this reason, measures of selection based entirely upon variance in mating success have been criticized (Banks

and Thompson, 1985; Sutherland 1985a; 1985b; Koenig and Albano, 1986). Although carefully controlled studies can reduce the error variance induced by chance (e.g., Houck *et al.*, 1985), accounting for inflation of the opportunity for selection by random effects remains a problem.

17.4   FITNESS SURFACES

17.4.1   Individual and Mean Fitness Surfaces

$W(z)$, the expected fitness of an individual with phenotype *z*, describes a *fitness surface* (or *fitness function*), relating fitness and character value. The *relative* fitness surface $w(z) = W(z)/\overline{W}$ is often more convenient than $W(z)$, and we use these two somewhat interchangeably. The nature of selection on a character in a particular population is determined by the local geometry of the individual fitness surface over that part of the surface spanned by the population (Figure 17.2). If fitness is strictly increasing (or decreasing) over some range of phenotypes, a population having its mean value in this interval experiences *directional selection*. If $W(z)$ contains a local maximum, a population with members within that interval experiences *stabilizing selection*. If the population is distributed around a local minimum, *disruptive selection* occurs. As is illustrated in Figure 17.2(d), when the local geometry of the fitness surface is complicated (e.g., multimodal) the simplicity of description offered by these three types of selection breaks down, as the population can experience all three simultaneously.
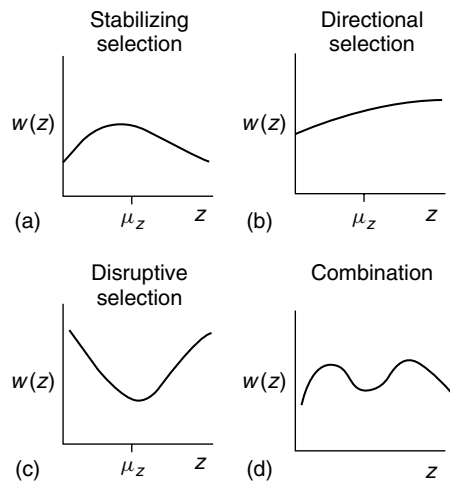
Figure 17.2   Selection is usually classified into three basic forms depending on the local geometry of $W(z)$: stabilizing (a), directional (b), and disruptive (c). As (d) illustrates, populations can simultaneously experience multiple forms of selection.

$W(z)$ may vary with genotypic and environmental backgrounds. In some situations (e.g., predators with search images, sexual selection, dominance hierarchies, truncation selection) the fitness of a phenotype depends on the frequency of other phenotypes in the population. In this case, fitnesses are said to be *frequency-dependent*.

Mean population fitness $\overline{W}$ is also a fitness surface, describing the expected fitness of the population as a function of the distribution $p(z)$ of phenotypes in that population,

$$\overline{W} = \int W(z) p(z) \, dz. \tag{17.29}$$

Mean fitness can be thought of as a function of the parameters of the phenotypic distribution. For example, if z is normally distributed, mean fitness is a function of the mean μ_z and variance σ_z^2 for that population.

To stress the distinction between the $W(z)$ and $\overline{W}$ fitness surfaces, the former is referred to as the *individual fitness surface*, the latter as the *mean fitness surface*. Knowing the individual fitness surface allows one to compute the mean fitness surface for any specified $p(z)$, but the converse is not true. The importance of the mean fitness surface is that it provides one way of describing how the population changes under selection – under the infinitesimal model, the derivative of $\overline{W}$ with respect to μ_z describes the change in mean (17.31b).

17.4.2 Measures of Selection on the Mean

A general way of detecting selection on a character is to compare the (fitness-weighted) phenotypic distribution before and after some episode of selection. Growth or other ontogenetic changes, immigration, and environmental changes can also change the phenotypic distribution, and great care must be taken to account for these factors. Another critical problem in detecting selection on a character is that selection on phenotypically correlated characters can also change the distribution. Keeping this important caveat in mind, we first examine measures of selection on a single character, as these form the basis for our discussion in Section 17.5 on measuring selection on multiple characters. Typically, selection on a character is measured by changes in the mean and variance, rather than changes in the entire distribution. While often only the mean is examined, considerable selection can occur without any significant change in the mean (e.g., stabilizing selection).

Two measures of within-generation change in phenotypic mean have been previously introduced: the directional selection differential S and the selection intensity i. A third measure is the directional selection gradient

$$\beta = \frac{S}{\sigma_z^2}. \tag{17.30}$$

As detailed in Section 17.4.6, β is the slope of the linear regression of fitness on phenotype. These three measures (i, S, β) are interchangeable for selection acting on a single character, but their multivariate extensions have very different behaviors (Section 17.5). In particular, the multivariate extension of β is the measure of choice, as it measures the amount of direct selection on a character, while S (and hence i) confounds direct selection with indirect effects due to selection on phenotypically correlated traits (17.48).

Recalling that $h^2 = \sigma_A^2/\sigma_z^2$, the response is often written by breeders as

$$R = h^2 \sigma_z i = \sigma_A h i. \tag{17.31a}$$

The response can also be expressed in terms of the directional selection gradient,

$$R = \sigma_A^2 \beta. \tag{17.31b}$$

17.4.3   Measures of Selection on the Variance

Similar measures can be defined to quantify the change in the phenotypic variance. At first glance this change seems best described by $\sigma_{z*}^2 - \sigma_z^2$, where σ_{z*}^2 is the phenotypic variance following selection. The problem with this measure is that directional selection reduces the variance. Lande and Arnold (1983) showed that

$$\sigma_{z*}^2 - \sigma_z^2 = \sigma[w, (z - \mu_z)^2] - S^2, \tag{17.32}$$

implying that directional selection decreases the phenotypic variance by S^2. With this in mind, Lande and Arnold suggest a corrected measure, the *stabilizing selection differential*,

$$C = \sigma_{z*}^2 - \sigma_z^2 + S^2, \tag{17.33}$$

that describes selection acting directly on the variance. As we will see below, the term 'stabilizing selection' differential may be slightly misleading, so following Phillips and Arnold (1989) we refer to C as the *quadratic selection differential*. Correction for the effects of directional selection is important, as claims of stabilizing selection based on a reduction in variance following selection can be due entirely to reduction in variance caused by directional selection. Similarly, disruptive selection can be masked by directional selection. Provided that selection does not act on characters phenotypically correlated with the one under study, C provides information on the nature of selection on the variance. Positive C indicates selection to increase the variance (as would occur with disruptive selection), while negative C indicates selection to reduce the variance (as would occur with stabilizing selection). As we discuss shortly, $C < 0$ ($C > 0$) is *consistent* with stabilizing (disruptive) selection, but not *sufficient*. A further complication in interpreting C is that if the phenotypic distribution is skewed, selection on the variance changes the mean. This causes a nonzero value of S that in turn inflates C (Figure 17.3).

Analogous to S equaling the covariance between z and relative fitness, (17.32) implies that C is the covariance between relative fitness and the squared deviation of a character from its mean,

$$C = \sigma[w, (z - \mu)^2]. \tag{17.34}$$

As was the case with S, the opportunity for selection I bounds the maximum possible within-generation change in variance (Arnold, 1986). Expressing C as a covariance and using the standard definition of a correlation gives $C = \rho_{w,(z-\mu)^2} \sigma_w \sigma[(z - \mu)^2]$. Since $\rho_{w,(z-\mu)^2}^2 \leq 1$, we have

$$C^2 \leq \sigma_w^2 \sigma^2[(z - \mu)^2] = I \cdot (\mu_{4,z} - \sigma_z^4).$$

Thus,

$$|C| \leq \sqrt{I(\mu_{4,z} - \sigma_z^4)}. \tag{17.35a}$$
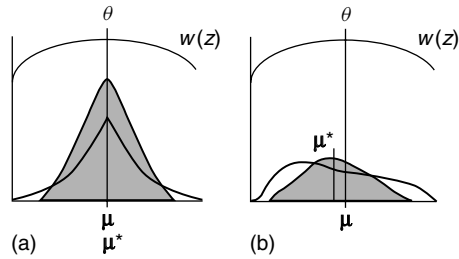
Figure 17.3 Even when a population is under strict stabilizing selection, the mean can change if the phenotypic distribution is skewed. A standard fitness function for stabilizing selection is $W(z) = 1 - b(\theta - z)^2$. O'Donald (1968) found that, even if the population mean is at the optimum value ($\mu_z = \theta$), S is nonnegative if the skew is nonzero ($\mu_{3,z} \neq 0$) as $S = -(b\mu_{3,z})/(1 - b\sigma_z^2)$. (a) If phenotypes are distributed symmetrically about the mean ($\mu_{3,z} = 0$), the distribution after selection (stippled) has the same mean when $\mu_z = \theta$. (b) If, however, the distribution is skewed, more of the distribution lies on one side of the mean than the other. Since the distribution is 'unbalanced', the longer tail experiences more selection than the shorter tail, changing the mean.

If z is normally distributed, the fourth central moment $\mu_{4,z} = 3\sigma_z^4$, giving

$$|C| \leq \sigma_z^2 \sqrt{2I}. \tag{17.35b}$$

The quadratic analog of β, the *quadratic* (or *stabilizing*) *selection gradient* γ, was suggested by Lande and Arnold (1983),

$$\gamma = \frac{\sigma[w, (z - \mu)^2]}{\sigma_z^4} = \frac{C}{\sigma_z^4}. \tag{17.36}$$

As with β, γ is the measure of choice when dealing with multiple characters because it accounts for the effects of (measured) phenotypically correlated characters (Section 17.5.5).

17.4.4 Gradients and the Geometry of Fitness Surfaces

A conceptual advantage of β and γ is that they describe the average local geometry of the fitness surface when phenotypes are normally distributed. When z is normal and individual fitnesses are not frequency-dependent, β can be expressed in terms of the geometry of the *mean* fitness surface,

$$\beta = \frac{\partial \ln \overline{W}}{\partial \mu_z} = \frac{1}{\overline{W}} \frac{\partial \overline{W}}{\partial \mu_z}, \tag{17.37a}$$

so that β is proportional to the slope of the $\overline{W}$ surface with respect to population mean. β can also be expressed as a function of the *individual* fitness surface. Lande and Arnold (1983) showed, provided z is normally distributed, that

$$\beta = \int \frac{\partial w(z)}{\partial z} p(z) \, \mathrm{d}z, \tag{17.37b}$$

implying that β is the average slope of the individual fitness surface, the average being taken over the population being studied. Likewise, if z is normal,

$$\gamma = \int \frac{\partial^2 w(z)}{\partial z^2} p(z)\,\mathrm{d}z, \tag{17.37c}$$

which is the average curvature of the individual fitness surface (Lande and Arnold, 1983). Thus, β and γ provide a measure of the geometry of the individual fitness surface averaged over the population being considered.

A final advantage of β and γ is that they appear as the measures of phenotypic selection in equations describing selection response. Recall from (17.31b) that under the assumptions leading to the breeders' equation, $R = \Delta\mu = \sigma_A^2 \beta$. Similarly, for predicting changes in additive genetic variance under the infinitesimal model, the expected change in variance from a single generation of selection is

$$\Delta\sigma_A^2 = \frac{\sigma_A^4}{2}(\gamma - \beta^2), \tag{17.38}$$

which decomposes the change in variance into changes due to selection on the variance (γ) and changes due to directional selection (β^2).

While the distinction between differentials and gradients seems trivial in the univariate case (only a scale difference), when considering multivariate traits, gradients have the extremely important feature of removing the effects of phenotypic correlations.

17.4.5 Estimating the Individual Fitness Surface

We can decompose the fitness W of an individual with character value z into the sum of its expected fitness $W(z)$ and a residual deviation e,

$$W = W(z) + e.$$

The residual variance in fitness for a given z, $\sigma_e^2(z)$, measures the variance in fitness among individuals with phenotypic value z. Estimation of the individual fitness surface is thus a generalized regression problem, the goal being to choose a candidate function for $W(z)$ that minimizes the average residual variance $\mathrm{E}_z[\sigma_e^2(z)]$. Since the total variance in fitness σ_W^2 equals the sum of the within-group (phenotype) and between-group variance in fitness,

$$\frac{\sigma_W^2 - \mathrm{E}_z[\sigma_e^2(z)]}{\sigma_W^2}$$

is the fraction of individual fitness variation accounted for by a particular estimate of $W(z)$, and provides a measure for comparing different estimates. In the limiting case where fitness is independent of z (and any characters phenotypically correlated with z), $W(z) = \overline{W}$, so that the between-phenotypic variance is zero while $\sigma_e^2(z) = \sigma_W^2$.

There are at least two sources of error contributing to e. First, there can be errors in measuring the actual fitness of an individual (these are almost always ignored). Second, the *actual* fitness of a particular individual can deviate considerably from the *expected value* for its phenotype due to chance effects and selection on other characters besides those being considered. Generally, these residual deviations are heteroscedastic, with σ_e^2 often depending on the value of $W(z)$ (Mitchell-Olds and Shaw, 1987; Schluter,

1988). For example, suppose fitness is measured by survival to a particular age. While $W(z) = p_z$ is the probability of survival for an individual with character value z, the fitness for a particular individual is either 0 (does not survive) or 1 (survives), giving $\sigma_e^2(z) = p_z(1 - p_z)$. Unless p_z is constant over z, the residuals are heteroscedastic. Note in this case that even after removing the effects attributable to differences in phenotypes, there still is substantial variance in individual fitness.

Inferences about the individual fitness surface are limited by the range of phenotypes in the population. Unless this range is very large, only a small region of the fitness surface can be estimated with any precision. Estimates of the fitness surface at the tails of the current phenotypic distribution are extremely imprecise, yet potentially very informative, suggesting what selection pressures populations at the margin of the observed range of phenotypes may experience. A further complication is that the fitness surface changes with the environment, so that year-to-year estimates can differ (e.g., Kalisz, 1986) and cannot be lumped together to increase sample size.

17.4.6   Linear and Quadratic Approximations of $W(z)$

The individual fitness surface $W(z)$ can be very complex and a wide variety of functions may be chosen to approximate it. The simplest and most straightforward approach is to use a low-order polynomial (typically linear or quadratic).

Consider first the simple linear regression of *relative* fitness w as a function of phenotypic value z. Since the directional selection gradient is

$$\beta = \frac{S}{\sigma_z^2} = \frac{\sigma(w, z)}{\sigma_z^2}, \tag{17.39a}$$

it follows from regression theory that β is the slope of the least-squares linear regression of relative fitness on z,

$$w = a + \beta z + e. \tag{17.39b}$$

Hence the best linear predictor of relative fitness is $w(z) = a + \beta z$. Since the regression passes through the expected values of w and z (1 and μ_z, respectively), this can be written as

$$w = 1 + \beta(z - \mu_z) + e, \tag{17.39c}$$

giving $w(z) = 1 + \beta(z - \mu_z)$. Assuming the fitness function is well described by a linear regression, β is the expected change in relative fitness given a unit change in z. The fraction of variance in individual fitness accounted for by this regression is

$$r_{z,w}^2 = \frac{\text{cov}^2(z, w)}{\text{var}(z) \cdot \text{var}(w)} = \hat{\beta}^2 \frac{\text{var}(z)}{\hat{I}}. \tag{17.40}$$

If the fitness surface shows curvature, as might be expected if there is stabilizing selection or disruptive selection, a *quadratic regression*,

$$w = a + b_1 z + b_2(z - \mu_z)^2 + e, \tag{17.41a}$$

is more appropriate. Since the regression passes through the mean of all variables,

$$w = 1 + b_1(z - \mu_z) + b_2[(z - \mu_z)^2 - \sigma_z^2] + e. \tag{17.41b}$$

The regression coefficients b_1 and b_2 nicely summarize the local geometry of the fitness surface. Evaluating the derivative of (17.41) at $z = \mu_z$ gives

$$\left.\frac{\partial w(z)}{\partial z}\right|_{z=\mu_z} = b_1 \quad \text{and} \quad \left.\frac{\partial^2 w(z)}{\partial z^2}\right|_{z=\mu_z} = 2b_2. \tag{17.42}$$

Hence b_1 is the slope and $2b_2$ the second derivative (curvature) of the best quadratic fitness surface around the population mean. $b_2 > 0$ indicates that the best-fitting quadratic of the individual fitness surface has an upward curvature, while $b_2 < 0$ implies the curvature is downward. Lande and Arnold (1983) suggest that $b_2 > 0$ indicates disruptive selection, while $b_2 < 0$ indicates stabilizing selection. Their reasoning follows from elementary geometry in that a *necessary* condition for a local minimum is that a function curves upward in some interval, while a necessary condition for a local maximum is that the function curves downward. Mitchell-Olds and Shaw (1987) and Schluter (1988) argue that this condition is not *sufficient*. Stabilizing selection is generally defined as the presence of a local maximum in $w(z)$ and disruptive selection by the presence of a local minimum, while b_2 indicates curvature, rather than the presence of local extrema.

Recalling the covariances given by (17.23) and (17.34), the least-squares values for b_1 and b_2 become

$$b_1 = \frac{(\mu_{4,z} - \sigma_z^4) \cdot S - \mu_{3,z} \cdot C}{\sigma_z^2 \cdot (\mu_{4,z} - \sigma_z^4) - \mu_{3,z}^2}, \tag{17.43a}$$

$$b_2 = \frac{\sigma_z^2 \cdot C - \mu_{3,z} \cdot S}{\sigma_z^2 \cdot (\mu_{4,z} - \sigma_z^4) - \mu_{3,z}^2}. \tag{17.43b}$$

Provided z is normally distributed before selection, the skew $\mu_{3,z} = 0$ and the kurtosis $\mu_{4,z} - \sigma_z^4 = 2\sigma_z^4$. In this case, $b_1 = \beta$ and $b_2 = \gamma/2$, giving the univariate version of the *Pearson–Lande–Arnold regression*,

$$w = 1 + \beta(z - \mu_z) + \tfrac{\gamma}{2}((z - \mu_z)^2 - \sigma_z^2) + e, \tag{17.44}$$

developed by Lande and Arnold (1983), motivated by Pearson (1903). The Pearson–Lande–Arnold regression provides a connection between selection differentials (directional and stabilizing) and quadratic approximations of the individual fitness surface.

An important point from (17.43a) is that if skew is present ($\mu_{3,z} \neq 0$), $b_1 \neq \beta$ and the slope term in the linear regression (the best *linear* fit) of $w(z)$ differs from the linear slope term in the quadratic regression (the best *quadratic* fit) of $w(z)$. This arises because the presence of skew generates a covariance between z and $(z - \mu_z)^2$. The biological significance of this can be seen by reconsidering Figure 17.3, in which the presence of skew in the phenotypic distribution results in a change in the mean of a population under strict stabilizing selection (as measured by the population mean being at the optimum of the individual fitness surface). Skew generates a correlation between z and $(z - \mu_z)^2$, so that selection acting only on $(z - \mu_z)^2$ generates a correlated change in z. From the Robertson–Price identity (17.23), the within-generation change in mean equals the covariance between-phenotypic value and relative fitness. Since covariances measure the amount of *linear* association between variables, in describing the change in mean, the correct measure is the slope of the best *linear* fit of the individual

fitness surface. If skew is present, using the linear slope b_1 from the quadratic regression to describe the change in mean is incorrect, as this quadratic regression removes the effects on relative fitness from a linear change in z due to the correlation between z and $(z - \mu_z)^2$.

17.5 MEASURING MULTIVARIATE SELECTION

Typically, investigation of the effects of natural selection involves measuring a suite of characters for each individual. In this setting, the phenotype of an individual is a vector $\mathbf{z} = (z_1, z_2, \ldots, z_n)^{\mathrm{T}}$ of the n measured character values. The major aim of a multivariate character analysis is to partition the total change in a trait into the effect from direct selection and the change due to selection on phenotypically correlated characters. Indeed, an unstated assumption of a univariate analysis is that there is no indirect selection on the trait induced by direct selection on unmeasured correlated characters. This assumption is also required for a multivariate analysis, as one cannot correct for the effects of any *unmeasured* selected characters phenotypically correlated to one (or more) of the traits under consideration.

Our development is based on the multiple regression approach of Lande and Arnold (1983). Similar approaches based on path analysis have also been suggested (Maddox and Antonovics, 1983; Mitchell-Olds, 1987; Crespi and Bookstein, 1988). We denote the mean vector and covariance matrix of $\mathbf{z}$ before selection by $\boldsymbol{\mu}$ and $\mathbf{P}$, and after selection (but before reproduction) by $\boldsymbol{\mu}^*$ and $\mathbf{P}^*$.

17.5.1 Changes in the Mean Vector: The Directional Selection Differential

The multivariate extension of the directional selection differential is the vector

$$\mathbf{S} = \boldsymbol{\mu}^* - \boldsymbol{\mu}, \tag{17.45a}$$

whose ith element is S_i, the differential for character z_i. As with the univariate case, the Robertson–Price identity (17.23) holds, so that the elements of $\mathbf{S}$ represent the covariance between character value and relative fitness, $S_i = \sigma(z_i, \omega)$. This immediately implies (17.26) that the opportunity for selection I bounds the range of S_i,

$$\frac{|S_i|}{\sigma_{z_i}} \leq \sqrt{I}. \tag{17.45b}$$

As is illustrated in Figure 17.4, $\mathbf{S}$ confounds the direct effects of selection on a character with the indirect effects due to selection on phenotypically correlated characters. Suppose that character 1 is under direct selection to increase in value while character 2 is not directly selected. As Figure 17.4 shows, if z_1 and z_2 are uncorrelated, there is no within-generation change in μ_2 (the mean of z_2). However, if z_1 and z_2 are positively correlated, individuals with large values of z_1 also tend to have large values of z_2, resulting in a within-generation increase in μ_2. Conversely, if z_1 and z_2 are negatively correlated, selection to increase z_1 results in a within-generation decrease in μ_2. Hence, a character not under selection can still experience a within-generation change in its phenotypic distribution due to selection on a phenotypically correlated character (indirect selection). Fortunately, the *directional selection gradient* $\boldsymbol{\beta} = \mathbf{P}^{-1}\mathbf{S}$ accounts for indirect selection

Figure 17.4 Selection on a character can result in a within-generation change in the mean of other phenotypically correlated characters not themselves under direct selection. Assume that character 1 is under simple truncation selection so that only individuals with $z_1 > T$ reproduce. (a) When z_1 and z_2 are uncorrelated, $S_2 = 0$. (b) When z_1 and z_2 are negatively correlated, $S_2 < 0$. (c) When z_1 and z_2 are positively correlated, $S_2 > 0$.

(a)

(b)

(c)
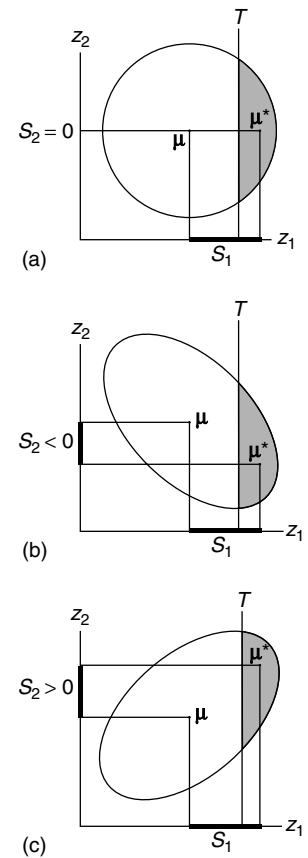
due to phenotypic correlations, providing a less biased picture of the nature of directional selection acting on $\mathbf{z}$.

17.5.2   The Directional Selection Gradient

From regression theory, the vector of partial regression coefficients for predicting the value of y given a vector of observations $\mathbf{z}$ is $\mathbf{P}^{-1}\boldsymbol{\sigma}(\mathbf{z}, y)$, where $\mathbf{P}$ is the covariance matrix of $\mathbf{z}$, and $\boldsymbol{\sigma}(\mathbf{z}, y)$ is the vector of covariances between the elements of $\mathbf{z}$ and the variable y. Since $\mathbf{S} = \boldsymbol{\sigma}(\mathbf{z}, \omega)$, it immediately follows that

$$\mathbf{P}^{-1}\boldsymbol{\sigma}(\mathbf{z}, \omega) = \mathbf{P}^{-1}\mathbf{S} = \boldsymbol{\beta} \qquad (17.46)$$

is the vector of partial regression for the best linear regression of relative fitness ω on phenotypic value $\mathbf{z}$, viz.,

$$\omega(\mathbf{z}) = 1 + \sum_{j=1}^{n} \beta_j (z_j - \mu_j) = 1 + \boldsymbol{\beta}^{\mathrm{T}}(\mathbf{z} - \boldsymbol{\mu}). \qquad (17.47)$$

Since β_j is a partial regression coefficient, it represents the change generated in relative fitness by changing z_j while holding all other character values in $\mathbf{z}$ constant – a one-unit

increase in z_j (holding all other characters constant) increases the expected relative fitness by β_j. Provided all characters under selection are measured (i.e., we can exclude the possibility of unmeasured characters influencing fitness that are phenotypically correlated with **z**), a character under no directional selection has $\beta_j = 0$ – when all other characters are held constant, the best linear regression predicts no change in expected fitness as we change the value of z_j. Thus, **β** accounts for the effects of phenotypic correlations only among the *measured* characters. Provided we measure all characters under selection and there are no confounding environmental or genetic effects, **β** provides an unbiased measure of the amount of directional selection acting on each character.

Since $\mathbf{S} = \mathbf{P}\boldsymbol{\beta}$, we have

$$S_i = \sum_{j=1}^{n} \beta_j P_{ij} = \beta_i P_{ii} + \sum_{j \neq i}^{n} \beta_j P_{ij}, \tag{17.48}$$

illustrating that the directional selection differential confounds direct selection on that character (β_i) with indirect contributions due to selection on phenotypically correlated characters (β_j). Equation (17.48) implies that, if two characters are phenotypically uncorrelated ($P_{ij} = 0$), selection on one has no within-generation effect on the phenotypic mean of the other.

17.5.3 Directional Gradients, Fitness Surface Geometry, and Selection Response

When phenotypes are multivariate normal (MVN), **β** provides a convenient descriptor of the geometry of both the individual and mean population fitness surfaces. Lande (1976; 1979) showed that

$$\boldsymbol{\beta} = \nabla_{\boldsymbol{\mu}}[\ln \overline{W}(\boldsymbol{\mu})] = \overline{W}^{-1} \cdot \nabla_{\boldsymbol{\mu}}[\overline{W}(\boldsymbol{\mu})], \tag{17.49}$$

which holds provided fitnesses are frequency-independent. Here $\nabla_{\boldsymbol{\mu}}$ is the *gradient* (or *gradient vector*) with respect to the vector of population means,

$$\nabla_{\boldsymbol{\mu}}[f] = \frac{\partial f}{\partial \boldsymbol{\mu}} = \begin{pmatrix} \dfrac{\partial f}{\partial \mu_1} \\ \vdots \\ \dfrac{\partial f}{\partial \mu_n} \end{pmatrix}.$$

The gradient at point $\boldsymbol{\mu}_0$ corresponds to a vector indicating the direction of steepest ascent of the function at that point (the multivariate slope of f at the point $\boldsymbol{\mu}_0$). Thus **β** is the gradient of mean population fitness with respect to the mean vector $\boldsymbol{\mu}$. Since **β** gives the direction of steepest increase in the mean population fitness surface, mean population fitness increases most rapidly when the change in the vector of means $\Delta\boldsymbol{\mu} = \boldsymbol{\beta}$. If fitnesses are frequency-dependent (individual fitnesses change as the population mean changes), then for $\mathbf{z} \sim \text{MVN}$,

$$\boldsymbol{\beta} = \nabla_{\boldsymbol{\mu}}[\ln \overline{W}(\boldsymbol{\mu})] + \int \nabla_{\boldsymbol{\mu}}[\omega(\mathbf{z})]\phi(\mathbf{z})\, d\mathbf{z}, \tag{17.50}$$

where the second term accounts for the effects of frequency-dependence and ϕ is the MVN density function (Lande, 1976). Here **β** does not point in the direction of steepest increase

in $\overline{W}$ unless the second integral is zero. If we instead consider the individual fitness surface, we can alternatively express $\boldsymbol{\beta}$ as the gradient of individual fitnesses averaged over the population distribution of phenotypes,

$$\boldsymbol{\beta} = \int \nabla_{\mathbf{z}}[\omega(\mathbf{z})]\phi(\mathbf{z})\, d\mathbf{z}, \tag{17.51}$$

which holds provided $\mathbf{z} \sim \text{MVN}$ (Lande and Arnold, 1983). Equation (17.51) holds regardless of whether fitness is frequency-dependent or frequency-independent.

Kingsolver *et al.* (2001) and Hoekstra *et al.* (2001) reviewed estimates of β (the univariate elements of $\boldsymbol{\beta}$) from selection in natural populations. They observed that the distribution for $|\beta|$ is close to exponential, with estimated selection gradients tending to be small (they found the median value for $|\beta|$ was 0.15). Interestingly, they found that the selection gradients for sexual selection (mate choice) tend to be larger than those for viability selection (survivorship). Most of the difference was due to a much higher frequency of values of $|\beta| < 0.1$ for viability selection. They further noted that estimated β values are highly correlated with their estimated selection intensities, i. Since i is a univariate measure that incorporates both direct and indirect effects of selection, while β removes any indirect effects from *measured* characters, they concluded that indirect selection is generally modest at best. The caveat with this conclusion is that correlated character under direct selection, but not included in the study, can falsely create an impression of direct selection on the actual traits measured.

17.5.4 Changes in the Covariance Matrix: The Quadratic Selection Differential

Motivated by the univariate case (17.34), define the multivariate *quadratic selection differential* to be a square ($n \times n$) matrix C whose elements are the covariances between all pairs of quadratic deviations $(z_i - \mu_{z_i})(z_j - \mu_{z_j})$ and relative fitness w:

$$C_{ij} = \sigma[w, (z_i - \mu_{z_i})(z_j - \mu_{z_j})]. \tag{17.52a}$$

Lande and Arnold (1983) showed that

$$\mathbf{C} = \sigma[w, (\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^{\mathrm{T}}] = \mathbf{P}^* - \mathbf{P} + \mathbf{SS}^{\mathrm{T}}. \tag{17.52b}$$

If no quadratic selection is acting, the covariance between each quadratic deviation and fitness is zero and $\mathbf{C} = \mathbf{0}$. In this case (17.52b) gives

$$P_{ij}^* - P_{ij} = -S_i S_j, \tag{17.53}$$

demonstrating that the $S_i S_j$ term corrects C_{ij} for the change in covariance caused by directional selection alone.

As was the case for $\mathbf{S}$, the fact that C_{ij} is a covariance immediately allows us to bound its range using the opportunity for selection. Since $\sigma^2(x, y) \leq \sigma^2(x)\,\sigma^2(y)$,

$$C_{ij}^2 \leq I\sigma^2[(z_i - \mu_{z_i})(z_j - \mu_{z_j})]. \tag{17.54a}$$

When z_i and z_j are bivariate normal, then (Kendall and Stuart, 1983)

$$\sigma^2[(z_i - \mu_{z_i})(z_j - \mu_{z_j})] = P_{ij}^2 + P_{ii}P_{jj} = P_{ij}^2(1 + \rho_{ij}^{-2}), \tag{17.54b}$$

where ρ_{ij} is the phenotypic covariance between z_i and z_j. Hence, for Gaussian-distributed phenotypes,

$$\left| \frac{C_{ij}}{P_{ij}} \right| \leq \sqrt{I} \sqrt{1 + \rho_{ij}^{-2}}, \tag{17.55}$$

which is a variant of the original bound based on I suggested by Arnold (1986).

17.5.5   The Quadratic Selection Gradient

Like **S**, **C** confounds the effects of direct selection with selection on phenotypically correlated characters. However, as was true for **S**, these indirect effects can also be removed by a regression. Consider the best quadratic regression of relative fitness as a function of multivariate phenotypic value,

$$w(\mathbf{z}) = a + \sum_{j=1}^{n} b_j z_j + \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \gamma_{jk} (z_j - \mu_j)(z_k - \mu_k) \tag{17.56a}$$

$$= a + \mathbf{b}^{\mathrm{T}} \mathbf{z} + \frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\gamma} (\mathbf{z} - \boldsymbol{\mu}). \tag{17.56b}$$

Using multiple regression theory, Lande and Arnold (1983) showed that the matrix $\boldsymbol{\gamma}$ of quadratic partial regression coefficients is given by

$$\boldsymbol{\gamma} = \mathbf{P}^{-1} \boldsymbol{\sigma} [w, (\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^{\mathrm{T}}] \mathbf{P}^{-1} = \mathbf{P}^{-1} \mathbf{C} \mathbf{P}^{-1}. \tag{17.57}$$

This is the *quadratic selection gradient* and (like $\boldsymbol{\beta}$) removes the effects of phenotypic correlations, providing a more accurate picture of how selection is operating on the multivariate phenotype.

The vector **b** of linear coefficients for the quadratic regression need not equal the vector $\boldsymbol{\beta}$ of partial regression coefficients obtained by assuming only a linear regression. As was true for the univariate case (17.43) when skew is present in at least one of the components of the phenotypic distribution, **b** is a function of both **S** and **C**, while $\boldsymbol{\beta}$ is only a function of **S**. When skew is absent, $\mathbf{b} = \boldsymbol{\beta}$, recovering the multivariate version of the Pearson–Lande–Arnold regression,

$$w(\mathbf{z}) = a + \boldsymbol{\beta}^{T} \mathbf{z} + \frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\gamma} (\mathbf{z} - \boldsymbol{\mu}). \tag{17.58}$$

Since the γ_{ij} are partial regression coefficients, they predict the change in expected fitness caused by changing the associated quadratic deviation while holding all other variables constant. Increasing $(z_j - \mu_j)(z_k - \mu_k)$ by one unit in such a way as to hold all other variables and pairwise combinations of characters constant, relative fitness is expected to increase by γ_{jk} for $j \neq k$ and by $\gamma_{jj}/2$ if $j = k$ (the difference arises because $\gamma_{jk} = \gamma_{kj}$, so that γ_{jk} appears twice in the regression unless $j = k$). The coefficients of $\boldsymbol{\gamma}$ thus describe the nature of selection on quadratic deviations from the mean for both single characters and pairwise combinations of characters. $\gamma_{ii} < 0$ implies fitness is expected to decrease as z_i moves away (in either direction) from its mean. As was discussed previously, this is a necessary, *but not sufficient*, condition for stabilizing selection on character i. Similarly, $\gamma_{ii} > 0$ implies fitness is expected to increase as z_i moves away

from its mean, again a necessary, but not sufficient, condition for disruptive selection. Kingsolver *et al.* (2001) examined the distribution of estimated γ_{ii} values reported in the literature. Surprisingly, they found this distribution to be symmetric about zero, implying that disruptive selection ($\gamma_{ii} > 0$) is reported as often as stabilizing selection ($\gamma_{ii} < 0$), which is quite contrary to the widely hide view that stabilizing selection is perhaps the most common mode of selection in natural populations. Further, most of the estimated γ_{ii} values are small, with the median value for $|\gamma|$ being 0.10. Thus, the widely held notion of strong stabilizing selection being common in nature is (presently) not supported by the published data.

Turning to combinations of characters, nonzero values of $\gamma_{jk}(j \neq k)$ suggest the
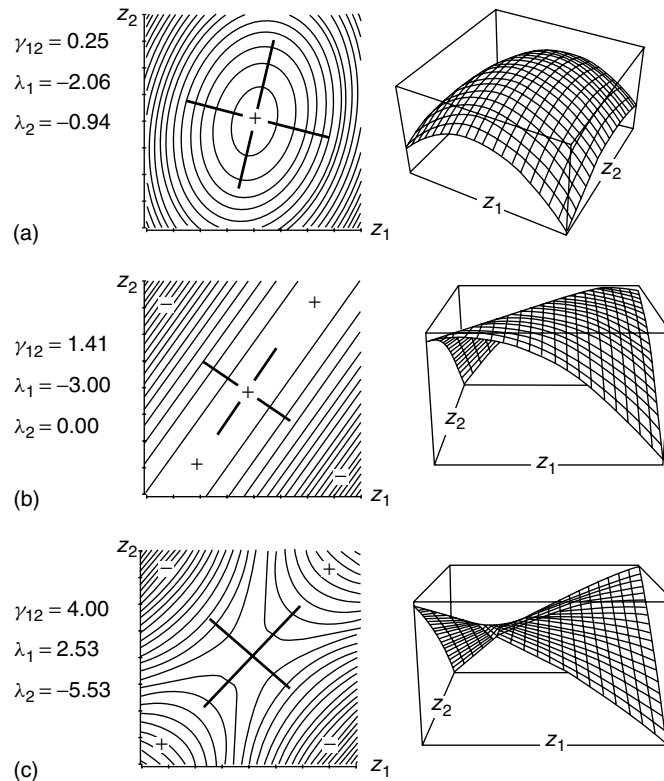
Figure 17.5 Three quadratic fitness surfaces, all of which have $\gamma_{11} = -2$, $\gamma_{22} = -1$ and $b = 0$. On the left are curves of equal fitness values, with peaks being represented by $+$ and valleys by $-$. Axes of symmetry of the surface (the canonical or principal axes of $\boldsymbol{\gamma}$) are denoted by the thick lines. These axes correspond to the eigenvectors of $\boldsymbol{\gamma}$. On the right are three-dimensional plots of individual fitness as a function of the phenotypic values of the characters z_1 and z_2. (a) $\gamma_{12} = 0.25$. This corresponds to stabilizing selection on both characters, with fitness falling off more rapidly (as indicated by the shorter distance between contour lines) along the z_1 axis than along the z_2 axis. (b) $\gamma_{12} = \sqrt{2} \simeq 1.41$, in which case $\boldsymbol{\gamma}$ is singular. The resulting fitness surface shows a ridge in one direction with strong stabilizing selection in the other. (c) When $\gamma_{12} = 4$, the fitness surface now shows a saddlepoint, with stabilizing selection along one of the canonical axes of the fitness surface and disruptive selection along the other.

presence of *correlation selection*: $\gamma_{jk} > 0$ suggests selection for a positive correlation between characters j and k, while $\gamma_{jk} < 0$ suggests selection for a negative correlation. Although it seems straightforward to infer the overall nature of selection by looking at these various pairwise combinations, this can give a very misleading picture about the geometry of the fitness surface (Figure 17.5). We discuss this problem and its solution in Section 17.5.8.

Finally, we can see the effects of phenotypic correlations on the quadratic selection differential. Solving for $\mathbf{C}$ by post- and premultiplying $\boldsymbol{\gamma}$ by $\mathbf{P}$ gives $\mathbf{C} = \mathbf{P}\boldsymbol{\gamma}\mathbf{P}$, or

$$C_{ij} = \sum_{k=1}^{n} \sum_{\ell=1}^{n} \gamma_{k\ell} P_{ik} P_{\ell j}, \tag{17.59}$$

showing that within-generation changes in phenotypic covariance, as measured by $\mathbf{C}$, are influenced by quadratic selection on phenotypically correlated characters.

17.5.6 Quadratic Gradients, Fitness Surface Geometry, and Selection Response

When phenotypes are multivariate normally distributed, $\boldsymbol{\gamma}$ provides a measure of the average curvature of the individual fitness surface,

$$\boldsymbol{\gamma} = \int \mathbf{H_z}[w(\mathbf{z})]\phi(\mathbf{z}) \, d\mathbf{z}, \tag{17.60a}$$

where $\mathbf{H_z}[f]$ denotes the Hessian of the function f with respect to $\mathbf{z}$ and is a multivariate measure of the quadratic curvature of a function (recall that the (i, j)th element of the Hessian is $\partial^2 f/\partial z_i \partial z_j$). This result, due to Lande and Arnold (1983), holds for both frequency-dependent and frequency-independent fitnesses. When fitnesses are frequency-independent (again provided $\mathbf{z} \sim$ MVN), $\boldsymbol{\gamma}$ provides a description of the curvature of the *log mean population* fitness surface, with

$$\mathbf{H_\mu}[\ln \overline{W}(\boldsymbol{\mu})] = \boldsymbol{\gamma} - \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}. \tag{17.60b}$$

This result is due to Lande (cited in Phillips and Arnold, (1989)) and points out that there are two sources for curvature in the mean fitness surface: $-\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}$ from directional selection and $\boldsymbol{\gamma}$ from quadratic selection.

The careful reader will undoubtedly have noted a number of similarities between direction and quadratic gradients. Table 17.1 summarizes these and other analogous features of measuring changes in means and covariances.

17.5.7 Estimation, Hypothesis Testing, and Confidence Intervals

Even if we can assume that a best-fit quadratic is a reasonable approximation of the individual fitness surface, we are still faced with a number of statistical problems. Unless we test for, and confirm, multivariate normality, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ must be estimated from separate regressions – $\boldsymbol{\beta}$ from the best linear regression, $\boldsymbol{\gamma}$ from the best quadratic regression. In either case, there are a large number of parameters to estimate – $\boldsymbol{\gamma}$ has $n(n + 1)/2$ terms and $\boldsymbol{\beta}$ has n terms, a total making of $n(n + 3)/2$. With 5, 10, and 25 characters, this corresponds to 20, 65, and 350 parameters. The number of

Table 17.1   Analogous features of directional and quadratic differentials and gradients.

Changes in means (directional selection)	Changes in covariances (Quadratic selection)

Differentials measure the covariance between relative fitness and phenotype

$$S_i = \sigma[w, z_i] \qquad\qquad C_{ij} = \sigma[w, (z_i - \mu_i)(z_j - \mu_j)]$$

The opportunity for selection bounds the differential

$$\frac{|S_i|}{\sigma(z_i)} \leq \sqrt{I} \qquad\qquad \left|\frac{C_{ij}}{P_{ij}}\right| \leq \sqrt{I}\sqrt{1 + \rho_{ij}^{-2}}$$

for any distribution of $\mathbf{z}$ provided $\mathbf{z} \sim$ MVN

Differentials confound direct and indirect selection

$$\mathbf{S} = \boldsymbol{\mu}^* - \boldsymbol{\mu} = \mathbf{P}\boldsymbol{\beta} \qquad\qquad \mathbf{C} = \mathbf{P}^* - \mathbf{P} + \mathbf{S}\mathbf{S}^\mathrm{T} = \mathbf{P}\boldsymbol{\gamma}\mathbf{P}$$

Gradients measure the amount of direct selection

$$\boldsymbol{\beta} = \mathbf{P}^{-1}\mathbf{S} \qquad\qquad \boldsymbol{\gamma} = \mathbf{P}^{-1}\mathbf{C}\mathbf{P}^{-1}$$

Gradients describe the slope and curvature of the mean population fitness surface, provided
  $z \sim$ **MVN and fitnesses are frequency-independent**

$$\beta_i = \frac{\partial \ln \overline{W}(\boldsymbol{\mu})}{\partial \mu_i} \qquad\qquad \gamma_{ij} = \frac{\partial^2 \ln \overline{W}(\boldsymbol{\mu})}{\partial \mu_i \partial \mu_j} + \beta_i \beta_j$$

Gradients describe the average curvature of the individual fitness surface, provided
  $\mathbf{z} \sim$ **MVN** (with $p(\mathbf{z})$ the density function),

$$\beta_i = \int \frac{\partial w(\mathbf{z})}{\partial z_i} p(\mathbf{z})\, d\mathbf{z} \qquad\qquad \gamma_{ij} = \int \frac{\partial^2 w(\mathbf{z})}{\partial z_i \partial z_j} p(\mathbf{z})\, d\mathbf{z}$$

Gradients appear as coefficients in fitness regressions

$$w(\mathbf{z}) = a + \mathbf{b}^\mathrm{T}(\mathbf{z} - \boldsymbol{\mu}) + \tfrac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^\mathrm{T}\boldsymbol{\gamma}(\mathbf{z} - \boldsymbol{\mu})$$

$$w(\mathbf{z}) = a + \boldsymbol{\beta}^\mathrm{T}(\mathbf{z} - \boldsymbol{\mu})$$

$\boldsymbol{\gamma} =$ the quadratic coefficient of the best

$\boldsymbol{\beta} =$ slope of best linear fit

quadratic fit; $\mathbf{b} = \boldsymbol{\beta}$ when $\mathbf{z} \sim$ MVN

Gradients appear as coefficients in evolutionary equations when $(\mathbf{z}, \mathbf{g}) \sim$ MVN

$$\Delta\boldsymbol{\mu} = \mathbf{G}\boldsymbol{\beta} \qquad\qquad \mathbf{G}^* - \mathbf{G} = \mathbf{G}(\boldsymbol{\gamma} - \boldsymbol{\beta}\boldsymbol{\beta}^\mathrm{T})\mathbf{G}$$

observations should greatly exceed $n(n + 3)/2$ in order to estimate these parameters with any precision.

A second problem is multicollinearity – if some of the characters being measured are highly correlated with each other, the phenotypic covariance matrix $\mathbf{P}$ can be nearly singular, so that small errors in estimating $\mathbf{P}$ result in large differences in $\mathbf{P}^{-1}$, which in turn gives a very large sampling variances for the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. One possibility is to use principal components (PCs) to extract a subset of the characters that explains most of the phenotypic variance of $\mathbf{P}$. Fitness regressions using the first few PCs as the characters can then be computed. This approach also reduces the problem of the number of parameters to estimate, but risks the real possibility of removing the most important characters. Further, PCs are often difficult to interpret biologically. While the first PC for morphological characters generally corresponds to a general measure of size, the others are typically much more problematic. Finally, this technique can spread the effects of selection on one character over several PCs, further complicating interpretation.

Another statistical concern with fitness regressions is that residuals are generally not normal – viability data typically have binominally distributed residuals, while count data

are often Poisson. Hence, many of the standard approaches for constructing confidence intervals on regression coefficients are not appropriate. Mitchell-Olds and Shaw (1987) suggest using the delete-one jackknife method for approximating confidence intervals in this case. Randomization tests and cross-validation procedures are other approaches. Multivariate tests of the presence of a single mode in the fitness surface are discussed by Mitchell-Olds and Shaw (1987).

17.5.8 Geometric Interpretation of the Quadratic Fitness Regression

In spite of their apparent simplicity, multivariate quadratic fitness regressions have a rather rich geometric structure. Scaling characters so that they have mean zero, the general quadratic fitness regression can be written as

$$w(\mathbf{z}) = a + \sum_{i=1}^{n} b_1 z_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} z_i z_j = a + \mathbf{b}^{\mathrm{T}} \mathbf{z} + \frac{1}{2} \mathbf{z}^{\mathrm{T}} \boldsymbol{\gamma} \mathbf{z}. \tag{17.61}$$

If $\mathbf{z} \sim \mathrm{MVN}$, then $\mathbf{b} = \boldsymbol{\beta}$ (the vector of coefficients of the best *linear* fit). Note that if we regard (17.61) as a second-order Taylor series approximation of $w(z)$, $\mathbf{b}$ and $\boldsymbol{\gamma}$ can be interpreted as the gradient and Hessian of individual fitness evaluated at the population mean (here $\boldsymbol{\mu} = \mathbf{0}$ by construction). The nature of curvature of (17.61) is determined by the matrix $\boldsymbol{\gamma}$. Even though a quadratic is the simplest curved surface, its geometry can still be very difficult to visualize.

We start our exploration of this geometry by considering the gradient of this best-fit quadratic fitness surface, which is

$$\nabla_{\mathbf{z}}[w(\mathbf{z})] = \mathbf{b} + \boldsymbol{\gamma} \mathbf{z}. \tag{17.62}$$

Thus, at the point $\mathbf{z}$ the direction of steepest ascent on the fitness surface (the direction in which to move in phenotype space to maximally increase individual fitness) is given by the vector $\mathbf{b} + \boldsymbol{\gamma} \mathbf{z}$ (when $\boldsymbol{\mu} = \mathbf{0}$). If the true individual fitness surface is indeed a quadratic, then the average gradient of individual fitness taken over the distribution of phenotypes is

$$\int \nabla_{\mathbf{z}}[w(\mathbf{z})] p(\mathbf{z}) \, \mathrm{d}\mathbf{z} = \mathbf{b} \int p(\mathbf{z}) \, \mathrm{d}\mathbf{z} + \boldsymbol{\gamma} \int \mathbf{z} p(\mathbf{z}) \, \mathrm{d}\mathbf{z} = \mathbf{b} \tag{17.63}$$

as the last integral is $\boldsymbol{\mu}$ (which is zero by construction). Hence, if the true fitness function is quadratic, the average gradient of individual fitness is given by $\mathbf{b}$, independent of the distribution of $\mathbf{z}$.

Solving $\nabla_{\mathbf{z}}[w(\mathbf{z})] = 0$, a point $\mathbf{z}_0$ is a candidate for a local extremum (stationary point) if $\boldsymbol{\gamma} \mathbf{z}_0 = -\mathbf{b}$. When $\boldsymbol{\gamma}$ is nonsingular,

$$\mathbf{z}_0 = -\boldsymbol{\gamma}^{-1} \mathbf{b} \tag{17.64a}$$

is the unique stationary point of this quadratic surface. Substituting into (17.61), the expected individual fitness at this point is

$$w_0 = a + \frac{1}{2} \mathbf{b}^{\mathrm{T}} \mathbf{z}_0 \tag{17.64b}$$

as obtained by Phillips and Arnold (1989). Since $\partial^2 w(\mathbf{z})/\partial z_i \partial z_j = \gamma_{ij}$, the Hessian of $w(\mathbf{z})$ is just $\boldsymbol{\gamma}$. Thus $\mathbf{z}_0$ is a local minimum if $\boldsymbol{\gamma}$ is positive definite (all eigenvalues are positive), a local maximum if $\boldsymbol{\gamma}$ is negative definite (all eigenvalues are negative), or a saddlepoint if the eigenvalues differ in sign. If $\boldsymbol{\gamma}$ is singular (has at least one zero eigenvalue) then there is no unique stationary point. An example of this is seen in Figure 17.5(b), where there is a ridge (rather than a single point) of phenotypic values having the highest fitness value. The consequence of a zero eigenvalue is that the fitness surface has no curvature along the axis defined by the associated eigenvector. If $\boldsymbol{\gamma}$ has k zero eigenvalues, then the fitness surface has no curvature along k dimensions. Ignoring fitness change along these dimensions, the remaining fitness space has only a single stationary point, which is given by (17.64a) for γ and $\mathbf{b}$ reduced to the $(n-k)$-dimensional subsurface showing curvature. While the curvature is completely determined by $\boldsymbol{\gamma}$, it is easy to be misled about the actual nature of the fitness surface if one simply tries to infer it by inspection of $\boldsymbol{\gamma}$, as Figure 17.5 illustrates.

17.5.8.1 Canonical Form of the Quadratic Fitness Surface

As Figure 17.5 shows, visualizing the individual fitness surface (even for two characters) is not trivial and can easily be downright misleading. The problem is that the cross-product terms (γ_{ij} for $i \neq j$) make the quadratic form difficult to interpret geometrically. Removing these terms by a change of variables so that the axes of new variables coincide with the axes of symmetry of the quadratic form (its canonical axes) greatly facilitates visualization of the fitness surface. Motivated by this, Phillips and Arnold (1989) suggest using two slightly different versions of the canonical transformation of $\boldsymbol{\gamma}$ to clarify the geometric structure of the best-fitting quadratic fitness surface. If we consider the matrix U whose columns are the eigenvalues of $\boldsymbol{\gamma}$, the transformation $\mathbf{y} = \mathbf{U}^T \mathbf{z}$ (hence $\mathbf{z} = \mathbf{Uy}$ since $\mathbf{U}^{-1} = \mathbf{U}^T$ as U is orthonormal, and $\mathbf{U}^T \boldsymbol{\gamma} \mathbf{U} = \boldsymbol{\Lambda}$, a diagonal matrix whose elements correspond to the eigenvalues of $\boldsymbol{\gamma}$) removes all cross-product terms in the quadratic form,

$$
\begin{aligned}
w(\mathbf{z}) &= a + \mathbf{b}^T \mathbf{U} \mathbf{y} + \frac{1}{2} (\mathbf{U}\mathbf{y})^T \boldsymbol{\gamma} (\mathbf{U}\mathbf{y}) \\[4pt]
&= a + \mathbf{b}^T \mathbf{U} \mathbf{y} + \frac{1}{2} \mathbf{y}^T (\mathbf{U}^T \boldsymbol{\gamma} \mathbf{U}) \mathbf{y} \\[4pt]
&= a + \mathbf{b}^T \mathbf{U} \mathbf{y} + \frac{1}{2} \mathbf{y}^T \boldsymbol{\Lambda} \mathbf{y} \\[4pt]
&= a + \sum_{i=1}^{n} \theta_i y_i + \frac{1}{2} \sum_{i=1}^{n} \lambda_i y_i^2,
\end{aligned}
\tag{17.65}
$$

where $\theta_i = \mathbf{b}^T \mathbf{e}_i$, $y_i = \mathbf{e}_i^T \mathbf{z}$, with λ_i and $\mathbf{e}_i$ the eigenvalues and associated unit eigenvectors of $\boldsymbol{\gamma}$. Alternatively, if a stationary point $\mathbf{z}_0$ exists (e.g., $\boldsymbol{\gamma}$ is nonsingular), the change of variables $\mathbf{y} = \mathbf{U}^T (\mathbf{z} - \mathbf{z}_0)$ further removes all linear terms (Box and Draper, 1987), so that

$$
w(\mathbf{z}) = w_0 + \frac{1}{2} \mathbf{y}^T \boldsymbol{\Lambda} \mathbf{y} = w_0 + \frac{1}{2} \sum_{i=1}^{n} \lambda_i y_i^2,
\tag{17.66}
$$

where $y_i = \mathbf{e}_i^T(\mathbf{z} - \mathbf{z}_0)$ and w_0 is given by (17.64b). Equation (17.65) is called the *A canonical form* and (17.66) the *B canonical form* (Box and Draper, 1987). Both forms represent a rotation of the original axis to the new set of axes (the canonical axes of $\boldsymbol{\gamma}$) that align them with axes of symmetry of the quadratic surface. The B canonical form further shifts the origin to the stationary point $\mathbf{z}_0$. Since the contribution to individual fitness from $\mathbf{b}^T\mathbf{z}$ is a hyperplane, its effect is to tilt the fitness surface. The B canonical form 'levels' this tilting, allowing us to focus entirely on the curvature (quadratic) aspects of the fitness surface.

The orientation of the quadratic surface is determined by the eigenvectors of $\boldsymbol{\gamma}$, while the eigenvalues of $\boldsymbol{\gamma}$ determine the nature and amount of curvature of the surface along each canonical axis. Along the axis defined by y_i, the individual fitness function has positive curvature (is concave) if $\lambda_i > 0$, has negative curvature (is convex) if $\lambda_i < 0$, or has no curvature (is a plane) if $\lambda_i = 0$. The amount of curvature is indicated by the magnitude of λ_i: the larger $|\lambda_i|$ the more extreme the curvature.

Returning to Figure 17.5, we see that the axes of symmetry of the quadratic surface are the canonical axes of $\boldsymbol{\gamma}$. For $\boldsymbol{\gamma}_{12} = 0.25$ (Figure 17.5a), $\lambda_1 = -2.06$ and $\lambda_2 = -0.94$, so that the fitness surface is convex along each canonical axis, with more extreme curvature along the y_1 axis. When $\gamma_{12} = \sqrt{2}$ (Figure 17.5b), one eigenvalue is zero while the other is -3, so that the surface shows no curvature along one axis (it is a plane), but is strongly convex along the other. Finally, when $\gamma_{12} = 4$ (Figure 17.5c), the two eigenvalues differ in sign, being -5.53 and 2.53. This generates a saddlepoint, the surface being concave along one axis and convex along with other, with the convex curvature being more extreme.

If $\lambda_i = 0$, the fitness surface along y_i has no curvature, so that the fitness surface is a ridge along this axis. If $\theta_i = \mathbf{b}^T\mathbf{e}_i > 0$ this is a rising ridge (fitness increases as y_i increases), if $\theta_i < 0$ it is a falling ridge (fitness decrease as y_i increase), and if $\theta_i = 0$ it is flat. Again returning to Figure 17.5 the effect of b is to tilt the fitness surface. Denoting values on the axis running along the ridge by y_1, if $\theta_1 > 0$ the ridge rises so that fitness increases as y_1 increases. Even if $\boldsymbol{\gamma}$ is not singular, it may be nearly so, with some of the eigenvalues being very close to zero. In this case, the fitness surface shows little curvature along the axes given by the eigenvectors associated with these near-zero eigenvalues. The fitness change a particular axis (here given by $\mathbf{e}_i$) is $\theta_i y_i + (\lambda_i/2)y_i^2$. If $|\theta_i| \gg |\lambda_i|$, the curvature of the fitness surface along this axis is dominated by the effects of linear (as opposed to quadratic) selection. Phillips and Arnold (1989) present a nice discussion of several other issues relating to the visualization of multivariate fitness surfaces, while Box and Draper (1987) review the statistical foundations of this approach.

17.5.9   Unmeasured Characters and Other Biological Caveats

Even if we are willing to assume that the best-fitting quadratic regression is a reasonable approximation of the individual fitness surface, there are still a number of important biological caveats to keep in mind. For example, the fitness surface can change in both time and space, often over short spatial/temporal scales (e.g., Kalisz, 1986; Stewart and Schoen, 1987; Scheiner, 1989; Jordan, 1991), so that one estimate of the fitness surface may be quite different from another estimation at a different time and/or location. Hence, considerable care must be used before pooling data from different times and/or sites to improve the precision of estimates. When the data are such that selection gradients can be estimated separately for different times or areas, space/time by gradient interactions can be tested for in a straightforward fashion (e.g. Mitchell-Olds and Bergelson, 1990).

Population structure can also influence fitness surface estimation in other ways. If the population being examined has overlapping generations, fitness data must be adjusted to reflect this (e.g. Stratton, 1992). Likewise, if members in the population differ in their amount of inbreeding, measured characters and fitness may show a spurious correlation if both are affected by inbreeding depression (Willis, 1996).

Perhaps the most severe caveat for the regression approach of estimating $w(z)$ is unmeasured characters – estimates of the amount of direct selection acting on a character are biased if that character is phenotypically correlated to unmeasured characters also under selection (Lande and Arnold, 1983; Mitchell-Olds and Shaw, 1987). Adding one or more of these unmeasured characters to the regression can change initial estimates of β and γ. Conversely, selection acting on unmeasured characters that are phenotypically *uncorrelated* with those being measured has no effect on estimates of β and γ.

17.6   MULTIPLE TRAIT SELECTION

17.6.1   Short-Term Changes in Means: The Multivariate Breeders' Equation

By combining (17.8) and (17.46), we can express the multivariate breeders equation as

$$\mathbf{R} = \mathbf{G}\mathbf{P}^{-1}\mathbf{S} = \mathbf{G}\boldsymbol{\beta}. \tag{17.67}$$

Presented in this form, the equation nicely demonstrates the relationship between the two main complications with selection on multiple characters: the *within-generation* change due to *phenotypic* correlations and the *between-generation* change (response to selection) due to *additive genetic* correlations. Cheverud (1984) makes the important point that although it is often assumed that a set of phenotypically correlated traits responds to selection in a coordinated fashion, this is not necessarily the case. Since $\boldsymbol{\beta}$ removes all the effects of phenotypic correlations and $\mathbf{R} = \mathbf{G}\boldsymbol{\beta}$, phenotypic characters will only respond as a group if they are all under direct selection or if they are *genetically* correlated.

17.6.2   The Effects of Genetic Correlations: Direct and Correlated Responses

While the use of $\boldsymbol{\beta}$ removes any further evolutionary effect of phenotypic correlations, additive genetic correlations strongly influence the response to selection. If n characters are under selection, and g_{ij} is the additive genetic covariance between traits i and j, then the response in character i to a single generation of selection is

$$R_i = \Delta\mu_i = \sum_{j=1}^{n} g_{ij}\beta_j = g_{ii}\beta_i + \sum_{j\neq i} g_{ij}\beta_j, \tag{17.68}$$

so that the response has a component due to direct selection on that character ($g_{ii}\beta_i$) and an additional component due to selection on all other genetically correlated characters.

If direct selection only occurs on character i, a genetically correlated character also shows a *correlated response to selection*. For example, for trait j, its response when only trait i is under selection is

$$R_j = g_{ij}\beta_i. \tag{17.69a}$$

Thus, the ratio of the expected change in two characters when only one is under direct selection is

$$\frac{R_j}{R_i} = \frac{g_{ij}}{g_{ii}} = \rho_{ij}^g \sqrt{\frac{g_{jj}}{g_{ii}}}. \tag{17.69b}$$

If both characters have the same additive genetic variance ($g_{ii} = g_{jj}$), the ratio of response simply reduces to ρ_{ij}^g, the correlation between additive genetic values.

17.6.3 Evolutionary Constraints Imposed by Genetic Correlations

One immediate consequence of (17.67) is that a character under selection does not necessarily change in the direction most favored by natural selection. For example, fitness may maximally increase if μ_2 decreases, so that $\beta_2 < 0$. However, if the sum of correlated responses on z_2 is sufficiently positive, then μ_2 will increase. Thus, a character may be dragged off its optimal fitness trajectory by a correlated response generated by stronger selection on other traits. As these other characters approach their equilibrium values, their β values approach zero, reducing their indirect effects on other characters, eventually allowing β_2 to dominate the dynamics (and hence ultimately μ_2 decreases).

Recall that the direction the mean vector must be changed to give the greatest change in mean fitness is β (as β is the gradient of $\overline{W}$ with respect to the mean vector; see (17.49)). However, generally $\mathbf{R} \neq \beta$, and thus the effect of the additive genetic covariance matrix $\mathbf{G}$ is to constrain the selection response from its optimal value. The mean vector changes in the direction most favored by selection if and only if

$$\mathbf{G}\beta = \lambda\beta, \tag{17.70}$$

which only occurs when β is an eigenvector (with associated eigenvalue λ) of $\mathbf{G}$. Note that even if $\mathbf{G}$ is a diagonal matrix (there is no correlation between the additive genetic values of the characters under selection), (17.70) is usually not satisfied. In fact, only when we can write $\mathbf{G} = \sigma_A^2 \mathbf{I}$ is (17.70) satisfied for arbitrary β. Thus, only when all characters have the same additive genetic variance and there is no additive genetic covariance between characters is the response to selection in the directional most favored by natural selection. Differences in the amounts of additive genetic variances between characters, in addition to nonzero additive genetic covariances, also impose constraints on character evolution.

17.6.4 Inferring the Nature of Previous Selection

Providing the breeders' equation holds over several generations, the cumulative response to m generations of selection is just

$$R^{(m)} = \sum_{t=1}^{m} R(t) = \sum_{t=1}^{m} \mathbf{G}(t)\beta(t). \tag{17.71a}$$

In particular, if the genetic covariance matrix remains constant, then

$$R^{(m)} = \mathbf{G}\left(\sum_{t=1}^{m} \beta(t)\right). \tag{17.71b}$$

Hence, if we observe a total response to selection of $\mathbf{R}_{\text{total}}$, then we can estimate the cumulative selection differential as

$$\boldsymbol{\beta}_{\text{total}} = \sum \boldsymbol{\beta}(t) = \mathbf{G}^{-1}\mathbf{R}_{\text{total}}. \tag{17.71c}$$

For example, if one observes a difference in the mean multivariate phenotype between two diverged populations, then provided we can assume that the genetic covariance matrix remains roughly constant, the history of past selection leading to the divergence (assuming it is entirely due to selection) can be inferred from (17.71c).

17.6.5 Changes in G under the Infinitesimal Model

17.6.5.1 Within-Generation Change in G

If the vector of breeding values g is multivariate normal with covariance matrix $\mathbf{G}$ before an episode of selection, then it can be shown (e.g., Lande and Arnold, 1983) that the covariance matrix $\mathbf{G}^*$ of breeding values after selection (but before reproduction) satisfies

$$\mathbf{G}^* - \mathbf{G} = \mathbf{G}\mathbf{P}^{-1}(\mathbf{P}^* - \mathbf{P})\mathbf{P}^{-1}\mathbf{G}. \tag{17.72a}$$

We show in (17.75) that

$$\mathbf{G}^* - \mathbf{G} = \mathbf{G}(\boldsymbol{\gamma} - \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}})\mathbf{G}. \tag{17.72b}$$

Thus, when the breeders' equation holds, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are sufficient to describe how phenotypic selection changes (within a generation) the additive genetic covariance matrix. Equation (17.72b) shows that the within-generation change in $\mathbf{G}$ has a component from directional selection and a second due from quadratic selection, as

$$\mathbf{G}^* - \mathbf{G} = -\mathbf{G}\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{G} + \mathbf{G}\boldsymbol{\gamma}\mathbf{G} = -\mathbf{R}\mathbf{R}^{\mathrm{T}} + \mathbf{G}\boldsymbol{\gamma}\mathbf{G}. \tag{17.72c}$$

In terms of the change in covariance for two particular characters, this can be factored as

$$G_{ij}^* - G_{ij} = -\left(\sum_{k=1}^{n} \beta_k G_{ik}\right)\left(\sum_{k=1}^{n} \beta_k G_{jk}\right) + \sum_{k=1}^{n}\sum_{\ell=1}^{n} \gamma_{k\ell} G_{ik} G_{\ell j}$$

$$= -\Delta\mu_i \cdot \Delta\mu_j + \sum_{k=1}^{n}\sum_{\ell=1}^{n} \gamma_{k\ell} G_{ik} G_{\ell j}. \tag{17.73a}$$

Thus the within-generation change in the additive genetic variance of character i is given by

$$G_{ii}^* - G_{ii} = -(\Delta\mu_i)^2 + \sum_{k=1}^{n}\sum_{\ell=1}^{n} \gamma_{k\ell} G_{ik} G_{i\ell}. \tag{17.73b}$$

17.6.5.2 The Response (Between-Generation Change) in G

While (17.72) and (17.73) describe how the covariance structure of breeding values changes within a generation, the between-generation change must consider how recombination alters any gametic-phase disequilibrium generated by selection. Recall our

treatment of the change in the additive variance in the univariate case (Section 17.2.2), in which all changes in variances occur due to gametic-phase disequilibrium d_t changing the additive variance. In this case, $\sigma_A^2(t) = \sigma_A^2(0) + d_t$, and hence $\sigma_z^2(t) = \sigma_z^2(0) + d_t$. Moving to multiple characters, disequilibrium is now measured by the matrix $\mathbf{D}_t = \mathbf{G}_t - \mathbf{G}_0$. Thus D_{ij} is the change in the additive genetic covariance between characters i and j induced by disequilibrium. Likewise, the phenotypic covariance matrix is given by $\mathbf{P}_t = \mathbf{P}_0 + \mathbf{D}_t$. Tallis (1987) and Tallis and Leppard (1988) extend Bulmer's result to multiple characters, showing that

$$\Delta \mathbf{D}_t = \tfrac{1}{2}(\mathbf{G}_t \mathbf{P}_t^{-1}(\mathbf{P}_t^* - \mathbf{P}_t)\mathbf{P}_t^{-1}\mathbf{G}_t - \mathbf{D}_t). \tag{17.74}$$

The first term represents the within-generation change in $\mathbf{G}$, the second the decay in $\mathbf{D}$ due to recombination.

We can also express (17.74) in terms of the quadratic and directional selection gradients, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Recalling the definitions of $\mathbf{C}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, we have

$$\mathbf{P}^* - \mathbf{P} = \mathbf{C} - \mathbf{s}\mathbf{s}^{\mathrm{T}}.$$

Hence

$$\begin{aligned}
\mathbf{P}^{-1}(\mathbf{P}^* - \mathbf{P})\mathbf{P}^{-1} &= \mathbf{P}^{-1}(\mathbf{C} - \mathbf{s}\mathbf{s}^{\mathrm{T}})\mathbf{P}^{-1} \\
&= \mathbf{P}^{-1}\mathbf{C}\mathbf{P}^{-1} - (\mathbf{P}^{-1}\mathbf{s})(\mathbf{P}^{-1}\mathbf{s})^{\mathrm{T}} \\
&= \boldsymbol{\gamma} - \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}.
\end{aligned} \tag{17.75}$$

Substituting into (17.74) gives

$$\Delta \mathbf{D}_t = \tfrac{1}{2}(\mathbf{G}_t(\boldsymbol{\gamma}_t - \boldsymbol{\beta}_t\boldsymbol{\beta}_t^{\mathrm{T}})\mathbf{G}_t - \mathbf{D}_t). \tag{17.76a}$$

Finally, recalling (17.72b), we can also express this as

$$\Delta \mathbf{D}_t = \frac{1}{2}(\mathbf{G}_t^* - \mathbf{G}_t) - \frac{\mathbf{D}_t}{2}. \tag{17.76b}$$

Equation (17.76b) shows that only half of the disequilibrium deviation in $\mathbf{G}$ from the previous generation ($\mathbf{D}_t$) is passed on to the next generation, as is half the within-generation change induced by selection in that generation. This is expected since the infinitesimal mode assumes unlinked loci and hence each generation recombination removes half of the present disequilibrium. The disequilibrium component converges to

$$\hat{\mathbf{D}} = \hat{\mathbf{G}}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\mathrm{T}})\hat{\mathbf{G}} = (\mathbf{G}_0 + \hat{\mathbf{D}})(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\mathrm{T}})(\mathbf{G}_0 + \hat{\mathbf{D}}), \tag{17.77}$$

where have place carets over $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ to remind the reader that these depend on $\mathbf{P}$, and hence on $\mathbf{D}$.

17.6.6 Effects of Drift and Mutation

17.6.6.1 Drift

If $\mathbf{G}_t$ changes strictly by drift, the multivariate analog to (17.16) is

$$\mathbf{G}_t = \left(1 - \frac{1}{2N_e}\right)^t \mathbf{G}_0. \tag{17.78}$$

This assumes that only additive genetic variance is present. Hence, ignoring the effect of disequilibrium, under the infinitesimal model the expected change after t generations of selection and drift is

$$\boldsymbol{\mu}_t - \boldsymbol{\mu}_0 = \sum_{k=0}^t \mathbf{G}_t \boldsymbol{\beta} = \mathbf{G}_0 \boldsymbol{\beta} \sum_{k=0}^t \left(1 - \frac{1}{2N_e}\right)^k$$

$$\simeq 2N_e (1 - e^{t/2N_e}) \mathbf{G}_0 \boldsymbol{\beta}. \tag{17.79}$$

17.6.6.2 Mutation

Let $\mathbf{U}$ be the matrix of per-generation input from mutations to the additive genetic variances and covariance. The expected equilibrium covariance matrix under drift and mutation becomes

$$\hat{\mathbf{G}} = 2N_e \mathbf{U}. \tag{17.80}$$

Under the assumptions of the infinitesimal model (selection is sufficiently weak that the dynamics of changes in allele frequencies are completely driven by drift and mutation), the additive genetic variance–covariance matrix at generation t is

$$\mathbf{G}_t = \hat{\mathbf{G}} + (\mathbf{G}_0 - \hat{\mathbf{G}}) e^{-t/2N_e}, \tag{17.81}$$

giving the cumulative response to t generations of constant directional selection $\boldsymbol{\beta}$ as

$$\mathbf{R}_t = \boldsymbol{\mu}_t - \boldsymbol{\mu}_0 = [t\hat{\mathbf{G}} + 2N_e (1 - e^{-t/2N_e})(\mathbf{G}_0 - \hat{\mathbf{G}})] \boldsymbol{\beta}. \tag{17.82a}$$

Note if there is no mutational input $\mathbf{U} = \mathbf{0}$, then the total response by generation t is given by

$$\mathbf{R}_t = 2N_e (1 - e^{-t/2N_e}) \boldsymbol{\beta} \mathbf{G}_0 = 2N_e (1 - e^{-t/2N_e}) \mathbf{R}_0, \tag{17.82b}$$

giving, at the limit,

$$\mathbf{R}_\infty = 2N_e \mathbf{R}_0, \tag{17.82c}$$

$2N_e$ times the initial response, as was found (Robertson, 1960) for univariate selection.

17.7 PHENOTYPIC EVOLUTION MODELS

The above machinery for selection response has been widely used by evolutionary biologists to model the dynamics of complex traits under selection. Under the assumptions

of the infinitesimal model, one only needs to specify the nature of selection, the starting mean and the initial covariance matrix. Further genetic details do not enter (unless one wishes to include mutation, when the additional composite parameter, σ_m^2, enters). Such analyses are often called *phenotypic evolution models*, given that the genetic details can be handed with a few composite parameters. As with the entire field of evolutionary quantitative genetics, this is a rather large enterprise. We will concern ourselves here with two applications: testing whether an observed pattern seen in the fossil record is due to selection or drift, and examining the consequences of stabilizing selection on mean population fitness. The goal is to give the reader a feel for the flavor of such modeling. A much more detailed treatment can be found in Walsh and Lynch (2003).

17.7.1 Selection versus Drift in the Fossil Record

One of the more interesting applications of phenotypic selection models is in trying to make inferences about the forces that shaped morphological evolution on extinct organisms. Obviously, one cannot perform the required crosses to estimate heritabilities on traits of interests. However, one might reasonably use the average heritability values for similar traits in (hopefully close) relatives of the extinct species. A variety of tests for whether an observed pattern of morphological change in the fossil record is consistent with drift (as opposed to having to invoke selection) have been proposed, and we briefly discuss a few here. An excellent overview of the many pitfalls of this general approach is given by Reyment (1991). Despite these concerns, this approach offers a powerful method of analysis to gain insight into the evolution of species long extinct.

17.7.1.1 Lande's Test

Lande (1976) offered the first test for whether an observed amount of morphological divergence was consistent with simple random genetic drift. If the mean is changing strictly under genetic drift, then if the distribution of breeding and environmental values is bivariate normal, the resulting distribution for the mean at generation t, $\mu(t)$, is normal with mean $\mu(0)$ (the initial mean) and variance $t\sigma_A^2/N_e$. Letting the observed divergence in means be $\Delta\mu = \mu(t) - \mu(0)$, then the largest effective population size consistent (at the 5% level) with this divergence being due strictly to drift is

$$N_e^* = \frac{(1.96)^2 h^2 t}{(\Delta\mu/\sigma_z)^2} = 3.84 \frac{th^2}{(\Delta\mu/\sigma_z)^2}. \tag{17.83}$$

This follows since the probability that a unit normal exceeds 1.96 is just 5%, and we have written $\sigma_A^2 = h^2\sigma_z^2$. Thus, provided that the heritability remains roughly constant, one can apply (17.83) to see if the maximal effective population size required for this amount of drift is exceeded by our reasonable estimate of the actual population size. If this is the case, then we can rule out drift as the sole agent for the observed divergence.

17.7.1.2 The Turelli–Gillespie–Lande and Bookstein Tests

Turelli *et al.* (1988) noted that Lande's test has the hidden assumption that the additive genetic variance is independent of the population size. Recall from (17.18) that the

equilibrium additive genetic variance under mutation and drift is $\sigma_A^2 = 2N_e\sigma_m^2$, where σ_m^2 is the polygenic mutation rate. Hence, after a sufficient number of generations, the additive variance should scale with N_e (if no other evolutionary forces are acting), in which case the drift variance becomes

$$\sigma^2[\mu(t)] = t\frac{\sigma_A^2}{N_e} = t\frac{2N_e\sigma_m^2}{N_e} = 2t\sigma_m^2. \tag{17.84}$$

The variance expression given by (17.84) transforms Lande's test from a critical maximal effective population size to one based on a critical minimum mutational variance. In particular, (17.83) is replaced by

$$\sigma_m^2 = \frac{(\Delta\mu)^2}{2t(1.96)^2} = 0.130\frac{(\Delta\mu)^2}{t}. \tag{17.85}$$

Drift is rejected at the 5 % of the suspected value of σ_m^2 is less than this value.

Turelli *et al.* also make the important point that the test for drift is really a two-sided test. While the tests (17.83) and (17.85) ask whether the observed divergence is too large under drift (as might be expected under directional selection), we can also ask if the observed divergence is *too small* under drift, as would be expected if stabilizing selection constrains the change on the mean. Noting that for a unit normal random variable U that $\Pr(|U| < 0.03) = 0.025$ and $\Pr(|U| > 2.24) = 0.025$ suggests the following two-sided test: if

$$\sigma_m^2 < \frac{(\Delta\mu)^2}{2t(2.24)^2} = 0.0858\frac{(\Delta\mu)^2}{t} \tag{17.86a}$$

the rate of evolution is too fast to be explained by drift, while if

$$\sigma_m^2 > \frac{(\Delta\mu)^2}{2t(0.03)^2} = 509\frac{(\Delta\mu)^2}{t} \tag{17.86b}$$

the rate of divergence is too slow for drift.

If one has access to a sample of generation means (as opposed to a starting and end points), then Bookstein (1989) notes that a more powerful test can be obtained (from the theory of random walks) by considering $\Delta\mu^*$, the largest absolute deviation from the starting mean anywhere along the series of t generation means. Here, the rate of evolution is too fast for drift if

$$\sigma_m^2 < \frac{(\Delta\mu^*)^2}{2t(2.50)^2} = 0.080\frac{(\Delta\mu^*)^2}{t}, \tag{17.87a}$$

while it is too slow for drift when

$$\sigma_m^2 > \frac{(\Delta\mu^*)^2}{2t(0.56)^2} = 1.59\frac{(\Delta\mu^*)^2}{t}. \tag{17.87b}$$

Note by comparison with (17.86) that the tests for too fast a divergence are very similar, while Bookstein's test for too slow a divergence (a potential signature of stabilizing selection) is much less stringent than the Turelli–Gillespie–Lande test.

17.7.2 Stabilizing Selection

17.7.2.1 The Nor-Optimal Fitness Function

A widely used fitness function in the evolutionary genetics literature is the *nor-optimal* or *normalizing* fitness function,

$$W(z) = \exp\left(-\frac{z^2}{2\omega^2}\right). \tag{17.88}$$

This is proportional to a normal density function with mean zero and variance ω^2, and implies that fitness decreases with increasing values of z^2, with the maximal fitness occurring for an individual with trait value $z = 0$. Typically, one scales the trait so that the optimal value occurs as zero, but more generally we can replace z^2 with $(z - \mu_0)^2$, where the optimal fitness now occurs at μ_0. Besides being a reasonable model for stabilizing selection, the main reason for the popularity of (17.88) in the literature is that if the phenotypic distribution is normal (with mean μ and variance σ^2) before selection, then the distribution of trait values after selection is normal with mean and variance

$$\mu^* = \frac{\omega^2\mu}{\sigma^2 + \omega^2} \text{ and } \sigma_*^2 = \frac{\omega^2\sigma^2}{\sigma^2 + \omega^2}. \tag{17.89a}$$

The expected response to selection (measured by the change in mean) in generation t is

$$R(t) = h^2(t)S(t) = h^2(t)[\mu^*(t) - \mu(t)] = -h^2(t)\frac{\mu(t)\sigma^2(t)}{\sigma^2(t) + \omega^2}. \tag{17.89b}$$

Selection thus drives the mean toward the optimal value of $\mu = 0$. When μ is nonzero, there is both directional selection (nonzero S) and stabilizing selection.

Turning to the dynamics of the variance, the change in the phenotypic variance (due to stabilizing selection generating disequilibrium) is given by (17.12),

$$\Delta d(t) = d(t + 1) - d(t)$$
$$= -\left(\frac{d(t)}{2} + \frac{h^2(t)}{2}[\sigma^2(t) - \sigma_*^2(t)]\right)$$
$$= -\left(\frac{d(t)}{2} + \frac{h^2(t)\sigma^4(t)}{2(\sigma^2(t) + \omega^2)}\right). \tag{17.89c}$$

Recalling from (17.11) that $\sigma_A^2(t) = \sigma_a^2 + d(t)$ and $\sigma_z^2(t) = \sigma_z^2 + d(t)$, at equilibrium ($\Delta d(t) = 0$) (17.89c) reduces to a quadratic equation whose admissible solution is given by

$$\hat{\sigma}_z^2 = \sigma_z^2\left(1 - \frac{2 + h^2 + \omega^2/\sigma_z^2 - \sqrt{8h^2 + (2 + h^2 + \omega^2/\sigma_z^2)^2}}{4}\right). \tag{17.89d}$$

Here σ_z^2 and h^2 are the values for the base population before selection. Figure 17.6 shows the reduction in the phenotypic variance at equilibrium. As expected, the reduction is more severe with higher heritabilities and/or stronger selection (smaller values of ω^2/σ^2).
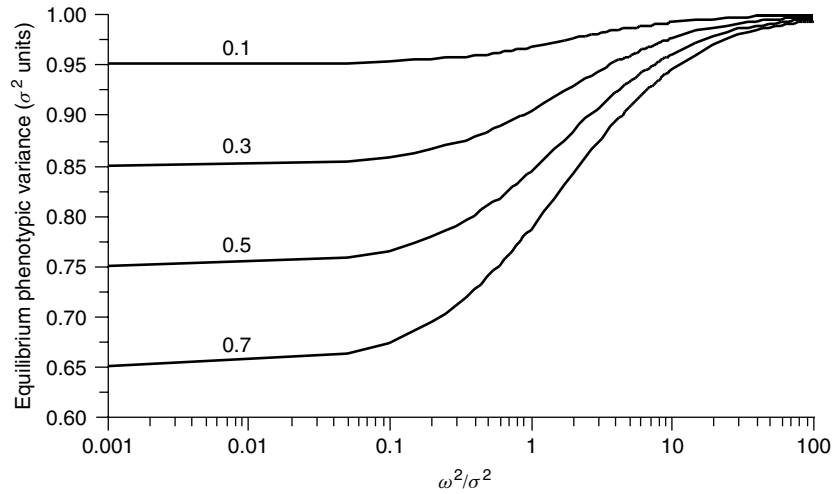
Figure 17.6 The equilibrium reduction in phenotypic variance under the infinitesimal model with nor-optimal selection (17.88). The curves correspond to initial heritability values ranging from 0.1 to 0.7. Stabilizing selection creates negative disequilibrium which reduces the additive genetic (and hence phenotypic) variance below its initial value. The reduction in phenotypic variance is greatest when the initial heritability is large and/or selection is strong (the width of the selection function is narrow relative to the initial width of the phenotypic distribution, i.e., $\omega^2 \ll \sigma^2$).

17.7.2.2 The Cost of Stabilizing Selection

Charlesworth (1984) examined the so-called cost of stabilizing selection by computing the selective load – the reduction in the mean fitness from the maximal value ($W = 1$ in this case). The load is given by $L = 1 - \overline{W}$ and measures the fraction of the population suffering selective mortality in a particular generation. The mean population fitness is obtained by integrating the fitness function over the distribution of phenotypes,

$$\overline{W} = \int p(z) W(z) \, dz = \sqrt{\frac{\omega^2}{\sigma^2 + \omega^2}} \exp\left[-\frac{\mu^2}{2(\sigma^2 + \omega^2)}\right], \tag{17.90a}$$

as found by Charlesworth. After selection has driven the phenotypic mean to zero, the load becomes

$$L = 1 - \sqrt{\frac{\omega^2}{\sigma^2 + \omega^2}} = 1 - \sqrt{\frac{\Lambda}{1 + \Lambda}}, \quad \text{where } \Lambda = \omega^2/\sigma^2. \tag{17.90b}$$

Charlesworth's expression for the load (17.90b) ignores the reduction in the phenotypic variance due to the generation of linkage disequilibrium (17.89c). A more precise expression for the load (when the population is at equilibrium) is

$$L = 1 - \sqrt{\frac{\omega^2}{\sigma_*^2 + \omega^2}} = 1 - \sqrt{\frac{\Lambda}{\sigma_*^2/\sigma^2 + \Lambda}}, \tag{17.90c}$$
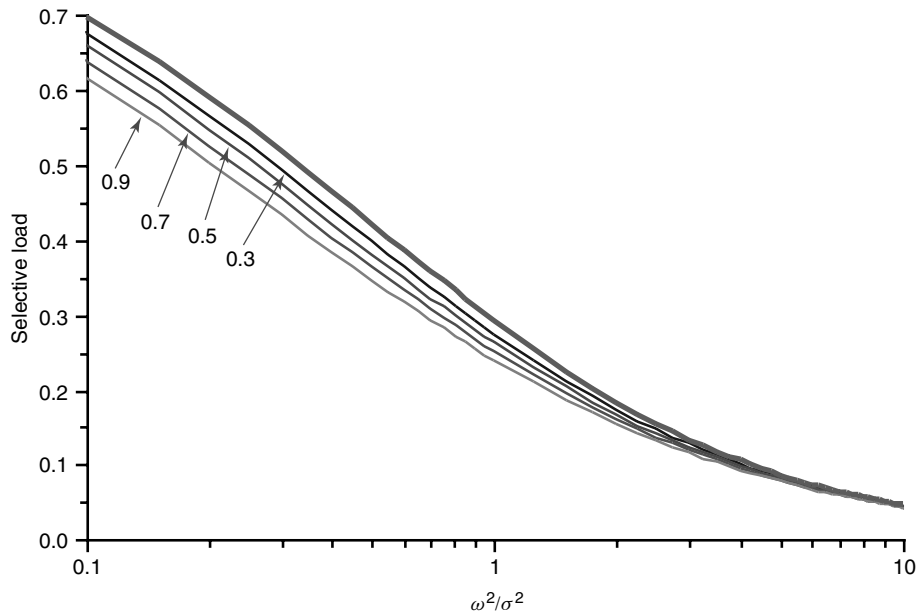
Figure 17.7   The selective load for an equilibrium population under nor-optimal selection. The solid upper curve is Charlesworth's approximation, while the four lower curves correspond to initial heritabilities ranging from 0.3 to 0.9.

where σ_*^2 is the equilibrium phenotypic variance. Figure 17.7 plots the corrected load for various values of initial heritability and different strengths of selection ω^2/σ^2. As expected, the selective loads are reduced when the disequilibrium is taken into account, as the negative disequilibrium generated by selection reduces the phenotypic variance, and this narrowing of the phenotypic variance increases the mean fitness. The reduction in the load is most dramatic for strong selection ($\omega^2 \ll \sigma^2$) and high heritabilities.

17.8   THEOREMS OF NATURAL SELECTION: FUNDAMENTAL AND OTHERWISE

Our final comments in this brief overview of evolutionary quantitative genetics concern the ultimate quantitative trait: individual fitness. What general statements, if any, can we make about the behavior of multilocus systems under selection? One oft-quoted is Fisher's *fundamental theorem of natural selection*, which states that 'the rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time'. This simple statement from Fisher's (1930) has generated a tremendous amount of work, discussion, and sometimes heated arguments. Fisher claimed his result was exact, a true theorem. The common interpretation of Fisher's theorem, that the rate of increase in fitness equals the additive variance in fitness, has been referred to by Karlin as 'neither fundamental nor a theorem' as it requires rather special conditions, especially when multiple loci influence fitness.

Since variances are nonnegative, the classical interpretation of Fisher's theorem (namely, $\Delta \overline{W} = \sigma_A^2(W)/\overline{W}$) implies that mean population fitness never decreases in a constant environment. As we discuss below, this interpretation of Fisher's theorem holds *exactly* only under restricted conditions, but is often a good approximate descriptor. However, an important corollary holds under very general conditions (Kimura, 1965; Nagylaki, 1976; 1977b; Ewens, 1976; Ewens and Thomson, 1977; Charlesworth, 1987): in the absence of new variation from mutation or other sources such as migration, selection is expected to eventually remove all additive genetic variation in fitness. If the population is at equilibrium, the average excess (in fitness) for each allele is zero as all segregating alleles have the same marginal fitness and hence $\sigma_A^2 = 0$ (Fisher, 1941).

17.8.1   The Classical Interpretation of Fishers' Fundamental Theorem

We first review the 'classical' interpretation and then discuss what Fisher actually seems to have meant.

To motivate Fisher's theorem for one locus, consider a diallelic (alleles A and a) locus with constant fitnesses under random mating. Let p denote the frequency of A, and denote the fitnesses for the three genotypes as W_{AA}, W_{Aa}, and W_{aa}. The mean population fitness is given by

$$\overline{W} = p^2 W_{AA} + 2p(1-p)W_{Aa} + (1-p)^2 W_{aa}, \tag{17.91a}$$

and the change in allele frequency can be expressed in terms of Wright's formula as

$$\Delta p = \frac{p(1-p)}{2\overline{W}} \frac{d\overline{W}}{dp} = \frac{p(1-p)}{2} \frac{d \ln \overline{W}}{dp}. \tag{17.91b}$$

Turning now to fitness, if the allele frequency change Δp is small, a first-order Taylor series approximation gives

$$\Delta \overline{W} = \overline{W}(p + \Delta p) - \overline{W}(p) \simeq \overline{W}(p) + \frac{\partial \overline{W}}{\partial p} \Delta p - \overline{W}(p) \tag{17.92a}$$

Applying (17.91b),

$$\Delta \overline{W} \simeq \frac{\partial \overline{W}}{\partial p} \Delta p = \frac{p(1-p)}{2\overline{W}} \left(\frac{\partial \overline{W}}{\partial p} \right)^2. \tag{17.92b}$$

From (17.91a),

$$\frac{\partial \overline{W}}{\partial p} = 2p W_{AA} + 2(1-2p)W_{Aa} + 2(p-1)W_{aa}$$

$$= 2[p(W_{AA} - W_{Aa}) + (1-p)(W_{Aa} - W_{aa})] = 2(\alpha_A - \alpha_a),$$

where the last equality follows the average effects of alleles A and a on fitness,

$$\alpha_A = p W_{AA} + (1-p)W_{Aa} - \overline{W} \quad \text{and} \quad \alpha_a = p W_{AA} + (1-p)W_{Aa} - \overline{W}.$$

The quantity $\alpha = \alpha_A - \alpha_a$ is the *average effect of an allelic substitution*; the difference in the average effects of these two alleles gives the mean effect on fitness from replacing

a randomly chosen a allele with an A allele. The additive genetic variance is related to α by $\sigma_A^2 = 2p(1 - p)\alpha^2$, giving

$$\Delta\overline{W} \simeq \frac{p(1 - p)(2\alpha)^2}{2\overline{W}} = \frac{\sigma_A^2(W)}{\overline{W}}. \tag{17.93}$$

Hence, the change in mean population fitness is proportional to the additive genetic variance in fitness.

As an example of the implication of Fisher's theorem, consider a diallelic locus with overdominance in fitness ($W_{AA} = 1$, $W_{Aa} = 1 + s$, $W_{aa} = 1$), giving the additive and dominance variance in fitness as

$$\sigma_A^2(W) = 2p(1 - p)s^2(1 - 2p)^2 \quad \text{and} \quad \sigma_D^2(W) = [2p(1 - p)s]^2.$$

As plotted in Figure 17.8, these variances change dramatically with p. The maximum genetic variance in fitness occurs at $p = 1/2$, but none of this variance is additive, and heritability in fitness is zero. It is easily shown that $\Delta p = 0$ when $p = 1/2$, and at this frequency $\sigma_A^2(W) = 0$, as the corollary of Fisher's theorem predicts. Thus, even though *total* genetic variation in fitness is maximized at $p = 1/2$, no change in $\overline{W}$ occurs as the *additive* genetic variance in fitness is zero at this frequency.

Even if Fisher's theorem holds exactly, its implication for character evolution can often be misinterpreted, as the following example illustrates. Suppose that locus A completely determines a character under stabilizing selection. Let the genotypes AA, Aa, and aa have discrete phenotypic values of $z = -1, 0$, and 1, respectively (so that this locus is strictly additive), and let the fitness function be $W(z) = 1 - sz^2$. If we assume no environmental variance, this generates very nearly the same fitnesses for each genotype as those given above, as the fitnesses can be normalized as $1 : (1 - s)^{-1} : 1$, where $(1 - s)^{-1} \simeq 1 + s$ for small s. The additive genetic variance for the trait z is maximized at $p = 1/2$, precisely the allele frequency at which the additive genetic variance in fitness $\sigma_A^2(W) = 0$. This stresses that Fisher's theorem concerns additive genetic variance in *fitness*, not in the *character*. In this example, the transformation of the phenotypic character value z to fitness W takes a character that is completely additive and introduces dominance when fitness is considered.

The above derivation of Fisher's theorem was only approximate. Under what conditions does the classical interpretation actually hold? While it is correct for a single locus with random mating, a single locus with no dominance under nonrandom mating (Kempthorne, 1957), and multiple additive loci (no dominance or epistasis: Ewens, 1969), it is generally compromised by nonrandom mating and departures from additivity (such as dominance or epistasis). Even when the theorem does not hold exactly, how good an approximation is it? Nagylaki (1976; 1977a; 1977b; 1991; 1992; 1993) has examined ever more general models of fitnesses when selection is weak (the fitness of any genotype can be expressed as $1 + as$, with s small and $|a| \ll 1$) and mating is random. Selection is further assumed to be much less than the recombination frequency c_{min} for the closest pair of loci ($s \ll c_{min}$). Under these fairly general conditions, Nagylaki shows that the evolution of mean fitness falls into three distinct stages. During the first $t < 2 \ln s / \ln(1 - c_{min})$ generations, the effects of any initial disequilibrium are moderated, first by reaching a point where the population evolves approximately as if it were in linkage equilibrium and then reaching a stage where the linkage disequilibria remain relatively constant. At this point, the change
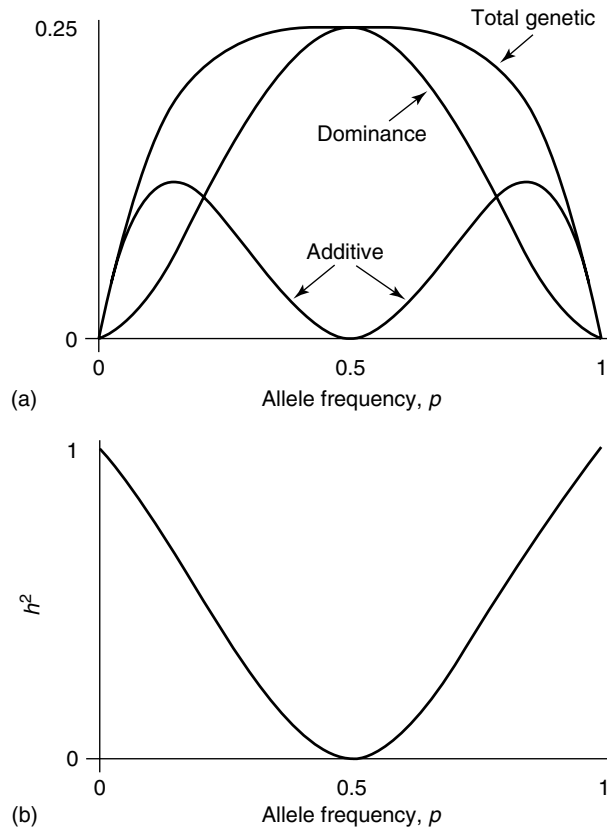
Figure 17.8 Genetic variances in a model with overdominance. Here, the heterozgyote has fitness $1 + s$, while both homozygotes have fitness 1. As (a) shows, the additive genetic variance is zero at allele frequency 0.5, but this is also the value where the total genetic variance is maximized. (b) shows that the heritability goes to zero as the allele frequency approaches 0.5. Thus, even though the genetic variance in fitness has maximal value at allele frequency 0.5, there is no response to selection, as none of the genetic variance is additive.

in mean fitness is

$$\Delta \overline{W} = \frac{\sigma_A^2(W)}{\overline{W}} + O(s^3), \tag{17.94}$$

where $O(s^3)$ means that terms of order s^3 have been ignored. Since additive variance is expected to be of order s^2, Fisher's theorem is expected to hold to a good approximation during this period. However, as gametic frequencies approach their equilibrium values, additive variance in fitness can be much less than order s^2, in which case the error terms of order s^3 can be important. During the first and third phases, mean fitness can decrease, but the fundamental theorem holds during the central phase of evolution.

17.8.2 What did Fisher Really Mean?

Fisher warned that his theorem 'requires that the terms employed should be used strictly as defined', and part of the problem stems from what Fisher meant by 'fitness'. Price

(1972b) and Ewens (1989; 1992) have argued that Fisher's theorem is always true, because Fisher held a very narrow interpretation of the change in mean fitness (see also Edwards, 1990; 1994; Frank, 1995; Lessard and Castilloux, 1995). Rather than considering the *total* rate of change in fitness, they argue that Fisher was instead concerned only with the *partial* rate of change, that due to changes in the contribution of individual alleles (specifically, changes in the average excesses/effects of these alleles). In particular, Ewens (1994) states:

> I believe that the often-made statement that the theorem concerns changes in mean fitness, assumes random-mating populations, is an approximation, and is not correct in the multi-locus setting, embodies four errors. The theorem relates the so-called partial increase in mean fitness, makes no assumption about random mating, is an exact statement containing no approximation, and finally is correct (as a theorem) no matter how many loci are involved.

What exactly is meant by the partial increase in fitness? A nice discussion is given by Frank and Slatkin (1992), who point out that the change in mean fitness over a generation is also influenced by the change in the 'environment', E. Specifically,

$$\Delta \overline{W} = (\overline{W}'|E') - (\overline{W}|E), \tag{17.95a}$$

where the prime denotes the fitness/environment in the next generation, so that the change in mean fitness compounds both the change in fitness and the change in the environment. We can partition out these components by writing

$$\Delta \overline{W} = [(\overline{W}'|E) - (\overline{W}|E)] + [(\overline{W}'|E') - (\overline{W}'|E)], \tag{17.95b}$$

where the first term in brackets represents the change in mean fitness under the initial 'environment' while the second represents the change in mean fitness due to changes in the environmental conditions. Fisher's theorem relates solely to changes in the first component, $(\overline{W}'|E) - (\overline{W}|E)$, which he called the change in fitness due to natural selection. Price (1972b). notes that Fisher had a very broad interpretation of 'environment', referring to both physical and genetic backgrounds. In particular, Fisher

> regarded the natural selection effect on fitness as being limited to the additive or linear effects of changes in gene (allele) frequencies, while everything else – dominance, epistasis, population pressure, climate, and interactions with other species – he regarded as a matter of the environment.

Hence the change in fitness referred to by Fisher is solely that caused by changes in allele frequencies and nothing else. Changes in gamete frequencies caused by recombination and generation of disequilibrium beyond those that influence individual allele frequencies are consigned to the 'environmental' category by Fisher. Given this, it is perhaps not too surprising that the strict interpretation of Fisher's theorem holds under very general conditions as nonrandom mating and nonadditive effects generate linkage disequilibrium, changing gametic frequencies in ways not simply predictable from changes in allele frequencies.

Nagylaki (1993) suggests that the statement $\Delta \overline{W} = \sigma_A^2(W)/\overline{W}$ be referred to as the *asymptotic fundamental theorem of natural selection*, while Fisher's more narrow (and correct) interpretation be referred to as the *Fisher–Price–Ewens theorem of natural*

selection. This distinction seems quite reasonable given the considerable past history of confusion.

17.8.3 Heritabilities of Characters Correlated with Fitness

The corollary of Fisher's theorem, that selection drives the additive variance in fitness to zero, makes a general prediction. Characters strongly genetically correlated with fitness should thus show reduced h^2 relative to characters less correlated with fitness (Robertson, 1955), reflecting the removal of additive variance by selection (which is partly countered by new mutational input). Because of the past action of natural selection, traits correlated with fitness are expected to have reduced levels of additive variance, while characters under less direct selection are expected to retain relatively greater amounts of additive variance. How well does this prediction hold up? Many authors have noticed that characters expected to be under selection (e.g., life-history characters, such as clutch size) tend, on average, to have lower heritabilities than morphological characters measured in the same population/species (reviewed by Robertson, 1955; Roff and Mousseau, 1987; Mousseau and Roff, 1987; Charlesworth, 1987). However, some notable exceptions are also apparent (Charlesworth, 1987). The difficulty with these general surveys is knowing whether a character is, indeed, highly genetically correlated with lifetime fitness. Clutch size, for example, would seem to be highly correlated with total fitness, but if birds with large clutch sizes have poorer survivorship, the correlation with lifetime fitness may be weak. Negative genetic correlations between components of fitness allow significant additive variance in each component at equilibrium, even when additive variance in total fitness is zero (Robertson, 1955; Rose, 1982).

17.8.4 Robertson's Secondary Theorem of Natural Selection

While the fundamental theorem predicts the rate of change of mean population fitness, we are often more interested in predicting the rate of change of a particular *character* under selection. Using a simple regression argument, Robertson (1966; 1968; see also Crow and Nagylaki, 1976; Falconer, 1985) suggested the *secondary theorem of natural selection*,

$$R = \mu'_z - \mu_z = \sigma_A(z, w),$$

namely, that the rate of change in a character equals the additive genetic covariance between the character and fitness (the covariance between the additive genetic value of the character and the additive genetic value of fitness). If this assertion holds, it provides a powerful connection between the fundamental theorem and the response of a character under selection. For example, it predicts that the character mean should not change when additive variance in *fitness* equals zero, regardless of the additive variance in the *character*.

To examine the conditions under which the secondary theorem holds, we first examine a single–locus model in some detail before considering what can be said for increasingly general multilocus models. Assuming random mating, a single generation of selection changes the frequency of allele A_i from p_i to $p'_i = p_i(1 + s_i)$, where s_i is the average excess in relative fitness for allele A_i. To map these changes in allele frequencies into changes in mean genotypic values, decompose the genotypic value of $A_i A_j$ as $G_{ij} = \mu_G + \alpha_i + \alpha_j + \delta_{ij}$, where α_i is the average effect of A_i on character value and

δ_{ij} the dominance deviation. Putting these together, the contribution from this locus to the change in mean after a generation of selection is

$$
\begin{aligned}
R &= \sum_{i,j} G_{ij} p_i' p_j' - \sum_{i,j} G_{ij} p_i p_j \\
&= \sum_{i,j} G_{ij} p_i (1 + s_i) p_j (1 + s_j) - \sum_{i,j} G_{ij} p_i p_j \\
&= \sum_{i,j} (\alpha_i + \alpha_j + \delta_{ij}) p_j p_i s_i + \sum_{i,j} (\alpha_i + \alpha_j + \delta_{ij}) p_i p_j s_i s_j.
\end{aligned}
\tag{17.96}
$$

The careful reader will note that we have already made an approximation by using G_{ij} instead of G_{ij}' in the very first sum. Since α_i and δ_{ij} are functions of the allele frequencies, they change as p_i changes, but we assume these are much smaller than the change in p_i itself (so we have assumed $\alpha_i' \simeq \alpha_i$ and $\delta_{ij}' \simeq \delta_{ij}$). A little more algebra (Nagylaki, 1989; 1991) simplifies the two sums in (17.96) to give the expected contribution to response as

$$
R = \sum_j \alpha_j s_j p_j + \sum_{i,j} \delta_{ij} p_i s_i p_j s_j.
\tag{17.97}
$$

The first sum is the expected product of the average effect α_j of an allele on character value times the average excess s_i of that allele on relative fitness. Since (by definition), $E[\alpha_j] = E[s_i] = 0$, the first sum is just the covariance between the average effect on the character with the average excess on fitness, in other words, the additive genetic covariance between *relative* fitness and character. Hence we can express (17.96) as

$$
R = \sigma_A(z, w) + B = \frac{\sigma_A(z, W)}{\overline{W}} + B.
\tag{17.98}
$$

If the character has no dominance (all $\delta_{ij} = 0$), the correction term B vanishes, recovering Robertson's original suggestion. If $\sigma_D^2(z)$ denotes the dominance variance in the trait value, Nagylaki (1989) shows that

$$
|B| \leq \frac{\sigma_D(z) \cdot \sigma_A^2(w)}{2},
\tag{17.99}
$$

where $\sigma_A^2(w)$ is the additive variance in relative fitness. Assuming no epistasis, Nagylaki (1989; 1991) shows that (17.98) holds for n loci, with the error term being

$$
B = \sum_{k=1}^{n} \sum_{i,j}^{n_k} \delta_{ij}(k) p_i(k) s_i(k) p_j(k) s_j(k),
\tag{17.100}
$$

where k indexes the loci, the kth of which has n_k alleles. This also holds when linkage disequilibrium is present, but the dominance deviations and average excesses in fitness are expected to be different than those under linkage equilibrium. Nagylaki (1991) shows that (17.100) is bounded by

$$
|B| \leq \left(\sum_{k=1}^{n} \sigma_{D(k)}(z) \right) \cdot \frac{\sigma_A^2(w)}{2}
\tag{17.101a}
$$

where $\sigma_{D(k)}^2(z)$ is the dominance variance in the character contributed by locus k. If all loci underlying the character are identical (the *exchangeable model*), this bound reduces to

$$|B| \leq \frac{\sigma_D(z) \cdot \sigma_A^2(w)}{2\sqrt{n}}. \tag{17.101b}$$

Hence, for fixed amounts of genetic variances, as the number of loci increases, the error in the secondary theorem becomes increasingly smaller.

Finally, the most general statement on the validity of the secondary theorem is due to Nagylaki (1993) for weak selection and random mating, but arbitrary epistasis and linkage disequilibrium. Similar to the weak selection analysis of Fisher's theorem discussed above, after a sufficient time the change in mean fitness is given by

$$R = \sigma_A(z, w) + O(s^2). \tag{17.102}$$

As with the fundamental theorem, when gametic frequencies approach their equilibrium values, terms in s^2 can become significant and mean response can differ significantly from Robertson's prediction.

17.9   FINAL REMARKS

Quantitative genetics is a very broad enterprise, comprising a number of distinct subfields – plant breeding, animal breeding, human genetics, and evolutionary quantitative genetics. The focus of this review has been on some of the unique aspects of evolutionary quantitative genetics that are likely to be less well known, but nonetheless potential very useful, to practitioners of these other subfields. For example, methods for estimating the nature of natural selection on traits are certainly of interest to plant and animal breeders, as they try to account for natural selection in the face of artificial selection. Likewise, evolutionary quantitative geneticists are also starting to draw upon some of the machinery developed in other subfields (such as best linear unbiased predictor estimation of breeding values). This is an encouraging trend, as each subfield of quantitative genetics has developed much useful machinery that deserves to be much more widely known (e.g. Walsh, 2002).

Acknowledgments

REFERENCES

Arnold, S.J. (1986). Limits on stabilizing, disruptive, and correlational selection set by the opportunity for selection. *American Naturalist* **128**, 143–146.

Arnold, S.J. and Wade, M.J. (1984a). On the measurement of natural and sexual selection: theory. *Evolution* **38**, 709–719.

Arnold, S.J. and Wade, M.J. (1984b). On the measurement of natural and sexual selection: applications. *Evolution* **38**, 720–734.

Banks, M.J. and Thompson, D.J. (1985). Lifetime mating success in the damselfly *Coenagrion puella*. *Animal Behaviour* **33**, 1175–1183.

Bookstein, F.L. (1989). Comment on a rate test. *Evolution* **43**, 1569–1570.

Box, G.E.P. and Draper, N.R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley, New York.

Bulmer, M.G. (1971). The effect of selection on genetic variability. *American Naturalist* **105**, 201–211.

Bulmer, M.G. (1974). Linkage disequilibrium and genetic variability. *Genetical Research* **23**, 281–289.

Bulmer, M.G. (1976). Regressions between relatives. *Genetical Research* **28**, 199–203.

Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, New York.

Bürger, R. (2000). *The Mathematical Theory of Selection, Recombination, and Mutation*. Wiley, Chichester.

Charlesworth, B. (1984). The cost of phenotypic evolution. *Paleobiology* **10**, 319–327.

Charlesworth, B. (1987). *Sexual Selection: Testing the Alternatives*, J.W. Bradbury and M.B. Andersson, eds. Wiley, New York, pp. 21–40.

Charlesworth, B. (1994). *Evolution in Age-Structured Populations*, 2nd edition. Cambridge University Press, Cambridge.

Cheverud, J.M. (1984). Quantitative genetics and developmental constraints on evolution by selection. *Journal of Theoretical Biology* **110**, 155–171.

Clutton-Brock, T.H. (ed) (1988). *Reproductive Success: Studies of Individual Variation in Contrasting Breeding Systems*, University of Chicago Press, Chicago, Ill.

Crespi, B.J. and Bookstein, F.L. (1988). A path-analytic model for the measurement of selection on morphology. *Evolution* **43**, 18–28.

Crow, J.F. (1958). Some possibilities for measuring selection intensities in man. *Human Biology* **30**, 1–13.

Crow, J.F. (1989). In *Evolution and Animal Breeding*, W.G. Hill and T.F.C. Mackay, eds. CAB International, Wallingford, CT, 91–97.

Crow, J.F. and Nagylaki, T. (1976). The rate of change of a character correlated with fitness. *American Naturalist* **110**, 207–213.

Downhower, J.F., Blumer, L.S. and Brown, L. (1987). Opportunity for selection: an appropriate measure for evaluating variation in the potential for selection. *Evolution* **41**, 1395–1400.

Edwards, A.W.F. (1990). Fisher, $\overline{W}$, and the fundamental theorem. *Theoretical Population Biology* **38**, 276–284.

Edwards, A.W.F. (1994). The fundamental theorem of natural selection. *Biological Review* **69**, 443–474.

Ewens, W.J. (1969). A generalized fundamental theorem of natural selection. *Genetics* **63**, 531–537.

Ewens, W.J. (1976). Remarks on the evolutionary effect of natural selection. *Genetics* **83**, 601–607.

Ewens, W.J. (1989). An interpretation and proof of the fundamental theorem of natural selection. *Theoretical Population Biology* **36**, 167–180.

Ewens, W.J. (1992). An optimizing principle of natural selection in evolutionary population genetics. *Theoretical Population Biology* **42**, 333–346.

Ewens, W.J. (1994). In *Frontiers in Mathematical Biology*, S. Levin, ed. Springer-Verlag, New York, pp. 186–197.

Ewens, W.J. and Thomson, G. (1977). Properties of equilibria in multi-locus genetic systems. *Genetics* **87**, 807–819.

Falconer, D.S. (1985). A note on Fisher's 'average effect' and 'average excess'. *Genetical Research* **46**, 337–347.

Falconer, D.S. and Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*, 4th edition. Longman, Harlow.

Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

Fisher, R.A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford. Reprinted in 1958 by Dover Publications, New York.

Fisher, R.A. (1941). Average excess and average effect of a gene substitution. *Annals of Eugenics* **11**, 53–63.

Frank, S.A. (1995). George Price's contributions to evolutionary genetics. *Journal of Theoretical Biology* **175**, 373–388.

Frank, S.A. and Slatkin, M. (1992). Fisher's fundamental theorem of natural selection. *Trends in Ecology and Evolution* **7**, 92–95.

Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hoang, A., Hill, C.E., Beerli, P. and Kingsolver, J.G. (2001). Strength and tempo of directional selection in the wild. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 9157–9160.

Houck, L.D., Arnold, S.J. and Thisted, R.A. (1985). A statistical study of mate choice: sexual selection in a plethodontid salamander (*Desmognathus ochrophaeus*). *Evolution* **39**, 370–386.

Jordan, N. (1991). Multivariate analysis of selection in experimental populations derived from hybridization of two ecotypes of the annual plant *Diodia teres*. *Evolution* **45**, 1760–1772.

Kalisz, S. (1986). Variable selection on the timing of germination in *Collinsia verna (Scrophulariaceae)*. *Evolution* **40**, 479–491.

Keightley, P.D. and Hill, W.G. (1987). Directional selection and variation in finite populations. *Genetics* **117**, 573–582.

Kempthorne, O. (1957). *An Introduction to Genetic Statistics*. Iowa State University Press, Ames, IA.

Kendall, M. and Stuart, A. (1983). *The Advanced Theory of Statistics, Volume 3: Design and Analysis, and Time Series*, 4th edition. Griffin, London.

Kimura, M. (1965). Attainment of quasi-linkage equilibrium when gene frequencies are changing by natural selection. *Genetics* **52**, 875–890.

Kingsolver, J.G., Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hill, C.E., Hoang, A., Gilbert, P. and Beerli, P. (2001). The strength of phenotypic selection in natural populations. *American Naturalist* **157**, 245–261.

Koenig, W.D. and Albano, S.S. (1986). On the measurement of sexual selection. *American Naturalist* **127**, 403–409.

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution* **30**, 314–334.

Lande, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. *Evolution* **33**, 402–416.

Lande, R. (1982). A quantitative genetic theory of life history evolution. *Ecology* **63**, 607–615.

Lande, R. and Arnold, S.J. (1983). The measurement of selection on correlated characters. *Evolution* **37**, 1210–1226.

Lenski, E.E. and Service, P.M. (1982). The statistical analysis of population growth rates calculated from schedules for survivorship and fecundity. *Ecology* **63**, 655–662.

Lessard, S. and Castilloux, A.-M. (1995). The fundamental theorem of natural selection in Ewens' sense (case of fertility selection). *Genetics* **141**, 733–742.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

Maddox, G.D. and Antonovics, J. (1983). Experimental ecological genetics in *Plantago*: a structural equation approach to fitness components in *Plantago aristata* and *Plantago patagonica*. *Ecology* **64**, 1092–1099.

Mitchell-Olds, T. (1987). Analysis of local variation in plant size. *Ecology* **68**, 82–87.

Mitchell-Olds, T. and Bergelson, J. (1990). Statistical genetics of *Impatiens capensis*. II. Natural selection. *Genetics* **124**, 417–421.

Mitchell-Olds, T. and Shaw, R.G. (1987). Regression analysis of natural selection: statistical inference and biological interpretation. *Evolution* **41**, 1149–1161.

Mousseau, T.A. and Roff, D.A. (1987). Natural selection and the heritability of fitness components. *Heredity* **59**, 181–197.

Nagylaki, T. (1976). The evolution of one- and two-locus systems. *Genetics* **83**, 583–600.

Nagylaki, T. (1977a). *Selection in One- and Two-Locus Systems*. Springer-Verlag, Berlin.

Nagylaki, T. (1977b). The evolution of one- and two-locus systems. II. *Genetics* **85**, 347–354.

Nagylaki, T. (1989). Rate of evolution of a character without epistasis. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 1910–1913.

Nagylaki, T. (1991). Error bounds for the fundamental and secondary theorems of natural selection. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 2402–2406.

Nagylaki, T. (1992). Rate of evolution of a quantitative character. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 8121–8124.

Nagylaki, T. (1993). The evolution of multilocus systems under weak selection. *Genetics* **134**, 627–647.

O'Donald, P. (1968). Measuring the intensity of selection. *Nature* **220**, 197–198.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London, Series A* **2001**, 1–66.

Phillips, P.C. and Arnold, S.J. (1989). Visualizing multivariate selection. *Evolution* **43**, 1209–1222.

Price, G.R. (1970). Selection and covariance. *Nature* **227**, 520–521.

Price, G.R. (1972a). Extension of covariance selection mathematics. *Annals of Human Genetics* **35**, 485–490.

Price, G.R. (1972b). Fisher's 'fundamental theorem' made clear. *Annals of Human Genetics* **36**, 129–140.

Prout, T. (1965). The estimation of fitness from genotypic frequencies. *Evolution* **19**, 546–551.

Prout, T. (1969). The estimation of fitness from population data. *Genetics* **63**, 949–967.

Reyment, R.A. (1991). *Multidimensional Palaeobiology*. Pergamon Press, Oxford.

Robertson, A. (1955). *Population Genetics: The Nature and Causes of Genetic Variability in Populations*, Cold Spring Harbor Symposia on Quantitative Biology 20. The Biological Laboratory, Cold Spring Harbor, NY, pp. 225–229.

Robertson, A. (1960). A theory of limits in artificial selection. *Proceedings of the Royal Society of London, Series B* **153**, 234–249.

Robertson, A. (1966). A mathematical model of the culling process in dairy cattle. *Animal Production* **8**, 95–108.

Robertson, A. (1968). In *Population Biology and Evolution*, R.C. Lewontin, ed. Syracuse University Press, Syracuse, NY, pp. 5–16.

Roff, D.A. (1997). *Evolutionary Quantitative Genetics*. Chapman & Hall, New York.

Roff, D.A. and Mousseau, T.A. (1987). Quantitative genetics and fitness: lessons from *Drosophila*. *Heredity* **58**, 103–118.

Rose, M. (1982). Antagonistic pleiotropy, dominance and genetic variance. *Heredity* **48**, 63–78.

Scheiner, S.M. (1989). Variable selection along a successional gradient. *Evolution* **43**, 548–562.

Schluter, D. (1988). Estimating the form of natural selection on a quantitative trait. *Evolution* **42**, 849–861.

Stewart, S.C. and Schoen, D.J. (1987). Patterns of phenotypic variability and fecundity selection in a natural population of *Impatiens pallida*. *Evolution* **41**, 1290–1301.

Stratton, D.A. (1992). Life-cycle components of selection in *Erigeron annuus*: I. Phenotypic selection. *Evolution* **46**, 92–106.

Sutherland, W.J. (1985a). Chance can produce a sex difference in variance in mating success and explain Bateman's data. *Animal Behaviour* **33**, 1349–1352.

Sutherland, W.J. (1985b). *Oxford Surveys in Evolutionary Biology, Vol. 2*, R. Dawkins and M. Ridley, eds. Oxford University Press, New York, pp. 90–101.

Sved, J.A. (1989). *Evolution and Animal Breeding*, W.G. Hill and T.F.C. Mackay, eds. CAB International, Wallingford, CT, pp. 113–120.

Tallis, G.M. (1987). Ancestral covariance and the Bulmer effect. *Theoretical and Applied Genetics* **73**, 815–820.

Tallis, G.M. and Leppard, P. (1988). The joint effects of selection and assortative mating on multiple polygenic characters. *Theoretical and Applied Genetics* **75**, 278–281.

Trail, P.W. (1985). The intensity of selection: intersexual and interspecific comparisons require consistent measures. *American Naturalist* **126**, 434–439.

Travis, J. and Henrich, S. (1986). Some problems in estimating the intensity of selection through fertility differences in natural and experimental populations. *Evolution* **40**, 786–790.

Turelli, M. and Barton, N.H. (1990). Dynamics of polygenic characters under selection. *Theoretical Population Biology* **38**, 1–57.

Turelli, M., Gillespie, J.H. and Lande, R. (1988). Rate tests for selection on quantitative characters during macroevolution and microevolution. *Evolution* **42**, 1085–1089.

Walsh, B. (2002). *Quantitative Genetics, Genomics, and Plant Breeding*, M.S. Kang, ed. CAB International, Wallingford, CT, pp. 23–32.

Walsh, B. and Lynch, M. (2003). *Evolution and Selection of Quantitative Traits*. Sinauer, Sunderland, MA. Draft chapters can be found on the book website at `http://nitro.biosci.arizona.edu/zbook/volume`$_2$`/vol2.html`..

Willis, J.H. (1996). Measures of phenotypic selection are biased by partial inbreeding. *Evolution* **50**, 1501–1511.

Yule, G.U. (1902). Mendel's laws and their probable relation to intra-racial heredity. *New Phytologist* **1**, 193–207, 222–238.

Part 4

Animal and Plant Breeding

18

Quantitative Trait Loci in Inbred Lines

R.C. Jansen

Groningen Bioinformatics Centre, University of Groningen, Groningen, The Netherlands

Quantitative traits result from the influence of multiple genes (quantitative trait loci) and environmental factors. Detecting and mapping the individual genes underlying such 'complex' traits is a difficult task. Fortunately, populations obtained from crosses between inbred lines are relatively ideal for this – at least far more ideal than livestock and human populations – and true multigenic models are now available and have been applied successfully. In this chapter we will introduce the reader to statistical tools for segregation analysis and genetic mapping with the aid of molecular markers.

18.1 INTRODUCTION

18.1.1 Mendelian Factors and Quantitative Traits

The breeding experiments of Mendel in the late nineteenth century led to the foundation of modern genetics. He crossed white- and purple-flowering pea plants, and proposed a simple theory to explain the observed frequencies of white- and purple-flowering plants from generation to generation. With today's knowledge, we can conclude that *one* gene with *qualitative* effect encoded the flower colour studied by Mendel. If many genes and/or genes with quantitative effects encode a trait of interest, then the effects of the individual genes can hardly be distinguished. Therefore dissection of truly quantitative variation into the underlying 'Mendelian factors' is difficult on the basis of phenotypic information only. In this chapter we will see what modern statistical and molecular tools we have in our toolbox for studying quantitative or complex traits in inbred line crosses. These tools are of prime importance for plant breeding, livestock improvement and medical research using model organisms, because application of advanced tools offers great opportunities for understanding and manipulation of complex biological processes. Genes underlying quantitative or complex traits are commonly called *quantitative trait loci* (hereafter QTLs).

We start with a very limited description of inbred line genetics and breeding – all you need to know about inbred lines, seen through a statistician's spectacles – and the reader can generalise if he or she wishes. The next sections deal with what we think are the essentials for proper statistical modelling and analysis, first for so-called segregation analysis and second for genetic mapping of QTLs with the aid of genetic markers. The reader will see that similar statistical models and algorithms will be used for segregation analysis as for QTL mapping with the aid of molecular markers. We will introduce much of the relevant statistical machinery in the section on the simpler 'marker-free' case of segregation analysis.

There is a large amount of literature, and within the scope of this chapter it is hardly possible to refer to all statistical methods and applications published. The author has made his own selection, but hopes that this text, the bibliography and references give the reader sufficient entries to find his own favourite way in the world of QTLs in inbred lines. We sometimes get the feeling that the statistical theory of genetics is presented with a semblance of mystery, as if it were very special and complex. We believe there is no reason for this in the case of QTLs in inbred lines – almost all the theory is closely related to standard analysis of variance (ANOVA) and regression analysis. Nevertheless, we will make numerous cautionary remarks – standard methods can still be greatly abused.

18.1.2   The Genetics of Inbred Lines

In order to understand the genetics of inbred lines, we first need to define what one commonly understands by the terms *inbred* and *line*, and then we can discuss what 'sort of genetics' is involved. Here, we consider only inbred line crosses of diploid organisms. The reader may generalise to similar theoretical results for polyploid organisms and biparental crosses between outbreeding lines. In all cases we start with two parental lines, say P_1 and P_2, and intercross them in order to generate new genetic combinations.

In the genetics literature a *line* is a set of genetically related individuals, which are maintained under one and the same breed identification and/or commercial name. Chromosomes occur in homologous pairs, one originating from the mother, the other from the father. A line P_1 is *homozygous* if at any given gene the maternal and paternal states or *alleles* are identical, say a_1a_1. Thus, any offspring from crossing or mating within the P_1 pool will have the same P_1 genotype. The genotype of P_2 is denoted by a_2a_2. How do you get a line to become homozygous? For instance, in plants by following a single-seed-descent strategy over multiple generations: each generation you randomly take, grow and self a single individual in order to generate the seeds of the next generation of the line. By expectation, the number of heterozygous loci, a_1a_2, will be halved each generation. After eight to ten generations one can stop as the breeding process has led to an (almost) homozygous 'inbred line'. Of course, this scheme is typical of plants and cannot be used in animals. For model animals (mice, rats, hamsters) and livestock animals (chickens) simple designs of brother–sister mating can be used. Various plant species (apple, strawberry, pine) and most animal species (cattle, horses) can hardly be manipulated this way, because it would take too much time and money, because the organism would suffer from severe inbreeding depression, and, last but not least, because it would be immoral (humans). Although we focus here on biparental crosses between divergent inbred lines, most of the theory also applies to the case of a biparental cross between divergent outbreeding lines. In the latter case, there can be up to four different alleles per locus. We present theory for the simpler two-allele situation and thus leave the extension to four alleles to the reader.

There follows a short description of a number of mating designs. All start with a cross between two divergent inbred lines P_1 and P_2, which generates heterozygous filial F_1-offspring ($a_1a_1 \times a_2a_2 \to a_1a_2$). In the backcross (BC) design, the F_1 is (back)crossed to one of the parents, say female F_1 to male P_1 ($a_1a_2 \times a_1a_1 \to a_1a_2$ and a_1a_1 for each locus). In the doubled haploids (DH) design, either male or female gametes of the F_1 are artificially doubled by some kind of treatment, giving rise to homozygous recombinant offspring in one step ($a_1a_2 \to a_1a_1$ and a_2a_2 per locus). In the BC and DH design one can only trace segregation and recombination events in either the male or the female F_1 gamete. In the filial F_2 design, the F_1 is selfed or two F_1 individuals are crossed, in order to generate offspring, resulting from recombination events in both male and female gamete production ($a_1a_2 \times a_1a_2 \to a_1a_1$, a_1a_2 and a_2a_2 in the ratio $1:2:1$ per locus). Finally, in the recombinant inbred line (RIL) design one first generates the F_2 progeny and then each F_2 enters individually a single-seed-descent inbreeding programme ($a_1a_2 \times a_1a_2 \to a_1a_1$ or a_2a_2 per locus). From here on we us the notation A, H and B for a_1a_1, a_1a_2 and a_2a_2, respectively.

18.1.3 Phenotype, Genotype and Environment

The basic model of quantitative genetics is

$$P = G + E,$$

or, in words, the phenotype (trait) is the sum of genetic and environmental factors. By taking two divergent lines, we have a high chance that the two lines have different states (alleles) at all the genes underlying the traits of interest. Unfortunately, the trait difference between the two parents reflects the total effect of the their genes and not their individual gene effects. In statistical terms, the effects of the genetic factors are *confounded*, the factors are *aliased* or *collinear*. The genetic reproduction mechanism will yield us offspring with new allele combinations, generated by independent segregation of different chromosomes and by recombination within chromosomes. Therefore in the offspring, genes on different chromosomes are independent stochastic variables (*unlinked* genes). In contrast, the genes on the same chromosome will show statistical dependence or *linkage association*, although this may be negligible if they are far apart, and we speak of *linked* genes. Unlinked genes will become orthogonal factors (in infinite populations) and closely linked genes become factors with a high degree of collinearity. Any recombination between two flanking genes increases our chance of dissecting their effects. The distance between two loci is usually reported in units of *morgans* (M) or *centimorgans* (cM) using Haldane's mapping function. According to Haldane, a recombination frequency of r between two loci corresponds to a map distance of m morgans, with

$$m = -\frac{1}{2}\ln(1 - 2r);$$

conversely, a distance of m morgans corresponds to a recombination frequency r of

$$r = \frac{1}{2}(1 - \exp(-2m)).$$

For small distances the relation is more or less proportional ($m = 0.01$ M $\sim r = 0.01$, in which case one in 100 individuals of a DH population will show up as recombinant).

18.2   SEGREGATION ANALYSIS

18.2.1   Visualisation of Quantitative Variation in a Histogram

Our first step, once we have collected the phenotypic data, will most probably be to rank the data and draw a histogram to visualise the results for the trait under study. This is 'just' descriptive statistics, but it is an important step, as conclusions drawn from descriptive statistics often have more practical impact than those from inferential statistics.

Figure 18.1 shows typical histograms. What can we learn from these pictures? And, are there any pitfalls? What we know from lessons in genetics is that homologous chromosomes segregate randomly, creating opportunity for 2^n different egg or sperm cells per meiosis and an equal amount of possible homozygous genotypes for an organism with n chromosomes pairs. Nature is even cleverer than that, and devised the crossover mechanism as the result of which the total number of possible genotypes is much bigger even than 2^n so that usually no two individuals in the population have the same genotype. What we see in Figure 18.1 is a 'mixture' of as many different genotypes as there are individuals.

Figure 18.1(a) clearly shows a bimodal distribution for a segregating DH population. If this is the case for our trait, we will probably be excited! We recognise that the mixture of genotypes falls into two different groups, which can be labelled A and B, respectively. Of course, it is likely that the two labels correspond to the two *genotypes* A and B at a yet unmapped major gene. One might say that the action of this gene is semiqualitative or semiquantitative, i.e. somewhere between pure qualitative, as with Mendel's pea genes, and pure quantitative, as with many complex traits such as body weight or crop yield. The action of our major gene may or may not be modified by other genes of minor effect, but the information in the histogram is too limited for us to come to conclusions on this aspect. In some cases trait values of a number of individuals for each of the parents are available. If a parent line is homozygous, all individuals of that line are genetically identical and any variation observed among them must therefore be environmental in origin. This
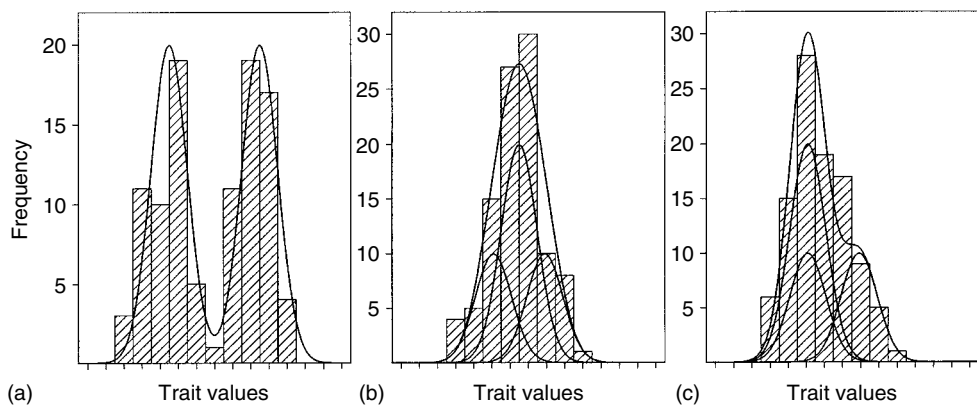
Figure 18.1   Mixture distributions plotted on top of the histograms (a) for a DH population, (b–c), for F_2 populations. The component distributions for each of the genotypes A, H and B are also plotted in (b) and (c).

would make it possible to compare within-parents variation, which is supposed to be environmental and not genetic, to within-A and within-B variation, which is the sum of environmental and possibly residual genetic variation.

Figure 18.1(b) shows a histogram with a unimodal distribution for a segregating F_2 population. It is not unlikely that our first reaction will be 'a real pity, my trait is complex, probably encoded by many genes of small action'. If many genes encode our trait, then we would expect a 'normal' distribution, as shown in Figure 18.1(b). Again, comparing with parental data, we can get an idea of the amount of genetic variation relative to environmental variation. However, our questions 'how many genes of what sizes of effects', will not be answered unless we invoke the aid of molecular techniques – this will be the topic of the next sections. Are there any potential pitfalls? The reader may feel misled, but the histogram in Figure 18.1(b) actually represents the case of a single additive major gene explaining half of the total variation! The segregating population consists of a mixture of genotypes A, H (heterozygote) and B in the ratio $1 : 2 : 1$. Thus even if we observe a unimodal distribution, we may still hope that the underlying genetics is not too complex.

Figure 18.1(c) shows another histogram with a unimodal distribution for a segregating F_2 population, but this time the distribution is skewed. This may give us a hint of dominant major gene action! Suppose individuals with genotypes A and H group together (i.e. have the same mean value) and underlie the main body of the distribution. The other individuals, with genotype B, underlie the 'shoulder' of the distribution. There is little by way of a proof in the picture, but we may think there is a lot of evidence for our hypothesis of a single dominant gene. Again, there is a chance of a serious pitfall. We may have made the implicit assumption that the error distribution or 'component distribution' is symmetric or even 'normal'. Many traits have a natural lower bound (often zero), and variances often increase with the mean. In reality our histogram was generated under a (polygenic) model with a skewed residual error distribution. Thus, seeing a skewed distribution may lead us to believe erroneously that there is a dominant major gene. Some researchers are aware of this pitfall and transform their data, for instance they take the natural logarithm of all trait values, thereby changing the scale of the x axis to a logarithmic scale on which the distribution looks more symmetrical. Is this a good procedure? Not necessarily, because there is the danger of throwing out the baby with the bathwater: we may loose power for detecting dominant gene action. However, there may sometimes be good biological reasons to log transform. To check the type of distribution one can best look at parental (non-mixture) data and use standard statistical techniques to help find an optimal data transformation – see Atkinson (1985), Jansen and Den Nijs (1993) and Jansen *et al.* (1993) for illustrative examples in the case of mixture models. Choice of mixing distribution becomes more critical when the model becomes more complex (e.g. if models include gene interactions). It is well known, for instance, that significant interactions can arise as statistical artefacts: the interactions try to compensate for wrong model assumptions. Still, distribution checks are only occasionally reported in genetic analysis.

Quantitative data are not always continuous data. Typically, quantification of disease scores may lead to count data (e.g. number of spots on a leaf), categorical or ordinal data (e.g. a disease score $1-5$) or proportions (e.g. number of affected individuals per DH line being tested on multiple seedlings). In many cases the distribution is close to normal after appropriate data transformation (e.g. log, square root, or probit). The type of transformation should preferably be chosen on the basis of non-mixture data. See McCullagh and Nelder (1989) for generalised linear models (GLMs) for many types of distribution.

Finally we refer the readers to Allard (1999) for several illustrative graphs (his Figures 18.1 and 18.2), showing qualitative and quantitative variation resulting from segregation for one, two, or more genes with or without dominance and/or in the presence or absence of environmental noise.

18.2.2   Plotting Mixture Distributions on Top of the Histogram

Histograms may show bimodality or skewness and as such they can indicate segregation of major genes. Still we cannot unambiguously assign individuals to genotypes. In statistical terms, a QTL is a latent unobserved categorical variable. Let us look again at Figure 18.1. How can we model, fit and plot the mixture distributions on top of the histogram?

Let us start with a relatively simple model for Figure 18.1(b): a model with a single QTL with *additive* allele effect and normal error in an F_2. Let y_i denote the trait value of the ith individual. We do not know its QTL genotype. There are three possible genotypes, each with its own probability of occurrence and component probability density function (PDF):

			P	*PDF*
1	y_i	A	0.25	$\phi(y_i; \mu_A, \sigma^2)$
2	y_i	H	0.50	$\phi(y_i; \mu_H, \sigma^2)$
3	y_i	B	0.25	$\phi(y_i; \mu_B, \sigma^2)$

Here

$$\phi(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(y - \mu)^2/2\sigma^2),$$

is the PDF of the normal (Gaussian) distribution with mean μ and variance σ^2, and

$$\mu_H = \tfrac{1}{2}(\mu_A + \mu_B),$$

because we assumed additivity of allele effects. Taking the weighted sum over the three components, we get the mixture PDF

$$f_{\text{mix}}(y_i) = \frac{1}{4}\phi(y_i; \mu_A, \sigma^2) + \frac{1}{2}\phi\left(y_i; \frac{\mu_A + \mu_B}{2}, \sigma^2\right) + \frac{1}{4}\phi(y_i; \mu_B, \sigma^2).$$

The simultaneous mixture likelihood of observations $y_1, y_2, \ldots, y_n$ is the product of the individual likelihood contributions:

$$L = \prod_{i=1}^{n} f_{\text{mix}}(y_i).$$

In order to shorten the notation we will often use $\phi_A(\cdot)$, $\phi_H(\cdot)$ and $\phi_B(\cdot)$ to denote the distributions for genotypes A, H and B, respectively.

Suppose – for the moment – that we know the parameter values. How do we show the fit of the model to the data? That is fairly easy: calculate $f_{\text{mix}}(\cdot)$ for values ranging from the minimum to the maximum trait value, multiply by the number of individuals n, and by the width w of the groups of the histogram to account for scaling, and finally plot this on top of the histogram. It will become a little more complicated if we want to show the histogram on the original scale of measurement, while we think that it is more appropriate to fit the mixture of normal distributions on a different scale. Suppose that

we fit the normal mixture model to the transformed trait values $T(y_i)$, i.e. on a different scale than the original scale of measurement. The protocol is now as follows: calculate $f_{\text{mix}}(T(\cdot))$ for values ranging from the minimum to the maximum trait value on the original scale of measurement, multiply by the Jacobian $T'(\cdot)$ to calculate the likelihood on the original scale of measurement, next multiply by n and by w, and finally plot this on top of the histogram. The Jacobian $T'(\cdot)$ is the first-order derivative of the transformation function – e.g. if $T(y) = \ln(y)$, then $T'(y) = 1/y$.

18.2.3 Fitting Mixture Distributions

What remains is the question of how we estimate the parameters of the finite (normal) mixture distribution. The standard maximum likelihood (ML) approach consists of finding the parameter values which maximise the simultaneous (log) likelihood. To achieve this we can take the first-order derivatives of the log likelihood l and set these to zero:

$$0 = \frac{\partial}{\partial\theta} l = \frac{\partial}{\partial\theta} \log\left(\prod_{i=1}^{n} f_{\text{mix}}(y_i)\right) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_{\text{mix}}(y_i).$$

We continue with our example of a single QTL with additive effect in an F_2. Now note that

$$\begin{aligned}
\frac{\partial}{\partial\theta} \log f_{\text{mix}}(y_i) &= \frac{1}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta}[0.25\phi_A(y_i) + 0.50\phi_H(y_i) + 0.25\phi_B(y_i)] \\
&= \frac{0.25}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta}\phi_A(y_i) + \frac{0.50}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta}\phi_H(y_i) + \frac{0.25}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta}\phi_B(y_i) \\
&= \frac{0.25\phi_A(y_i)}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta} \log\phi_A(y_i) + \frac{0.50\phi_H(y_i)}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta} \log\phi_H(y_i) \\
&\quad + \frac{0.25\phi_B(y_i)}{f_{\text{mix}}(y_i)} \frac{\partial}{\partial\theta} \log\phi_B(y_i) \\
&= P(A|y_i)\frac{\partial}{\partial\theta} \log\phi_A(y_i) + P(H|y_i)\frac{\partial}{\partial\theta} \log\phi_H(y_i) + P(B|y_i)\frac{\partial}{\partial\theta} \log\phi_B(y_i),
\end{aligned}$$

which you can recognise as a sum of weighted 'normal' likelihood contributions, where the weights are conditional probabilities of the genotype given the observed phenotype $P(A|y)$, $P(H|y)$ and $P(B|y)$, which depend on the unknown θ (to shorten the notation we write $P(\cdot)$ instead of $P_\theta(\cdot)$). Unfortunately the likelihood equations cannot be solved analytically. But there is a simple – thus popular – algorithm, called the *expectation-maximisation* (EM) algorithm – for short the EM algorithm. Dempster *et al.* (1977) considered the mixture problem as one of many examples in which data are incomplete. They interpreted mixture data as incomplete data by regarding an observation on the mixture as missing its component of origin. See their Section 4.3 on finite mixtures. In our context the information on QTL genotype is missing (with three components A, H and B). The basic idea of the iterative EM algorithm is to replace the incomplete observation y_i by its three complete observations (y_i, A), (y_i, H) and (y_i, B), weighting the three complete observations by specified or updated (conditional) probabilities. Iteration consists of two steps:

(E step) Specify or update weights $P(A|y_i)$, $P(H|y_i)$ and $P(B|y_i)$.

(M step) Update estimates of μ_A, μ_B and σ^2.

In the E step, conditional probabilities are calculated by using the current parameter estimates. The M step consists of weighted regression analysis on the triplicate data set, which can be done with most statistical packages, and which requires no more than a routine for weighted least squares. The explicit solution of the M step can be written as

$$\hat{\mu}_A = \frac{\sum_{i=1}^{n}\left[P(A|y_i)y_i + P(H|y_i)\frac{1}{2}y_i\right]}{\sum_{i=1}^{n}\left[P(A|y_i) + P(H|y_i)\frac{1}{2}\right]}, \quad \hat{\mu}_B = \frac{\sum_{i=1}^{n}\left[P(B|y_i)y_i + P(H|y_i)\frac{1}{2}y_i\right]}{\sum_{i=1}^{n}\left[P(B|y_i) + P(H|y_i)\frac{1}{2}\right]},$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\left[P(A|y_i)(y_i - \hat{\mu}_A)^2 + P(H|y_i)\left(y_i - \frac{1}{2}(\hat{\mu}_A + \hat{\mu}_B)\right)^2 + P(B|y_i)(y_i - \hat{\mu}_B)^2\right].$$

Setting parameters equal to (well-chosen) initial values conveniently starts the algorithm.

We will now use the above rather simple example to introduce the general concept of data completion (augmentation) and parameter estimation via iterative reweighted least squares. Hopefully this will help the reader to understand the more complicated cases that will appear in later sections. Let $\mathbf{y}^{(c)}$ denote the $1 \times 3n$ vector of augmented trait data $(y_1, y_1, y_1, y_2, y_2, y_2, \ldots, y_n, y_n, y_n)'$, where the superscript (c) is used for 'complete'. Furthermore, let $\boldsymbol{\beta}$ denote the $1 \times p$ vector of regression parameters, $\mathbf{X}^{(c)}$ the corresponding design matrix of size $3n \times p$, $\mathbf{e}^{(c)}$ the vector or residuals, and finally $\mathbf{W}^{(c)}$ is the diagonal matrix of conditional probabilities $(P(A|y_1), P(H|y_1), P(B|y_1), P(A|y_2), P(H|y_2), P(B|y_2), \ldots, P(A|y_n), P(H|y_n), P(B|y_n))'$. Note that for the ith individual the model is

$$\begin{pmatrix} y_i \\ y_i \\ y_i \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{pmatrix}, \quad \text{weighted by} \begin{pmatrix} P(A|y_i) \\ P(H|y_i) \\ P(B|y_i) \end{pmatrix},$$

or in matrix notation $\mathbf{y}^{(c)} = \mathbf{X}^{(c)}\boldsymbol{\beta} + \mathbf{e}^{(c)}$ with weight matrix $\mathbf{W}^{(c)}$. The M step is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{(c)\prime}\mathbf{W}^{(c)}\mathbf{X}^{(c)})^{-1}\mathbf{X}^{(c)\prime}\mathbf{W}^{(c)}\mathbf{y}^{(c)}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{Y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}})'\mathbf{W}^{(c)}(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}}).$$

18.2.4 Wanted: QTLs!

We want to discover all about the QTLs underlying quantitative variation for our trait(s) of interest. How many genes are involved? Where are they located on the chromosomes? What type of (inter)action do they show? It is clear from the Figure 18.1(b), that for truly quantitative variation the effects of the individual genes can hardly be distinguished. But there are clever ways to tackle this problem: use molecular markers, and this is the topic

of the following sections! Thus, we now proceed to apply the mixture model tools in a new context where we have partial information, provided by molecular markers, on the underlying genotypes.

18.3 DISSECTING QUANTITATIVE VARIATION WITH THE AID OF MOLECULAR MARKERS

18.3.1 Molecular Markers

Since the early 1980s new ways to unravel complex traits have emerged (Botstein *et al.*, 1980; Beckmann and Soller, 1983). The magic phrase is 'molecular marker'. Remember that a QTL is – in a statistical sense – a categorical variable whose values remain unobserved. A molecular marker is also a locus on the genome, but the genotype can be observed with molecular tools. From the statistical point of view it hardly matters how this molecular technique works; the only important point is that a marker is a categorical variable with observable state. Nowadays hundreds, not to say thousands, of molecular markers are available or will become available over the next few years, all with more or less *known* positions on the genome. In this chapter the focus will be on dense or ultradense marker maps. Which means that there is almost always a marker almost on top of any QTL. By 'on top' we mean that in the segregating population no recombination events between the given marker and QTL have appeared.

Markers are often just non-functional or selectively neutral sites. So why should it help to collect information about such loci? It will not help if our population is similar to a random mating population in *linkage equilibrium*, which means that genotypes at different loci are statistically independent. But in this chapter we deal with inbred line crosses for which linkage disequilibrium is at its maximum, i.e. there is maximum statistical correlation between the genotypes at linked loci. For example, between an observable marker that does not affect the phenotype and an unobservable QTL that does affect the trait. Say the homozygous parents P_1 and P_2 have marker-QTL genotype $m_1a_1//m_1a_1$ and $m_2a_2//m_2a_2$, respectively (chromosomes occur in pairs, and the symbol '//' separates the alleles of the first and second chromosome of the pair; the marker has alleles m_1 and m_2; the QTL has alleles a_1 and a_2). The F_1 genotype is $m_1a_1//m_2a_2$. It will produce a mixture of non-recombinant gametes (m_1a_1 or m_2a_2) with probability $1 - r$, and recombinant gametes (m_1a_2 or m_2a_1) with probability r. Figure 18.2 is based on the same data as shown in Figure 18.1(a), and shows the effect of tight linkage between marker and QTL on the mixture distribution ($r = 0.01$). There would be weak or no association if the marker and QTL were far apart. Thus, if the marker is close to or on top of a QTL, then the two marker genotype groups (say, A for m_1m_1 and B for m_2m_2) will clearly show different means for the trait. The basic and simple idea is to reverse this statement – if the two marker classes clearly show different means, then there is evidence for 'QTL activity' in the neighbourhood. Note that we have made the reverse statement a little less strong, i.e. we cannot claim that there is only one *isolated* QTL involved – there is much more to say about this, but we will postpone this to one of the later sections of this chapter.

Whether two marker classes indeed show different means (or not) can be tested with standard Analysis Of Variance ANOVA and regression tools. Suppose that we fit a single QTL at a fully informative marker in an F_2 population. The one-way classification model
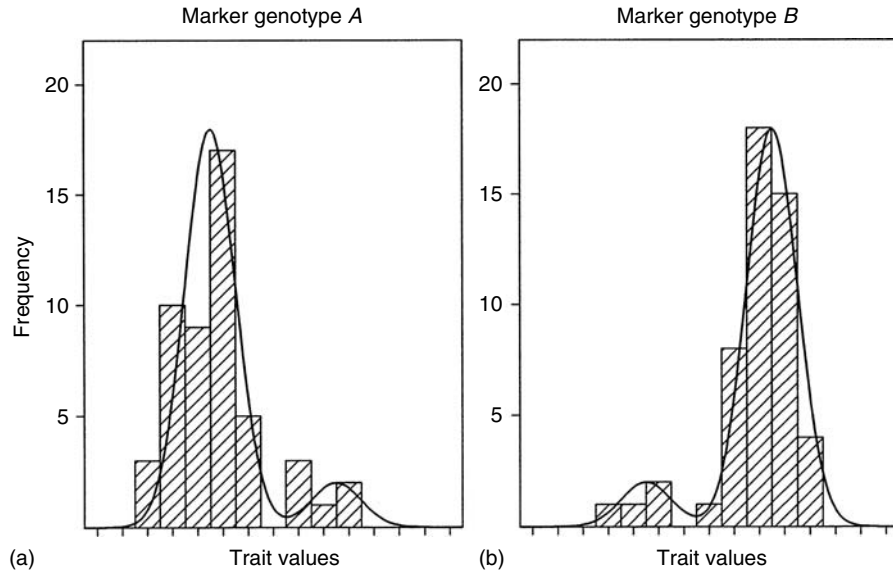
Figure 18.2 Mixture distributions plotted on top of the histograms per marker genotype for the same data as shown in Figure 18.1(a). The marker is close to a QTL, but a few recombinations have occurred. Therefore, the distribution given the marker genotype is a mixture over recombinant (probability r) and non-recombinant individuals (probability $1 - r$).

reads $y_{ij} = \mu + \beta_i + \varepsilon_{ij}$ for the jth individual of the ith genotype ($i = 1, 2, 3$; A, H and B at the marker). This model can easily be extended to two or more marker factors (Cowen, 1989; Stam, 1991). Assuming that all scores of marker factors are known – i.e. their data are complete – the general model for regression on multiple markers is $\mathbf{y}^{(c)} = \mathbf{X}^{(c)}\boldsymbol{\beta} + \mathbf{e}^{(c)}$. A characteristic of ANOVA and regression models is that they often are overparameterised, containing more parameters than needed to represent the effects. Usually, setting the sum of allele effects to zero or, equivalently, working in terms of main effect and allele substitution effects compensates for this. The parameters are commonly estimated by the least-squares method, i.e.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{(c)\prime}\mathbf{X}^{(c)})^{-1}\mathbf{X}^{(c)\prime}\mathbf{y}^{(c)}$$

and

$$\hat{\sigma}^2 = \frac{1}{n - p}(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}})'(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}}).$$

Note that in the latter formula one divides by $n - p$, the degrees of freedom (df) of the residual error, and not by n; n is the population size, p is the number of regression parameters in $\boldsymbol{\beta}$.

18.3.2 Mixture Models

What happens if some marker scores are missing? ANOVA and regression models do not allow for missing values in the explanatory variables. We could eliminate any individual with one or more missing marker scores and then perform an analysis on the remainder

of the data set. But in real experiments most individuals are missing at least some marker scores, so this approach does not seem to be very attractive. Of course, we can think of filling in the gaps *prior to ANOVA analysis* by using the map information. How does this work? Look at the markers flanking the marker with a missing score. If the markers are close together, we can be pretty sure how to fill in the score and then we can apply ANOVA and regression! This is an *ad hoc* procedure, which works well for ultradense maps, but there are more sophisticated methods available for sparse maps. As we have seen in Section 18.2, data completion can be integrated with analysis of quantitative variance – and the type of models we need are mixture models.

We should not be surprised if up to 5 % of the marker scores are missing in a real experiment. Often some markers or some individuals are difficult to score. This can be caused by various technical failures such as a given individual's bad DNA sample or problems with the fingerprint image. The statistician assumes that the scores are missing at random, or at least that there is no tendency that one score, say A, is relatively more frequently missing than B or H. Another type of incomplete data occurs when one or more markers are *dominant* and scored in, for instance, an F_2 population. Such a dominant marker still has three different states A, H and B, like a codominant marker. However, for technical reasons either A and H cannot be discriminated, which is often labelled as 'not-B' or just D, or H and B cannot be discriminated, labelled as 'not-A' or just C. A dominant marker is, in statistical terms, an observable categorical variable with two observable states. We can therefore use such a variable straightforwardly in ANOVA. On the other hand, we may want to complete the marker data by again using information from neighbouring markers. For instance, a 'not-A' observation means that the genotype is B or H, and the first one is more likely if we know that an adjacent marker has score B. How successfully can we complete data? This can be quantified by what one calls the *multilocus information content*; this will be defined later. We finish by remarking that the distinction between dominant and codominant markers is not always so absolute. Some types of marker may yield A, H, B, C and D, that is there are five categories (e.g. AFLP$^{©}$ markers).

Currently, marker maps are still not ultradense, and this brings us to the third type of incomplete data. Suppose that we have complete marker data, but that our marker map is sparse, that is the markers are spread coarsely over the genome. Our data will show many recombinants between any pair of flanking markers and, if there is a QTL in between, then it is likely that none of the markers is 'on top' of the QTL. In other words, a flanking marker can have the same genotype as the QTL, but not for all individuals at the same time. Therefore, ANOVA on a flanking marker is no longer identical to ANOVA on the QTL. ANOVA may still work relatively well, but the results are biased or less powerful, because we have built-in errors by ignoring one or more recombinations between the marker and the QTL. Does statistics offer a solution to this problem? Any locus with incomplete data can be added to the model, whether it is a marker (as discussed above) or a QTL (which is a locus with all its scores missing). The statistical trick of data completion will work in any case, and it will help us to exploit the full multilocus information content. Basically, there is no obstacle to using multiple-QTL models. Over the past decade most theoretical papers on QTL mapping have been devoted to the sparse map situation. The older papers dealt with single-QTL mixture models only (see Weller, 1986; Jensen, 1989; Lander and Botstein, 1989; Simpson, 1989), methods that became known as *interval mapping* (IM). Jansen (1992) developed a general approach for fitting multiple-QTL models, an approach that became later known as *MQM* (multiple-quantitative trait loci mapping), and

was elaborated in a number of papers (Jansen, 1993b; 1994b; Jansen and Stam, 1994). Zeng (1994) published a similar strategy and called it *composite interval mapping* (CIM).

Let us now work out an example. For convenience, but without loss of generality, we will zoom in on one F_2 individual with trait value y_i and five loci with scores

$$A\ A\ D\ U\ A,$$

where D means *not-B* and U means *genotype unknown*. We suppose that the loci 1, 2, 3 and 5 are markers, and that the fourth locus is a QTL, which is located at a map position within the interval between loci 3 and 5. All genotype scores of the QTL are unknown. The observed data can be completed for the missing locus information, giving rise to six different complete genotypes:

		1	2	3	**4**	5	Simultaneous genotype probability	Component density
1	y_i	A	A	A	**A**	A	$\frac{1}{4}(1-q)^2(1-r)^2(1-s)^2(1-t)^2$	$\phi_A(y_i)$
2	y_i	A	A	A	**H**	A	$\frac{1}{4}(1-q)^2(1-r)^2\,2(1-s)s(1-t)t$	$\phi_H(y_i)$
3	y_i	A	A	A	**B**	A	$\frac{1}{4}(1-q)^2(1-r)^2s^2t^2$	$\phi_B(y_i)$
4	y_i	A	A	H	**A**	A	$\frac{1}{4}(1-q)^2\,2(1-r)r(1-s)s(1-q)^2$	$\phi_A(y_i)$
5	y_i	A	A	H	**H**	A	$\frac{1}{4}(1-q)^2\,2(1-r)r((1-s)^2+s^2)(1-t)t$	$\phi_H(y_i)$
6	y_i	A	A	H	**B**	A	$\frac{1}{4}(1-q)^2\,2(1-r)r(1-s)st^2$	$\phi_B(y_i)$

Suppose you know the frequencies of recombination q, r, s and t between the loci. The multilocus genotype probabilities can be derived straightforwardly: $\frac{1}{4}$ for 'genotype A at the first locus, then $(1-q)^2$ for no recombination between A on the first locus and the A at the second locus, etc. For the moment, our model will include only one explanatory variable – the fourth locus is a QTL – and we can assign component densities $\phi_A(y_i)$, $\phi_H(y_i)$ and $\phi_B(y_i)$. By multiplying each of the six genotype probabilities with the corresponding component densities, and then summing these products over the six genotypes, we can see that the contribution of this individual to the *multilocus* or *multipoint* simultaneous likelihood is

$$f_{\text{mix}}(y_i, AADUA) = P(AAAAA)\phi_A(y_i) + P(AAAHA)\phi_H(y_i) + P(AAABA)\phi_B(y_i)$$
$$+ P(AAHAA)\phi_A(y_i) + P(AAHHA)\phi_H(y_i) + P(AAHBA)\phi_B(y_i).$$

The crux is that we complete the explanatory variable(s), here the fourth locus, so that we can perform an ANOVA-like analysis. There is no need to complete the data for the other non-explanatory loci, and we can rewrite the above expression as

$$f_{\text{mix}}(y_i, AADUA) = [P(AAAAA) + P(AAHAA)]\phi_A(y_i)$$
$$+ [P(AAAHA) + P(AAHHA)]\phi_H(y_i)$$
$$+ [P(AAABA) + P(AAHBA)]\phi_B(y_i).$$

The genotype probabilities can easily be calculated recursively. The regression model for the ith individual is

$$\begin{pmatrix} y_i \\ y_i \\ y_i \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{pmatrix}, \text{ weighted by } \begin{pmatrix} P(A|y_i) \\ P(H|y_i) \\ P(B|y_i) \end{pmatrix},$$

in which

$$P(A|y_i) = \frac{[P(AAAAA) + P(AAHAA)]\phi_A(y_i)}{f_{\mathrm{mix}}(y_i)},$$

and $P(H|y_i)$ and $P(B|y_i)$ are calculated analogously. It is easy to add extra QTLs to the model. For instance, we can insert a second QTL (a locus with only U scores) in the first marker interval. Doing so, the PDF of the ith individual is

$$f_{\mathrm{mix}}(y_i, AUADUA) = \sum_{\substack{u=A,H,B \\ v=A,H,B}} [P(AuAAAvA) + P(AuAAHvA)]\phi_{uv}(y_i).$$

The general regression model, for all individuals together and for any number of markers and QTLs, reads $\mathbf{y}^{(c)} = \mathbf{X}^{(c)}\boldsymbol{\beta} + \mathbf{e}^{(c)}$, with weight matrix $\mathbf{W}^{(c)}$. The parameters can be conveniently estimated via the EM algorithm by iterating in two steps, similar to what we have seen for the marker-free case of segregation analysis:

(E step) Specify or update weights $P(\text{multilocus genotype } |y_i)$.

(M step) Update estimates of $\boldsymbol{\beta}$ and σ^2.

In the E-step conditional multipoint genotype probabilities are calculated by using the current parameter estimates. The M step is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{(c)\prime}\mathbf{W}^{(c)}\mathbf{X}^{(c)})^{-1}\mathbf{X}^{(c)\prime}\mathbf{W}^{(c)}\mathbf{y}^{(c)}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}((\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}})'\mathbf{W}^{(c)}(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}}).$$

In our small example presented above, data augmentation for the given individual gave rise to six different complete genotypes. However, different individuals may have different amounts of missing genetic (QTL and marker) information. In this augmentation approach there is no restriction on multiplying any other individual with any other value (3, 6, 10, 100, etc.) within the same analysis. For instance, 'triplicate' the jth individual with observed data $(y_j, AAAUA)$ to complete data $\{(y_j, AAAAA), (y_j, AAAHA), (y_j, AAABA)\}$.

Use of the EM algorithm (generally) yields Maximum Likelihood (ML) estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$. Jansen (1994b) proposed to use restricted ML, i.e. adjust the ML estimate $\hat{\sigma}^2$ for the df used. Divide by the df of residual error instead of by n, just as in ANOVA and regression analysis:

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}})'\mathbf{W}^{(c)}(\mathbf{y}^{(c)} - \mathbf{X}^{(c)}\hat{\boldsymbol{\beta}}).$$

As we will see in a later section, this can lead to several appealing advantages during selection and inference, at least if multipoint linkage information is high.

We would like to emphasise that we model the *simultaneous* likelihood of the trait and of multiple markers and QTLs. Genotypes are *multilocus* and genotype probabilities, are *unconditional* (Jansen, 1992; 1993b; 1994b). This is in contrast to most literature on inbred line analysis, in which one usually calculates the likelihood of the trait *conditional* on the two markers flanking the interval study (e.g. Lander and Botstein, 1989). If the flanking marker scores are missing or incomplete for a given individual, then the nearest informative marker in the same direction is taken. This 'conditional marker-QTL-marker approach' works well for simple single-QTL models, but is definitely less 'transparent' for multiple-QTL models. Only unconditional multilocus likelihoods can form the basis for direct comparison of the fit of several competing models.

How do we define multilocus information content at a given map location? In the ideal case, one of the conditional probabilities $P(A|y_i)$, $P(H|y_i)$, or $P(B|y_i)$ is 1 and the two others are 0. This happens only if the QTL is on top of a complete-data marker. In the worst case of no marker data, the three probabilities are 0.25, 0.50 and 0.25. One simple way to visualise the quality of the data works as follows. Take for each individual the most likely genotype and average the corresponding probabilities over all individuals. If the average is close to 1, then our data are excellent at the given map position. If it is close to 0.5, then the multimarker information is really poor. This is just our own definition of information content, and others exist. See, for instance, the definition, based on variances of conditional probabilities, in Spelman *et al.* (1996). Is it a major problem if the information content is substantially lower than 1? This has not been studied in much detail in the QTL literature, but it may not be much of a problem: Redner and Walker (1984) pointed out that small to moderate proportions of completely informative individuals are sufficient for good ML estimation in mixture problems.

18.3.3 Alternative Regression Mapping

Let us look again at our simple example in which the fourth locus was the QTL and

$$f_{\text{mix}}(y_i, AADUA) = [P(AAAAA) + P(AAHAA)]\phi_A(y_i)$$
$$+ [P(AAAHA) + P(AAHHA)]\phi_H(y_i)$$
$$+ [P(AAABA) + P(AAHBA)]\phi_B(y_i).$$

Note that $f_{\text{mix}}(y_i, AADUA) = f_{\text{mix}}(y_i|AADUA)P(AADUA)$. Conditional on observed genotype, the expected trait value of the ith individual, μ_i, is given by

$$P(AADUA)\mu_i = [P(AAAAA) + P(AAHAA)]\mu_A$$
$$+ [P(AAAHA) + P(AAHHA)]\tfrac{1}{2}(\mu_A + \mu_B)$$
$$+ [P(AAABA) + P(AAHBA)]\mu_B$$
$$= [P(AAAAA) + P(AAHAA) + \tfrac{1}{2}P(AAAHA) + \tfrac{1}{2}P(AAHHA)]\mu_A$$
$$+ [P(AAABA) + P(AAHBA) + \tfrac{1}{2}P(AAAHA) + \tfrac{1}{2}P(AAHHA)]\mu_B,$$

which we can write as

$$\mu_i = p_1\mu_A + p_2\mu_B.$$

Haley and Knott (1992) and Martinez and Curnow (1992) proposed to use the regression model for the ith individual

$$y_i = [p_1 p_2] \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} + \varepsilon_i,$$

and parameters can be estimated easily by the least-squares approach, if p_1 and p_2 are assumed to be known. Here we refer to this method as 'regression mapping'. In an ML framework, their approach would be equivalent to approximating the mixture density $f_{\text{mix}}(y_i)$ by a normal distribution $\phi(y_i; \mu_i, \tau^2)$. Figure 18.2 visualises the worst case of a major QTL for which the approximation can only be poor. In many cases, however, this approach works reasonably well. Where the EM algorithm often requires five to seven iterations of iterative weighted least squares, the regression approach obviously only needs one step.

18.3.4 Highly Incomplete Marker Data

Let us now briefly discuss problems with *highly* incomplete genetic data (e.g. if many marker scores are missing or if we postulate multiple QTLs). The PDF is then a sum over many different candidate genotypes, in which case computation may become time consuming. In the above example, the complete genotypes *AAABA* (no. 3) and *AAHBA* (no. 6) have negligibly small probabilities compared to the most likely genotype *AAAAA*. We can set some threshold for simply ignoring relatively unlikely genotypes, to save us a lot of computer time in genuine multilocus problems. An alternative approach is to use so-called Markov Chain Monte Carlo (MCMC) sampling techniques. Such techniques have been developed for the more complex situation of outbred line crosses (e.g. Jansen, 1996). The basic MCMC idea is simple: if you cannot enumerate all possible genotypes, then just sample a representative set. We refer to our Bibliographic Notes (Section 18.5) and to **Chapters 19** and **26** for more details.

18.3.5 ANOVA and Regression Tests

Now suppose that we fit a single QTL at a fully informative marker in an F_2 population. As before, the one-way classification model reads $y_{ij} = \mu + \beta_i + \varepsilon_{ij}$ for the jth individual of the ith genotype ($i = 1, 2, 3$; A, H and B). The ANOVA table is shown in Table 18.1. We refer to Soller *et al.* (1976) and Soller and Genizi (1978) for some of the early references.

The test statistic

$$F_{\text{QTL}} = \frac{\text{MS between}}{\text{MS within}},$$

Table 18.1 Analysis of variance table for a single QTL at a fully informative marker in an F_2 population.

Source	df	SS	MS	E(MS)
Between QTL genotypes	2	$\sum_{i=1}^{3} n_i (\bar{y}_i - \bar{y})$	s_{QTL}^2	$\sigma_{\text{e}}^2 + \frac{1}{2} \sum_{i=1}^{3} n_i \beta_i^2$
Within QTL genotypes	$\sum_{i=1}^{3} (n_i - 1) = n - 3$	$\sum_{i=1}^{3} n_i (y_{ij} - \bar{y}_i)^2$	s_{e}^2	σ_{e}^2
Total	$\left(\sum_{i=1}^{3} n_i \right) - 1 = n - 1$	$\sum_{i=1}^{3} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$		

has an F distribution with 2 and $n - 3$ df under the null hypothesis of no segregating QTL (i.e. a ratio of two independent χ^2 variables, $(\chi_2^2/2)/(\chi_{n-3}^2/(n-3))$). The ANOVA model can easily be extended to two or more factors, but the interpretation of the table is now complicated by the possible collinearity of the putative QTLs. Each sum of squares represents the variation accounted for by the given QTL, having eliminated all the effects of the QTLs above it in the table, but ignoring any terms of QTLs below it (e.g. McCullagh and Nelder, 1989). Thus, in the case of strong collinearity the order of fitting terms in the model may affect the sum of squares value of a given QTL. Alternatively, we can tabulate sum of squares values representing the variation accounted for by a given QTL having eliminated all the effects of the other QTLs postulated in the model (no matter whether they are tabulated above or below the given QTL).

18.3.6   Maximum Likelihood Tests

Let us now look at the same one-way QTL classification, but use ML procedures for estimation of the parameters of the normal distributions involved. We still assume that we fit a single QTL at a fully informative marker. The *full* model (fitting the QTL) reads $y_{ij} = \mu + \beta_i + \varepsilon_{ij}$ with residual variance σ_{full}^2, whereas the *reduced* model (omitting the QTL) simply reads $y_{ij} = \mu + \varepsilon_{ij}$ with residual variance $\sigma_{\text{reduced}}^2$. Suppose that the marker data are complete. ML estimates can be obtained via least-squares analysis as usual. Note that

$$\hat{\sigma}_{\text{full}}^2 = \frac{\text{SS within}}{n}$$

and

$$\hat{\sigma}_{\text{reduced}}^2 = \frac{\text{SS total}}{n}.$$

Then, the log likelihood of the full model, l_{full}, is

$$l_{\text{full}} = \sum \left(\log \frac{1}{\sqrt{2\pi\hat{\sigma}_{\text{full}}^2}} - \frac{(y_{ij} - x_i'\hat{\beta})^2}{2\hat{\sigma}_{\text{full}}^2} \right) = -\frac{1}{2}n \log(2\pi\hat{\sigma}_{\text{full}}^2) - \frac{1}{2}n;$$

and the log likelihood of the reduced model, l_{reduced}, is

$$l_{\text{reduced}} = \sum \left(\log \frac{1}{\sqrt{2\pi\hat{\sigma}_{\text{reduced}}^2}} - \frac{(y_{ij} - \hat{\mu})^2}{2\hat{\sigma}_{\text{reduced}}^2} \right) = -\frac{1}{2}n \log(2\pi\hat{\sigma}_{\text{reduced}}^2) - \frac{1}{2}n.$$

The likelihood ratio (*LR*) test statistic is

$$LR = 2 \log \frac{L_{\text{full}}}{L_{\text{reduced}}} = 2(l_{\text{full}} - l_{\text{reduced}}) = n \log \frac{\hat{\sigma}_{\text{reduced}}^2}{\hat{\sigma}_{\text{full}}^2} = n \log \frac{\text{SS total}}{\text{SS within}}.$$

It is asymptotically χ^2 distributed with 2 df (Wilks, 1938). In general, in a combined test for p parameters, the LR test is χ^2 distributed with p df. The models can easily be extended to two or more factors (QTLs). In all cases one tests on the basis of *LR* between a full model and a reduced model.

Some readers may not be familiar with the calculation of *LR* test statistic values on a natural logarithm scale. In applied genetics literature one frequently uses the logarithm with 10 as base, in which case one uses the notation *log-odds* (*LOD*) instead of *LR*. It is defined as

$$LOD = \log_{10} \frac{L_{\text{full}}}{L_{\text{reduced}}}.$$

We can easily convert a score on one scale to a score on the other scale via

$$LR = 2 \times \log_e(10) \times LOD \approx 4.6 \times LOD.$$

For large *n* the *F* test is approximately distributed as $\chi^2(2)/2$ under the null hypothesis of no QTL (Lynch and Walsh, 1998), in which case $2LR \sim F$. Thus, in general QTL detection via ANOVA/regression is not identical to QTL detection via ML – slightly different outcomes may be expected, even for complete data!

18.3.7 Analysis-of-deviance Tests

We will now describe a procedure which can be used for complete data and also for *incomplete* data, i.e. all cases in which we do not fit a QTL at a fully informative marker (Jansen, 1994b). We refer the readers to the textbook by McCullagh and Nelder (1989) on *Generalized Linear Models* Generalized Linear Models for more information.

In comparing a sequence of models, an unbiased estimate of the residual variance can be obtained from the full 'most complex' model. This estimate is used for all models in the sequence to make the comparison fair, between the *full* model and any *reduced* model and amongst reduced models. We can then generate an ANOVA-like table, which in the GLM literature is commonly called an *analysis-of-deviance table*, where the term *deviance* is used for the difference $2(l_{\text{full}} - l_{\text{reduced}})$. The procedure has three steps (Jansen, 1994b):

1. Calculate ML estimates $\hat{\beta}_{\text{full}}$ and $\hat{\sigma}^2_{\text{full}}$. Adjust $\hat{\sigma}^2_{\text{full}}$ for bias, i.e. divide by the residual df and not by *n*.

2. Calculate ML estimates $\hat{\beta}_{\text{reduced}}$ given 'known' residual error as estimated in the full model ($\hat{\sigma}^2_{\text{full}}$ in step 1).

3. Calculate the deviance (*LR* test statistic).

Let us look again at the simple one-way ANOVA for complete data. The log likelihood of the *full* model (QTL model) reads, as before,

$$l_{\text{full}} = \sum \left(\log \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\text{full}}}} - \frac{(y_{ij} - x_i'\hat{\beta})^2}{2\hat{\sigma}^2_{\text{full}}} \right).$$

The real difference comes for the log likelihood of the *reduced* model (the no-QTL model), which now reads

$$l_{\text{reduced}} = \sum \left(\log \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\text{full}}}} - \frac{(y_{ij} - \hat{\mu})^2}{2\hat{\sigma}^2_{\text{full}}} \right).$$

Next we calculate the deviance

$$Deviance = 2(l_{\text{full}} - l_{\text{reduced}}) = \frac{\text{SS total} - \text{SS within}}{\hat{\sigma}^2_{\text{full}}} = 2 \times \frac{\text{MS between}}{\text{MS within}},$$

where the LR test has an F distribution with 2 and $n - 3$ df. In general,

$$Deviance = p \times \frac{\text{MS between}}{\text{MS within}} = pF,$$

where p is the difference between the parameter numbers in the two models being compared. The maximum achievable likelihood is

$$l_{\text{max}} = -0.5n \log(2\pi\hat{\sigma}^2_{\text{full}}),$$

which we obtain if we use as many parameters as we have individuals. If our model contains all the important parameters, the deviance $2(l_{\text{max}} - l_{\text{model}})$ will be close to the difference between the number of individuals, n, and the number of parameters in the model, p. One can use this as a measure for goodness of fit (McCullagh and Nelder, 1989).

We have described the analysis-of-deviance approach for a simple one-factor model, but this can easily be extended to two or more factors. The analysis-of-deviance approach is a unified approach, which can be applied to many types of data (McCullagh and Nelder, 1989). It can also be applied to incomplete data, provided that at least a proportion of the individuals are completely informative at the region under study (Jansen, 1994b). Of course, with rather incomplete information the fitting of many parameters becomes dangerous and therefore undesirable; one can then use the ML approach. We close this section with a remark on the use of non-segregating data. If we have data from parental or F_1 populations, we can estimate the environmental variance, $\hat{\sigma}^2_{\text{e}}$, in the non-segregating populations. It could be an option to use $\hat{\sigma}^2_{\text{e}}$ instead of $\hat{\sigma}^2_{\text{full}}$ during the selection procedure, or at least to define the maximum achievable likelihood as

$$l_{\text{max}} = -0.5n \log(2\pi\hat{\sigma}^2_{\text{e}}).$$

18.3.8   How Many Parameters Can We Fit Safely?

To obtain a good estimate of the residual error variance in ANOVA and regression analysis, it is necessary to have at least 10 residual error df, and many statisticians would take $12-20$ df as their preferred lower limit. Increasing the residual error df makes little difference to the significance threshold in F tests. Say we have a DH population of 100 individuals and suppose our organism has 10 chromosomes of length 200 cM each (e.g. maize). We can then spend as many as 80 df on explanatory variables, for instance by postulating eight QTLs per chromosome equally spaced every $\sim$20 cM.

How many parameters can we fit in the ML approach? In the case of large numbers of parameters, residual error variance can be severely underestimated and asymptotic relations such as the χ^2 approximations do not necessarily hold. Therefore the number of parameters should not be too large, preferably less than $2\sqrt{(\text{number of observations})}$ (Jansen and Stam, 1994). In a DH population of 100 individuals, we have df for fitting approximately $2\sqrt{100} = 20$ QTLs, considerably less than in the ANOVA and regression framework. A critic could well state that we are wasting our resources by not asking

enough questions of our data. Fortunately, this serious disadvantage of the ML approach can be overcome by the general analysis-of-deviance approach described above, by which we can fit up to 80 QTLs (Jansen and Stam, 1994).

18.4   QTL DETECTION STRATEGIES

18.4.1   Model Selection and Genome Scan

Having read the previous sections you have almost all the statistical tools in hand for building the QTL model and for estimating its parameters. You may think the major hurdles have been overcome, and that all that remains – mapping of all QTLs – must be more or less straightforward from the statistical point of view. Here are some cautionary remarks. There are several different statistical techniques for mapping and they do not necessarily all lead to the same answer. Moreover, by no means should the results from statistical analysis be considered as a proof of how nature shaped our trait – any model is wrong because it is a simplification and the possibility of ending up with erroneous conclusions must be kept in mind. The more complex the model, the greater the dangers.

Suppose that we have formulated and fitted a number of different models; which one should we adhere to? Probably the model with a good fit to the data and the lowest number of parameters. In statistics this is called the *principle of parsimony*. Unfortunately there is no general theory for testing one model against another, and as such no claim such as 'this is the significantly best model at a confidence level of 95 %' is possible.

In the ideal case all genetic variance of the trait is explained by detected QTLs only. In practice, a number of QTLs may be missed (error type 2) and at the same time a number of false positives may occur, indicating QTLs at map positions (or regions) where actually no QTLs are present (error type 1). The actual balance between the cost of false positives and the benefit of detected QTLs depends on the aim of the experiment (e.g. map-based cloning or marker-assisted breeding). Nevertheless, one often strives to keep the probability of a type 1 error below 5 %. At the same time one should minimise the probability of an error of type 2. Approaches for QTL mapping generally comprise one or both of the following two steps:

- *Model selection*: Compare different QTL models and select the best one(s).

- *Genome scan*: Plot QTL likelihood along the genome using the selected model.

In our opinion the first step is the more critical one, whereas the second step is merely a good way of visualising QTL results along the genetic map.

In the model selection step we will try to relate the trait variable to one or more explanatory loci (possibly at marker positions) or 'factors'. Different sets of factors can be considered as competing statistical models and many statistical criteria and selection approaches can be used. It is important to note that in genetic mapping one has a major advantage over many other model selection applications. In most applications there is no theory explaining the fact that two or more factors are correlated. But markers and QTLs are known to be located on the genome map, and thus genetic linkage theory tells us that we can replace one factor in the model by another one from the same region without much changing the fit of the model. When two or more explanatory factors are strongly

correlated to one other, it is difficult to disentangle their individual effects – this is known as *multicollinearity* – but this is less of a problem in QTL analysis and in some sense simplifies the difficult task of model selection.

In this section we will review several major techniques in use for model selection and inference in QTL mapping: single-marker analysis; IM; CIM and MQM. With our cautionary remarks in mind, one can safely apply the procedures as given, without being too concerned with the actual significance levels.

18.4.2   Single-marker Analysis and Interval Mapping

We have depicted QTL mapping as a problem of selection among many possible models. Despite the potential benefits of genuine multifactor approaches, its statistical complications lead some users to adopt single-factor ANOVA analyses. How does it work? Suppose we have our known map with markers covering part or all of the genome. We then calculate the QTL likelihood (*F* or *LR* score) at each marker position and plot it along the map. In statistical terms, we produce the likelihood profile along the map. Next we look at our plot and we think there is evidence for QTL activity in regions where the QTL likelihood peaks and exceeds a significance threshold. If the markers are sparse, we can also calculate QTL likelihood at any position within a marker interval to get a smooth QTL likelihood curve, by using IM (see our Figure 18.3; Lander and Botstein, 1989) or the regression mapping approach (Haley and Knott, 1992; Martinez and Curnow, 1992). In this section we will discuss several features of this single-factor type of analysis.

What can we do in order to avoid reporting too many false positives (i.e. non-existing QTLs)? The significance of an effect plays a dominant role in the genetics literature. The common rule is that reviewers will accept a QTL which is significant at a 95 % experimen-twise confidence level ($\alpha = 0.05$). We do multiple tests, and what we obviously need is
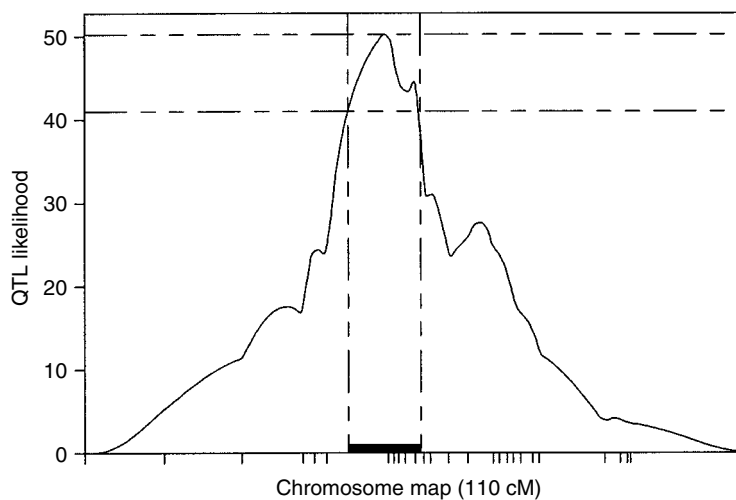
Figure 18.3   QTL likelihood profile obtained with interval mapping for the case of an isolated QTL. Marks below the *x* axis indicate positions of markers along the chromosome under study. The bar indicates the 2 *LOD* support interval for the QTL. The genome-wide significance threshold is ∼14.

therefore a chromosome-wide or genome-wide significance threshold. Test scores at linked markers are strongly correlated. No unifying analytical solutions for the genome-wide distribution of the test statistic in a general inbred line cross are known. Several authors developed (complex) formulae for specific cases (Lander and Botstein, 1989; Rebaï *et al.*, 1994; Doerge and Rebaï, 1996), while others published tables resulting from extensive simulations (van Ooijen, 1999). We can also use simple permutation strategies (Churchill and Doerge, 1994) or bootstrap strategies (Jansen, 1993b; 1994b; Visscher *et al.*, 1996b). Suppose that we use one of these methods. Let us briefly look at permutation and the bootstrap. In both approaches new sets of data will be generated and analysed over many runs. In each run the 'old' marker data will be completed with 'new' trait data. The artificial data set is then analysed for QTLs and the maximum test score over the genome is calculated and stored. The entire procedure is repeated many (say, 1000–10 000) times in order to generate an empirical cumulative distribution for the test statistic. In the permutation approach, observed trait data are reshuffled over the individuals under the null hypothesis of no QTL, thereby breaking down any existing marker – QTL associations. Under the null hypothesis any of the trait data for any of the individuals could have come equally well from either of the marker or QTL classes. In the parametric bootstrap approach new trait data are generated from a standard normal distribution postulating no QTL.

Suppose that we have detected a 'genuine' QTL. If we replicated the experiment, the gene's location would of course remain identical, but the size of its effect can easily change significantly due to QTL by environment interaction. Therefore we believe that the gene's location and the sign of its effect are probably the most relevant parameters. Thus, we question the value of putting standard errors on effects, but we would certainly like to report a 95 % confidence region for the QTL location. The inverse of Fisher's information matrix is the standard tool in a statistician's ML toolbox for calculation of standard errors. Computation of this matrix is a difficult task in the case of mixture models, but at relatively low cost the matrix can be approximated by using first – order partial derivatives, which are directly available from the EM algorithm (Redner and Walker, 1984; Jansen and Stam, 1994). But there is a more direct way to construct a confidence region for the QTL location. We usually calculate and plot the QTL likelihood profile at each map position along the map. A clear peak in this profile is taken as the most likely QTL position. Suppose we have a clear peak in the profile. According to standard ML theory, a 95 % confidence region for QTL location is then bounded by the map positions where the profile is $\chi^2_{0.05}(1) = 3.84$ less than at the peak (equivalent to 0.84 *LOD*). Let us call this region a *support region* (Figure 18.3). This region is a 95 % confidence region if and only if the confidence interval falls within one marker interval. What is the critical problem? If the confidence region spans more than one interval, then we actually compare several different statistical models. But ML properties only hold within one model and not across models. In other words, we have no general theory to guarantee that the support interval is a 95 % confidence region. Various simulation studies have shown that a 2 *LOD* support interval is a safe choice in most cases yielding at least 95 % confidence regions (e.g. van Ooijen, 1992).

Here are some more warnings. Most of them are pretty obvious from the statistical point of view, knowing that what can go wrong in ANOVA and regression analysis is often due to multicollinearity. Suppose that you have used the appropriate genome-wide significance threshold and detected a significant QTL at a certain map position where the test score peaks highest. You may yell 'Hurray, I have found a QTL!'. But you have only found a statistical association and not a gene, and there are at least four traps that

may lead you to erroneous inference. The first trap is that there are actually two or more linked QTLs with effects of equal sign (QTLs are in coupling phase), in which case it is not unlikely that the analysis will reveal a single QTL in the middle of the two true QTLs. This is known as the *detection of a ghost QTL* (Martinez and Curnow, 1992), an error of type 1. The second trap is that an unlinked major QTL has inflated the test score. Incidental association can arise due to deviations from expected segregation ratios for any pair of loci on the genome. This is especially probable in the case of severe segregation distortion, or in small populations where larger deviations from expected segregation ratios may arise as the result of natural sampling variation. Of course, this is another example of a ghost QTL, but the general user less often anticipates it. The third trap is even nastier than the other two, although it can be considered as the multi-QTL extension. What we test with IM is nothing more or less than an *average* effect of all QTLs in the region under study. With IM there is no way to dissect their effects; any effect detected can be the sum of many possibly small QTL effects instead of the effect of an isolated QTL. If we repeat our experiment, the effects of the QTLs can be modulated in different ways and therefore the average effect may give rise to a peak of the test score at another map location. Furthermore, if two or more QTLs have effects of different sign (QTLs are in repulsion phase), then the joint effect can be close to zero. In other words, in such cases (major) QTLs will remain undetected. Finally, the fourth trap has to do with variable information content. Suppose that the information content is relatively low in a region containing a QTL. What may happen is that the peak is shifted towards more informative regions (Knott and Haley, 1992).

18.4.3   Composite Interval Mapping

Is there a way to avoid the traps described above? In fact we have already seen almost all the essentials for such an approach. In Section 18.3 we mentioned the solution: tabulate the variation accounted for by a given QTL, having eliminated all the effects of the other QTLs postulated in the model. (Recall the DH example, in which we postulated up to eight QTLs per chromosome, equally spaced every ∼20 cM.) As in Section 18.4.2, we calculate the QTL likelihood at each marker position, plot it along the map and look for peaks. If the marker map is sparse, we can calculate QTL likelihood at multiple positions within a marker interval to get a smooth QTL likelihood profile. In principle we can postulate QTLs at any position we like – on top of markers or inside marker intervals (Jansen, 1992; Kao *et al.*, 1999). In the practice of working with dense maps we often place them on top of markers. We perform a genome scan by moving a QTL along the chromosomes while using a pre-identified set of markers as cofactors (Jansen, 1992; 1993b; Zeng, 1994). Or, in other words, for the sparse map case, we combine IM with multiple regression on markers. Stam (1991), and later Rodolphe and Lefort (1993) and Zeng (1993), demonstrated that in regression the effect of a QTL is absorbed only by its flanking cofactors, at least if progeny size is large. So, suppose we test for the presence of a QTL at a certain map position with a cofactor at some distance on the left and one on the right-hand side of the QTL. The test is now asymptotically unaffected by any QTLs located to the left of the left cofactor and to the right of the right cofactor (Stam, 1991).

Let us first pay some attention to what is known about setting the genome-wide significance threshold. Zeng (1994) studied the sparse map case and proposed to use all markers as cofactors, except the ones flanking the interval under study. Thus, the

model for the ith individual in a DH population is

$$y_i = \mu + \sum_{j=1}^{M} x_{ij}\beta_j + x_{i0}\beta_0 + \varepsilon_i,$$

where summation is over M marker cofactors not flanking the interval under study, x_{ij} are 0–1 indicator variables and β_j are substitution effects for the jth marker cofactor ($j = 1, \ldots, M$), or for the QTL ($j = 0$). Zeng (1994) used ML and called the method *composite interval mapping*. Simulation work with BC populations demonstrated that $\chi^2_{\alpha/M}(2)$ can be used as an upper bound for the $100\alpha\%$ genome-wide threshold on a genome with M marker intervals, unless the number of parameters is too large (Zeng, 1994; see his Figure 1). The $\chi^2_{\alpha/M}(2)$ relation does not hold if the number of parameters exceeds $2\sqrt{}$(number of observations) (Jansen and Stam, 1994). With the analysis-of-deviance approach (Jansen and Stam, 1994), however, there is no limitation on the number of parameters and $2F_{a/M}$ (2, df) can be used as an upper bound, where df are the degrees of freedom for estimating the residual error variance.

Although attractive in properly disentangling the effects of multiple (linked) QTLs, the precision mapping approach has some serious drawbacks. Suppose there is a QTL located in the interval currently under study. What can happen? First, any linked cofactor in the (wide) neighbourhood of our testing position will absorb at least part of the effect of the QTL, because we use conditional tests. Second, in models with many cofactors it is even likely that a linear combination of unlinked cofactors explains a considerable part of the variation of our QTL (Jansen, 1994b). Third, the genome-wide threshold is a function of the number of intervals M, which means that the significance threshold increases with the number of cofactors. Putting the pieces together, the conclusion is that with too many cofactors we will simply end up with high precision (we will make no type 1 errors), but no power (we will miss most QTLs). In a later paper Kao *et al.* (1999) proposed an alternative version of CIM, which they called *multiple-interval mapping* (MIM), and which included a stepwise selection phase. Such strategies will be discussed in more detail in the next subsection.

18.4.4 Multiple-QTL Mapping

Parsimony of parameters is one of the basic paradigms in statistics. A parsimonious model does not include parameters which are unnecessary or which (even worse) decrease the chance (power) of reaching our goals (Draper and Smith, 1981; McCullagh and Nelder, 1989). In our context, we do not want the model to include cofactors in regions where there are no QTLs. There is no unique best statistical method for finding the 'important' cofactors, but there are several good and general criteria, penalising the use of extra parameters (e.g. Akaike's information criterion (AIC)). But selection of a useful set of cofactors from a large set of possible cofactors to form a parsimonious model is a non-trivial task with both statistical and computational problems. Below we will discuss forward selection, backward elimination and stepwise regression. We think that selection is the most important step in QTL analysis. Subsequent to cofactor selection, we perform a genome scan by moving a QTL along the chromosomes while using a preselected set of markers as cofactors (Jansen, 1993b; 1994b; Jansen and Stam, 1994). If the moving QTL gets close to a cofactor, say within a preset window of 10 or 20 cM, then we drop
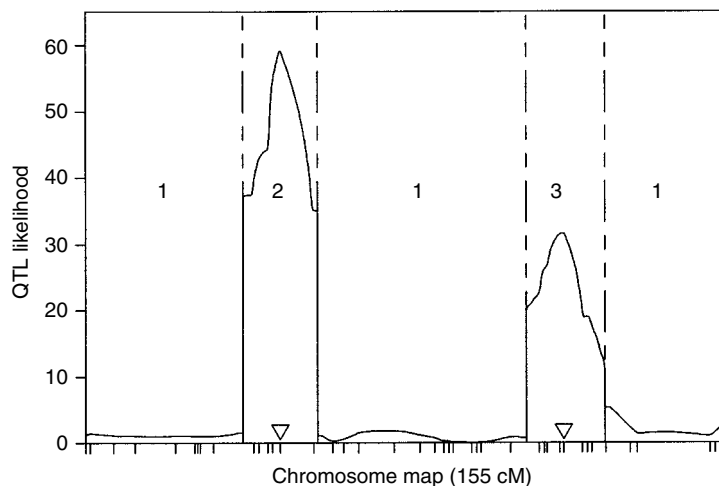
Figure 18.4 QTL likelihood (deviance) profile obtained with MQM mapping for the case of two linked QTLs. Marks below the *x* axis indicate positions of markers along the chromosome under study. Triangles indicate positions of selected cofactors. Different tests for the presence of a QTL are performed in regions indicated by '1', '2' and '3', respectively. In region '1' the test is conditional on the cofactors at 36 and 76 cM. In region '2' it is conditional on the cofactor at 76 cM, and in region '3' it is conditional on the cofactor at 36 cM. Regions '2' and '3' are chosen (i) symmetrically around their respective cofactor, and (ii) wide enough to span their respective 2*LOD* support interval for the QTL (support intervals not shown). The genome-wide significance threshold is ∼14.

the cofactor (Figure 18.4). This approach is called *multiple-QTL mapping*. Zeng (1994) also suggested the use of a thinned set of cofactors (e.g. via stepwise regression), and elaborated on it in detail in Kao *et al.* (1999).

We mentioned three selection strategies: forward, backward and stepwise. In the forward selection approach, at each stage the best new cofactor satisfying the selection criterion is added until no further candidates remain. This approach is often used in QTL analysis, on top of the IM approach. Once we have found one or more QTLs with IM, we add them to our model and rerun the genome scan. Although a natural procedure, we do not recommend it for two reasons. First, because we do not avoid the traps described above for the single-QTL analysis. Second, the approach does not exploit power to the full: each test is based on the ratio between variance explained by the factor under study and the unexplained variance. In a forward selection approach unexplained variance contains environmental plus as yet unexplained genetic variance, which decreases power relative to a backward elimination procedure.

The backward elimination procedure starts with a multiple regression model, using a full set of cofactors (putative QTLs/markers) evenly spread over the genome. The unimportant or least important cofactors are dropped one by one until all remaining cofactors are essential given the selection criterion. This is a satisfactory approach, especially if we wish to see all the variables in the model in order 'not to miss any QTL'. The full model gives an unbiased estimate of the maximum amount of variance explainable by (non-interacting) cofactors (QTLs). In order to exclude redundant cofactors, the selection criterion should be stringent, but not so stringent that important cofactors (those flanking the QTLs) are

thereby excluded. We can use one of the two approaches that we described above: the ML approach (Jansen, 1993b; Zeng, 1994; Kao *et al.*, 1999) or the analysis-of-deviance approach (Jansen and Stam, 1994; Jansen, 1994b).

For the ML approach, Jansen (1993b) proposed to maximise the log likelihood (l) minus the number of free parameters (k) in the model; this is equivalent to minimising AIC, given by $-2(l - k)$. In general, a penalty in the range of k to $3k$ may provide plausible initial models (McCullagh and Nelder, 1989). In 'ordinary' regression with adequate df to estimate σ^2, a penalty of k is equivalent to the use of (about) the 16 % point of the F test for the comparison of two nested models, which differ only by the inclusion of one free parameter; a penalty of $3k$ is equivalent to the use of (about) the 2 % point (McCullagh and Nelder, 1989). Note that we can also compare *non-nested* models by using the AIC. And we can even compare several models fitted on different scales, having multiplied $f_{\mathrm{mix}}(T(\cdot))$ by its Jacobian $T'(\cdot)$. Finally, we recall here that there are two disadvantageous features of the ML approach. First, the number of parameters should remain relatively small – less than $2\sqrt{(\text{number of observations})}$. Second, there is the danger of overfitting and thereby underestimating the error variance.

In the analysis-of-deviance approach, one can use partial F-tests conditional on the other cofactors in the current model. Note that the same approach is valid for regression mapping (Haley and Knott, 1992; Martinez and Curnow, 1992). The partial F-test values are calculated for each cofactor. The cofactor with the lowest partial F-test value is removed if its effect is less significant than a preset significance level. This process is repeated until all remaining cofactors have a partial F-test value exceeding the threshold. Jansen (1994b) used a 2 % threshold in simulations. It was demonstrated that this penalty is stringent, since no or only a few cofactors are generally selected in the case of no QTLs segregating. Moreover, it was shown that this penalty is still not too stringent, since cofactors are selected for those QTLs that considerably affect the test statistic in their nearby region; the effects of such QTLs are satisfactorily eliminated by selected cofactors. Because of this feature, the IM thresholds, which were obtained for the case that no QTLs are segregating, are also suitable for MQM mapping. These thresholds can be used when no QTLs are segregating, since in that case no or only a few cofactors will be selected; moreover, these thresholds can still be used when there are QTLs segregating, the effects of which are eliminated by cofactors. Detection and unravelling of the separate QTL effects in the case of linked loci is much easier in MQM mapping than in IM (Jansen, 1994b).

One of the disadvantages of the backward elimination procedure is that for a cofactor 'once out' means 'always out'. Backward elimination followed by a stepwise procedure, including new cofactors and dropping old ones, may help to overcome this – at the cost of more computation. Alternatively, replacing important cofactors by neighbours, not present in the initial set of cofactors, can also help in fine-tuning the model.

We emphasise one important difference between MQM, on the one hand, and IM and CIM, on the other hand, that arises because of MQM's use of a fixed residual error for both full and reduced models. Suppose we compare two nested models, in the simplest case a model with a QTL versus a model with no QTL. In regression mapping and MQM the *LR* test or the equivalent F test is based on the fit of the full model versus that of a reduced model with the residual error fixed at the value estimated from the full model. If the QTL has a major effect, then the no-QTL model will fit badly and the test score will be high. If the (putative) QTL has a minor or no effect, then the residual variance estimate of the full model is fully satisfactory for the no-QTL model. What happens in

CIM? The residual error variance in the no-QTL model is not fixed, but it is estimated and it will absorb the QTL variation. As a result the test score will not be as high as in the MQM and regression mapping approaches. A minor detail? Not really, because the effect on the test score can be considerable. For instance, in Figure 18.3 the test score is 50 for IM and 66 for MQM (the latter is not shown). Only one QTL was simulated on the whole genome. Thus in this case the difference between IM and MQM is solely due to the way the residual variance is treated.

18.4.5 Uncritical use of Model Selection Procedures

We would like to emphasise the following warning often posted in the statistical literature: *the uncritical use of the results of selection procedures can be very dangerous.* Here is one example of what can go wrong. Suppose we fit two cofactors, which are not far apart on the map; say that there is only one recombinant individual in a DH population. What can happen? The model for the ith individual is

$$y_i = \mu + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i,$$

where x_{ki} and β_{ki} are the 0–1 indicator variables and effect variables for the two closely linked cofactors, respectively. For all individuals but one, $x_{1i} = x_{2i}$, that is either

$$y_i = \mu + \varepsilon_i$$

or

$$y_i = \mu + \beta_1 + \beta_2 + \varepsilon_i$$

holds. In the latter case only the sum $\beta = \beta_1 + \beta_2$ is relevant. Suppose the single recombinant individual has $x_{1i} = 0$ and $x_{2i} = 1$, in which case

$$y_i = \mu + \beta_2 + \varepsilon_i.$$

For a given value of μ and β, we can obtain a perfect fit for the ith individual by setting

$$\beta_2 = y_i - \mu,$$
$$\beta_1 = \beta - \beta_2.$$

If there are no real QTLs in this region, then one would expect that $\beta = \beta_1 + \beta_2 = 0$. But we can still get large values $\beta_1 = -\beta_2$, which we might take as evidence for two linked QTLs with opposite effects. In some cases this may be true, but more frequently we probably just misinterpret what is actually no more than a statistical artefact. Other artificial patterns of near collinearity between sets of cofactors can arise. Moreover, any real data set will contain erroneous trait and marker scores (genotypic misscores, phenotypic outliers and so forth), for which our models may want to compensate for by an increase in the number of parameters. Several things can be done to avoid serious problems – on the biological and on the statistical side of the problem. First, make sure that the trait and marker data are very well checked! We prefer that labels U, C or D replace suspicious marker scores. Second, avoid simultaneous fitting of closely linked cofactors. For instance, in a DH population of 100 individuals we should make sure that simultaneously fitted cofactors stay at least 10–20 cM apart. If we do so, then there are 10–20 recombinants expected between any pair of cofactors. Third, we can set the

genome-wide significance threshold by an empirical (permutation or bootstrap) strategy, in which we mimic the complete selection protocol at each run. See also Piepho and Gauch (2001), Hwang and Nettleton (2002), Broman and Speed (2002), Jansen *et al.* (2002) and Bogdan *et al.* (2004) for further discussions about the best criterion for model selection (e.g. AIC or Bayesian information criterion). .

18.4.6   Final Comments

What if the trait is encoded by many QTLs, say 40, each of small effect? It will definitely be hard to decompose the phenotypic variation into the underlying QTL components. The number of recombinants is the limiting factor, and the only way out is to generate more of them by increasing the population size or by choosing another type of cross with more informative meioses. See Beavis (1996) and Visscher *et al.* (2000) for some illustrative examples of the hard job of dissecting polygenic variation.

Last but not least, we would like to state that the proof of the QTL pudding is not in the eating of the results from statistical QTL analysis. It is in the eating of results from cross-validation and/or new biological experiments.

18.5   BIBLIOGRAPHIC NOTES

18.5.1   Statistical Approaches

During the past two decades, a number of important papers have been published on QTL analysis with sparse marker maps. We list some of the early and most influential papers: Lander and Botstein (1989), in which the name 'interval mapping' was coined; Haley and Knott (1992) and Martinez and Curnow (1992), with the 'regression mapping' approach; Jansen (1992; 1993b; 1994b) and Jansen and Stam (1994), in which he general GLM framework for QTL analysis and the MQM approach was developed; Zeng (1993; 1994) and Kao *et al.* (1999), with CIM and MIM; and finally, Churchill and Doerge (1994) and Doerge and Churchill (1996), with the permutation test for setting the genome-wide significance threshold. There are various software packages available, for example, Map Manager, MapMaker/QTL, MapQTL, Multi-QTL, PlabQTL, QTL Cartographer, and MCQTL (further information can easily be found on the Worldwide Web). The textbook *Genetics and Analysis of Quantitative Traits* by Lynch and Walsh (1998) also provides a wealth of information for the interested reader. The book brings together basic biology and many methods of statistical analysis. From a statistical viewpoint, there is not much special to QTL modelling in inbred lines. If we look in our statistical toolbox, we will see, for instance, tools for interaction models, for mixed models with random and fixed effects, for multivariate instead of univariate analysis, and for Bayesian analysis. In this chapter, QTL analysis in inbred lines is accomplished by embedding it in a more or less standard statistical framework of GLMs. For general discussion of linear models, refer to the textbook *Applied Regression Analysis* by Draper and Smith (1981), and the textbook *Generalized Linear Models* by McCullagh and Nelder (1989). GLMs provide a unified analysis of diverse types of trait data (e.g. normal, binary, ordinal) via a simple approach of iterative reweighted least squares. Jansen (1993a; 1994a) embedded mixture models in the GLM framework, which is of prime importance to segregation analysis

and QTL analysis, where one deals with mixtures of different genotypes. For extensive rigorous statistical treatment of mixture models, we refer to the review by Redner and Walker (1984), the textbook by Titterington *et al.* (1985) and Dempster *et al.*'s (1977) paper on the EM algorithm. Here we have focused on QTL analysis in inbred lines for normally distributed traits; see, for instance, Hackett and Weller (1995) for QTL analysis in inbred lines with ordinal trait data and Visscher *et al.* (1996a) for QTL analysis with binary data. In most experiments, two or more traits are scored on the same individual. Jiang and Zeng (1995), Korol *et al.* (1995) and Lange and Whittaker (2001) are good entries for multitrait QTL analysis. For correlated traits, a higher resolution can be obtained by multitrait analysis than by separate univariate analyses. Bayesian modelling of mixtures with an unknown number of underlying components has made enormous progress (Richardson and Green, 1997) and is beginning to have an impact on various applications, including genetics. Undoubtedly, the popularity is largely due to the (MCMC) algorithms that enable the estimation of complex genetic models. Bayesian methods are particularly good for missing data augmentation (which we have emphasised here for the EM and MCEM algorithm). With ultradense maps, we can obtain nearly complete data, and the use of Bayesian methods in the relatively simple case of inbred line genetics is perhaps overkill. However, we have seen that model selection – how many QTLs and on which map positions – is still a hard problem, and Bayesian methods are becoming more prominent in this regard. We refer to Green (1995), certainly one of the pioneering papers on Bayesian model selection. Satagopan *et al.* (1996) described the first Bayesian method for inbred line crosses. In their approach, models with different (fixed) numbers of QTLs were compared via Bayes factors (similar to the LR test). Sillanpää and Arjas (1998) allowed for a variable number of linked QTLs on the chromosome under study. The effects of unlinked QTLs were taken into account by fixed preselected cofactors (as in CIM and MQM). Stephens and Fisch (1998) allowed for a variable number of linked and unlinked QTLs. See Boer *et al.* (2002) for a semi-Bayesian and Xu (2003) and Ter Braak *et al.* (2005) for Bayesian methods, estimating polygenic QTL effects with the aid of all markers of the entire genome. There is much more to be said about the potential of Bayesian methodology, and we refer the interested reader to **Chapters 19** and **20** for further details.

18.5.2   Learning More about Important Genetic Parameters

The advent of the molecular markers and the development of statistical QTL tools have opened up new ways for studying the complexity of traits in more detail. Several issues are of prime importance to plant breeders, mouse geneticists and others studying inbred line crosses: how do the genes talk to each other (interaction between QTLs)? are they influenced by the environment (interaction between QTL and environment)? and are QTL findings consistent across multiple crosses (interaction between QTL and genetic background)? Here we list a number of key papers. QTL by QTL epistatic interaction is the first important type of interaction. See, for instance, Chase *et al.* (1997), Fijneman *et al.* (1996), Kao *et al.* (1999), Carlborg *et al.* (2000) and Carlborg and Andersson (2002). One of the major obstacles with epistatic interactions is the increase in the number of parameters, which is generally not compensated by a similar increase in the population size. Jannink and Jansen (2001) developed a one-dimensional genome search strategy for detecting epistatic QTL in diallel crosses; by combining information from the different crosses, QTL of high interaction with genetic background can be identified in a first

step, and QTL by QTL interaction in a second step. Boer *et al.* (2002) developed a one-dimensional genome strategy for searching QTL by QTL interaction in a single population, bounding the dimension of the search via penalised likelihood methods. Yi *et al.* (2003) use Bayesian methodology and variable selection methods to develop strategies for searching for multiple QTLs with complex interaction patterns. QTL by environment interaction is the second important type of interaction. In breeding programs, this interaction needs to be characterised so that target QTLs can be effectively utilised in them. We refer to Jansen (1992), Jansen *et al.* (1995), Tinker and Mather (1995) and Jiang and Zeng (1995) for early work on QTL by environment. Wang *et al.* (1999) combine QTL by environment and additive QTL by QTL interaction and propose the use of fixed effect terms for the QTL, and random effect terms for cofactors and interactions. A wide range of models for 'traditional' genotype by environment interaction (GEI) had been developed in quantitative genetics, e.g. modelling the interaction as products of genotype sensitivities and environmental characterisations (e.g. van Eeuwijk *et al.*, 1996). These traditional models and the QTL models were later integrated by several authors e.g. Korol *et al.* (1998), Malosetti *et al.* (2004) and Piepho and Pillen (2004). The third important type of interaction is that between QTL and genetic background. In a single cross, inferences about QTLs and their estimated effects are limited to the particular cross. A breeder, however, usually produces many crosses instead of one. A multipopulation analysis would allow us to extend inferences across populations. Rebaï and Goffinet (1993) dealt with QTL analysis of diallel crosses between inbred lines, Xu (1998) with multiple unrelated families, and Liu and Zeng (2000) with diverse cross designs involving multiple inbred lines. Jansen *et al.* (2003) developed an approach based on haplotyped putative QTL alleles, where the haplotype information was recovered from the marker data. A similar approach was taken by Crepieux *et al.* (2004). Recent studies in mouse (Li *et al.*, 2005) and in maize (Blanc *et al.*, 2006) illustrate the potential benefits of analysing the different populations jointly: one can extract a lot more relevant QTL information from the whole than from the individual parts. We refer the readers to **Chapter 19** for a more general theory about QTL analysis in complex populations.

18.5.3 QTL Analysis in Inbred Lines on a Large Scale

We conclude by pointing out a booming field of application for QTL mapping: genetical genomics (Jansen and Nap, 2001). Recent biomolecular technologies allow for the high-throughput profiling of thousands of transcripts, proteins and/or metabolites. Costs have already dropped down to a level where it is possible to profile hundreds of individuals. A logical step – at least for a quantitative geneticist – is then to profiling segregation of populations. The abundance of each of thousands of genes, proteins and metabolites can be treated as a quantitative trait. Therefore, the statistical analysis tools described in this chapter are useful to map eQTLs underlying variation in expression of genes, pQTLs underlying variation in protein abundance, and mQTLs underlying variation in metabolite abundance. Running thousands of QTL analyses imposes new challenges, e.g. how to handle the dramatically increased problem of multiple testing that may lead to many false positive QTLs, or how to exploit the results of all these QTL analyses to narrow down organisms-level phenotypic QTLs to the genes and to reconstruct regulatory, developmental and metabolic pathways. Storey and Tibshirani (2003) proposed the 'q value' to control the false discovery rate (FDR). De la Fuente *et al.* (2004), Zhu *et al.* (2004) and Bing and Hoeschele (2005) have developed methods to uncover gene

regulatory networks in genetical genomics studies, from either expression profiles or QTL profiles. Schadt *et al.* (2003), Klose *et al.* (2003) and Keurentjes *et al.* (2006) reported the first detailed studies of eQTL, pQTL and mQTL in inbred line crosses, respectively. We refer the readers to Doerge (2002), Jansen (2003) and Rockman and Kruglyak (2006) for reviews, and to **Chapter 9** for more general theory about eQTL analysis of gene expression on a large scale.

Acknowledgments

REFERENCES

Allard, R.W. (1999). *Principles of Plant Breeding*. Wiley, New York.

Atkinson, A.C. (1985). *Plots, Transformations and Regression*. Clarendon Press, Oxford.

Beavis, W.D. (1996). QTL analyses: power, precision and accuracy. In *Molecular Analysis of Complex Traits*, A.H. Paterson, ed. CRC Press, Boca Raton, FL.

Beckmann, J.S. and Soller, M. (1983). Restriction fragment length polymorphisms in genetic improvement methodologies, mapping and costs. *Theoretical and Applied Genetics* **67**, 35–43.

Bing, N. and Hoeschele, I. (2005). Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170**, 533–542.

Blanc, G., Charcosset, A., Mangin, B., Gallais, A. and Moreau, L. (2006). Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. *Theoretical and Applied Genetics* **113**, 206–224.

Boer, M.P., Ter Braak, C.J.F. and Jansen, R.C. (2002). A penalized likelihood method for mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* **162**, 951–960.

Bogdan, M., Ghosh, J.K. and Doerge R.W. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* **167**, 989–999.

Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980). Construction a genetic map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**, 314–331.

Broman, K.W. and Speed, T.D. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society, Series B* **64**, 641–656.

Carlborg, O. and Andersson, L. (2002). Use of randomization testing to detect multiple epistatic QTLS. *Genetical Research* **79**, 175–184.

Carlborg, O., Andersson, L. and Kinghorn, B. (2000). The use of a genetic algorithm for simultaneous mapping of multiple interaction quantitative trait loci. *Genetics* **155**, 2003–2010.

Chase, K., Adler, F.R. and Lark, K.G. (1997). Epistat: a computer program for identifying and testing interactions between pairs of quantitative trait loci. *Theoretical and Applied Genetics* **94**, 724–730.

Churchill, G.A. and Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Cowen, N.M. (1989). Multiple linear regression analysis of RFLP data sets used in mapping QTLs. In *Development and Application of Molecular Markers to Problems in Plant Genetics*, T. Helentjaris and B. Burr, eds. Cold Spring Harbor Laboratory, Cold Spring Harbor, pp. 113–116.

Crepieux, S., Lebreton, C., Servin, B. and Charmet, G. (2004). Quantitative trait loci (QTL) detection in multicross inbred designs: recovering QTL identical-by-descent status information from marker data. *Genetics* **168**, 1737–1749.

De La Fuente, A., Bing, N., Hoeschele, I. and Mendez, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20**, 3565–3574.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Doerge, R.W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* **3**(1), 43–52.

Doerge, R.W. and Churchill, G.A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294.

Doerge, R.W. and Rebaï, A. (1996). Significance thresholds for QTL interval mapping tests. *Heredity* **76**, 459–464.

Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis*. Wiley, New York.

van Eeuwijk, F.A., Denis, J.B. and Kang, M.S. (1996). Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. In: *Genotype-by-Environment Interaction*. Kang, M.S. and Gauch, H.G. (eds), CRC Press Inc., Boca Raton, Florida, pp 15–50.

Fijneman, R.J.A., de Vries, S.S., Jansen, R.C. and Demant, P. (1996). Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility of lung cancer in the mouse. *Nature Genetics* **14**, 465–467.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Hackett, C.A. and Weller, J.I. (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**, 1252–1263.

Haley, C.S. and Knott, S.A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Hwang, J.T.G. and Nettleton, D. (2002). Investigating the probability of sign inconsistency in the regression coefficients of markers flanking quantitative trait loci. *Genetics* **160**, 1697–1705.

Jannink, J.L. and Jansen, R.C. (2001). Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* **157**, 445–454.

Jansen, R.C. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* **85**, 252–260.

Jansen, R.C. (1993a). Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics* **49**, 227–231.

Jansen, R.C. (1993b). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.

Jansen, R.C. (1994a). Maximum likelihood in a finite mixture model by exploiting the GLM facilities of Genstat. *Genstat Newsletter* **30**, 25–27.

Jansen, R.C. (1994b). Controlling the type 1 and type 2 errors in mapping quantitative trait loci. *Genetics* **138**, 871–881.

Jansen, R.C. (1996). A general Monte Carlo method for mapping quantitative trait loci. *Genetics* **142**, 305–311.

Jansen, R.C. (2003). Studying complex biological systems using multifactorial perturbation. *Nature Reviews Genetics* **4**, 145–151.

Jansen, R.C. and Den Nijs, A.P.M. (1993). A statistical mixture model for estimating the proportion of unreduced pollen grains in perennial ryegrass (lolium perenne) via the size of pollen grains. *Euphytica* **70**, 205–215.

Jansen, R.C., Jannink, J.L. and Beavis, W.D. (2003). Mapping quantitative trait loci in plant breeding populations: use of parental haplotype sharing. *Crop Science* **43**, 829–834.

Jansen, R.C., van Ooijen, J.W., Stam, P., Lister, C. and Dean, C. (1995). Genotype by environment interaction in genetic mapping of multi quantitative trait loci. *Theoretical and Applied Genetics* **91**, 33–37.

Jansen, R.C. and Nap, J.P. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics* **17**, 388–391.

Jansen, R.C. and Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447–1455.

Jansen, R.C., Ter Braak, C.J.F., Boer, M.P. and Maliepaard, C.M. (2002). Discussion on 'statistical modelling and analysis of genetic data'. *Journal of the Royal Statistical Society, Series B* **64**, 737–775.

Jansen, R.C., Reinink, K. and van der Heijden, G.W.A.M. (1993). Analysis of grey level histograms by using statistical methods for mixtures of distributions. *Pattern Recognition Letters* **14**, 585–590.

Jensen, J. (1989). Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker gene loci. *Theoretical and Applied Genetics* **78**, 613–618.

Jiang, C. and Zeng, Z.-B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**, 1111–1127.

Kao, C.-H., Zeng, Z.-B. and Teasdale, R.D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

Keurentjes, J.J.B., Fu, J., De Vos, C.H.R., Lommen, A., Hal, R.D., Bino, R.J., Van Der Plas, L.H.W., Jansen, R.C., Vreugdenhil, D. and Koornneef, M. (2006). The genetics of plant metabolism. *Nature Genetics* **38**, 842–849.

Klose, J., Nock, C., Hermann, M., Stuhler, K., Marcus, K., Bluggel, M., Krause, E., Schalkwyk, L.C., Rastan, S., Brown, S.D.M., Bussow, K., Himmelbauer, H. and Lehrach, H. (2003). Genetic analysis of the mouse brain proteome. *Nature Genetics* **30**, 385–393.

Knott, S.A. and Haley, C.S. (1992). Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics* **132**, 1211–1222.

Korol, A.B., Ronin, Y.I. and Kirzhner, V.M. (1995). Interval mapping of quantitative trait loci employing correlated trait complexes. *Genetics* **140**, 1137–1147.

Korol, A.B., Ronin, Y.I. and Nevo, E. (1998). Approximate analysis of QTL-environment interaction with no limits on the number of environments. *Genetics* **148**, 2015–2028.

Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Lange, C. and Whittaker, J.C. (2001). Mapping quantitative trait loci using generalized estimating equations. *Genetics* **159**, 1325–1337.

Li, R.H., Lyons, M.A., Wittenburg, H., Paigen, B. and Churchill, G.A. (2005). Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping. *Genetics* **169**, 1699–1709.

Liu, Y. and Zeng, Z.-B. (2000). A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. *Genetical Research* **75**, 345–355.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Malosetti, M., Voltas, J., Romagosa, I., Ullrich, S.E. and Van Eeuwijk, F.A. (2004). Mixed models including environmental covariables for studying QTL by environment interaction. *Euphytica* **137**, 139–145.

Martinez, O. and Curnow, R.N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480–488.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

van Ooijen, J.W. (1992). Accuracy of mapping quantitative trait loci in autogamous species. *Theoretical and Applied Genetics* **84**, 803–811.

van Ooijen, J.W. (1999). LOD significance thresholds for QTL analysis in experimental populations of diploid species. *Heredity* **83**, 613–624.

Piepho, H.-P. and Gauch, H.G. Jr. (2001). Marker pair selection for mapping quantitative trait loci. *Genetics* **157**, 433–444.

Piepho, H.P. and Pillen, K. (2004). Mixed modelling for QTL x environment interaction analysis. *Euphytica* **137**, 147–153.

Rebaï, A. and Goffinet, B. (1993). Power of tests for QTL detection using replicated progenies desired from a diallel cross. *Theoretical and Applied Genetics* **86**, 1014–1022.

Rebaï, A., Goffinet, B. and Mangin, B. (1994). Approximate thresholds for interval mapping tests for QTL detection. *Genetics* **138**, 235–240.

Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26**, 195–239.

Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B* **59**, 731–792.

Rockman, M.V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews Genetics* **7**, 862–872.

Rodolphe, F. and Lefort, M. (1993). A multi-marker model for detecting chromosomal segments displaying QTL activity. *Genetics* **134**, 1277–1288.

Satagopan, J.M., Yandell, B.S., Newton, M.A. and Osborn, T.C. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144**, 805–816.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B. and Friend S.H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302.

Sillanpää, M.J. and Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**, 1373–1388.

Simpson, S.P. (1989). Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. *Theoretical and Applied Genetics* **77**, 815–819.

Soller, M., Brody, T. and Genizi, A. (1976). On the power of experimental design for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47**, 35–39.

Soller, M. and Genizi, A. (1978). The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. *Biometrics* **34**, 47–55.

Spelman, R.J., Coppieters, W., Karim, L., Van Arendonk, J.A.M. and Bovenhuis, H. (1996). Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. *Genetics* **144**, 1799–1808.

Stam, P. (1991). Some aspects of QTL analysis In *Proceedings of the Eighth Meeting of the Eucarpia Section 'Biometrics in Plant Breeding'*, Brno.

Stephens, D.A. and Fisch, R.D. (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* **54**, 1334–1347.

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–9445.

Ter Braak, C.J.F., Boer, M.P. and Bink, M.C.A.M. (2005). Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. *Genetics* **170**, 1435–1438.

Tinker, N.A. and Mather, D.E. (1995). Methods for QTL analysis with progeny replicated in multiple environment. *Journal of Agricultural Genomics* **1**(1.), http://www.ncgr.org/research/jag/.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Visscher, P.M., Haley, C.S. and Knott, S.A. (1996a). Mapping QTLs for binary traits in backcross and F_2 populations. *Genetical Research* **68**, 55–63.

Visscher, P.M., Thompson, R. and Haley, C.S. (1996b). Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**, 1013–1020.

Visscher, P.M., Whittaker, J.C. and Jansen, R.C. (2000). Mapping multiple QTL of different effects: comparison of a simple sequential testing strategy and MQM. *Molecular Breeding* **6**, 11–24.

Wang, D.L., Zhu, J., Li, Z.K. and Paterson, A.H. (1999). Mapping QTLs with epistatic effects and QTL × environment interactions by mixed linear model approaches. *Theoretical and Applied Genetics* **99**, 1255–1264.

Weller, J.I. (1986). Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**, 627–640.

Wilks, S.S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* **9**, 60–62.

Xu, S.-Z. (1998). Mapping quantitative trait loci using multiple families of line crosses. *Genetics* **148**, 517–524.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**, 789–801.

Yi, N.J., Xu, S.Z. and Allison, D.B. (2003). Bayesian model choice and search strategies for mapping interacting quantitative trait loci. *Genetics* **165**, 867–883.

Zeng, Z.-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 10972–10976.

Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

Zhu, J., Lum, P.Y., Lamb, J., GuhaThakurta, D., Edwards, S.W., Thieringer, R., Berger, J.P., Wu, M.S., Thompson, J., Sachs, A.B. and Schadt, E.E. (2004). An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenetic and Genome Research* **105**, 363–374.

Mapping Quantitative Trait Loci in Outbred Pedigrees

I. Höschele

Virginia Bioinformatics Institute and Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

In this chapter, we present statistical methods for mapping quantitative-trait loci (QTLs) in outbred or complex pedigrees. Such pedigrees exist primarily in livestock populations, also in human populations, and occasionally in experimental animal or plant populations. The main focus of this chapter is on linkage mapping, but methods for linkage disequilibrium (LD) and combined linkage/LD mapping are also discussed. We describe least-squares and maximum likelihood (ML) methods for estimating QTL effects, and variance-components analysis by approximate (residual) ML for estimating QTL variance contributions. We describe Bayesian QTL mapping, its prior distributions and other distributional assumptions, its implementation via Markov chain Monte Carlo (MCMC) algorithms, its inferences, and contrast it with frequentist methodology. Genotype sampling algorithms using genotypic peeling, allelic peeling or descent graphs are described. Genotype samplers are a critical component of MCMC algorithms implementing ML and Bayesian analyses for complex pedigrees. Lastly, fine-mapping methods including chromosome dissection and LD mapping using current and historical recombinations, respectively, are outlined, and recently developed methods for joint LD and linkage mapping of disease genes and QTL are discussed.

19.1 INTRODUCTION

In this chapter, the focus is on statistical methods for mapping of quantitative-trait loci (QTLs) in populations which have not been formed recently by line crossing, and which have pedigree information available over multiple generations. Methods suitable for QTL mapping in (inbred) line crosses are described in **Chapter 18**. Pedigrees are used for QTL mapping in livestock and human populations, where the development and crossing of inbred lines is not feasible. Therefore, the methods discussed here have applications primarily in mapping genes for quantitative traits of economic importance in livestock and complex disease risk factors in humans. Examples include milk production in dairy

cattle and cholesterol measures in humans. Occasionally, complex pedigrees also arise in experimental animal populations and plants.

In outbred populations, which have not been formed recently, and for a moderate-resolution marker map (e.g. a 10 cM map), disequilibrium measures between markers and QTLs must be expected to be zero. For a moderate-resolution map, distances between QTLs and markers often exceed 1 cM, and therefore any disequilibrium resulting from an event in the not very recent history of a population will have eroded over time. As a consequence, QTL effects cannot be estimated across the population but rather must be estimated within parents, or phase-known QTL genotypes must be inferred for each parent. Consider a parent which is heterozygous at marker M with alleles M_1 and M_2. Suppose we have a large number of offspring from this parent (a half-sib design in cattle). Then if we compare the phenotypes of the two offspring groups inheriting the alternative marker alleles, we may find (1) that M_1 offspring have a higher average phenotype, (2) that M_2 offspring have a higher average phenotype or (3) that there is no detectable difference between the two offspring groups. If we assume that there is a biallelic QTL located near this marker, then in case (1) the allele increasing the phenotype is linked with the M_1 allele, while in (2) it is linked with M_2. In case (3), the parent is homozygous at the QTL. In several of the QTL mapping methods – maximum likelihood (ML), Bayesian – to be presented later, QTLs are often assumed to be biallelic. Although in outbred populations the number of QTL alleles is unknown and may exceed two, the assumption of a biallelic QTL is often invoked because it is difficult to infer the exact number of QTL alleles, and because it may be a good approximation. The latter conjecture results from the fact that a QTL must be detectable, and therefore must have at least one allele with an effect on phenotype clearly distinguishable from the effects of the other alleles. Other QTL mapping methods, such as least-squares (LS) or variance-components analyses, do not make any specific assumption about the number of QTL alleles.

The power of detecting a QTL is limited in outbred pedigrees by the degree of informativeness of the markers and of the QTL. The informativeness of a single marker is measured by its polymorphism information content (PIC) (Botstein *et al.*, 1980; Dekkers and Dentine, 1991; Hoeschele and Romano, 1993). PIC combines heterozygosity of parents with fraction of offspring for which the inheritance at a marker is known. Inheritance is unknown if an offspring has the same marker genotype as both of its parents, or as one of the parents with the other parent unknown. Moreover, these considerations apply to a single marker only. When using multiple linked markers, phases or haplotypes will be unknown and will have to be inferred. For example, a base or founder individual with observed genotype 12/45 at two linked markers ('/' separating the loci) may have phases $1 - 4/2 - 5$ or $1 - 5/2 - 4$ ('/' now separates haplotypes across loci); that is to say, the founder's haplotype pair is either $1 - 4$ and $2 - 5$ or $1 - 5$ and $2 - 4$. If an individual is a descendant, then we need to know not only phase but also parental origin of its haplotypes, i.e. its paternal and maternal haplotypes. These are known if the individual has a known, ordered genotype at each locus. A single-locus genotype is ordered if the parental origin of its two alleles is known (it is 12 (45) if 1 (4) was inherited from the father, or 21 (54) if 2 (5) was inherited from the father).

The degree of heterozygosity at a QTL is also influential in its detection. When the frequency of a detectable QTL allele is very low or very high, a pedigree is likely to contain only families which are not segregating for this allele, and hence the QTL will not be detected. On the other hand, with the additive variance at a biallelic QTL being

$2p(1 - p)a^2$ (p is allele frequency, a is substitution effect), the same amount of variance is explained with p near 0.5 and a smaller a value, and with p low or high and a larger a value, and this may improve the detection of a QTL due to the larger a effect.

When analyzing a pedigree consisting of multiple families, we have two options: (1) analyze each family separately; or (2) analyze all families jointly. Option (1) is not available when family size is small, as is the case in human populations. In certain livestock populations (dairy cattle), however, a single half-sib family may have several hundreds or even thousands of members. However, selecting just one or a few very large families may prevent us from detecting a QTL, unless this QTL happens to be segregating in these few families. Furthermore, analysis of single families limits sample size and hence power of QTL detection.

The complexity of statistical methods for QTL mapping in outbred populations depends on the structure of the population. For analysis of individual large families, or for joint analysis of a small number of such families and when ignoring genetic ties among families, simpler methods used for analysis of (inbred) line crosses (see **Chapter 18**) can be adapted quite easily, e.g. by choosing only informative markers and allowing for offspring with uncertain marker allelic inheritance. In contrast, multigenerational pedigrees, potentially with substantial amounts of missing data, require much more sophisticated methods of analysis.

19.2 LINKAGE MAPPING VIA LEAST SQUARES OR MAXIMUM LIKELIHOOD AND FIXED EFFECTS MODELS

19.2.1 Least Squares

In least-squares (LS) analysis, QTL allelic or genotypic effects are treated as fixed effects. Here we consider LS analysis for half-sib designs. Different markers will be informative in different families. However, different families can be analyzed jointly if a marker map and all-marker interval mapping are used, where a certain QTL position is evaluated in all families irrespective of where the flanking markers are located (i.e. different flanking markers may be selected in different families as the markers must be informative). For an analysis across half-sib families, the following linear regression models have been employed (see also Knott *et al.*, 1994):

$$y_{ij} = \mu_i + p_{ij}\alpha_{i1} + (1 - p_{ij})\alpha_{i2} + e_{ij} = \mu_i^* + p_{ij}\alpha_i + e_{ij}, \qquad e_{ij} \sim \text{iid}N(0, \sigma_e^2),$$
(19.1)

where

$$\alpha_{ik} = \sum_m q_m \mu_{ik,m} \quad \text{for} \quad k = 1, 2; \qquad \alpha_i = \alpha_{i1} - \alpha_{i2}; \qquad \mu_i^* = \mu_i + \alpha_{i2},$$

in which α_{ik} is the additive effect of allele k ($k = 1, 2$) in parent i, $\mu_{ik,m}$ is the expected phenotypic deviation of individuals with genotype $Q_{ik}Q_m$, q_m is the population frequency of allele m, p_{ij} is the probability that offspring j of parent i inherited QTL allele 1 given the marker information or $p_{ij} = E(x_{ij}|\mathbf{M}_{\text{obs}})$ with $x_{ij} = 1$ (0) if $Q_{i1}(Q_{i2})$ was inherited,

$\mathbf{M}_{\mathrm{obs}}$ is a vector of observed marker data, and μ_i and μ_i^* represent (fixed) within-family means for different reparameterizations of the model. The normality assumption in (19.1) is often not explicitly stated, but is needed for the derivation of test statistics and is further discussed below. Model (19.1) can also be reparameterized using the restriction $\alpha_{i1} = -\alpha_{i2}$ for all i, or

$$y_{ij} = \mu_i^{**} + (1 - 2p_{ij})\alpha_i^* + e_{ij}, \tag{19.2}$$

where

$$\alpha_i^* = 0.5(\alpha_{i2} - \alpha_{i1}); \quad \mu_i^{**} = \mu_i + 0.5(\alpha_{i1} + \alpha_{i2}).$$

Note that $|(1 - 2p_{ij})| = 1$ if $p_{ij} = 1$ or $p_{ij} = 0$, and $0 \le |(1 - 2p_{ij})| < 1$ otherwise. Consequently, the average (over all offspring) $|(1 - 2p_{ij})|$ should be calculated as a measure of the informativeness of the flanking markers.

The p_{ij} are calculated by assuming a certain map position of the QTL on a chromosome, using informative flanking markers in each family i, considering all possible linkage phases of the parent in family i, and all possible haplotypes of the offspring. For example, suppose we have a pair of flanking markers and a founder parent (i) with observed genotype $M_1/M_2 = 12/34$. The two possible phase-known (or ordered) genotypes for this parent are denoted by $G_i = 1 - 3/2 - 4$ and $G_i = 1 - 4/2 - 3$. The probability of phase-known genotype (G) of i, given observed marker genotypes on i and on n_i offspring, is computed as

$$\Pr(G_i|\mathbf{M}) = \left[\Pr(G_i)\Pr(M_i|G_i) \prod_{j=1}^{n_i} \sum_{G_{ij}} \Pr(G_{ij}|G_i)\Pr(M_{ij}|G_{ij}) \right] \Big/ \left[\sum_{G_i} \text{numerator} \right], \tag{19.3}$$

where G_i is the multilocus, phase-known marker genotype of parent i, M_i is its observed marker genotype ($M_1/M_2 = 12/34$), G_{ij} is the multilocus, ordered marker genotype of offspring j of i, n_i is the number of offspring of i, $\Pr(M_i|G_i)$ is 1 or 0 depending on whether G_i is consistent with M_i or not, and $\Pr(M_{ij}|G_{ij})$ is 1 or 0 depending on consistency of G_{ij} with M_{ij}. Also, $\Pr(G_i)$ is equal to the population frequency of G_i, and $\Pr(G_{ij}|G_i)$ is equal to the product of the transmission probability of the haplotype that offspring j inherited from parent i and the population frequency of the other haplotype of j. The G_i with the highest probability from (19.3) is chosen and treated as the true linkage phase of parent i, which can be justified only if the number of offspring is sufficiently large so that most likely the correct phase has been identified.

Consider now an offspring j with observed marker genotype 12/35 with possible paternal haplotypes $1 - 3$ and $2 - 3$. Assuming that the phase-known genotype of parent i is $G_i = 1 - 3/2 - 4$, the probabilities of the offspring's ordered, multilocus genotypes are $Pr(G_{ij} = 1 - 3/2 - 5|G_i, M_{ij}) = 0.5(1 - r)q_{12}q_{25}/c$ and $\Pr(G_{ij} = 2 - 3/1 - 5|G_i, M_{ij}) = 0.5rq_{11}q_{25}/c$, respectively, assuming that the other parent of j offspring has no genotype data, letting r denote the recombination rate of the two markers and q_{mk} frequency of allele k at locus m, and with $c = 0.5(1 - r)q_{12}q_{25} + 0.5rq_{11}q_{25}$. Now we postulate the presence of a QTL between the markers. Let $Q_i^1(Q_i^2)$ be the QTL allele in parent i associated with marker haplotype $1 - 3$ ($2 - 4$). Let Q_j^f be the QTL allele that offspring j inherited from father i. The probability that Q_j^f is a copy of Q_i^1,

conditional on observed marker genotypes and i's phase-known genotype, is the p_{ij} value required in (19.1) and (19.2), and is computed as

$$\Pr(Q_j^f \Leftarrow Q_i^1 | G_i, \mathbf{M}) = \sum_{G_{ij}} \Pr(G_{ij} | G_i, M_{ij}) \Pr(Q_j^f \Leftarrow Q_i^1 | G_i, G_{ij})$$

$$= \frac{(1-r)q_{12}}{(1-r)q_{12} + rq_{11}} \cdot \frac{(1-r_{1q})(1-r_{2q})}{1-r}$$

$$+ \frac{rq_{11}}{(1-r)q_{12} + rq_{11}} \cdot \frac{r_{1q}(1-r_{2q})}{r}, \tag{19.4}$$

where r_{kq} denotes recombination rate between marker k ($k = 1, 2$) and the QTL. For calculating the offspring's ordered marker genotype probabilities, additional, informative markers should be used, if any of the flanking markers are not fully informative. For example, there is an additional marker to the left of the first marker (marker 1), and the QTL is still postulated between the other markers (markers 2 and 3). Parent i has genotype $M_i = 12/12/34$, and assume that its phase-known genotype is $G_i = 1 - 1 - 3/2 - 2 - 4$. The offspring has genotype $M_{ij} = 13/12/35$, which implies two possible ordered genotypes, $G_{ij} = 1 - 1 - 3/3 - 2 - 5$ and $G_{ij} = 1 - 2 - 3/3 - 1 - 5$. Now we consider all three markers for calculating p_{ij} instead of only markers 2 and 3 as above. The probabilities of the two G_{ij} are $(1 - r_{12})(1 - r_{23})q_{22}/c$ and $r_{12}r_{23}q_{21}/c$, with $c = (1 - r_{12})(1 - r_{23})q_{22} + r_{12}r_{23}q_{21}$. For p_{ij} values calculated with these probabilities, the quantity $\sum_{ij} |1 - 2p_{ij}|$ is maximized. Note that if the q_{mk} were unknown or both parents were genotyped at marker 2 with their genotypes equal to that of the offspring, markers 1 and 3 (both informative) would be used as flanking markers instead of 2 and 3. As an example, letting marker recombination rates equal 0.1, $q_{22} = 0.1$ and $q_{21} = 0.3$, the p_{ij} values are 0.75 and 0.25 when using only markers 2 and 3, and 0.964 and 0.036 when using markers 1, 2 and 3 as described above.

The null hypothesis of no QTL, $H_0: \alpha_1 = \ldots = \alpha_m = 0$ for m families, is tested against the alternative, $H_A: \alpha_i \neq 0$ for some i, using the test statistic $\{\hat{\alpha}' \mathbf{C}^{\alpha\alpha} \hat{\alpha}/m\}/\{SSE/(n - f)\}$, where SSE is residual sum of squares, n is number of offspring across families, f is number of estimable fixed effects in the model ($f = 2m$), α is a vector containing the α_i, $\hat{\alpha}$ is a subvector of LS estimates of the estimable fixed effects (within-family means and α), and $\mathbf{C}^{\alpha\alpha}$ is the submatrix of the inverse of the coefficient matrix of the normal equations, used to compute the LS estimates, pertaining to α. The distribution of this test statistic is either assumed to be a standard F distribution with m and $n - f$ degrees of freedom (df), or preferably is found using data permutation (Churchill and Doerge, 1994). The F statistic can be converted into an approximate likelihood-ratio statistic using the equation $LR = n \log_e(SSE_{\text{red}}/SSE_{\text{full}})$, where red (full) indicates the model under $H_0(H_A)$, to obtain a likelihood profile for the chromosome.

Such regression models are approximate, because residuals in models (19.1) and (19.2) are not truly iid$N(0, \sigma_e^2)$. Although this assumption is typically made in analyses based on models similar to (19.1) and (19.2), estimation and testing of QTL effect and position appears to be little affected (e.g. Xu, 1995), even if the standard F or t distribution is invoked for testing. Rather than relying on this finding, however, it is prudent to perform data permutation for estimating the exact threshold. A first reason why residual variance is heterogeneous is that it increases with a decrease in the quantity

$|1 - 2p_{ij}|$. Residual variance in models (19.1) and (19.2) equals $\sigma_\varepsilon^2 + \text{var}(x_{ij}|\mathbf{M}_{\text{obs}})$, where $\text{var}(x_{ij}|\mathbf{M}_{\text{obs}}) = p_{ij}(1 - p_{ij})\alpha_i^2$. Xu (1998a) used iteratively reweighted LS to account for heterogeneous variance, but found little difference between LS and this modification. Residual variance is also larger in families with more than average and/or larger than average other QTLs segregating (if these are not accounted for in the model). Next, models (19.1) and (19.2) do not account for the maternally inherited QTL alleles, which causes residuals to have a mixture distribution rather than a single normal distribution. Finally, models (19.1) and (19.2) do not separate polygenic effects out of the residual term. Since the half-sib family means are included as fixed effects, the residual variance contains 75 % of the additive genetic variance. If offspring are not related through their dams, residuals are uncorrelated.

The LS analysis, as described above, fits a single marked QTL. Extension to multiple linked or unlinked QTLs is straightforward and has been described by Spelman *et al.* (1996), Uimari *et al.* (1996) and Zhang *et al.* (1998).

19.2.2 Maximum Likelihood

With QTL effects treated as fixed, one usually assumes a biallelic QTL (although the allele number at any QTL which cannot be genotyped is unknown). We now denote marker information by $\mathbf{M}$ and the vector of QTL genotypes on a pedigree by $\mathbf{G}$. Considering a single QTL and polygenic background, the most general likelihood can be written as

$$L(\mathbf{y}|\mathbf{M}) = \sum_{\mathbf{G}} \Pr(\mathbf{G}|\mathbf{M})f(\mathbf{y}|\mathbf{G}), \qquad \mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V}), \quad \mathbf{V} = \mathbf{Z}\mathbf{A}\mathbf{Z}'\sigma_u^2 + \mathbf{R}\sigma_e^2, \quad (19.5)$$

where $\mathbf{u}$ is vector of polygenic effects with $\text{var}(\mathbf{u}) = \mathbf{A}\sigma_u^2$, $\mathbf{A}$ is the known, additive genetic relationship matrix, σ_u^2 is polygenic variance, $\text{var}(\mathbf{e}) = \mathbf{R}\sigma_e^2$ is the residual variance–covariance matrix with $\mathbf{R}$ being a known, diagonal matrix, σ_e^2 is residual variance, and $\mathbf{Z}$ is an incidence matrix relating phenotypes in $\mathbf{y}$ to individuals. The difficulty with this likelihood is the summation over $\mathbf{G}$ and the evaluation of the multivariate normal density $f(\mathbf{y}|\mathbf{G})$, as $\mathbf{V}$ is not diagonal in the presence of polygenic effects. In an alternative formulation of the likelihood,

$$L(\mathbf{y}|\mathbf{M}) = \sum_{\mathbf{G}} \Pr(\mathbf{G}|\mathbf{M}) \int f(\mathbf{u})f(\mathbf{y}|\mathbf{u}, \mathbf{G}) \, d\mathbf{u}, \qquad (19.6)$$

there is the problem of performing jointly summation with respect to $\mathbf{G}$ and integration with respect to $\mathbf{u}$. Note that for likelihoods (19.5) and (19.6), when marker genotypes are only partially known, $\Pr(\mathbf{G}|\mathbf{M})$ is replaced by $\Pr(\mathbf{M}\mathbf{G}|\mathbf{M}_{\text{obs}})$, where $\mathbf{M}_{\text{obs}}$ is a vector of observed marker data, and $\mathbf{M}\mathbf{G}$ is a vector of complete, multilocus marker-QTL genotypes. Deterministic evaluation of likelihood (19.6) becomes tractable only for simple designs such as the half-sib design with each offspring having a different mother and all parents unrelated. Then,

$$L(\mathbf{y}|\mathbf{M}) = \prod_{i=1}^{F} \sum_{G_i} \Pr(G_i) \int_{-\infty}^{\infty} f(u_i) \prod_{j=1}^{N_i} \sum_{G_{ij}} \Pr(G_{ij}|G_i, M_i, M_{ij})f(y_{ij}|u_i, G_{ij}) \, du_i,$$
$$(19.7)$$

where F is the number of fathers, G_i and M_i are QTL genotype and marker data on father i respectively, N_i is the number of offspring of father i, u_i is the polygenic effect of father i, and G_{ij} and M_{ij} are QTL genotype and marker data on offspring j of father i. This likelihood can be evaluated deterministically using numerical integration (e.g. Morton and MacLean, 1974; Knott *et al.*, 1991). Alternatively, polygenic effects of fathers have been treated as fixed, e.g. by Weller (1986) and by Jansen *et al.* (1998), who use deterministic and stochastic expectation–maximization algorithms to compute ML parameter estimates. With polygenic effects of fathers treated as fixed, the likelihood simplifies to

$$L(\mathbf{y}|\mathbf{M}) = \prod_{i=1}^{F} \sum_{G_i} \Pr(G_i) \prod_{j=1}^{N_i} \sum_{G_{ij}} \Pr(G_{ij}|G_i, M_i, M_{ij}) f(y_{ij}|u_i, G_{ij}). \qquad (19.8)$$

Marker information $\mathbf{M}$ may be on a single marker (e.g. Weller, 1986; Knott and Haley, 1992), or preferably on multiple linked markers to obtain more accurate estimates of QTL effects and location.

Several approximations to likelihoods (19.5), (19.6) and (19.7) have been proposed, e.g. modal estimation (Hoeschele, 1988; Le Roy *et al.*, 1989; Elsen and Le Roy, 1990; Knott *et al.*, 1991; Elsen *et al.*, 1997). In modal estimation, the joint probability density of the vector of phenotypes ($\mathbf{y}$) and the vector of polygenic effects ($\mathbf{u}$) is maximized with respect to $\mathbf{u}$ and other unknowns. For complex pedigrees, likelihood (19.6) can be evaluated using Markov chain Monte Carlo (MCMC) algorithms for sampling genotypes and/or polygenic effects. For details, see **Chapter 33** and Section 19.6 below.

19.3 LINKAGE MAPPING VIA RESIDUAL MAXIMUM LIKELIHOOD AND RANDOM EFFECTS MODELS

In this method, QTL genotype and allelic effects are treated as random, and variance components due to QTL and polygenic effects are estimated. This variance-components method is not restricted to the analysis of individual large families, or to the joint analysis of several large families without relationship ties across families. It is therefore much more general than the regression analysis, and in theory it can be applied to any complex pedigree. In some sense, it is approximate and does not fully utilize the available information. For a discussion on the random treatment of QTL effects and the approximations involved; see Section 19.3.4 below. The variance-components method is based on fewer parametric assumptions than other methods such as ML and Bayesian analyses (often assuming biallelic QTLs), as it does not require us to specify the number of alleles and to estimate their frequencies, and it has been shown to be quite robust to the actual number of alleles at a QTL (Xu and Atchley, 1995; Grignola and Hoeschele, 1997). It has been recognized as a very useful QTL mapping method not only for livestock but also for human pedigrees, in particular with the development of computer packages such as SOLAR (Almasy and Blangero, 1998).

19.3.1 Identity-by-descent Probabilities of Alleles

Random effects at an individual QTL can either be effects of genotypes or of alleles. When random allelic effects are considered, then the covariance matrix among the $2n$

allelic effects at a QTL of the n individuals in a pedigree is needed, which is equal to an identity-by-descent (IBD) matrix times the QTL allelic variance or half of the additive variance contributed by this QTL. An element of the IBD matrix is $\Pr(Q_i^k \equiv Q_{i'}^{k'}|\mathbf{G})$, where Q_i^k is allele k ($k = 1, 2$) in individual i, $\mathbf{G}$ is a set of observed genotypes on the pedigree, and $\equiv$ denotes IBD. These IBD probabilities can be computed recursively, and recursive rules have been stated for various situations, including (1) no observed genotypes available (e.g. DeBoer and Hoeschele, 1993), (2) (partially) observed genotypes available (Davis *et al.*, 1996), and (3) (partially) observed genotypes available for a single locus linked to the locus under consideration (Wang *et al.*, 1995). In all three cases, we have two boundary conditions, which apply to any founder individuals i and i'. These are

$$\Pr(Q_i^k \equiv Q_i^k|\mathbf{G}) = 1,$$

$$\Pr(Q_i^k \equiv Q_{i'}^{k'}|\mathbf{G}) = 0, \qquad \text{for } i \neq i' \text{ or } k \neq k'. \tag{19.9}$$

For case (1) and for allele Q_i^1 in descendant i with father f, where we define $k = 1$ (2) as the allele inherited from the father (mother), we have the recurrence equation

$$\Pr(Q_i^1 \equiv Q_{i'}^{k'}) = \Pr(Q_i^1 \Leftarrow Q_f^1)\Pr(Q_f^1 \equiv Q_{i'}^{k'}) + \Pr(Q_i^1 \Leftarrow Q_f^2)\Pr(Q_f^2 \equiv Q_{i'}^{k'}), \tag{19.10}$$

where i' is an individual which is not a direct descendant of i, $\Pr(Q_i^1 \Leftarrow Q_f^1)$ is the probability that parent f passed on allele Q_f^1 to offspring i, and $\Pr(Q_i^1 \Leftarrow Q_f^1) = \Pr(Q_i^1 \Leftarrow Q_f^2) = 0.5$. Note that in general and with inbreeding, $\Pr(Q_i^1 \Leftarrow Q_f^1)$ is not equal to $\Pr(Q_i^1 \equiv Q_f^1)$, as the latter equals $\Pr(Q_i^1 \Leftarrow Q_f^1) + \Pr(Q_i^1 \Leftarrow Q_f^2)\Pr(Q_f^1 \equiv Q_f^2)$.

If all individuals are genotyped, which is a special case of cases (2) and (3), Davis *et al.* (1996) and Wang *et al.* (1995) expand recurrence (19.10) to also include the alleles in the mother (m), because the parental origins of the alleles in an offspring may be unknown (both parents and the offspring have the same genotype; one parent and the offspring have the same genotype with the other parent untyped). Then the recurrence becomes

$$\Pr(Q_i^k \equiv Q_{i'}^{k'})$$
$$= \Pr(Q_i^k \Leftarrow Q_f^1|\mathbf{G})\Pr(Q_f^1 \equiv Q_{i'}^{k'}|\mathbf{G}) + \Pr(Q_i^k \Leftarrow Q_f^2|\mathbf{G})\Pr(Q_f^2 \equiv Q_{i'}^{k'}|\mathbf{G})$$
$$+ \Pr(Q_i^k \Leftarrow Q_m^1|\mathbf{G})\Pr(Q_m^1 \equiv Q_{i'}^{k'}|\mathbf{G}) + \Pr(Q_i^k \Leftarrow Q_m^2|\mathbf{G})\Pr(Q_m^2 \equiv Q_{i'}^{k'}|\mathbf{G}). \tag{19.11}$$

For case (2), probabilities of the type $\Pr(Q_i^k \Leftarrow Q_f^1|\mathbf{G})$ can be evaluated simply by first evaluating the two probabilities that allele Q_i^k was inherited from the father or the mother respectively, or $\Pr(Q_i^k \Leftarrow (Q_f^1, Q_f^2)|\mathbf{G})$ and $\Pr(Q_i^k \Leftarrow (Q_m^1, Q_m^2)|\mathbf{G})$. When all individuals are typed, then these parental origin probabilities take on one out of only three pairs of values, which are (0,1), (1,0), and (0.5, 0.5). Next, the probabilities of either parental allele being passed on to offspring i, given that Q_i^k originated from this parent (f), are evaluated, e.g. $\Pr(Q_i^k \Leftarrow Q_f^1|Q_i^k \Leftarrow (Q_f^1, Q_f^2), \mathbf{G})$. These probabilities again take on one of only three pairs of values, (0,1), (1,0), or (0.5, 0.5). Then if all individuals are genotyped, $\Pr(Q_i^k \Leftarrow Q_f^1|\mathbf{G}) = \Pr(Q_i^k \Leftarrow (Q_f^1, Q_f^2)|\mathbf{G})\Pr(Q_i^k \Leftarrow Q_f^1|Q_i^k \Leftarrow (Q_f^1, Q_f^2), \mathbf{G})$.

For case (3), and still assuming that all individuals are typed (in this case for markers, not for the locus of interest), probabilities in (19.11) now pertain to a QTL. Equation (19.11) still holds with $\mathbf{G}$ replaced by $\mathbf{M}$ to indicate that genotype data are now available

on a marker. Then, for example, $\Pr(Q_i^k \Leftarrow Q_f^1|\mathbf{G})\Pr(Q_f^1 \equiv Q_{i'}^{k'}|\mathbf{G})$ in (19.11) is replaced by $\Pr(Q_i^k \Leftarrow Q_f^1|\mathbf{M})\Pr(Q_f^1 \equiv Q_{i'}^{k'}|\mathbf{M})$, where

$$\Pr(Q_i^k \Leftarrow Q_f^1|\mathbf{M}) = \Pr(Q_i^k \Leftarrow Q_f^1|M_i^k \Leftarrow M_f^1, \mathbf{M})\Pr(M_i^k \Leftarrow M_f^1|\mathbf{M})$$

$$+ \Pr(Q_i^k \Leftarrow Q_f^1|M_i^k \Leftarrow M_f^2, \mathbf{M})\Pr(M_i^k \Leftarrow M_f^2|\mathbf{M})$$

$$= (1-r)\Pr(M_i^k \Leftarrow M_f^1|\mathbf{M}) + r\Pr(M_i^k \Leftarrow M_f^2|\mathbf{M}), \quad (19.12)$$

and r is the recombination rate between the marker locus and the QTL. Note that to apply (19.11) with (19.12) QTL alleles are identified by their linkage with marker alleles rather than by their parental origin. Consider a mating between individuals 2 (father) and 1 producing offspring 4, and a mating between individuals 2 and 3 producing offspring 5. Assume that all five individuals have marker genotype $M_i^1 M_i^2 = 12$ ($i = 1, \ldots, 5$). Then, the additive relationship between individuals 4 and 5 at the QTL is

$$\text{cov}_{QTL}(4, 5) = 0.5[\Pr(Q_4^1 \equiv Q_5^1|\mathbf{M}) + \Pr(Q_4^1 \equiv Q_5^2|\mathbf{M})$$

$$+ \Pr(Q_4^2 \equiv Q_5^1|\mathbf{M}) + \Pr(Q_4^2 \equiv Q_5^2|\mathbf{M})], \quad (19.13)$$

where, e.g.

$$\Pr(Q_4^1 \equiv Q_5^1|\mathbf{M}) = \Pr(Q_4^1 \Leftarrow Q_2^1|M_4^1 \Leftarrow M_2^1)\Pr(M_4^1 \Leftarrow M_2^1|\mathbf{M})\Pr(Q_5^1 \equiv Q_2^1|\mathbf{M})$$

$$+ \Pr(Q_4^1 \Leftarrow Q_2^2|M_4^1 \Leftarrow M_2^1)\Pr(M_4^1 \Leftarrow M_2^1|\mathbf{M})\Pr(Q_5^1 \equiv Q_2^2|\mathbf{M}),$$

$$\Pr(Q_5^1 \equiv Q_2^1|\mathbf{M}) = \Pr(Q_5^1 \Leftarrow Q_2^1|M_5^1 \Leftarrow M_2^1)\Pr(M_5^1 \Leftarrow M_2^1|\mathbf{M})$$

$$+ \Pr(Q_5^1 \Leftarrow Q_2^2|M_5^1 \Leftarrow M_2^1)\Pr(M_5^1 \Leftarrow M_2^1|\mathbf{M})\Pr(Q_2^2 \equiv Q_2^1|\mathbf{M}).$$

In (19.13), QTL alleles in descendants 4 and 5 are identified by marker alleles; e.g. allele Q_4^1 (Q_4^2) is the QTL allele linked with marker allele $M_4^1 = 1$ ($M_4^2 = 2$) in individual 4, where the marker allele is of unknown parental origin. Alternatively, we can identify QTL alleles in descendants by parental origin ($Q_4^{f=1}$, $Q_4^{m=2}$). Then we need to distinguish four cases: (a) marker allele 1 in individual 4 is paternal and marker allele 1 in 5 is paternal; (b) marker allele 1 in individual 4 is paternal and marker allele 1 in 5 is maternal; (c) marker allele 1 in individual 4 is maternal and marker allele 1 in 5 is paternal; and (d) marker allele 1 in individual 4 is maternal and marker allele 1 in 5 is maternal. Each case has probability 0.25. Then

$$\text{cov}_{\text{QTL}}(4, 5) = 0.5[\Pr(Q_4^f \equiv Q_5^f|(a), \mathbf{M})\Pr((a)|\mathbf{M}) + \Pr(Q_4^f \equiv Q_5^f|(b), \mathbf{M})\Pr((b)|\mathbf{M})$$

$$+ \Pr(Q_4^f \equiv Q_5^f|(c), \mathbf{M})\Pr((c)|\mathbf{M}) + \Pr(Q_4^f \equiv Q_5^f|(d), \mathbf{M})\Pr((d)|\mathbf{M})],$$

$$(19.14)$$

where, e.g.

$$\Pr(Q_4^f \equiv Q_5^f|(a), \mathbf{M}) = \Pr(Q_4^f \Leftarrow Q_2^1|M_4^1 \Leftarrow M_2^1)\Pr(Q_5^f \equiv Q_2^1|(a), \mathbf{M})$$

$$+ \Pr(Q_4^f \Leftarrow Q_2^2|M_4^1 \Leftarrow M_2^1)\Pr(Q_5^f \equiv Q_2^2|(a), \mathbf{M}),$$

$$\Pr(Q_5^f \equiv Q_2^1|(a), \mathbf{M}) = \Pr(Q_5^f \Leftarrow Q_2^1|M_5^1 \Leftarrow M_2^1)$$

$$+ \Pr(Q_5^f \Leftarrow Q_2^2|M_5^1 \Leftarrow M_2^1)\Pr(Q_2^2 \equiv Q_2^1|\mathbf{M}).$$

Both (19.13) and (19.14) lead to the same result, which is $0.5[(1-r)^2 + r^2] + r(1-r)$.

   When parent f is not genotyped, (19.11) no longer holds (Wang *et al.*, 1995). Independence of events in (19.9) and (19.11) such as $Q_i^k \Leftarrow Q_f^1$ and $Q_f^1 \equiv Q_{i'}^{k'}$ no longer holds, implying that the corresponding component in (19.11) must be replaced with $\Pr(Q_i^k \Leftarrow Q_f^1, Q_f^1 \equiv Q_{i'}^{k'}|\mathbf{M})$. For a numerical example showing the discrepancy between (19.11) modified as just described and the original (19.11) for the single-marker case, see Wang *et al.* (1995).

   In general, if some pedigree members are not genotyped, the matrix of IBD probabilities pertaining to the $2n$ alleles or meioses of the n pedigree members at the QTL, conditional on observed marker data, is (Wang *et al.*, 1995)

$$\mathbf{Q}_{\mathbf{M}_{\mathrm{obs}}} = \sum_{\mathbf{M}_i} \mathbf{Q}_{\mathbf{M}_i} \Pr(\mathbf{M}_i|\mathbf{M}_{\mathrm{obs}}), \tag{19.15}$$

where $\mathbf{M}_{\mathrm{obs}}$ is the observed, incomplete marker data, $\mathbf{M}_i$ is a particular set of phase-known (ordered) marker genotypes on the pedigree, which is consistent with $\mathbf{M}_{\mathrm{obs}}$, and $\mathbf{Q}_{\mathbf{M}_i}$ is the IBD matrix conditional on $\mathbf{M}_i$. Note that $\mathbf{Q}_{\mathbf{M}_i}$ can be evaluated recursively with QTL alleles of descendants identified by parental origin as in (19.10).

   When flanking markers, or more generally multiple linked markers, are used to trace inheritance at the QTL, the independence assumption used in (19.9) and (19.11) does not hold even if parent f is genotyped for all markers, but the linkage phase of the markers is unknown. Consider an example with two linked marker loci flanking a QTL. The father has genotypes 12/12 with possible linkage phases $L_f = 1 - 1$ and $L_f = 1 - 2$. A first offspring ($O1$) has genotypes 13/13 with paternal haplotype $1 - 1$, and a second offspring ($O2$) has genotypes 13/23 with paternal haplotype $1 - 2$. We are interested in computing $\Pr(Q_{O1}^f \equiv Q_{O2}^f|\mathbf{M})$ or the IBD probability for the paternally inherited QTL alleles of the offspring:

$$\Pr(Q_{O1}^f \equiv Q_{O2}^f|\mathbf{M}) = \Pr(Q_{O1}^f \Leftarrow Q_f^1, Q_{O2}^f \equiv Q_f^1|\mathbf{M}) + \Pr(Q_{O1}^f \Leftarrow Q_f^2, Q_{O2}^f \equiv Q_f^2|\mathbf{M})$$

$$= \Pr(Q_{O1}^f \Leftarrow Q_f^1, Q_{O2}^f \Leftarrow Q_f^1|\mathbf{M}) + \Pr(Q_{O1}^f \Leftarrow Q_f^2, Q_{O2}^f \Leftarrow Q_f^2|\mathbf{M})$$

(because individual 2 is noninbred)

$$= \Pr(Q_{O1}^f \Leftarrow Q_f^1, Q_{O2}^f \Leftarrow Q_f^1|L_f = 1 - 1, \mathbf{M})\Pr(L_f = 1 - 1|\mathbf{M})$$

$$+ \Pr(Q_{O1}^f \Leftarrow Q_f^1, Q_{O2}^f \Leftarrow Q_f^1|L_f = 1 - 2, \mathbf{M})\Pr(L_f = 1 - 2|\mathbf{M})$$

$$+ \Pr(Q_{O1}^f \Leftarrow Q_f^2, Q_{O2}^f \Leftarrow Q_f^2|L_f = 1 - 1, \mathbf{M})\Pr(L_f = 1 - 1|\mathbf{M})$$

$$+ \Pr(Q_{O1}^f \Leftarrow Q_f^2, Q_{O2}^f \Leftarrow Q_f^2|L_f = 1 - 2, \mathbf{M})\Pr(L_f = 1 - 2|\mathbf{M}) \tag{19.16}$$

where, e.g.

$$\Pr(Q_{O1}^f \Leftarrow Q_f^1, Q_{O2}^f \Leftarrow Q_f^1|L_f = 1 - 1, \mathbf{M})$$

$$= \Pr(Q_{O1}^f \Leftarrow Q_f^1|L_f = 1 - 1, \mathbf{M})\Pr(Q_{O2}^f \Leftarrow Q_f^1|L_f = 1 - 1, \mathbf{M}).$$

The recombination rate between markers is $r = 0.16484$, the recombination rate between left (right) marker and QTL is $r_1 = 0.08$ ($r_2 = 0.101$), and probabilities of linkage phases

are taken as $\Pr(L_f = 1 - 1|\mathbf{M}) = 0.8$ and $\Pr(L_f = 1 - 2|\mathbf{M}) = 0.2$. Then,

$$
\begin{aligned}
\Pr(&Q_{O1}^f \equiv Q_{O2}^f|\mathbf{M}) \\
&= \frac{(1 - r_1)(1 - r_2)}{1 - r} \cdot \frac{(1 - r_1)r_2}{r} \cdot 0.8 + \frac{(1 - r_1)r_2}{r} \cdot \frac{(1 - r_1)(1 - r_2)}{1 - r} \cdot 0.2 \\
&\quad + \frac{r_1 r_2}{1 - r} \cdot \frac{r_1(1 - r_2)}{r} \cdot 0.8 + \frac{r_1(1 - r_2)}{r} \cdot \frac{r_1 r_2}{1 - r} \cdot 0.2 = 0.562\,465\,5. \quad (19.17)
\end{aligned}
$$

However, now let, e.g.

$$
\begin{aligned}
\Pr(Q_{O1}^f \Leftarrow Q_f^1|\mathbf{M}) &= \Pr(Q_{O1}^f \Leftarrow Q_f^1|L_f = 1 - 1, \mathbf{M})\Pr(L_f = 1 - 1|\mathbf{M}) \\
&\quad + \Pr(Q_{O1}^f \Leftarrow Q_f^1|L_f = 1 - 2, \mathbf{M})\Pr(L_f = 1 - 2|\mathbf{M}); \quad (19.18)
\end{aligned}
$$

then

$$
\begin{aligned}
\Pr(Q_{O1}^f \equiv Q_{O2}^f|\mathbf{M}) = 0.562\,465\,5 &\neq \Pr(Q_{O1}^f \Leftarrow Q_f^1|\mathbf{M})\Pr(Q_{O2}^f \Leftarrow Q_f^1|\mathbf{M}) \\
&\quad + \Pr(Q_{O1}^f \Leftarrow Q_f^2|\mathbf{M})\Pr(Q_{O2}^f \Leftarrow Q_f^2|\mathbf{M}). \quad (19.19)
\end{aligned}
$$

When parental linkage phases are unknown, IBD probabilities are in the form of (19.16) irrespective of whether QTL alleles are identified by parental origin or by linkage with alleles at one of the markers. Consequently, in the presence of multiple linked markers with unknown phases of parents, the QTL IBD matrix must generally be evaluated using the form of (19.15). Equation (19.15) must usually be approximated via a Monte Carlo method using samples ($\mathbf{M}_i$) from the conditional distribution $\Pr(\mathbf{M}_i|\mathbf{M}_{\text{obs}})$. Consequently, efficient genotype sampling algorithms for complex pedigrees and partial genotype data are required (see Section 19.6 below).

19.3.2 Mixed Linear Model with Random QTL Allelic Effects

The residual likelihood is based on a linear mixed model including allelic effects at q QTLs and residual polygenic effects, or

$$
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{Z}\mathbf{T}\sum_{i=1}^{q} \mathbf{w}_i + \mathbf{e}, \quad (19.20)
$$

with

$$
\mathbf{u} \sim N(0, \mathbf{A}\sigma_u^2), \qquad \mathbf{w}_i \sim N(0, \mathbf{Q}_i\sigma_{w_i}^2), \qquad \mathbf{e} \sim N(0, \mathbf{R}\sigma_e^2),
$$

$$
\text{cov}(\mathbf{w}_i, \mathbf{w}_j') = 0 \,\forall i, j, \qquad \text{cov}(\mathbf{w}_i, \mathbf{u}') = 0 \,\forall i,
$$

$$
E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \qquad \text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{A}\mathbf{Z}'\sigma_u^2 + \mathbf{Z}\mathbf{T}\sum_{i=1}^{q} \mathbf{Q}_i\mathbf{T}'\mathbf{Z}'\sigma_{w_i}^2 + \mathbf{R}\sigma_e^2,
$$

where $\mathbf{y}$ is vector of m phenotypic observations, $\mathbf{X}$ is a design-covariate matrix, $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{Z}$ is an incidence matrix relating phenotypes to individuals, $\mathbf{u}$ is a vector of $n(n \leq m)$ residual polygenic effects, $\mathbf{T}$ is an incidence matrix relating an individual to its two allelic effects at a QTL, q is the number of QTLs fitted

simultaneously, $\mathbf{w}_i$ is vector of $2n$ allelic effects at QTL i, and $\mathbf{e}$ is vector of m residuals. The assumptions of zero covariances between $\mathbf{u}$ and $\mathbf{w}$ vectors above rely on (1) fitting at most one QTL per marker interval, (2) defining the polygenic effects in $\mathbf{u}$ as the sum of the (very small) allelic effects at all loci that are not fitted as marked QTLs in the model and that are not linked with the marked QTLs and (3) assuming that there is linkage equilibrium with respect to all QTLs in the population.

The residual likelihood is the likelihood of a vector of error contrasts (Patterson and Thompson, 1971), $\mathbf{K}'\mathbf{y}$ ($\mathbf{K}'\mathbf{X} = 0$, $\text{rank}(\mathbf{K}) = m - \text{rank}(\mathbf{X})$). Use of the likelihood of error contrasts instead of the likelihood of $\mathbf{y}$ is advantageous in situations where the model contains many fixed effects. For $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ and $\mathbf{K}'\mathbf{y} \sim N(0, \mathbf{K}'\mathbf{V}\mathbf{K})$, the likelihood of $\mathbf{K}'\mathbf{y}$ is

$$\log f(\mathbf{K}'\mathbf{y}|\sigma) \propto -0.5 \log |\mathbf{K}'\mathbf{V}\mathbf{K}| - 0.5(\mathbf{K}'\mathbf{y})'(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{y})$$

$$\propto -0.5 \log |\mathbf{V}| - 0.5 \log |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - 0.5(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \quad (19.21)$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

assuming that $\mathbf{X}$ is reparameterized to full rank. This likelihood can be evaluated very efficiently for simple designs, e.g. half-sib families (Grignola *et al.*, 1996; Xu, 1998b). For large complex pedigrees, likelihood evaluation is more challenging and requires the use of algorithms such as derivative-free (e.g. Graser *et al.*, 1987, Grignola and Hoeschele, 1997) or derivative-intense (e.g. Meyer and Smith, 1996; Neumaier and Groeneveld, 1999) methods. In residual or restricted maximum likelihood (REML), likelihood (19.21) is maximized with respect to all variance components. More information about the REML method is given in **Chapter 20**.

19.3.3 Mixed Linear Model with Random QTL Genotypic Effects

A mixed model fitting QTL genotypic effects represents an alternative to the allelic effects model. While the latter has been developed in animal genetics, the former model is utilized in the field of human genetics and has been implemented in the computer package SOLAR (Almasy and Blangero, 1998). It includes dominance and two-locus epistasis, although the covariance structure is approximate under inbreeding. The genotypic effects mixed model with Q QTLs is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{Z}\sum_{i=1}^{Q} \mathbf{g}_i + \mathbf{e}, \quad (19.22)$$

where $\mathbf{g}_i$ is a vector of genotypic effects of the n members of the pedigree for QTL i. Under this model a typical element of $\mathbf{V} = \text{var}(\mathbf{y})$ is (ignoring epistasis)

$$\text{cov}(y_j, y_k) = V_{jk} = \sum_{i=1}^{Q} (\pi_{i,jk(\mathbf{M}_{\text{obs}})}\sigma_{a(i)}^2 + \phi_{i,jk(\mathbf{M}_{\text{obs}})}\sigma_{d(i)}^2) + A_{jk}\sigma_u^2, \quad (19.23)$$

where $\sigma_{a(i)}^2$ and $\sigma_{d(i)}^2$ are additive and dominance variance at QTL i, σ_u^2 is polygenic variance, A_{jk} is the additive genetic relationship between relatives j and k for the

unmarked portion of the genome, and $\pi_{i,jk(\mathbf{M}_{\text{obs}})}$ and $\phi_{i,jk(\mathbf{M}_{\text{obs}})}$ are additive and dominance relationships for QTL i conditional on observed marker data. More specifically,

$$\pi_{i,jk(\mathbf{M}_{\text{obs}})} = 0.5(\Pr(Q_{i,j}^1 \equiv Q_{i,k}^1 | \mathbf{M}_{\text{obs}}) + \Pr(Q_{i,j}^2 \equiv Q_{i,k}^2 | \mathbf{M}_{\text{obs}}) \tag{19.24}$$

$$+ \Pr(Q_{i,j}^1 \equiv Q_{i,k}^2 | \mathbf{M}_{\text{obs}}) + \Pr(Q_{i,j}^2 \equiv Q_{i,k}^1 | \mathbf{M}_{\text{obs}})),$$

$$\phi_{i,jk(\mathbf{M}_{\text{obs}})} = \Pr(Q_{i,j}^1 \equiv Q_{i,k}^1, Q_{i,j}^2 \equiv Q_{i,k}^2 | \mathbf{M}_{\text{obs}}) + \Pr(Q_{i,j}^1 \equiv Q_{i,k}^2, Q_{i,j}^2 \equiv Q_{i,k}^1 | \mathbf{M}_{\text{obs}}).$$
$$\tag{19.25}$$

With inbreeding, additional coefficients and components of covariance arise (e.g. a covariance between additive and dominance effects), which are typically ignored (e.g. DeBoer and Hoeschele, 1993).

In matrix notation and assuming single records or $\mathbf{Z} = \mathbf{I}$ in (19.22) for simplicity,

$$\text{var}(\mathbf{y}) = \mathbf{V} = \sum_{i=1}^{Q} (\mathbf{\Pi}_{i(\mathbf{M}_{\text{obs}})} \sigma_{a_i}^2 + \mathbf{\Phi}_{i(\mathbf{M}_{\text{obs}})} \sigma_{d_i}^2) + \mathbf{A}\sigma_u^2 + \mathbf{R}\sigma_e^2, \tag{19.26}$$

where

$$\mathbf{\Pi}_{i(\mathbf{M}_{\text{obs}})} = [\pi_{i,jk(\mathbf{M}_{\text{obs}})}], \; \mathbf{\Phi}_{i(\mathbf{M}_{\text{obs}})} = [\phi_{i,jk(\mathbf{M}_{\text{obs}})}]. \tag{19.27}$$

Two-QTL epistasis can formally be incorporated in a straightforward manner. For example, for QTLs i and i' and additive-by-additive epistasis, we add the component $\mathbf{\Pi}_{i(\mathbf{M}_{\text{obs}})} \odot \mathbf{\Pi}_{i'(\mathbf{M}_{\text{obs}})} \sigma_{aa_{ii'}}$, where $\odot$ denotes the Hadamard product (with typical element equal to the product of the two elements in the same position in the original two matrices).

The log-likelihood function of the phenotypes, assuming multivariate normality and a set of map locations of the QTLs, is:

$$\ln f(\mathbf{y}|\boldsymbol{\beta}, (\sigma_{a_i}^2, \sigma_{d_i}^2, \sigma_{aa_{ii'}}; i, i' = 1, \ldots, Q; i < i'), \sigma_u^2, \sigma_e^2;$$

$$(\mathbf{\Pi}_{i(\mathbf{M}_{\text{obs}})}, \mathbf{\Phi}_{i(\mathbf{M}_{\text{obs}})}; i = 1, \ldots, Q), \mathbf{A}, \mathbf{R}). \tag{19.28}$$

Estimation of $\mathbf{\Pi}_{i(\mathbf{M}_{\text{obs}})}$, $i = 1, \ldots, Q$, at a specific, marked location in the genome requires (1) estimation of $\mathbf{\Pi}_m$ for a number (M) of linked markers ($m = 1, \ldots, M$), and (2) estimation of the elements of $\mathbf{\Pi}_{i(\mathbf{M}_{\text{obs}})}$ using multiple regression. For step (1), if marker genotypes are observed on all individuals, $\mathbf{\Pi}_m$ is estimated with the algorithm of Davis *et al.* (1996). If some marker genotypes are missing, these are imputed using Monte Carlo, $\mathbf{\Pi}_m$ is computed for each imputed vector, and subsequently a weighted average of the $\mathbf{\Pi}_m$ matrices is computed (Almasy and Blangero, 1998).

For step (2), Fulker *et al.* (1995) developed a method performing approximate multipoint calculations for sib pairs. This method was generalized by (Almasy and Blangero, 1998) to other types of relatives. For a given relationship R (e.g. half-sibs), the correlation between IBD values of marker q and QTL m, for a given QTL location (determining recombination rate r with the marker), is calculated using the following results:

$$\rho_{qm}(R, r) = \frac{\text{cov}(\pi_q, \pi_m)}{\sqrt{\text{var}(\pi_q)\,\text{var}(\pi_m)}} = \frac{\text{cov}(\pi_q, \pi_m)}{\text{var}(\pi)},$$

$$E(\pi) = E(\pi_q) = E(\pi_m), \tag{19.29}$$

$$\mathrm{var}(\pi) = \mathrm{var}(\pi_q) = \mathrm{var}(\pi_m) = E(\pi^2) - [E(\pi)]^2,$$

$$\mathrm{cov}(\pi_q, \pi_m) = \sum_{i,j} \mathrm{Pr}(\pi_q = i, \pi_m = j)[i - E(\pi)][j - E(\pi)].$$

As an example, for half-sibs, probabilities of sharing genes IBD are

$\mathrm{Pr}(\pi_q, \pi_m)$	$\pi_q = 0(i_q = 0)$	$\pi_q = 0.5(i_q = 1)$	$\mathrm{Pr}(\pi_m)$
$\pi_m = 0 \ (i_m = 0)$	$0.5 - r + r^2$	$r - r^2$	0.5
$\pi_m = 0.5 \ (i_m = 1)$	$r - r^2$	$0.5 - r + r^2$	0.5
$\mathrm{Pr}(\pi_q)$	0.5	0.5	1

with $E(\pi) = 0.5 \times 0 + 0.5 \times 0.5 = 0.25$, $\mathrm{var}(\pi) = 0.5(0 - 0.25)^2 + 0.5(0.5 - 0.25)^2 = 0.0625$, and $\mathrm{cov}(\pi_q, \pi_m) = (0.5 - r + r^2)(0 - 0.25)^2 + 2(r - r^2)(0 - 0.25)(0.5 - 0.25) + (0.5 - r + r^2)(0.5 - 0.25)^2 = 0.0625 - 0.25(r - r^2)$. Note that $\mathrm{cov}(\pi_q, \pi_m) = 0.0625 - 0.25(r - r^2) = 0.0625(1 - 2r)^2 = \mathrm{var}(\pi)(1 - 2r)^2$ (analogous results hold for other types of relatives, as shown by Almasy and Blangero (1998)). As an example, $\mathrm{Pr}(\pi_m = 0, \pi_q = 0) = 0.5 - r + r^2 = (2 \times 0.5 \times 0.5)(1 - r)^2$ (both half-sibs nonrecombinant for different alleles) $+ (2 \times 0.5 \times 0.5)r^2$ (both half-sibs recombinant for different alleles). Then, multiple regression with M markers is used to estimate the IBD value at a QTL for a given location, or

$$\pi_q = \alpha + \sum_{m=1}^{M} \beta_m \pi_m, \tag{19.30}$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_\eta \end{bmatrix} = \begin{bmatrix} \mathrm{var}(\pi_1) & \cdots & \mathrm{cov}(\pi_1, \pi_\eta) \\ \vdots & \vdots & \vdots \\ \mathrm{cov}(\pi_\eta, \pi_1) & \cdots & \mathrm{var}(\pi_\eta) \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{cov}(\pi_1, \pi_q) \\ \vdots \\ \mathrm{cov}(\pi_\eta, \pi_q) \end{bmatrix}$$

and

$$\alpha = E(\pi_q) - \sum_{m=1}^{\eta} \beta_m E(\pi_m).$$

In reality, the π_m values are unknown and estimated in step 1. For each marker, from the estimated π_m values for a particular type of relatives (e.g. half-sibs), $\mathrm{var}(\pi)$ can be estimated empirically and used to obtain an empirical estimate of $\mathrm{cov}(\pi_m, \pi_q)$. These are then used to estimate $\boldsymbol{\beta}$ in (19.30). Subsequently, α is evaluated using the mean of the unobserved true values of π_q for a given relative type, and the empirical mean of the estimated π_m values for each marker m.

19.3.4 Relationship with Other Likelihood Methods

In the REML method as described, a likelihood of error contrasts is maximized conditionally on predetermined covariance matrices for each QTL, which depend on the observed marker data and QTL positions. Hence we denote the logarithm of the likelihood by $L(\mathbf{K}'\mathbf{y}|\mathbf{Q}_{\mathbf{M}_{\mathrm{obs}}})$. This likelihood is an approximation, which does not fully capture all the

available information. A major source of approximation is that the complete marker data are inferred conditionally on observed marker data, but not on phenotypic data (if markers are linked to QTLs, their genotype probabilities depend on phenotypes). Therefore, other log likelihoods need to be considered. A first alternative log likelihood is

$$L^*(\mathbf{K}'\mathbf{y}) = \log\left(\sum_{\mathbf{M}} P(\mathbf{M})P(\mathbf{M}_{\text{obs}}|\mathbf{M})e^{L(\mathbf{K}'\mathbf{y}|\mathbf{Q_M})}\right), \tag{19.31}$$

which takes account of the distribution of the complete multilocus marker genotypes given the phenotypes and observed marker data. This likelihood is similar to the likelihood of Schork (1993). In Schork's method, originally proposed by Goldgar (1990), a covariance matrix is constructed for a chromosomal region (as opposed to a specific QTL point) based on the proportion of IBD sharing of two sibs in this region. This proportion depends on the combination of IBD and recombination states at a number of markers. While Goldgar replaces the IBD proportion for the region with its expected value given the observed marker data (analogous to $L(\mathbf{K}'\mathbf{y}|\mathbf{Q_{M_{\text{obs}}}})$), Schork takes account of the distribution of the indicator variable representing the combination of IBD and recombination states, pertaining to two sibs, of the markers in the region. Another, alternative log likelihood is

$$L^{**}(\mathbf{K}'\mathbf{y}) = \log\left(\sum_{\mathbf{M},\mathbf{G}} P(\mathbf{M},\mathbf{G})P(\mathbf{M}_{\text{obs}}|\mathbf{M})e^{L(\mathbf{K}'\mathbf{y}|\mathbf{G})}\right), \tag{19.32}$$

where $\mathbf{G}$ is a matrix of ordered QTL genotypes on the pedigree. Likelihood (19.32) takes account of the distribution of the complete marker-QTL genotypes given the phenotypes and in that sense fully utilizes the available information. While likelihood (19.31) is based on $\mathbf{w}_i \sim N(0, \mathbf{Q}_{i|\mathbf{M}}\sigma^2_{w_i})$, with $\mathbf{Q}_{i|\mathbf{M}}$ being the covariance matrix for QTL i conditional on a complete marker genotype configuration $\mathbf{M}$ on the pedigree (19.32) is a likelihood conditional on a particular QTL genotype configuration on the pedigree. Likelihood (19.32) assumes that there are $2n_b$ alleles (n_b is the number of founders of the pedigree) with $w_{ik} \sim N(0, \sigma^2_{w_i})$ for founder allele k at locus i, and all possible ways in which the founder alleles can be passed on to offspring are considered (e.g. Hoeschele *et al.*, 1997; Xu, 1996). It may also assume any other number of alleles (if there are two alleles and their effects are treated as fixed, then the likelihood described in Section 19.3 is obtained). The likelihood described by Xu (1996), which he refers to as the full likelihood or the distribution method, can be considered as an exact or approximate (e.g. in the accounting for inbreeding) reparameterization of log likelihood (19.32). We may describe log likelihood (19.32) as the *distribution* log likelihood, while log likelihood $L(\mathbf{K}'\mathbf{y}|\mathbf{Q_{M_{\text{obs}}}})$ may be described as the *expectation* log likelihood (Gessler and Xu, 1996).

In contrast to log likelihood (19.32), log likelihood $L(\mathbf{K}'\mathbf{y}|\mathbf{Q_{M_{\text{obs}}}})$ does not require us to specify the number of alleles at a QTL. The vector of allelic ($\mathbf{w}$) or genotypic ($\mathbf{g}$) QTL effects for all individuals in the pedigree may be written as (R. Fernando, personal communication) $\mathbf{w} = \mathbf{S}_w\mathbf{a}$ or $\mathbf{g} = \mathbf{S}_g\mathbf{a}$, where, e.g. for the case of additive gene action, $\mathbf{a}$ is a vector of (fixed) effects of the different allelic forms at the QTL, and $\mathbf{S}_w$ and $\mathbf{S}_g$ are matrices of random design variables relating effects of allelic forms to effects of alleles or genotypes of individuals (e.g. for a biallelic QTL, $\mathbf{a} = [a, -a]'$, $\mathbf{S}_w$ has two columns, with each row being either [1,0] or [0,1], and $\mathbf{S}_g$ has two columns, with each row being either [2,0] or [1,1], or [0,2]). This is analogous to the random

residual term in, e.g. a completely randomized design, which is a function of fixed experimental unit effects and design random variables. In this setting, QTL allelic (**w**) or genotypic effects (**g**) are random because the association of allelic forms with alleles and genotypes of individuals is random. The distributional assumption $\mathbf{w}_i \sim N(0, \mathbf{Q}_i \sigma_{wi}^2)$ in (19.20) is based on recurrence equations of the type (dropping subscript i for QTL) $w_o^1 = \Pr(Q_o^1 \Leftarrow Q_f^1)w_f^1 + \Pr(Q_o^1 \Leftarrow Q_f^2)w_f^2 + \varepsilon_o^1$, where w_o^1 is the effect of allele Q_o^1, which offspring o inherited from its father f. Because either $w_o^1 = w_f^1$ or $w_o^1 = w_f^2$, residual ε_o^1 is discrete and hence not normally distributed. Consequently, the normality assumption is another source of approximation in the variance-components method. In conclusion, the expectation log likelihood $L(\mathbf{K}'\mathbf{y}|\mathbf{Q}_{\mathbf{M}_{obs}})$ is less parametric and hence more robust than the distribution log likelihood (19.32), while the distribution log likelihood may have slightly more power when its assumptions are not violated.

19.4   LINKAGE MAPPING VIA BAYESIAN METHODOLOGY

19.4.1   General

In Bayesian analysis, we first set up a probability model by defining the joint distribution of all observed and unobserved (or known and unknown) quantities in a problem. For the simpler problem of mapping a Mendelian disease gene, observed quantities (data) are observed genotypes at a marker and at the disease gene, and an unknown quantity is the recombination rate between these two loci. Obtaining inferences about unknown quantities is based on Bayes's theorem, or

$$\Pr(\text{unknown quantities} \mid \text{known quantities}) = \frac{\Pr(\text{unknown quantities, known quantities})}{\Pr(\text{known quantities})}$$

$$= \frac{\Pr(\text{unknown quantities})\Pr(\text{known quantities} \mid \text{unknown quantities})}{\Pr(\text{known quantities})}. \tag{19.33}$$

In (19.33), Pr(.) represents a probability for a discrete quantity (e.g. genotype) and a probability density for a continuous quantity (e.g. recombination rate). In the frequentist setting, probability refers to hypothetical repeated sampling and is a long-run frequency; e.g. the Type I error rate represents the long-run frequency of incorrectly rejecting a null hypothesis over repeated experiments in which the null hypothesis is true. In contrast, in the Bayesian framework, probability refers to the uncertainty attached to the true values of the unknown quantities (note that unknown quantities can be parameters, random effects and missing data in the frequentist setting), unconditionally (prior) or conditionally (posterior) on the observed data. In (19.33), Pr(unknown quantities) represents the probability density of the prior distribution of the unknowns, while Pr(unknown quantities | known quantities) represents the probability density of the posterior distribution of the unknowns. The prior describes the degree of uncertainty about or belief in any values of the unknowns unconditionally on the current, observed data, while the posterior reflects the degree of uncertainty about or belief in any values of the unknowns conditionally on the current, observed data.

One criticism of Bayesian analysis focuses on the prior distribution of the unknowns, with the argument that different investigators, using the same data but their own different

priors, may arrive at different inferences about the unknowns. One may disagree with this criticism for the following reasons. First, if priors are chosen carefully (so that they do not lead to improper posterior distributions), use of different priors often leads to practically identical inferences, if there is sufficient information in the data about the unknowns. If different priors lead to different answers, we learn that we must collect more data to obtain strong and reliable inferences. Therefore, a careful analyst will conduct a robustness study evaluating the effect of different priors on the posterior inferences. Secondly, Bayesian analysis provides us with the opportunity to incorporate biologically meaningful prior information. In linkage analysis, frequentist inferences are based on the null hypothesis that there are no QTLs segregating in the entire genome. However, this null hypothesis is usually irrelevant, because we know from the heritability of a trait and from previous QTL mapping experiments that there are QTLs segregating. We can construct a prior reflecting this information (see below) and vary this prior to investigate whether it impacts our results. Another example of prior biological information is the knowledge from an annotated genome sequence and mutation database (Rannala and Reeve, 2001).

19.4.2 Bayesian Mapping of a Monogenic Trait

Vieland (1998) provides a nice introduction to Bayesian linkage analysis for the case of a single-marker and a single trait gene. The Bayesian analysis summarizes evidence in favor of linkage between two loci by the posterior probability of linkage (*PPL*), or the probability of linkage given the data. For this simpler problem, the data ($\mathbf{D} = [\mathbf{M}, \mathbf{T}]$) consist of the observed genotypes at the marker ($\mathbf{M}$) and at the trait locus ($\mathbf{T}$), and we are asking whether these two loci are linked. Let $H_0: r = 0.5$ represent the null hypothesis of no linkage and $H_L: 0 \leq r < 0.5$ the alternative hypothesis of linkage. Then

$$PPL = \Pr(H_L|\mathbf{D}) = \frac{\Pr(H_L)\Pr(\mathbf{D}|H_L)}{\Pr(H_L)\Pr(\mathbf{D}|H_L) + [1 - \Pr(H_L)]\Pr(\mathbf{D}|H_0)]},$$

where

$$\Pr(H_L)\Pr(\mathbf{D}|H_L) = \Pr(H_L) \int_{r \geq 0}^{r < 0.5} \Pr(\mathbf{D}|r)\Pr(r|H_L)\,dr. \qquad (19.34)$$

The prior probability of linkage is obtained as the product $\Pr(M$ and T locus on the same chromosome) $\Pr(\delta < \delta_{min})$, where δ represents genetic distance related to r by a mapping function, and δ_{min} represents the minimum distance for which $r = 0.5$. The two probabilities are derived under the assumption that both loci can be located anywhere in the genome with equal probability. If all N_c chromosomes were of equal length, then $\Pr(M$ and T locus on the same chromosome) $= 1/N_c$, but unequal length of chromosomes can be incorporated (e.g. Hoeschele and VanRaden, 1993). Then, given that both loci are on the same chromosome c, $\ell_M \sim U(0, L_c)$ and $\ell_T \sim U(0, L_c)$, where ℓ denotes a map position. From the joint probability density $\Pr(\ell_M, \ell_T)$ being equal to the product of the two uniform densities, $\Pr(\ell_M, \delta)$ is obtained by change of variables, $\Pr(\delta)$ by integration, and $\Pr(\delta < \delta_{min})$ and $\Pr(r)$ are obtained via a mapping function (for details, see Hoeschele and VanRaden, 1993).

Evidence in favor of linkage can be accumulated using Bayes's theorem. Let $\mathbf{D}_1$ and $\mathbf{D}_2$ represent a first and a later data set, which are distributed independently of each other. Then, in (19.34), $\mathbf{D}$ is replaced by $\mathbf{D}_2$, and the priors $\Pr(H_L)$ and $\Pr(r|H_L)$ are replaced by the posteriors $\Pr(H_L|\mathbf{D}_1)$ and $\Pr(r|H_L, \mathbf{D}_1)$. These replacements follow from

the equations

$$PPL = \Pr(H_L|\mathbf{D}_1, \mathbf{D}_2) = \frac{\Pr(H_L)\Pr(\mathbf{D}_1, \mathbf{D}_2|H_L)}{\Pr(H_L)\Pr(\mathbf{D}_1, \mathbf{D}_2|H_L) + [1 - \Pr(H_L)]\Pr(\mathbf{D}_1, \mathbf{D}_2|H_0)}$$

$$= \frac{\Pr(H_L|\mathbf{D}_1)\Pr(\mathbf{D}_2|H_L)}{\Pr(H_L|\mathbf{D}_1)\Pr(\mathbf{D}_2|H_L) + [1 - \Pr(H_L|\mathbf{D}_1)]\Pr(\mathbf{D}_2|H_0)},$$

(19.35)

where

$$\Pr(H_L)\Pr(\mathbf{D}_1, \mathbf{D}_2|H_L) \propto \Pr(H_L|\mathbf{D}_1)\Pr(\mathbf{D}_2|H_L)$$

$$= \Pr(H_L|\mathbf{D}_1) \cdot \int_{r \geq 0}^{r < 0.5} \Pr(\mathbf{D}_2|r)\Pr(r|H_L)\, dr.$$

Vieland (1998) also relates *PPL* to the standard *logarithm of odds* (*LOD*) score (Barnard, 1949). The antilog of the *LOD* score is the frequentist likelihood ratio LR

$$LR = \frac{\Pr(\mathbf{D}|r = r_{\text{sup}})}{\Pr(\mathbf{D}|r = 0.5)},$$

(19.36)

where r_{sup} is the value of r which maximizes $\Pr(\mathbf{D}|r)$. *LR* represents the relative probabilities of the data $\mathbf{D}$ given the two different r values. The corresponding posterior odds ratio is equal to the prior odds ratio times the likelihood ratio, or

$$\frac{\Pr(r = r_{\text{sup}}|\mathbf{D})}{\Pr(r = 0.5|\mathbf{D})} = \frac{\Pr(r = r_{\text{sup}})}{\Pr(r = 0.5)} \cdot \frac{\Pr(\mathbf{D}|r = r_{\text{sup}})}{\Pr(\mathbf{D}|r = 0.5)},$$

(19.37)

and it represents the likelihood of the two different r values given the data. Consider the alternative, composite posterior odds ratio,

$$\frac{\Pr(r < 0.5|\mathbf{D})}{\Pr(r = 0.5|\mathbf{D})} = \frac{\Pr(H_L|\mathbf{D})}{\Pr(H_0|\mathbf{D})} = \frac{\Pr(H_L)}{\Pr(H_0)} \cdot \frac{\Pr(\mathbf{D}|H_L)}{\Pr(\mathbf{D}|H_0)},$$

(19.38)

with $\Pr(H_L|\mathbf{D})$ defined in (19.34). *LR* does not involve the prior distribution of r, while the composite odds ratio does. *LR* (or the odds ratio in (19.37)) does not use all the available information in situations where a meaningful prior distribution can be specified, while the composite odds ratio in (19.38) does.

PPL should not be confused with another quantity, which is the probability of actual linkages among all significant results, or 1 minus the false positive rate (Genin *et al.*, 1995; Vieland, 1998; Southey and Fernando, 1998; Morton, 1955; 1998). If linkage is declared whenever *LOD* > 3, then

$$\Pr(H_L|LOD \geq 3) = \frac{\Pr(H_L)\Pr(LOD \geq 3|H_L)}{\Pr(H_L)\Pr(LOD \geq 3|H_L) + [1 - \Pr(H_L)]\Pr(LOD \geq 3|H_0)}$$

$$= \frac{\Pr(H_L)(1 - \beta)}{\Pr(H_L)(1 - \beta) + [1 - \Pr(H_L)] \cdot \alpha},$$

(19.39)

where α and β are Type I and Type II error probabilities. This quantity is called the *reliability* in statistics (e.g. Stuart and Ord, 1987). Note that *PPL* is the posterior

probability of linkage or the probability of linkage given the observed data, while reliability is a frequentist quantity representing the long-run frequency or probability of linkage given that a particular action has been taken (acceptance of linkage), which depends on a preset value for α but not on the number of tests performed. Southey and Fernando (1998) discuss this quantity in the context of multiple QTL mapping, and Weller *et al.* (1998) attempt to control the false positive rate, or 1 minus the probability of linkage given acceptance of linkage.

19.4.3 Bayesian QTL Mapping

19.4.3.1 General

Mapping QTLs is much more complex than mapping Mendelian disease genes, because multiple QTLs, polygenic background and gene interactions may all act on the trait under consideration, because the number of QTLs is unknown, and because genotypes at the QTLs are not observed but rather phenotypes, pedigree and marker genotypes are used to infer QTL genotypes. For the mapping of QTLs via Bayesian analysis, we begin by defining the known and unknown quantities in this problem. The known quantities include the vector of phenotypes ($\mathbf{y}$) and the observed marker genotypes ($\mathbf{M}_{\text{obs}}$) at multiple linked markers on one or several chromosomes (and the pedigree). The unknown quantities include the number of QTLs (q); the ordered genotypes of all individuals at all QTLs ($\mathbf{G}$); the ordered genotypes of all individuals at all markers ($\mathbf{M}$); QTL locations modeled as linkage status ($\ell_Q = \{\ell_i\}_{i=1,\ldots,q}$) (representing whether QTL i is located on a particular marked chromosome ($\ell_i = c$), or in the residual, unmarked genome ($\ell_i = 0$)); map positions ($\mathbf{t}_Q = \{t_{i(Q)}\}_{i=1,\ldots,q}$); map positions of markers if treated as unknown (conditional or unconditional on order) ($\mathbf{t}_M$); allele frequencies at the QTLs ($\mathbf{p}_Q = \{p_{i(Q)}\}_{i=1,\ldots,q}$) and at the markers if treated as unknown ($\mathbf{p}_M$); QTL genotypic effects (additive effects $\mathbf{a} = \{a_i\}_{i=1,\ldots,q}$; dominance effects $\mathbf{d} = \{d_i\}_{i=1,\ldots,q}$; gene interactions); systematic environmental effects ($\mathbf{b}$); parameters of the residual distribution of the phenotypes ($\mathbf{t}_e$); and parameters of the polygenic background variation ($\mathbf{u}$). Inferences about all unknown quantities are based on the joint posterior probability density, or

$$\Pr(q, \mathbf{G}, \mathbf{M}, \ell_Q, \mathbf{t}_Q, \mathbf{t}_M, \mathbf{p}_Q, \mathbf{p}_M, \mathbf{a}, \mathbf{d}, \mathbf{b}, \mathbf{t}_e, \mathbf{u}|\mathbf{y}, \mathbf{M}_{\text{obs}})$$
$$\propto \Pr(q)\Pr(\ell_Q|q)\Pr(\mathbf{t}_Q|\ell_Q, q)\Pr(\mathbf{t}_M)\Pr(\mathbf{p}_Q|q)\Pr(\mathbf{p}_M)$$
$$\Pr(\mathbf{G}, \mathbf{M}|q, \ell_Q, \mathbf{t}_Q, \mathbf{t}_M, \mathbf{p}_Q, \mathbf{p}_M)\Pr(\mathbf{M}_{\text{obs}}|\mathbf{M})$$
$$\Pr(\mathbf{a}, \mathbf{d}|\mathbf{p}_Q, q)\Pr(\mathbf{b})\Pr(\mathbf{t}_e)\Pr(\mathbf{u})\Pr(\mathbf{y}|\mathbf{G}, \mathbf{a}, \mathbf{d}, \mathbf{b}, \mathbf{t}_e, \mathbf{u}). \qquad (19.40)$$

Details on current implementations of the Bayesian analysis can be found in several contributions (e.g. Heath, 1997; 1999; Hoeschele, 1999; Satagopan *et al.*, 1996; Satagopan and Yandell, 1996; Sillanpää and Arjas, 1998; 1999; Stephens and Smith, 1993; Stephens and Fisch, 1998; Thomas *et al.*, 1997; Thomas and Cortessis, 1992; Uimari and Hoeschele, 1997). In the posterior probability density (19.40) and in the list of unknowns in the Bayesian QTL mapping problem, we have left out one set of unknowns of importance, the number of alleles at each QTL. In all current implementations of the Bayesian analysis, the number of alleles at a QTL is not treated as an unknown, but rather as a predetermined number na, set equal to the extreme values of either $na = 2$ (biallelic

QTL) or $na = 2n_f$, with n_f being the number of founders in the pedigree (e.g. Hoeschele *et al.*, 1997). Although the number of alleles should be treated as unknown, doing so adds another level of complexity to the analysis. Next we describe briefly the most important features of the sampling scheme, the prior distributions, the conditional distributions of the knowns and the inferences.

Inferences about the unknowns are based on samples of all unknowns obtained from the joint posterior distribution via Markov chain Monte Carlo MCMC algorithms (e.g. Gilks *et al.*, 1996). Groups of unknowns are sampled jointly and conditionally on all other groups of unknowns. The sampling scheme depends somewhat on how the genotypes are sampled (see Section 19.6). If genotypes are sampled one locus at a time, then a rough outline of the sampling scheme is as follows:

1. For each marker locus (one locus at a time), sample ordered genotypes, map positions (if treated as unknown) and allele frequencies (if treated as unknown).

2. For each QTL, sample linkage status and map position, additive and dominance effects, ordered QTL genotypes and allele frequency.

3. Sample polygenic effects.

4. Sample classification effects and regression parameters.

5. Sample parameters of the error distribution of phenotypes.

6. Sample QTL number (i.e. attempt move to a different QTL number).

Most of these steps are either Metropolis–Hastings (MH) or Gibbs steps, while step (6) requires a reversible jump step (e.g. Green, 1995; Heath, 1997; 1999; Hoeschele, 1999; Satagopan and Yandell, 1996; Stephens and Smith, 1993). In some implementations, the reversible jump steps are facilitated by proposing to add or drop a QTL unconditionally on its genotypes, and by proposing QTL effects from their prior distributions, where small effects are more likely than large effects.

19.4.3.2   *Metropolis–Hastings Algorithm and the Composite Model Space Framework*

Very briefly, the Metropolis–Hastings MH (e.g. Chib and Greenberg, 1995) algorithm works as follows. Let $\boldsymbol{\theta}_i$ be subvector i of unknowns. Its conditional distribution is $\Pr(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}, \mathbf{y}, \mathbf{M}_{obs})$, where $\boldsymbol{\theta}_{-i}$ contains all unknowns except those in $\boldsymbol{\theta}_i$, which are evaluated at their current sample values. Let $\mathbf{x}$ be a candidate for the new sample value of $\boldsymbol{\theta}_i$ drawn from a proposal or candidate generating distribution with probability density $q(\mathbf{x}; \boldsymbol{\theta}_i)$. Proposal $\mathbf{x}$ is accepted with probability $\alpha(\mathbf{x}; \boldsymbol{\theta}_i) = \min(1, R)$, where

$$R = \frac{\Pr(\mathbf{x} | \boldsymbol{\theta}_{-i}, \mathbf{y}, \mathbf{M}_{obs}) q(\boldsymbol{\theta}_i; \mathbf{x})}{\Pr(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}, \mathbf{y}, \mathbf{M}_{obs}) q(\mathbf{x}; \boldsymbol{\theta}_i)}. \tag{19.41}$$

If the proposal is accepted, the new sample vector is $\boldsymbol{\theta}_i = \mathbf{x}$, else it is set equal to the previous sample vector. As an example, $q(\mathbf{x}; \boldsymbol{\theta}_i)$ is a univariate uniform or normal distribution centered at the previous sample value and with its spread chosen by trial and error to result in a suggested, moderate acceptance rate of, say, 15–40 % (to minimize autocorrelations between sample values: too high an acceptance rate results from a very

small spread, too low an acceptance rate from a high spread, and both lead to high autocorrelations, as in the latter case the new sample value is often equal to the old one). Note that the Gibbs sampler is a special case of the MH algorithm, where each sample is accepted, as it is drawn directly from $\Pr(\theta_i | \theta_{-i}, \mathbf{y}, \mathbf{M}_{obs})$, which can be done easily if this conditional distribution is standard and low-dimensional.

The MH algorithm is suitable for problems with fixed dimension of the parameter space or problems with a fixed model, i.e. for QTL mapping models with a fixed number of QTLs (usually a single QTL) and without epistasis. However, the number of QTLs should be treated as an unknown, because this is a parameter of interest, to avoid biases in the inference about QTL positions and effects that may occur in a single QTL model when there are truly multiple linked QTL, and to improve detection power by explaining more of the trait variability. Adding an additional QTL into a current model increases the dimension of the parameter space, as there are additional parameters for QTL position, effects and genotypes. The reversible jump algorithm of Green (1995; 2003) was developed as a general extension of the MH algorithm for Bayesian model selection by varying the dimension of the space of the unknowns. This algorithm has been successfully applied to QTL mapping with an unknown number of QTLs with applications including outbred populations (e.g., Heath, 1997; Uimari and Hoeschele, 1997; Thomas *et al.*, 1997; Uimari and Sillanpää, 2001; Yi and Xu, 2001; Narita and Sasaki, 2004). However, the reversible jump algorithm is complicated, and consequently it can be difficult to design effective move types (from the current model to one with a QTL added or dropped) producing good mixing, and the computational demand can be high.

Fortunately, Godsill (2001; 2003) established the composite model space framework, a modification of the product space approach of Carlin and Chib (1995). In this framework, MCMC simulations are performed on a space of fixed dimension. Application of standard Gibbs samplers to this space produces several Bayesian variable selection methods for linear models, e.g. the method of Kuo and Mallick (1998) and the stochastic search variable selection (SSVS) of George and McCulloch (1993; 1996; 1997), while the application of MH algorithms produces reversible jump algorithms (see Godsill, 2001; 2003; Green, 2003; Yi, 2004 for details). Application of the composite model space framework to QTL mapping is straightforward when only the markers are considered as potential QTL positions. If the QTL positions are unknown, then the QTL model is still a linear model conditional on a set of QTL positions. For this case, Yi (2004) provides a general discussion about how to apply the composite model space framework to multiple QTL mapping in inbred line crosses, and it should be equally useful when applied to outbred populations.

Representing a QTL model as a linear regression model with each of Q regressors (x_{iq}) corresponding to a marker or QTL position, assuming for simplicity only one (additive) effect per QTL (β_q), and introducing a zero/one indicator variable for each effect (λ_q) representing inclusion/exclusion in the model (as is commonly done in Bayesian variable selection for linear models), we have

$$y_i = \mu + \sum_{q=1}^{Q} \lambda_q x_{iq} \beta_q + e_i; \quad e_i \sim (0, \sigma^2).$$

The vector of unknown parameters (θ) includes the λ_q's, the x_{iq}'s (if nonmarker QTL positions are allowed), the β_q's, μ, σ^2, and the QTL positions p_q (if not restricted

to markers). The number of QTL (Q) is fixed at a chosen value (see Yi, 2004 for a discussion). For a given QTL model as specified by a set of λ_q values, we partition $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_{\lambda=0}$ and $\boldsymbol{\theta}_{\lambda=1}$ corresponding to parameters excluded and included in the current model, respectively, and with prior independence assumptions (about the parameters at different QTL and other parameters) the parameter prior is partitioned into $f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_{\lambda=0}) \cdot f(\boldsymbol{\theta}_{\lambda=1})$, where $f(\boldsymbol{\theta}_{\lambda=0})$ is a "pseudoprior". Because the parameters in $\boldsymbol{\theta}_{\lambda=0}$ do not contribute to the likelihood function of the data, or $f(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta}_{\lambda=1})$, the posterior of $\boldsymbol{\theta}_{\lambda=0}$ equals its prior. Samplers, which simply update the model indicator variables conditional on all parameters, require sampling from pseudopriors, and these samplers are suboptimal as the pseudopriors do not contain current knowledge about the corresponding parameters. Alternatively, λ_q and β_q can be updated jointly using a Gibbs sampler, and in a linear model such as marker regression, indicators λ_q can be sampled from their marginal posterior with all parameters integrated out (Broman and Speed, 2002). While it is therefore simple to design an efficient sampler for linear (marker) models, for QTL models (with nonmarker QTL positions allowed) there is the added difficulty of having to generate genotypes and position for a QTL currently not in the model. These unknowns are typically sampled from their priors, which may lead to inefficient mixing. Yi (2004) provides some discussion of this problem, and the reader should consult Godsill (2001; 2003) and Yi (2004) for further discussion of alternative samplers for the composite model space.

Lastly, the prior chosen for the β_q may influence the mixing of the sampler, and it requires hyper-parameters to be fixed at prespecified values, estimated empirically, or treated as additional unknowns (theoretically preferable but possibly causing convergence problems). For a discussion of alternative priors and their relationship to standard model selection criteria in the context of linear (fixed) models, see, e.g. Chipman *et al.* (2001). Briefly, we can formulate a prior for all Q parameters in the maximal (full) model, or $f(\boldsymbol{\beta}) = N_Q(0, \boldsymbol{\sigma}^2 \boldsymbol{\Sigma})$, coupled with an inverse gamma (IG) prior for σ^2, or $f(\sigma^2) = IG(\nu/2, \nu\xi/2)$, where in the absence of prior knowledge we can obtain a 'noninformative' prior for σ^2 by setting $\nu = 3$ and ξ to the sample variance (or ν and ξ can be chosen such that the prior assigns most of its probability to the interval between the variance under the full model and the total sample variance). In this normal-inverse-gamma prior, the prior for $\boldsymbol{\beta}$ is dependent on σ^2 and instead, we may set $f(\boldsymbol{\beta}) = N_Q(0, \boldsymbol{\Sigma})$. The usual choices for $\boldsymbol{\Sigma}$ are $c(\mathbf{X}'\mathbf{X})^{-1}$ or $c\mathbf{I}$, or a compromise where all or some of the off-diagonal elements of $\mathbf{X}'\mathbf{X}$ are set to 0. Frequently, priors of the form $f(\boldsymbol{\beta}_\lambda, \lambda) = f(\boldsymbol{\beta}_\lambda|\lambda) f(\lambda)$ are used, where $\boldsymbol{\beta}_\lambda$ is of dimension $Q_\lambda = \Sigma_q \lambda_q$. Then $f(\boldsymbol{\beta}_\lambda) = N_{Q\lambda}(0, \boldsymbol{\Sigma}_\lambda)$ with $\boldsymbol{\Sigma}_\lambda = c(\mathbf{X}'_\lambda \mathbf{X}_\lambda)^{-1}$ or $c\mathbf{I}_{Q\lambda}$. As noted by several authors (e.g. George and Foster, 2000), this prior provides a consistent description of uncertainty in the sense that it is the conditional distribution of the nonzero elements of $\boldsymbol{\beta}$ given λ when $f(\boldsymbol{\beta}) = N_Q(0, \boldsymbol{\Sigma})$. For $f(\lambda)$, usually an independence prior of the form $f(\lambda_q|w) = w^{Q\lambda}(1 - w)^{Q-Q\lambda}$ is chosen (see Yi, 2004; Yi *et al.*, 2005 for details on this prior in the context of multiple QTL mapping). Another alternative prior specification is the SSVS prior $f(\boldsymbol{\beta}|\lambda) = N_Q(0, \mathbf{D}_\lambda \mathbf{R} \mathbf{D}_\lambda)$ with $\mathbf{D}_\lambda$ diagonal and its qth diagonal element equal to v_{0q} when $\lambda_q = 0$ or 1 when $\lambda_q = 1$. With $\mathbf{R} = c\mathbf{I}_Q$, $f(\beta_q|\lambda_q) = \lambda_q \cdot N(0, c) + (1 - \lambda_q) \cdot N(0, c v_{0q})$. The obvious problem with any of the above priors is the necessary specification of values for the hyper-parameters including c, v_{0q} and w. Choices of values are discussed by various authors, e.g. Chipman *et al.* (2001), who also discuss the connection between choices of values for c and w and standard model selection criteria including BIC, AIC and RIC.

19.4.3.3 Prior Distributions

Concerning the prior distributions, different types of unknowns are independent a priori, as shown in (19.40). The prior distribution for the number of QTLs, q, is usually Poisson with predetermined mean λ chosen to reflect the belief that there is a relatively small number of QTLs, which can be separated from the polygenic background. The analysis is performed for different values of λ to evaluate its effect on the number of QTLs detected. The joint probability of the ordered genotypes $(\mathbf{G}, \mathbf{M})$ is discussed in Section 19.6. The linkage indicators ℓ_i of different QTLs are independent a priori, $\Pr(\ell_i = c) = t_c/T$, where T is the total genetic length of the genome, t_c is the length of marked chromosome c, $\Pr(\ell_i = 0) = 1 - \sum_{c=1}^{C} t_c/T$ is the prior probability that QTL i is unlinked (i.e. located in the unmarked portion of the genome), and C is the number of marked chromosomes in the analysis. QTL positions, conditional on linkage status, are independent a priori, and uniform on $[0, t_c]$ if $\ell_i = c$. QTL allele frequencies are independent beta(α_q, β_q) with $\alpha_q = \beta_q = 1$ resulting in the $U(0,1)$ distribution as a noninformative prior. Similarly, allele frequencies at different markers are independent Dirichlet$(\alpha_1, \ldots, \alpha_m)$ for m alleles and with $\alpha_k = 1$, $k = 1, \ldots, m$, as an uninformative prior. Priors for $\mathbf{b}$ and $\mathbf{t}_e$ are standard in the linear mixed model framework and under the assumption of normality where $\mathbf{t}_e = [\sigma_e^2]$ (e.g. Wang et al., 1993). Priors for the a_i and d_i in $\mathbf{a}$ and $\mathbf{d}$ should be chosen to reflect that QTLs with large effects are less likely than QTLs with small effects. Such priors can be put either on the QTL effects themselves or on the QTL variances. For the former, possible choices include independent truncated normal (on $[0, \infty]$) or exponential priors (Pong-Wong et al., 1999; Du and Hoeschele, 2000a). For the latter, independent exponential priors can be placed on the additive and dominance variances of a QTL, and from this prior a prior for the additive and dominance effects can be derived. More specifically,

$$\Pr(\mathbf{p}_Q, \mathbf{a}, \mathbf{d}|q) = \prod_{i=1}^{q} \Pr(a_i, d_i|p_i)\Pr(p_i), \tag{19.42}$$

where $\Pr(p_i)$ is uniform as described above, and $\Pr(a_i, d_i|p_i)$ is derived from the prior distribution of the QTL additive and dominance variances, $\Pr(\sigma_{ai}^2, \sigma_{di}^2)$, where $\Pr(\sigma_{ai}^2, \sigma_{di}^2) = \Pr(\sigma_{ai}^2)\Pr(\sigma_{di}^2)$ and $\Pr(\sigma_{ai}^2) = e^{-\sigma_{ai}^2/\lambda_a}/\lambda_a$, λ_a is the mean of the exponential distribution, and $\Pr(\sigma_{di}^2) = e^{-\sigma_{di}^2/\lambda_d}/\lambda_d$. Next, given σ_{ai}^2 and σ_{di}^2, there are four sets of (x_a, x_d) values which are equally likely: $(x_a, x_d) = (\sigma_{ai}, \sigma_{di})$, $(-\sigma_{ai}, \sigma_{di})$, $(\sigma_{ai}, -\sigma_{di})$, and $(-\sigma_{ai}, -\sigma_{di})$. Therefore, the probability density of a particular set is $f(x_a, x_d) = 0.25 f(\sigma_{ai}^2 = x_a^2, \sigma_{di}^2 = x_d^2)|\partial(\sigma_{ai}^2, \sigma_{di}^2)/\partial(x_a, x_d)|$, where $|\cdot|$ is the Jacobian of the transformation from $(\sigma_{ai}^2, \sigma_{di}^2)$ to $(x_a, x_d) = (\sigma_{ai}, \sigma_{di})$. Then $\Pr(a_i, d_i|p_i) = \Pr(x_a = h_a(a_i, d_i, p_i), x_d = h_d(a_i, d_i, p_i))|\partial(x_a, x_d)/\partial(a_i, d_i)|$, where $h_a(\cdot)$ and $h_d(\cdot)$ are deterministic functions obtained by solving $x_a = \sqrt{2p_i(1 - p_i)}[a_i + d_i(1 - 2p_i)]$ and $x_d = 2p_i(1 - p_i)d_i$ for a_i and d_i given p_i, and $|\cdot|$ is the Jacobian of the one-to-one transformation from (x_a, x_d) to (a_i, d_i) for a given p_i. Finally, we find that $\Pr(a_i, d_i|p_i) = \Pr(\sigma_{ai}^2 = [h_a(a_i, d_i, p_i)]^2, \sigma_{di}^2 = [h_d(a_i, d_i, p_i)]^2)[8p_i^3(1 - p_i)^3[a_i d_i + d_i^2(1 - 2p_i)]]$.

19.4.3.4 Modeling of Polygenic and QTL Effects

The residual polygenic variation can be described with an infinitesimal model by letting $\mathbf{u}$ consist of a vector of additive polygenic effects and additive polygenic variance. Under

the infinitesimal model, polygenic variation is assumed to be due to the very small effects of many unlinked genes, and hence polygenic values have a continuous, usually normal distribution. In inbred pedigrees with epistatic interactions involving dominance, the covariance between relatives depends on many parameters, resulting in inaccurate estimation and computational inefficiency (e.g. DeBoer and Hoeschele, 1993). Therefore, a finite polygenic model (FPM) has recently been considered as an alternative to the infinitesimal model. In an FPM, the distribution of polygenic values is discrete, and polygenic variation is explained with n unmarked loci, which are unlinked and biallelic. The FPM assumes additive gene action and a constant additive effect across loci, resulting in the fitting of polygenic number (the number of plus or favorable alleles across loci that an individual carries) instead of individual genotypes. Two versions of the FPM exist, the model of Thompson and Skolnick (1977), expanded upon by Cannings *et al.* (1978) and Lange (1997), and the model of Fernando *et al.* (1994). While the former model is not consistent with Mendelian inheritance (an offspring inherits a random sample of size n out of a pool of $2n$ alleles, not one allele per locus), in both models the conditional independence of sibs and ancestors given parents does not strictly hold (Stricker and Fernando, 1998).

Recently and with the advancement of MCMC and genotype sampling algorithms, finite locus models (FLMs) fitting individual genotypes rather than polygenic number have been investigated. These FLMs may include additive, dominance and two-locus epistatic interactions. Furthermore, FLMs differ in the number of loci, the treatment of additive, dominance and epistatic effects as constant or variable across loci, the number of two-locus interactions, and the prior distributions of the effects (Du *et al.*, 2000; Du and Hoeschele, 2000a; Pong-Wong *et al.*, 1999). Particularly in the presence of dominance or epistasis, estimates of the genetic variance components have been found to depend on the number of loci in the FLM, probably a small-sample problem, and this dependency is affected by the prior distribution of effects and the genotype sampling scheme (Du and Hoeschele, 2000a; Du *et al.*, 2000; Pong-Wong *et al.*, 1999). Modeling polygenic variation with an FLM seems more suitable in particular for linkage analysis, where an existing QTL is allowed to become unlinked and hence a member of the group of polygenic loci.

Returning to the QTLs, a first alternative to the biallelic QTL model is a model with $2n_f$ alleles, where n_f is the number of founders of the pedigree. Any Bayesian analysis, which specifies the number of alleles and considers all possible genotype configurations on a pedigree, may be referred to as a *distribution Bayesian analysis*, analogous to the distribution log likelihood of Section 19.3.4. Alternatively, the Bayesian analysis may be based on a model with $2n$ alleles, where n is the size of the pedigree, which have correlated effects (as described in Section 19.3). This analysis may be described as the *expectation* Bayesian method, analogous again to the expectation log likelihood of Section 19.3.4. Allelic effects in the latter model can be sampled jointly and efficiently using peeling (see Section 19.6) or data resampling (Garcia-Cortes *et al.*, 1995; Garcia-Cortes and Sorensen, 1996). While the biallelic QTL model may seem less realistic than the others, given that QTLs tend to occur in tight clusters, it greatly facilitates the incorporation of dominance and epistatic effects in complex pedigrees.

19.4.3.5 *Conditional Distributions of the Knowns*

We begin with the distribution of the observed marker data conditional on a sample of ordered marker genotypes. $\Pr(\mathbf{M}_{obs}|\mathbf{M})$ is 1 if $\mathbf{M}$ is consistent with $\mathbf{M}_{obs}$ and 0 otherwise.

$\Pr(\mathbf{y}|\mathbf{G}, \mathbf{a}, \mathbf{d}, \mathbf{b}, \mathbf{t}_e, \mathbf{u})$ is the penetrance function, which, conditional on the QTL genotypes ($\mathbf{G}$) and polygenic values ($\mathbf{u}$), factors into a product over individuals, or

$$\Pr(\mathbf{y}|\mathbf{G}, \mathbf{a}, \mathbf{d}, \mathbf{b}, \mathbf{t}_e, \mathbf{u}) = \prod_{i=1}^{N} \Pr(y_i|G_i, \mathbf{a}, \mathbf{d}, \mathbf{b}, \mathbf{t}_e, u_i). \qquad (19.43)$$

For a continuous trait (e.g. high-density lipoprotein (HDL) cholesterol level or protein yield), the typical assumption is that

$$y_i|G_i, \mathbf{a}, \mathbf{d}, \mathbf{b}, \mathbf{t}_e, u_i \sim N\left(\mathbf{x}_i'\mathbf{b} + \sum_{k=1}^{q} \mathbf{w}_{ki}'\mathbf{g}_k + u_i, \sigma_e^2\right), \qquad (19.44)$$

where $\mathbf{g}_k = [a_k, d_k]'$, $\mathbf{w}_{ki}' = [-1, -0.5]$, $[0, 0.5]$, or $[1, -0.5]$ corresponding to genotypes qq, Qq (or qQ) and QQ at QTL k, e_i is a residual, and σ_e^2 is residual variance.

However, the assumptions of normality and homoskedasticity of residual variance may not hold, and an analysis based on these assumptions is then not optimum (Coppieters *et al.*, 1998). A nonparametric QTL interval mapping approach for line crosses was developed by Kruglyak and Lander (1995) and extended to unrelated half-sibships (Coppieters *et al.*, 1998). Performing a power transformation of phenotypes prior to analysis may reduce the power of QTL detection (Demenais *et al.*, 1986). Common power transformations also have limited applicability. Therefore, George and Elston (1988) proposed a generalized modulus power transformation, performed simultaneously with QTL detection and estimation, capable of removing both skewness and kurtosis, and inducing normality from a wide range of distributions, at the expense of estimating two additional parameters. von Rohr and Hoeschele (2002) use an alternative approach, where the normal distribution is replaced with a skewed student t distribution. This method is based on Fernández and Steel (1998), who showed how to introduce skewness into any continuous, unimodal and symmetric distribution. This approach, when applied to t distributions, also requires the inclusion of two additional parameters or unknowns, one for skewness and one for the Degrees of Freedom df of the t distribution.

Most of the properties of the multivariate normal distribution also hold for the multivariate t distribution, except that uncorrelated random variables are not independent for the multivariate t distribution. If all residual phenotypes were assigned to a single multivariate t distribution, then the df parameter would not be estimable. The df parameter is estimable if phenotypes are assigned to multiple (more than one) clusters, with the k observations in a given cluster having a k-variate t distribution, and with the same df and spread parameters across clusters. Observations in the same cluster may be uncorrelated but are not independent. Observations in different clusters are independent. Lange *et al.* (1989) and von Rohr and Hoeschele (2002) assign each residual phenotype to a different cluster, i.e. residual phenotypes have univariate t distributions with the same df and spread parameters, or

$$e_i = y_i - \mathbf{x}_i'\mathbf{b} - \sum_{k=1}^{q} \mathbf{w}_{ki}'\mathbf{g}_k - u_i \sim t_1(0, \sigma_e^2, \nu_e). \qquad (19.45)$$

Then the joint probability of the residual phenotypes, or the penetrance, still factors into a product over individuals. When e_i has a t distribution with ν_e df, then

$$f(e_i|\sigma_e^2, v_e) = f(y_i|G_i, \mathbf{a}, \mathbf{d}, \mathbf{b}, u_i, \sigma_e^2, v_e) = \frac{\Gamma\left(\frac{v_e + 1}{2}\right)}{\Gamma\left(\frac{v_e}{2}\right) v_e^{0.5} \pi^{0.5} \sigma_e} \left[1 + \frac{e_i^2}{\sigma_e^2 v_e}\right]^{-0.5(v_e+1)},$$

(19.46)

and when e_i has a skewed t distribution with skewness parameter γ, then

$$f_\gamma(e_i|\sigma_e^2, v_e) = \frac{2}{\gamma + \gamma^{-1}} \left[f(e_i\gamma)I_{(-\infty,0)} + f(e_i\gamma^{-1})I_{(0,\infty)}\right]$$

$$= \frac{2}{\gamma + \gamma^{-1}} \cdot \frac{\Gamma\left(\frac{v_e + 1}{2}\right)}{\Gamma\left(\frac{v_e}{2}\right) v_e^{0.5} \pi^{0.5} \sigma_e}$$

$$\times \left[\left(1 + \frac{e_i^2}{\sigma_e^2 v_e}\right)\left(\gamma^2 I_{(-\infty,0)} + \gamma^{-2} I_{(0,\infty)}\right)\right]^{-0.5(v_e+1)}, \quad (19.47)$$

where $I_{(-\infty,0)} = 1$ if $-\infty < e_i \leq 0$ and 0 otherwise, and $I_{(0,\infty)} = 1$ if $0 < e_i < \infty$ and 0 otherwise. To incorporate the skewed t distribution in the Bayesian analysis, the normal penetrance function is replaced with the skewed t penetrance, the Gibbs-based sampling steps for the regression parameters and classification effects are replaced with MH steps, and additional MH steps for v_e and γ are added (von Rohr and Hoeschele, 2002). This MH-based scheme is anticipated to provide faster mixing than the alternative of a Gibbs sampler with data augmentation. Data augmentation results from writing the t distribution as the integral of a normal distribution for the residual phenotypes and a gamma distribution for the missing data w_i, or

$$f(e_i|\sigma_e^2, v_e) = \int_0^\infty f(w_i|v_e) f(e_i|\sigma_e^2/w_i) \, dw_i.$$

(19.48)

Inclusion of the missing data w_i into the sampling scheme leads to standard distributions for the regression and classification parameters, the residual variance, and for the w_i (γ), which can be sampled from with Gibbs steps (e.g. Fernández and Steel, 1998).

When $\mathbf{y}$ follows some other, possibly discontinuous (e.g. binomial, multinomial or Poisson) distribution, generalized linear (mixed) models can be used to replace the normal penetrance function (McCullagh and Nelder, 1989; see also **Chapter 18**).

19.4.3.6 Bayesian Inference

Inferences about unknowns of interest are based on the (dependent) samples of the unknowns obtained from their joint posterior distribution via an MCMC sampling scheme. Typical inferences are obtained as follows. $\Pr(q|\mathbf{y}, \mathbf{M}_{\text{obs}})$ $q = 1, \ldots, q_{\max}$, or the probability of q QTLs on all marked chromosomes, is obtained by counting the number of samples in which the number of QTLs with $\ell_i > 0$ is q, divided by the total number of samples. $\Pr(q_c|\mathbf{y}, \mathbf{M}_{\text{obs}})$ or the probability that there are q_c QTLs on chromosome c, or in a specific region of chromosome c, is obtained similarly. The probability that interval I_c on chromosome c contains at least one QTL is obtained as the number of

```
–|–––|––X––|––––X––––|–––|–––|––X––|–––––|–  16 000
–|–––|––X––|––––X––––|–––|–––|––X––|––X––|–   2000
–|–––|––X––|–––––––––|–––|–––|––X––|–––––|–    600
–|–––|––X––|––––X––––|–––|–––|–––––|––X––|–   1000
–|–––|––X––|–––––––––|–––|–––|––X––|––X––|–    400
–|–––|––1––|–0.95–|–––|–––|–0.95–|–0.17–|–   Intervals
     –|–––––|2: 0.95, 1: 0.05 |–––|–––|1: 0.88, 2: 0.12|–  Regions
Chromosome: 2 QTL: 0.03; 3 QTL: 0.87; 4 QTL: 0.10
```
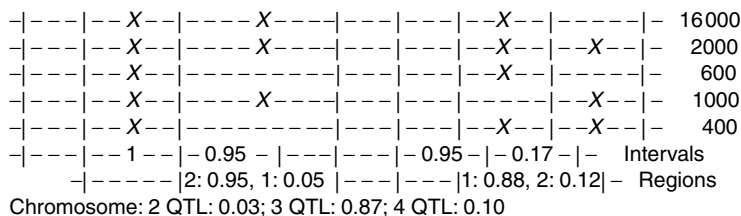
Figure 19.1 Number of samples (total of 20 000) with certain QTL patterns on a chromosome, with vertical lines depicting (eight) marker locations and X indicating the presence of a QTL in any of the seven marker intervals (lines 1–5); marginal posterior probabilities of a QTL in a marker interval (line 6); marginal posterior probabilities of q ($q = 1$ or $q = 2$) QTLs in a region (line 7); marginal posterior probabilities of q QTLs on the chromosome (line 8); marginal posterior probabilities are estimated from sample counts.

samples with at least one QTL in this interval over the total number of samples. These inferences are illustrated in Figure 19.1, which depicts results from a typical run for a single chromosome. For example, in 16 000 out of 20 000 cycles, marker intervals 2, 3 and 6 contained QTLs; marker interval 2 contained a QTL in each sample, while marker interval 3 contained a QTL in 95 % of all samples. With a (marginal posterior) probability of 0.95, there are two QTLs in the region formed by marker intervals 2 and 3. With a (marginal posterior) probability of 0.87, the chromosome contains three QTLs. Note that in this analysis we did not allow for multiple QTLs in the same marker interval, although in principle this could be done.

For any marker interval, region on a marked chromosome, marked chromosome or marked genome we can obtain the marginal posterior mean and distribution of the additive (and dominance, etc.) genetic variance explained by all QTLs in this region by summing, in each cycle, the additive variances of all QTLs in the region and estimating mean and distribution of these values. For unknowns such as overall (or chromosomal or regional) QTL variance, polygenic variance, and residual variance, posterior variances and correlations and highest posterior density regions can be estimated from the samples, and the corresponding empirical marginal posterior distributions can be depicted. For marginal posterior density plots, see, e.g. Uimari and Hoeschele (1997). If an unknown is sampled from a standard conditional distribution (Gibbs sampler), its marginal density can be estimated as the average density of the conditional distributions across cycles (referred to as *Rao–Blackwellization*). If the sampling distribution is not standard, various techniques are available for density estimation using the sample values for this unknown, e.g. average shifted histograms (Scott, 1992). Marginal posterior densities are asymptotically normal, so lack of normality indicates insufficient sample size. These quantities are Bayesian counterparts of frequentist standard errors and confidence intervals, and are computed to evaluate the precision of the inferences.

Why should a Bayesian approach be used for the linkage mapping of QTLs? Firstly, because it accounts for all uncertainties in the system (e.g. unknown number of QTLs, unknown genotypes, unknown QTL locations). Inferences about particular unknowns of interest are obtained conditionally on the observed data $(\mathbf{y}, \mathbf{M}_{obs})$ but not on particular values of any of the other unknowns (e.g. the additive genetic variance due to QTLs on a particular chromosome does not depend on QTL number and positions). In

addition to point estimates, marginal posterior distributions, which are exact small-sample distributions, can be obtained for any unknown of interest, as well as posterior summary statistics such as posterior variances and correlations. Secondly, the Bayesian mean estimator of the QTL variance in a marker interval can be interpreted as a multiple shrinkage estimator. Multiple shrinkage estimators have been shown to exhibit excellent mean-squared error performance in other problems such as wavelet estimation (Clyde *et al.*, 1997). In Bayesian linkage analysis, multiple shrinkage occurs because, the mean of the prior Poisson distribution of QTL number q is typically chosen to be much smaller than the number of marker intervals, implying that a priori many marker intervals do not contain a QTL. Moreover, QTL effects or variances are assumed to have exponential or truncated normal distributions, with values near zero much more likely than large values. Thirdly, Bayesian analysis provides a formal mechanism for the incorporation of meaningful, biological prior information as discussed earlier.

The Bayesian analysis provides us directly with any desired probability statement, conditional only on the observed data, from the joint posterior distribution of all unknowns. Examples include the probability of any number of detectable QTLs on all marked chromosomes, on a specific chromosome or in a specific chromosomal region, and the probability of at least one QTL in a specific marker interval. Recently, there has been some concern about the interpretation of QTLs (in a frequentist setting), which can be summarized as follows. A stringent Type I error rate is a necessary but not sufficient condition for most QTL findings to be real. This is so, because the proportion of false positives among all QTL findings, or 1 minus the reliability, depends on Type I error rate, Type II error rate or power, and the prior probability of the null hypotheses (a single null hypothesis may state that a given marker interval is empty). This is shown in (19.39), as reliability is also the probability that the QTL is real given a significant test. Moreover, reliability is equal to $t/(t + f)$, where f is the expected number of false positives (prior probability of null hypothesis $\times$ Type I error rate), and t is the expected number of true QTLs detected (prior probability of alternative hypothesis $\times$ power of test). If power is low, then the proportion of false positives may exceed the Type I error rate. Attempts to take this relationship into consideration in frequentist analyses are difficult, because the calculation of reliability seems infeasible due to too many assumptions pertaining to the definition and prior probability of the null hypotheses and power, which in turn depend on number and effects of QTLs. In the Bayesian analysis, there is no need a priori to specify a null hypothesis, and it provides us directly with any desired probability (belief) statement conditional only on the observed data, from the joint posterior distribution of all unknowns.

For additional information on Bayesian analysis, see **Chapter 20** as well as Box and Tiao (1973) and Berger (1985).

19.5   DETERMINISTIC HAPLOTYPING IN COMPLEX PEDIGREES

Haplotyping in a pedigree refers to the reconstruction of haplotypes from observed (marker) genotype data within the pedigree. It is an important computational step in the QTL analysis of outbred, pedigreed populations (e.g. (19.15)). A consistent haplotype configuration is an assignment of haplotypes to all individuals in the pedigree, which is

consistent with all the observed genotype data and the pedigree structure. The maternal and paternal haplotypes of an individual comprise its multilocus ordered genotype. For large, complex pedigrees and with a large number of linked loci, the number of consistent haplotype configurations is usually too large for an exhaustive search to be feasible. Methods that are capable of handling a larger number of loci are not guaranteed to find the most probable configuration (Lin and Speed, 1997).

Various algorithms for haplotyping in pedigrees have been developed. Some of these algorithms are purely logical and rule-based (Wijsman, 1987; O'Connell, 2000; Tapadar *et al.*, 2000; Qian and Beckmann, 2002), while others are based on likelihood or conditional probability computations (Lander and Green, 1987; Sobel and Lange, 1996; Sobel *et al.*, 1996; Lin and Speed, 1997; Thomas *et al.*, 2000; Abecasis *et al.*, 2002). The rule-based procedures are deterministic and fast, and hence applicable to large pedigrees and large numbers of linked loci, but they do not use the distance between markers and are not suitable for situations with substantial amounts of missing data. The likelihood or conditional probability–based algorithms are typically stochastic (Sobel and Lange, 1996; Sobel *et al.*, 1996; Lin and Speed, 1997; Thomas *et al.*, 2000), and although they can be applied to complex pedigrees, their computing time requirements can be unacceptable.

In the space of all consistent haplotype configurations on a pedigree (SACHC), typically most configurations have very small probabilities conditional on the observed genotype data, so that only a relatively small subset of configurations is relevant. Gao *et al.* (2004) and Gao and Hoeschele (2006) developed a deterministic method that identifies a subset of haplotype configurations with high conditional probabilities in SACHC. This (approximate) method uses information from the closest informative flanking markers and close relatives (parents, offspring). It is not guaranteed to find the most probable or the true configuration. This conditional enumeration haplotyping method is briefly described below.

We call the combination of a specific individual and a specific marker locus a *person-marker*. The genotypes of some person-markers in nonfounders can be ordered on the basis of their parents' genotypes. Let $\mathbf{U}$ denote all remaining person-markers in a pedigree with unordered heterozygous genotype. Assume that the size of $\mathbf{U}$ is t. Reconstructing a haplotype configuration for the entire pedigree consists of assigning an ordered genotype to each person-marker in $\mathbf{U}$. Let $\{M_1, M_2, \ldots, M_t\}$ be a specific order of the person-markers in $\mathbf{U}$. Let m_i denote an ordered genotype assigned to person-marker M_i. The joint probability of a haplotype configuration for $\mathbf{U}$, $(m_1, m_2, \ldots, m_t)$, conditional on the observed data ($\mathbf{D}$) is

$$\Pr(m_1, m_2, \ldots, m_t|\mathbf{D}) = \Pr(m_1|\mathbf{D})\Pr(m_2|m_1, \mathbf{D}) \cdots \Pr(m_t|m_1, \ldots, m_{t-1}, \mathbf{D}). \quad (19.49)$$

Let $p_i = \Pr(m_i|m_1, \ldots, m_{i-1}, \mathbf{D})$. Also, m_i is one of the ordered genotypes m_i^1 and m_i^2, where $m_i^1(m_i^2)$ has the larger (smaller) conditional probability $p_i^1(p_i^2)$ at person-marker M_i, and $p_i^j = \Pr(m_i^j|m_1, \ldots, m_{i-1}, \mathbf{D})$ for $j = 1, 2$, with $p_i^1 \geq 0.5$ and $p_i \leq p_i^1$. The conditional probabilities (p_i^j) are calculated using information from informative flanking markers in the individual and its parents and offspring. The conditional probability of the haplotype configuration $(m_1, m_2, \ldots m_t)$ can then be written as

$$\Pr(m_1, m_2, \ldots, m_t|\mathbf{D}) = \prod_{i=1}^{t} p_i.$$

Let T denote the largest conditional probability of all haplotype configurations for $\mathbf{U}$, and let R denote the ratio of the conditional probability of haplotype configuration $(m_1, m_2, \ldots, m_t)$ to T. For any $k \leq t$, let $Q_k = \prod_{j=1}^{k} \frac{p_j}{p_j^1}$. Because $p_j \leq p_j^1$ and $T \geq \prod_{i=1}^{t} p_j^1$ is very likely true, the following inequality is very likely to hold:

$$R = \frac{\Pr(m_1, m_2, \ldots, m_t)}{T} \leq \prod_{j=1}^{t} \frac{p_j}{p_j^1} \leq Q_k. \tag{19.50}$$

Given a relatively small threshold $10^\alpha (\alpha < 0$, e.g. $\alpha = -3)$, if we can find some $k \leq t$ such that $Q_k \leq 10^\alpha$, then $R \leq 10^\alpha$ and $\Pr(m_1, m_2, \ldots m_t|\mathbf{D})$ is very small relative to T. Then this configuration is deleted from SACHC.

Given a user-determined threshold λ for the conditional probabilities of ordered genotypes at individual loci, and a threshold 10^α for the conditional probabilities of haplotype configurations $(\alpha < 0$ and $10^\alpha < (1 - \lambda)/\lambda)$, the conditional enumeration method is implemented as follows: After the first $i - 1$ person-markers have been assigned ordered genotypes, for each assignment combination pertaining to the $i - 1$ person-markers, $m_1, m_2, \ldots, m_{i-1}$, we find the person-marker M_i with the highest conditional probability p_i^1 among all remaining person-markers in $\mathbf{U}$. We assign ordered genotypes to person-marker M_i as described below $(i = 1, 2, \ldots, t)$:

1. When $p_i^1 \geq \lambda (\lambda \geq 0.5$; e.g. $\lambda = 0.90)$, then assign only m_i^1 to person-marker $M - i$.

2. When $p_i^1 < \lambda$, then if assigning m_i^2 to person-marker M_i produces $Q_i \leq 10^\alpha$, we only assign m_i^1; otherwise, we retain both ordered genotypes, m_i^1 and m_i^2, for person-marker M_i.

After all person-markers in $\mathbf{U}$ have been processed with this algorithm, a set of haplotype configurations for the pedigree has been retained, and a smaller subset of configurations with highest likelihood can be obtained by eliminating additional configurations, if desired. When λ approaches 1, and α approaches $-\infty$ (10^α approaches 0), then the conditional enumeration haplotyping method approaches an exhaustive (exact) enumeration method.

This new haplotyping method was shown to be faster and to provide more information than existing methods (Gao *et al.*, 2004; Gao and Hoeschele, 2006). Gao and Hoeschele (2005) applied this method to the calculation of IBD matrices used in QTL analysis based on variance-components models by summing over a subset of haplotype configuration with high conditional probabilities instead of the exhaustive summation in (19.15). Simulated pedigree data were analyzed with the variance-components QTL analysis of George *et al.* (2000). The required IBD matrices were computed with our deterministic conditional enumeration method and with the MCMC method implemented in Loki (Heath, 1997; Thompson and Heath, 1999). Both methods gave essentially the same estimates of QTL positions and variance components, while the deterministic method was computationally much more efficient. It also allows the incorporation of linkage disequilibrium (LD) using the method of Meuwissen and Goddard (2000; see Section 19.7.2), which is not possible with Loki. The current implementation of the enumeration haplotyping method (Gao *et al.*, 2004; Gao and Hoeschele, 2006) does not allow for (substantial amounts of)

missing marker data, for which we have work under way that also needs to consider LD among dense markers in the founders of a pedigree.

19.6 GENOTYPE SAMPLING IN COMPLEX PEDIGREES

In this section, we consider the distribution of genotypes on a pedigree. In the most general case, these genotypes are multilocus, ordered genotypes including linked markers and QTLs. We are interested in obtaining genotype samples from the joint distribution of the genotypes of all pedigree members and at all loci, conditional on observed genotype data (on marker loci) ($\mathbf{M}$) and phenotypic data ($\mathbf{y}$). Such genotype samples are needed in implementations of ML and Bayesian QTL mapping methods when applied to complex pedigrees. The probability of a genotype configuration $\mathbf{G}$ (including the genotypes of all pedigree members at all loci) on the pedigree is

$$\Pr(\mathbf{G}|\mathbf{M}, \mathbf{y}) \propto \Pr(\mathbf{G})\Pr(\mathbf{M}|\mathbf{G}) f(\mathbf{y}|\mathbf{G}), \tag{19.51}$$

where $\Pr(\mathbf{G})$ is the joint probability of all genotypes in genotype configuration $\mathbf{G}$, $\Pr(\mathbf{M}|\mathbf{G})$ is 1 for any $\mathbf{G}$ that is consistent with $\mathbf{M}$ and 0 otherwise, and $f(\mathbf{y}|\mathbf{G})$ is the conditional probability density of the phenotypes. Conditional on $\mathbf{G}$, and assuming that polygenic effects are either absent or also conditioned on $\mathbf{G}$, the y_i are independent, or

$$f(\mathbf{y}|\mathbf{G}) = \prod_{i=1}^{N} f(y_i|\mathbf{G}_i), \tag{19.52}$$

where y_i is the phenotype and $\mathbf{G}_i$ is the vector of ordered genotypes of individual i. Furthermore, the probability of the multilocus genotypes can be factored into

$$\Pr(\mathbf{G}) = \prod_{i=1}^{N_f} \Pr(\mathbf{G}_i) \prod_{i=N_f+1}^{N} \Pr(\mathbf{G}_i|\mathbf{G}_{f_i}, \mathbf{G}_{m_i}), \tag{19.53}$$

where N_f is the number of founders, and f_i (m_i) is the father (mother) of i. The probability of the multilocus genotype of a founder can be factored into

$$\Pr(\mathbf{G}_i) = \prod_{j=1}^{L} \Pr(G_{ij}), \tag{19.54}$$

where L is the number of loci and $\Pr(G_{ij})$ is typically set equal to the Hardy–Weinberg frequency of the genotype of individual i at locus j. The genotype probability for a descendant is factored into the first-order Markov chain

$$\Pr(\mathbf{G}_i|\mathbf{G}_{f_i}, \mathbf{G}_{m_i}) = \Pr(G_{i1}|G_{f_i 1}, G_{m_i 1})$$
$$\times \prod_{j=2}^{L} \Pr\big(G_{ij}|G_{i(j-1)}, G_{f_i j}, G_{f_i (j-1)}, G_{m_i j}, G_{m_i (j-1)}\big). \tag{19.55}$$

As G_{ij} represents an ordered genotype, its conditional probability can be further partitioned into

$$\Pr\big(G_{ij}|G_{i(j-1)}, G_{f_i j}, G_{f_i(j-1)}, G_{m_i j}, G_{m_i(j-1)}\big)$$
$$= \Pr\big(H_{ij}^f|H_{i(j-1)}^f, G_{f_i j}, G_{f_i(j-1)}\big)\Pr\big(H_{ij}^m|H_{i(j-1)}^m, G_{m_i j}, G_{m_i(j-1)}\big), \quad (19.56)$$

where H_{ij}^f (H_{ij}^m) is the paternally (maternally) inherited allele of G_{ij}.

The fully conditional distribution of G_{ij}, given the observed data $(\mathbf{M}, \mathbf{y})$ and all other genotypes $(\mathbf{G}_{-ij})$, is obtained from the distribution of $\mathbf{G}$ in (19.51), using the factorizations in (19.52) through (19.55), and retaining all terms that depend on G_{ij}, or

$$\Pr(G_{ij}|\mathbf{G}_{-ij}, \mathbf{M}, \mathbf{y}) = \Bigg[\Pr(G_{ij}|G_{i(j-1)}, G_{f_i j}, G_{f_i(j-1)}, G_{m_i j}, G_{m_i(j-1)})$$
$$\times \Pr(G_{i(j+1)}|G_{ij}, G_{f_i(j+1)}, G_{f_i j}, G_{m_i(j+1)}, G_{m_i j})$$
$$\times \prod_{o_i} \Pr(G_{o_{ij}}|G_{o_{i(j-1)}}, G_{ij}, G_{i(j-1)}, G_{m_{o_i} j}, G_{m_{o_i}(j-1)})$$
$$\times \Pr(G_{o_{i(j+1)}}|G_{o_{ij}}, G_{i(j+1)}, G_{ij}, G_{m_{o_i}(j+1)}, G_{m_{o_i} j})$$
$$\times \Pr(M_{ij}|G_{ij})f(y_i|\mathbf{G}_i) \Bigg] \Bigg/ \sum_{G_{ij}}\text{numerator}, \quad (19.57)$$

where o_i is offspring o of individual i, and m_{o_i} is the mate of individual i of the marriage producing o_i (note that the numerator extends over the entire right-hand side in (19.57)).

The conditional distribution of G_{ij}, given the observed data $(\mathbf{M}, \mathbf{y})$ and all other genotypes except those of i's final offspring (of_i) at locus j, is also easily obtained from the distribution of $\mathbf{G}$ in (19.51) by retaining all terms that depend on G_{ij} and on the $G_{of_i j}$ of final offspring of_i (final offspring are offspring which are not parents). Then in (19.57) the contribution of final offspring to the product over o_i is replaced by

$$\prod_{of_i}\sum_{G_{of_i j}}[\Pr(G_{of_i j}|G_{of_i(j-1)}, G_{ij}, G_{i(j-1)}, G_{m_{of_i} j}, G_{m_{of_i}(j-1)})$$
$$\times \Pr(G_{of_i(j+1)}|G_{of_i j}, G_{i(j+1)}, G_{ij}, G_{m_{of_i}(j+1)}, G_{m_{of_i} j})\Pr(M_{of_i j}|G_{of_i j})f(y_{of_i}|\mathbf{G}_{of_i})].$$
$$(19.58)$$

The simplest genotype sampling scheme is a Gibbs sampler, which samples one genotype (of one individual at one locus) from the univariate, fully conditional distribution in (19.57), referred to as *person-by-person* and *locus-by-locus sampling* by Kong (1991). This sampler often does not work (well), because transitions between certain subsets of genotype configurations on a pedigree are performed infrequently. While the sampler is theoretically irreducible for a biallelic locus, it can be reducible for loci with more than two alleles (theoretical (practical) reducibility): from some genotype configuration, certain other configurations can never (cannot in acceptable CPU time) be reached; mating $??\times?? \rightarrow AB, 00$, with parental genotypes missing, has two parental genotype configurations: (A0, B0) and (B0, A0), with no transitions possible using the univariate

Gibbs sampler). In animal populations with parents having multiple progeny, the univariate Gibbs sampler is very sticky, and therefore modification (19.58) to (19.57) is performed. While Janss *et al.* (1995b) perform marginalization with respect to final offspring only for fathers (who typically have more offspring than mothers in animal populations), Du *et al.* (2000) and Du and Hoeschele (2000b) marginalized the genotype sampling distribution for fathers and mothers after observing stickiness of the sampler when mothers had 10 offspring on average.

A number of techniques have been proposed to obtain a genotype Gibbs sampler, which is theoretically and practically irreducible, including blocking or joint sampling of genotypes (Ott, 1989; Janss *et al.*, 1995b; Jensen and Kong, 1996; Heath, 1997; 1999; Hoeschele, 1999), sampling with modified transmission probabilities (Sheehan and Thomas, 1993), sampling from a series of distributions including the desired distribution and others which allow faster mixing (heated companion chains or Metropolis-coupled chains of Lin *et al.* (1993); simulated tempering of Geyer and Thompson (1995)), and a hybrid algorithm mixing the Gibbs cycles with Metropolis jumping steps (Lin, 1995; 1996), requiring the identification of noncommunicating subsets of genotype configurations in the sampling space of $\mathbf{G}$ using the method of Lin *et al.* (1994), which fails in certain, not uncommon situations (Lin, 1996; Jensen, 1997). Joint sampling is most general and most promising for the hardest problems (large complex pedigrees with loops and substantial amounts of missing genotype data on linked loci). Ideally, the genotypes of all individuals at all (linked) loci are sampled jointly, and different strategies to achieve or move toward that goal are currently under investigation.

Joint sampling of genotypes is relatively simple to perform in two cases: joint sampling of the genotype of a parent and its final offspring (Janss *et al.*, 1995b; Du *et al.*, 2000; Du and Hoeschele, 2000b); and joint sampling of the genotypes on a pedigree without loops at a single locus using peeling (Elston and Stewart, 1971; Fernando *et al.*, 1993) and 'reverse peeling' (Ott, 1989; Heath, 1997).

Joint sampling of genotypes for a pedigree without loops via reverse peeling is based on partitioning the probability of genotype configuration $\mathbf{G}$ given observed data ($\mathbf{y}$) via the method of composition (Tanner, 1993), or

$$
\begin{aligned}
\Pr(\mathbf{G}|\mathbf{y}) &= \Pr(G_1|\mathbf{y})\Pr(G_2|G_1, \mathbf{y})\Pr(G_3|G_2, G_1, \mathbf{y}) \ldots \Pr(G_n|G_1, \ldots, G_{n-1}, \mathbf{y}) \\
&= \Pr(G_1|\mathbf{y})\Pr(G_2|G_1, y_2, \ldots, y_n) \\
&\quad \times \Pr(G_3|G_2, G_1, y_3, \ldots, y_n) \ldots \Pr(G_n|G_1, \ldots, G_{n-1}, y_n).
\end{aligned} \tag{19.59}
$$

We sample from this distribution by first sampling G_1 from its marginal distribution, then G_2 given G_1, etc. Computation of the conditional probabilities in (19.59) is described next for a pedigree without loops. For an arbitrary individual i, we can partition the joint probability of the genotypes as follows:

$$
\begin{aligned}
\Pr(\mathbf{G}|\mathbf{y}) &\propto \Pr(\mathbf{G})\Pr(\mathbf{y}|\mathbf{G}) \\
&= [\Pr(\mathbf{G}_A)\Pr(\mathbf{y}_A|\mathbf{G}_A)\Pr(G_i|\mathbf{G}_A)]f(y_i|G_i)[\Pr(\mathbf{G}_P|G_i)\Pr(\mathbf{y}_P|\mathbf{G}_P)],
\end{aligned} \tag{19.60}
$$

where G_i is genotype of individual i, $\mathbf{G}_A$ ($\mathbf{G}_P$) is the set of genotypes on all individuals anterior (posterior) to i, and $\mathbf{y}$ is partitioned accordingly. Anterior (posterior) to i are all

individuals connected to i through its parents and full-sibs (offspring and mates). Based on this partitioning, we find the following marginal and partially conditional distributions of G_i:

$$\Pr(G_i|\mathbf{y}) = \left[\sum_{\mathbf{G}_A}\Pr(\mathbf{G}_A)\Pr(\mathbf{y}_A|\mathbf{G}_A)\Pr(G_i|\mathbf{G}_A)\right]f(y_i|G_i)\left[\sum_{G_P}\Pr(\mathbf{G}_P|G_i)\Pr(\mathbf{y}_P|\mathbf{G}_P)\right],$$

$$\Pr(G_i|\mathbf{G}_A,\mathbf{y}) = [\Pr(\mathbf{G}_A)\Pr(\mathbf{y}_A|\mathbf{G}_A)\Pr(G_i|\mathbf{G}_A)]f(y_i|G_i)\left[\sum_{G_P}\Pr(\mathbf{G}_P|G_i)\Pr(\mathbf{y}_P|\mathbf{G}_P)\right],$$

$$\Pr(G_i|\mathbf{G}_P,\mathbf{y}) = \left[\sum_{\mathbf{G}_A}\Pr(\mathbf{G}_A)\Pr(\mathbf{y}_A|\mathbf{G}_A)\Pr(G_i|\mathbf{G}_A)\right]f(y_i|G_i)[\Pr(\mathbf{G}_P|G_i)\Pr(\mathbf{y}_P|\mathbf{G}_P)].$$

$$(19.61)$$

In peeling, the conditional probabilities in (19.61) are calculated in a particular order (a 'peeling sequence'), for which all conditionals simplify such that each depends on the genotypes of at most two other individuals in the same nuclear family (a nuclear family consists of a pair of parents and their offspring). This computation is illustrated using the pedigree (without loops) in Table 19.1.

For pedigree 1 (without loops), we partition $\Pr(\mathbf{G}|\mathbf{y})$ as follows:

$$
\begin{aligned}
\Pr(\mathbf{G}|\mathbf{y}) = {} & \Pr(G_1|\mathbf{y})\Pr(G_2|G_1,\mathbf{y})\Pr(G_6|G_1,G_2,\mathbf{y})\Pr(G_5|G_1,G_2,\mathbf{y}) \\
& \times \Pr(G_7|G_1,G_2,G_5,G_6,\mathbf{y})\Pr(G_9|G_1,G_2,G_5,G_6,G_7,\mathbf{y}) \\
& \times \Pr(G_3|G_1,G_2,G_5,G_6,G_7,G_9,\mathbf{y})\Pr(G_4|G_1,G_2,G_5,G_6,G_7,G_9,G_3,\mathbf{y}) \\
& \times \Pr(G_8|G_1,G_2,G_5,G_6,G_7,G_9,G_3,G_4,\mathbf{y}) \qquad (19.62) \\
= {} & \Pr(G_1|\mathbf{y})\Pr(G_2|G_1,\mathbf{y}_{-1})\Pr(G_6|G_1,G_2,\mathbf{y}_{-1,2,5})\Pr(G_5|G_1,G_2,y_5) \\
& \times \Pr(G_7|G_6,\mathbf{y}_{-1,2,5,6})\Pr(G_9|G_6,G_7,y_9)\Pr(G_3|G_7,y_3,y_4,y_8) \\
& \times \Pr(G_4|G_7,G_3,y_4,y_8)\Pr(G_8|G_3,G_4,y_8).
\end{aligned}
$$

Table 19.1 Example pedigree without loops.

Individual	Father	Mother
1	0	0
2	0	0
3	0	0
4	0	0
5	1	2
6	1	2
7	3	4
8	3	4
9	6	7

This pedigree consists of three nuclear families, (1,2,5,6), (6,7,9) and (3,4,7,8). Individuals 6 and 7 are connectors, i.e. members of multiple nuclear families. Any family with only one connector can be peeled. Therefore, a peeling sequence for this pedigree starts either with the first or the last family. Here we start with the last family, (3,4,7,8). Then, conditional probabilities are computed in the reverse order of (19.62). Specifically, we obtain

$$\Pr(G_8|G_3, G_4, y_8) \propto p(G_8|G_3, G_4)f(y_8|G_8) \qquad \forall\, G_3, G_4, G_8,$$

$$\Pr(G_4|G_7, G_3, y_4, y_8) \propto p(G_4)f(y_4|G_4)p(G_7|G_3, G_4)$$

$$\times \sum_{G_8} \Pr(G_8|G_3, G_4, y_8) \qquad \forall\, G_3, G_4, G_7,$$

$$\Pr(G_3|G_7, y_3, y_4, y_8) \propto p(G_3)f(y_3|G_3)\sum_{G_4}\Pr(G_4|G_7, G_3, y_4, y_8) \qquad \forall\, G_3, G_7;$$

$$(19.63a)$$

at the end of family store for connector 7,

$$p(G_7|y_3, y_4, y_8) = \sum_{G_3}\Pr(G_3|G_7, y_3, y_4, y_8) \qquad \forall\, G_7$$

$$\Pr(G_9|G_6, G_7, y_9) \propto p(G_9|G_6, G_7)f(y_9|G_9) \qquad \forall\, G_6, G_7, G_9$$

$$\Pr(G_7|G_6, \mathbf{y}_{-1,2,5,6}) \propto p(G_7|y_3, y_4, y_8)f(y_7|G_7)\sum_{G_9}\Pr(G_9|G_6, G_7, y_9) \qquad \forall\, G_6, G_7;$$

$$(19.63b)$$

at the end of family store for connector 6,

$$p(G_6|\mathbf{y}_{-1,2,5,6}) = \sum_{G_7}\Pr(G_7|G_6, \mathbf{y}_{-1,2,5,6}) \qquad \forall\, G_6$$

$$\Pr(G_5|G_1, G_2, y_5) \propto p(G_5|G_1, G_2)f(y_5|G_5) \qquad \forall\, G_1, G_2, G_5$$

$$\Pr(G_6|G_1, G_2, \mathbf{y}_{-1,2,5}) \propto p(G_6|\mathbf{y}_{-1,2,5,6})p(G_6|G_1, G_2)f(y_6|G_6) \qquad \forall\, G_1, G_2, G_6$$

$$\Pr(G_2|G_1, y_{-1}) \propto p(G_2)f(y_2|G_2)\sum_{G_5}\Pr(G_5|G_1, G_2, y_5)$$

$$\times \sum_{G_6}\Pr(G_6|G_1, G_2, \mathbf{y}_{-1,2,5}) \qquad \forall\, G_1, G_2$$

$$\Pr(G_1|\mathbf{y}) \propto p(G_1)f(y_1|G_1)\sum_{G_2}\Pr(G_2|G_1, y_{-1}) \qquad \forall\, G_1, \qquad (19.63c)$$

where $\Pr(\cdot)$ denotes a conditional probability in (19.62), and $p(\cdot)$ denotes Hardy–Weinberg, transmission or other intermediate probabilities of genotypes. Sampling from these probabilities is performed in the reverse order of (19.63a–c).

A pedigree with loops is not peelable in the sense that a peeling sequence with only one connector for each nuclear family cannot be found. Some conditional sampling probabilities will then depend on the genotypes of more than two individuals. For such a

Table 19.2   Example pedigree with
inbreeding loop.

Individual	Father	Mother
1	0	0
2	0	0
3	1	2
4	1	3

pedigree, after peeling a number of nuclear families, a core of unpeelable families remains, with each family containing more than one connector. There are inbreeding loops and marriage loops. The former exist when there are inbred individuals in the pedigree, and the latter exist when several individuals from one family are mated with individuals from another family. The pedigree in Table 19.2 represents an inbreeding loop, while the pedigree in Table 19.3 represents a marriage loop, which are common in livestock species with large, extended families. An optimal peeling sequence can be found with the method of Fernández and Fernando (2002), which uses algorithms for determining an optimal order of elimination in sparse systems of equations. Given an optimal sequence, the conditional sampling probabilities will depend on the genotypes of a minimal number of individuals, but this approach is not feasible in large pedigrees with complex loops. Therefore, other approaches are taken.

The pedigree in Table 19.3, e.g. consists of four nuclear families, $(1,2,5^*,6^*)$, $(3,4,7^*,8^*)$, $(5^*,8^*,9)$ and $(7^*,6^*,10)$, where connectors are marked with *. Note that if the genotype of individual 5 (or 6, 7 or 8) was known, the pedigree would be peelable with one connector per nuclear family. In approximate peeling via loop cutting (e.g. Wang *et al.*, 1996), one individual (e.g. 5) is selected, and the pedigree is modified such that individual 5 in family $(5,8,9)$ is replaced with $5'$, an individual with the same phenotype as 5, but unrelated to it. With this change the pedigree becomes peelable; however, likelihood evaluation based on peeling of the modified pedigree is approximate. To improve the degree of approximation, Wang *et al.* (1996) suggest several strategies, including the introduction of artificial families rather than individuals into the pedigree to break loops (here family $(1',2',5',6')$ rather than just $5'$), which is shown to be equivalent to iterative

Table 19.3   Example pedigree with
marriage loop.

Individual	Father	Mother
1	0	0
2	0	0
3	0	0
4	0	0
5	1	2
6	1	2
7	3	4
8	3	4
9	5	8
10	7	6

peeling. Genotype sampling can be performed exactly by iterative peeling, followed by an MH step using the distribution of the sample as a proposal probability and the probability of the genotype sample under the exact pedigree (Fernández *et al.*, 2001). Alternatively, the genotypes of different individuals can be fixed (treated as known) in different cycles of sampling, with individuals being selected such that all loops are broken (Hoeschele, 1999; Heath, 1999). In the example, in one cycle the genotype of individual 5 is fixed, in another that of 6, then that of 7, etc., so that the genotypes of all individuals are sampled, but not in each cycle. It appears that this scheme cannot be guaranteed to be irreducible for loci with more than two alleles.

An alternative to genotypic peeling is allelic peeling. To derive allelic peeling, we first consider the anterior and posterior genotype probabilities of individual i (Fernando *et al.*, 1993). From (19.61), the anterior (A) and posterior (P) probabilities are

$$A(G_i) = \Pr(\mathbf{y}_A, G_i) = \sum_{\mathbf{G}_A} \Pr(\mathbf{G}_A)\Pr(\mathbf{y}_A|\mathbf{G}_A)\Pr(G_i|\mathbf{G}_A),$$

$$P(G_i) = \Pr(\mathbf{y}_P|G_i) = \sum_{\mathbf{G}_P} \Pr(\mathbf{G}_P|G_i)\Pr(\mathbf{y}_P|\mathbf{G}_P). \tag{19.64}$$

Assuming temporarily that there are no full-sibs (i.e. that each nuclear family contains only one offspring), the anterior set of genotype G_i can be partitioned into anterior sets for the two meioses or alleles that G_i is comprised of, paternal meiosis Q_i^f and maternal meiosis Q_i^m, so that $A(G_i) = A(Q_i^f)A(Q_i^m)$ as shown below, where f and m are the parents of i, and $A(Q_i^f)$ is the anterior probability of the paternal meiosis of individual i, Q_i^f. Formulae for recursive calculation of $A(G_i)$ and $P(G_i)$ for a pedigree without loops were given by Fernando *et al.* (1994). Their anterior probability can be rewritten as (in the absence of full-sibs)

$$A(G_i) = \left[\sum_{G_f} A(G_f)f(y_f|G_f)p(Q_i^f|G_f)\right]\left[\sum_{G_m} A(G_m)f(y_m|G_m)p(Q_i^m|G_m)\right]$$

$$= \Pr(\mathbf{y}_{A_i^f}, Q_i^f)\Pr(\mathbf{y}_{A_i^m}, Q_i^m) = A(Q_i^f) \cdot A(Q_i^m), \tag{19.65}$$

where A_i^f represents the set of all meioses anterior to the paternal meiosis of i or Q_i^f, and $\mathbf{y}_{A_i^f}$ is the set of all phenotypes pertaining to individuals with genotypes consisting of meioses, which are anterior to Q_i^f. Similarly, the posterior probability of G_i can be rewritten as

$$P(G_i) = \prod_j P_{ij}(G_i) = \prod_j \Pr(\mathbf{y}_{ij}|G_i)$$

$$= \prod_j 0.5[\Pr(\mathbf{y}_{ij}|Q_j^i = Q_i^f) + \Pr(\mathbf{y}_{ij}|Q_j^i = Q_i^m)]$$

$$= \prod_j 0.5[P_{ji}(Q_j^f = Q_i^f) + P_{ji}(Q_j^f = Q_i^m)], \tag{19.66}$$

where j is an offspring of i (we assume here that i is a father), $\mathbf{y}_{ij}$ is the set of phenotypes of all individuals posterior to i through offspring j (including the other parent of j), $P_{ji}(Q_j^f = Q_i^f)$ denotes the posterior probability of the meiosis ji or Q_j^f (the paternal meiosis of j), with Q_j^f evaluated at Q_i^f. Note that for equation (19.66) to hold we are assuming that $\Pr(Q_j^f \Leftarrow Q_i^f) = \Pr(Q_j^f \Leftarrow Q_i^m) = 0.5$, which would be straightforward to generalize.

Formulae for recursive calculation of $A(Q_i^f)$ and $P_{is}(Q_i^f)$ for a pedigree without loops are obtained next. Let s and d denote father and mother of i. The set of meioses anterior to Q_i^f includes Q_s^f and Q_s^m, all meioses anterior to Q_s^f, all meioses anterior to Q_s^m, and all meioses posterior to Q_s^f and Q_s^m except for those meioses through i. Then

$$
\begin{aligned}
A(Q_i^f) = \sum_{Q_s^f} &\Bigg[\sum_{\mathbf{Q}_{A_s^f}} \Pr(\mathbf{Q}_{A_s^f}) \Pr(\mathbf{y}_{A_s^f}|\mathbf{Q}_{A_s^f}) \Pr(Q_s^f|\mathbf{Q}_{A_s^f}) \\
&\times \sum_{Q_s^m} \Bigg[\sum_{\mathbf{Q}_{A_s^m}} \Pr(\mathbf{Q}_{A_s^m}) \Pr(\mathbf{y}_{A_s^m}|\mathbf{Q}_{A_s^m}) \Pr(Q_s^m|\mathbf{Q}_{A_s^m}) \Pr(y_s|Q_s^f, Q_s^m) \Pr(Q_i^f|Q_s^f, Q_s^m) \\
&\times \prod_{o \in C_s, o \neq i} \Bigg[\sum_{Q_o^f} \Pr(Q_o^f|Q_s^f, Q_s^m) P_{os}(Q_o^f) \Bigg] \Bigg] \Bigg] \\
= \sum_{Q_s^f} &\Bigg[A(Q_s^f) \sum_{Q_s^m} \Bigg[A(Q_s^m) \Pr(y_s|Q_s^f, Q_s^m) \Pr(Q_i^f|Q_s^f, Q_s^m) \\
&\times \prod_{o \in C_s, o \neq i} \Bigg[\sum_{Q_o^f} \Pr(Q_o^f|Q_s^f, Q_s^m) P_{os}(Q_o^f) \Bigg] \Bigg] \Bigg] \\
= 0.5 &\Bigg[A(Q_s^f = Q_i^f) \sum_{Q_s^m} \Bigg[A(Q_s^m) f(y_s|Q_s^f, Q_s^m) \\
&\times \prod_{o \in C_s, o \neq i} 0.5 \Big[P_{os}(Q_o^f = Q_s^f) + P_{os}(Q_o^f = Q_s^m) \Big] \Bigg] \\
&+ A(Q_s^m = Q_i^f) \sum_{Q_s^f} \Bigg[A(Q_s^f) \cdot f(y_s|Q_s^f, Q_s^m) \\
&\times \prod_{o \in C_s, o \neq i} 0.5 \Big[P_{os}(Q_o^f = Q_s^f) + P_{os}(Q_o^f = Q_s^m) \Big] \Bigg] \Bigg],
\end{aligned}
\tag{19.67}
$$

where C_s is the set of all offspring of s, and $A(Q_i^m)$ is defined analogously to (19.67). Note that $A(Q_s^f = Q_i^f)$ denotes that the anterior probability of meiosis Q_s^f is evaluated at $Q_s^f = Q_i^f$. From (19.67), the anterior probability of Q_i^f can be computed once the anterior probabilities of Q_s^f and Q_s^m and the posterior probabilities of the meioses pertaining to

all offspring of s other than i have been computed. Note that the last equality in (19.67) again only holds as a special case, where $\Pr(Q_i^f \Leftarrow Q_s^f) = \Pr(Q_i^f \Leftarrow Q_s^m) = 0.5$, and represents the average of the joint probabilities of the meioses anterior to Q_i^f and of Q_i^f conditional on $Q_i^f = Q_s^f$ and on $Q_i^f = Q_s^m$, respectively.

The set of meioses posterior to meiosis is, or Q_i^f, includes all meioses anterior to meiosis id or Q_i^m, all paternal (i is father) or maternal (i is mother) meioses of the offspring of i, and all meioses posterior to the latter. Then, assuming that i is a father, e.g.

$$
\begin{aligned}
P_{is}(Q_i^f) &= \sum_{Q_i^m} \sum_{Q_{A_i^m}} \Pr(\mathbf{Q}_{A_i^m}) \Pr(\mathbf{y}_{A_i^m} | \mathbf{Q}_{A_i^m}) \Pr(Q_i^m | \mathbf{Q}_{A_i^m}) \\
&\quad \times \Pr(y_i | Q_i^f, Q_i^m) \prod_{o \in C_i} \sum_{Q_o^f} \Pr(Q_o^f | Q_i^f, Q_i^m) \Pr(\mathbf{y}_{P_{oi}} | Q_o^f) \\
&= \sum_{Q_i^m} A(Q_i^m) \Pr(y_i | Q_i^f, Q_i^m) \prod_{o \in C_i} \sum_{Q_o^f} \Pr(Q_o^f | Q_i^f, Q_i^m) \Pr(\mathbf{y}_{P_{oi}} | Q_o^f) \\
&= \sum_{Q_i^m} A(Q_i^m) \Pr(y_i | Q_i^f, Q_i^m) \prod_{o \in C_i} 0.5 [P_{oi}(Q_o^f = Q_i^f) + P_{oi}(Q_o^f = Q_i^m)],
\end{aligned}
$$
(19.68)

where the last equality again holds only as a special case.

For genotypic reverse peeling, the conditional genotype probabilities in (19.59) are computed based on a genotypic peeling sequence for the nuclear families comprising the pedigree. The joint probability of all meioses in a pedigree can be partitioned analogously to (19.59), and the resulting conditional probabilities of the meioses can be computed using a peeling sequence for nuclear meiosis groups. A nuclear meiosis group consists of the two meioses of a father (mother) and all paternal (maternal) meioses of its offspring. The two meioses of an individual, which has a phenotype but no offspring, also form a nuclear meiosis group. Consider the example pedigree in Table 19.4 and assume that individuals 8 and 9 have phenotypes. This pedigree consists of nine nuclear meiosis groups, and in a particular peeling sequence, these are $(3f, 3m, 6f^*)$, $(4f, 4m, 6m^*)$, $(6f, 6m, 8m^*)$, $(8f^*, 8m)$, $(7f, 7m, 9m^*)$, $(9f^*, 9m)$, $(2f, 2m, 5m^*)$, $(5f^*, 5m, 8f, 9f)$,

Table 19.4 Example pedigree for allelic peeling.

Individual	Father	Mother	Meioses
1	0	0	$1f$, $1m$
2	0	0	$2f$, $2m$
3	0	0	$3f$, $3m$
4	0	0	$4f$, $4m$
5	1	2	$5f$, $5m$
6	3	4	$6f$, $6m$
7	0	0	$7f$, $7m$
8	5	6	$8f$, $8m$
9	5	7	$9f$, $9m$

$(1f, 1m, 5f)$, where connector meioses are marked with*. Conditional probabilities are computed in this sequence, e.g., for meiosis group $(3f, 3m, 6f^*)$,

$$\Pr(Q_3^m | Q_3^f, Q_6^f, y_3) \propto p(Q_3^m) f(y_3 | Q_3^f, Q_3^m) p(Q_6^f | Q_3^f, Q_3^m)$$

$$= \begin{cases} 0.5 p(Q_3^m) f(y_3 | Q_3^f, Q_3^m) & \forall\ Q_3^f = Q_6^f \\ 0.5 p(Q_3^m = Q_6^f) f(y_3 | Q_3^f, Q_3^m = Q_6^f) & \forall\ Q_3^f \neq Q_6^f, Q_3^m = Q_6^f \\ 0 & \forall\ Q_3^f \neq Q_6^f, Q_3^m \neq Q_6^f \end{cases},$$

$$\Pr(Q_3^f | Q_6^f, y_3) \propto p(Q_3^f) \sum_{Q_3^m} p(Q_3^m) f(y_3 | Q_3^f, Q_3^m) p(Q_6^f | Q_3^f, Q_3^m)$$

$$= \begin{cases} 0.5 p(Q_3^f = Q_6^f) \sum_{Q_3^m} p(Q_3^m) f(y_3 | Q_3^f = Q_6^f, Q_3^m) & \forall\ Q_3^f = Q_6^f \\ 0.5 p(Q_3^f) p(Q_3^m = Q_6^f) f(y_3 | Q_3^f, Q_3^m = Q_6^f) & \forall\ Q_3^f \neq Q_6^f; \end{cases} \tag{19.69a}$$

at end of meiosis group store for connector $6f$,

$$p(Q_6^f | y_3) = 0.5 p(Q_3^f = Q_6^f) \sum_{Q_3^m} p(Q_3^m) f(y_3 | Q_3^f = Q_6^f, Q_3^m)$$

$$+ \sum_{Q_3^f \neq Q_6^f} 0.5 p(Q_3^f) p(Q_3^m = Q_6^f) f(y_3 | Q_3^f, Q_3^m = Q_6^f) \forall\ Q_6^f; \tag{19.69b}$$

for meiosis group $(9f^*, 9m)$,

$$\Pr(Q_9^m | Q_9^f, y_9, y_7) = p(Q_9^m | y_7) f(y_9 | Q_9^f, Q_9^m) \qquad \forall\ Q_9^f, Q_9^m; \tag{19.69c}$$

at end of this meiosis group store for connector $9f$,

$$p(Q_9^f | y_9, y_7) = \sum_{Q_9^m} p(Q_9^m | y_7) f(y_9 | Q_9^f, Q_9^m) \quad \forall\ Q_9^f; \tag{19.69d}$$

and for meiosis group $(1f, 1m, 5f)$,

$$\Pr(Q_5^f | Q_1^f, Q_1^m, \mathbf{y}_{-1}) = p(Q_5^f | \mathbf{y}_{-1}) p(Q_5^f | Q_1^f, Q_1^m)$$

$$= \begin{cases} 0.5 p(Q_5^f | \mathbf{y}_{-1}) & \forall\ Q_5^f = Q_1^f \text{ or } Q_5^f = Q_1^m \\ 0 & \text{otherwise,} \end{cases}$$

$$\Pr(Q_1^m | Q_1^f, \mathbf{y}) = p(Q_1^m) f(y_1 | Q_1^f, Q_1^m) \sum_{Q_5^f} p(Q_5^f | \mathbf{y}_{-1}) p(Q_5^f | Q_1^f, Q_1^m)$$

$$= p(Q_1^m) f(y_1 | Q_1^f, Q_1^m) 0.5 [p(Q_5^f = Q_1^f | \mathbf{y}_{-1})$$

$$+ p(Q_5^f = Q_1^m | \mathbf{y}_{-1})] \qquad \forall\ Q_1^f, Q_1^m,$$

$$\Pr(Q_1^f | \mathbf{y}) = p(Q_1^f) \sum_{Q_1^m} p(Q_1^m) f(y_1 | Q_1^f, Q_1^m) \sum_{Q_5^f} p(Q_5^f | \mathbf{y}_{-1}) p(Q_5^f | Q_1^f, Q_1^m)$$

$$= p(Q_1^f) \sum_{Q_1^m} p(Q_1^m) f(y_1|Q_1^f, Q_1^m) 0.5[p(Q_5^f = Q_1^f|\mathbf{y}_{-1})$$

$$+ p(Q_5^f = Q_1^m|\mathbf{y}_{-1})] \qquad \forall Q_1^f. \qquad (19.69e)$$

Both genotypic and allelic peeling, as described above, do not work for pedigrees with loops. Moreover, allelic peeling does not work for pedigrees with full-sibs, that is to say, the presence of full-sibs creates loops for allelic peeling. For illustration, consider father 1 $(1f, 1m)$, mother 2 $(2f, 2m)$, and their joint offspring 3 $(3f, 3m)$ and 4 $(4f, 4m)$, which have phenotypes. This pedigree consists of four nuclear meiosis groups, $(1f, 1m, 3f^*, 4f^*)$, $(2f, 2m, 3m^*, 4m^*)$, $(3f^*, 3m^*)$ and $(4f^*, 4m^*)$, and each group contains more than one connector, hence allelic peeling cannot be performed. Earlier results such as (19.65) do not hold in the presence of full-sibs. Iterative genotypic peeling (Janss *et al.*, 1995a; Kerr and Kinghorn, 1996) provides an approximate approach to peeling in pedigrees with loops. Thallmann *et al.* (2001a,b) presented allelic peeling formulae in matrix notation, and Thallmann *et al.* (2001b) proposed iterative allelic peeling for pedigrees with full-sibs and loops. In the presence of full-sibs, it is expected that the iterative allelic peeling algorithm often produces results that are very close to the exact calculation. As pointed out by Thallmann *et al.* (2001b), allelic peeling produces exact results unless at least two full-sibs in a nuclear family have ambiguous ordered genotypes, and (nearly) exact results if one of the parents has a known unordered genotype. Computational efficiency is greatly improved by allelic peeling relative to genotypic peeling: Assuming a pedigree without full-sibs, for most calculations in allelic peeling, the number of computing operations is proportional to α^2, where α is the number of alleles, while it is proportional to $(\alpha^2)3 = \alpha^6$ for genotypic peeling.

An alternative to genotype sampling based on genotypic or allelic peeling is descent graph sampling. With genotypic or allelic peeling, we sample on the space of genetic descent states, which specify the paths of gene flow and actual founder alleles dropped down each path. Thompson (1994) suggested that a better (smaller) sampling space is that of descent graphs (also referred to as *segregation patterns* or *sets of meioses indicators*). Consider the example pedigree in Table 19.5 with partial genotype data, where $\{\cdot\}$ denotes an observed, unordered genotype and $\{00\}$ a missing genotype.

A particular descent graph for this pedigree can be represented graphically as shown in Figure 19.2, where pairs of circles represent pairs of alleles of individuals (individuals 1 and 2 at the top, 3 to 5 below and 6 at the bottom), and arrows indicate the gene flow. A numerical representation of this descent graph is [0,0,0,1,0,0], pertaining to the alleles in descendants 3, 4 and 6. A value of 0 (1) indicates grandpaternal (grandmaternal)

Table 19.5 Example pedigree with partial genotype data.

Individual	Father	Mother	Observed genotype
1	0	0	{00}
2	0	0	{00}
3	1	2	{12}
4	1	2	{12}
5	0	0	{00}
6	4	5	{11}

Figure 19.2   Descent graph for the pedigree in Table 19.5.
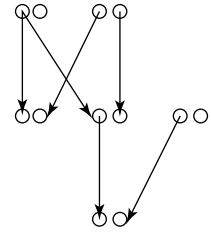
inheritance. Computationally, a descent graph can be stored and manipulated most efficiently using bit representation, with one 32-bit integer word encoding the meiosis indicators of 16 individuals, as suggested by Tier and Henshall (2001). For single (Tier and Henshall, 2001) or multiple unlinked loci (Du and Hoeschele, 2000a) with no genotype information available, genotype sampling can be performed by proposing small changes to the founder alleles and/or the descent graph in each cycle, and accepting or rejecting these in an MH step. Changes to the founder alleles are proposed by resampling a few alleles according to allele frequencies, and changes to the descent graph are proposed by swapping a few inheritance states. Typically, many MH steps would be performed before other unknowns are updated.

When genotype data are available as in the example pedigree, the goal is to obtain many descent graph samples, which are consistent with the observed data. Given a complete descent graph (a descent graph with each meiosis indicator specified), Sobel and Lange (1996) propose an algorithm to determine the founder origin of each meiosis in the pedigree and to obtain a list of permissible alleles for each founder meiosis. The genotype elimination algorithm of Lange and Goradia (1987) is based on checking the consistency of genotypes within nuclear families. In a pedigree without loops, repeated processing of the set of nuclear families eliminates all inconsistent genotypes and creates for each individual a list of consistent (with observed genotype data) genotypes. From the genotype lists, lists of consistent alleles for each meiosis in the pedigree can be obtained. Genotype elimination can be used to determine whether a descent graph is consistent with the observed genotype data. Du and Hoeschele (2000b) modify genotype elimination so that every inconsistent, complete or incomplete descent graph is detected in a pedigree without loops (an incomplete descent graph has some but not all meiosis indicators specified). These authors also provide two modifications of genotype elimination, one being more efficient computationally and the other also eliminating more inconsistent genotypes in the presence of loops. Lastly, Du and Hoeschele (2000b) describe an allele elimination algorithm which is based on checking the consistency of alleles within each nuclear meiosis group and creates for each meiosis a list of alleles directly. Allele elimination is computationally more efficient than genotype elimination and is identical to the method of Sobel and Lange (1996) for a complete descent graph, but it does not detect every illegal, incomplete descent graph, even in a pedigree without loops.

Sobel and Lange (1996) presented a descent graph sampler for pedigrees with partial genotype data, where meiosis indicators are sampled from their marginal distribution (not conditional on genotypes of the founders) using a Metropolis sampler with block-updating based on half- or full-sib relationship and with multiple updating steps through illegal transmission patterns to achieve an irreducible chain. Descent graph samplers, which perform meiosis grouping based on genetic relationship and observed genotype data to produce groups with very small descent graph spaces given observed data, and

block-updating should improve mixing by allowing larger updates relative to the sampler of Sobel and Lange (Stricker *et al.*, 2002).

19.7   FINE MAPPING OF QUANTITATIVE TRAIT LOCI

19.7.1   Fine Mapping Using Current Recombinations

Initial genome-wide linkage analysis typically assigns a QTL to a 10–20 cM region. To further reduce the region of likely QTL location using current recombinants, additional markers need to be placed in the initial region at, say, 0.5–2 cM intervals. Such an approach utilizes recombinant chromosomes from a heterozygous parent and has been termed the *chromosome dissection method* (Thoday, 1961; Soller and Andersson, 1998). Considering QTL position refinement in large half-sib families existing in some species (e.g. dairy cattle, fish, trees) by typing offspring, which are recombinant for the original marker bracket with additional markers, Thaller and Hoeschele (2000) proposed simple statistics for assigning a QTL to a marker subinterval of 1–5 cM using contrast mapping and substitution mapping. Simple statistics were proposed, because the goal was to assign a QTL to a subinterval within the original interval, but not to obtain a position estimate within that subinterval, due to the conjecture that such an estimate would be rather inaccurate.

Offspring are assigned to haplotype groups based on the subinterval containing the recombination event in their paternal haplotype (the father is the common parent). For contrast mapping, the contrast for a particular marker subinterval is obtained as the difference in average phenotype between all haplotype groups with grandpaternal origin of the subinterval and those with grandmaternal origin. The largest contrast is expected for the subinterval containing the QTL, and this approach is equivalent in power to sliding three-marker regression of haplotype group means (or individual phenotypes) on paternal (or maternal if mother is the common parent) marker alleles (Thaller and Hoeschele, 2000). Confidence in the identified subinterval was evaluated by bootstrapping (Thaller and Hoeschele, 2000).

In substitution mapping, the occurrence of a certain pair of haplotypes is sufficient for QTL assignment to a specific subinterval, if the pair of haplotypes has a certain segregation status combination. Offspring in the case of dairy cattle are sons of a father with many daughters, which are genotyped to determine whether a son is segregating (i.e. heterozygous) at the QTL. Assuming that the son did not inherit the mutant QTL allele from its mother, segregation indicates that the QTL allele was inherited from its father. Let '1' indicate a marker allele linked with the mutant QTL allele and '0' a marker allele linked with the normal QTL allele in the father. Consider two sons with paternal haplotypes 111000 and 000111. Both are recombinant in subinterval 3. If both sons are not segregating (i.e. did not inherit the mutant QTL allele from the father), then the QTL can be assigned to subinterval 3 (Thaller and Hoeschele, 2000).

19.7.2   Fine Mapping Using Historical Recombinations

Positional cloning or candidate positional cloning requires the assignment of a gene to a region of 0.3 cM or less (Falconer and Mackay, 1996), which will often not be feasible

with chromosome dissection methods due to the limited number of current recombinations in livestock or human populations. Positional cloning of monogenic diseases has been successful after assignment of the gene to such a small region by methods utilizing historical recombinations or Linkage Disequilibrium LD (more precisely, gametic phase disequilibrium: Falconer and Mackay, 1996; Crow and Kimura, 1970). For a (statistical) description of LD, see Weir (1996), and for simple disequilibrium measures and their application to the fine mapping of monogenic diseases, see Devlin and Risch (1995). For details on association analyses between a binary trait and genotype, see **Chapter 34**. There are different types of LD, and LD is influenced by multiple factors (such as selection, admixture, genetic drift, mutation, migration, coancestry, population expansion; e.g. Xiong and Guo, 1997). LD fine-mapping methods assume that LD is primarily due to the introduction of a variant on an ancestral haplotype via mutation (or migration), which is partially preserved in descendants of the current generation.

For linkage mapping, pedigree information is available on all members of the pedigree, except for founders which are assumed to be unrelated, and marker and phenotype data may be available across multiple generations. For fine mapping using historical recombinations, marker and phenotypic data are available only on the current generation(s), and there is no pedigree information relating current generation individuals or haplotypes back to one or multiple ancestral haplotypes carrying a unique, mutant trait allele. Similar to the first linkage mapping methods, the earliest LD methods used single-marker statistics. More recent methods utilize multiple linked markers (e.g. Terwilliger, 1995; Xiong and Guo, 1997; McPeek and Strahs, 1999; Service et al., 1999; Lam et al., 2000; Meuwissen and Goddard, 2000; 2001; Morris et al., 2000; Farnir et al., 2002; Garner and Slatkin, 2002; Liu et al., 2001; Meuwissen et al., 2002). Some methods consider, to some extent, the evolutionary history of the population (e.g. McPeek and Strahs, 1999; Xiong and Guo, 1997; Lam et al., 2000; Service et al., 1999), while others do not (e.g. Terwilliger, 1995; Meuwissen and Goddard, 2000; 2001; Morris et al., 2000; Liu et al., 2001). Modeling the population history allows some haplotypes to be more recently related than others. Assuming independent recombinational histories of all haplotypes in the current generation represents the limiting case of a rapidly growing population carrying the mutant allele, due to selection or chance (e.g. McPeek and Strahs, 1999), referred to as a *star-shaped genealogy*. Nearly all methods have been applied to relatively young and very rare diseases, where marker allele frequencies on normal chromosomes can be assumed to be constant across generations (and roughly equal to the population frequencies), and all the information about gene location is contained in the disease-causing chromosomes. This is clearly not true for a quantitative-trait gene, which may substantially increase in frequency due to selection or chance.

An imperfect relationship between genetic distance and LD leads to potential drawbacks common to all LD mapping methods. A trait locus may be in disequilibrium with some but not all markers in its vicinity. Therefore, methods utilizing multiple linked markers and combining linkage and LD seem to be required for obtaining reliable inferences in fine mapping of QTLs. Terwilliger (1995) therefore proposes to adapt his likelihood model to emulate the transmission/disequilibrium test (e.g. Spielman et al., 1993) for qualitative traits; Allison, 1997 for quantitative traits), which, given an appropriate population structure (Sham, 1998), can be considered as a test for both linkage and LD. For analysis of nuclear families, a test combining linkage and LD has been proposed by Abecasis et al. (2000). Other methods combining LD and Linkage for QTL mapping

(LDL mapping) include Zhao *et al.* (1998), Allison *et al.* (1999), Almasy *et al.* (1999), Fulker *et al.* (1999), Wu and Zeng (2001), Farnir *et al.* (2002), Meuwissen *et al.* (2002) and Fan and Jung (2003). Methods for LDL mapping that are more general and applicable to a wider range of experimental designs and population structures including (at least in principle) complex pedigrees have recently been developed (e.g. Perez-Enciso, 2003; Lou *et al.*, 2005; Gao and Hoeschele, 2005).

The ML methodology of Terwilliger (1995) and Xiong and Guo (1997) can in principle be extended to quantitative traits. The joint likelihood of phenotypes and marker haplotypes observed on the current generation is a weighted sum over the possible QTL alleles on the marker haplotypes and over all possible ancestral marker haplotypes. Indeed, Farnir *et al.* (2002) extended Terwilliger's ML method to a biallelic QTL mapped in a half-sib design. Unfortunately, Terwilliger's method does not account for dependencies among the tightly linked markers in a haplotype and employs a composite log likelihood (CL), which is the sum of the log likelihoods of the individual markers. A CL-based method can overstate the evidence and hence underestimate confidence intervals for position, it can give biased results, and asymptotic likelihood-ratio theory does not apply when the CL ignores dependencies (Garner and Slatkin, 2002). Dependencies among tightly linked markers in a haplotype can be accounted for as described by McPeek and Strahs (1999), Service *et al.* (1999), Morris *et al.* (2000), Perez-Enciso (2003) and Sillanpaa and Bhattacharjee (2005). In analogy with linkage multipoint mapping, LD or LDL multipoint mapping is performed by evaluating the likelihood over a fine grid of QTL positions along the chromosomal segment investigated, assuming known marker positions (note the need for reliable ordering of tightly linked markers and accurate position estimation). While some likelihood methods were developed for mapping disease genes (e.g. Terwilliger, 1995; McPeek and Strahs, 1999), others were devised to map QTL in a specific population type (Farnir *et al.*, 2002) or in a general setting not restricted to a specific experimental design or population (Lou *et al.*, 2005).

The likelihood methods have the disadvantage of making inferences about QTL position conditional on the values for the other (nuisance) parameters (e.g. allele and haplotype frequency parameters, QTL effects, ancestral haplotype, etc.) fixed at their ML estimates. Bayesian methods have been developed for LD mapping of disease genes (Morris *et al.*, 2000; Liu *et al.*, 2001), for LDL mapping of QTL (Perez-Enciso, 2003) and for LD fine mapping of QTL and disease genes in short chromosomal regions (Sillanpaa and Bhattacharjee, 2005). The Bayesian methods fully account for the uncertainty associated with the true values of the nuisance parameters, but they require implementation with MCMC algorithms, which may be slow to converge and computationally expensive. For ML LDL mapping of QTL, closed form solutions for parameter estimation exist; however, the quantification of uncertainty is based on the large sample, asymptotic variance–covariance matrix of the parameter estimates (Lou *et al.*, 2005). Another strong advantage of Bayesian methods is their ability to incorporate external genomic information into the prior distribution of gene location (Rannala and Reeve, 2001). While the method of Lou *et al.* (2005) integrates LD mapping with interval (linkage) mapping by considering haplotypes consisting of a QTL position and its flanking markers, the method of Perez-Enciso (2003) considers the entire multimarker haplotype by adapting the population model for LD decay of Morris *et al.* (2000). Both methods assume a biallelic QTL, i.e. they do not allow for more than a single ancestral mutation (while the number of alleles at a QTL cannot be reliably inferred via linkage mapping, it might be possible to

estimate this parameter via LDL mapping). Sillanpaa and Bhattacharjee (2005) consider each marker in a small, dense chromosomal segment as a potential QTL position with two or more alleles depending on the marker type. They incorporate covariances among consecutive markers via the prior for the indicator variables of the markers (which control inclusion/exclusion of each marker in a multiple regression model).

In analogy with linkage mapping, ML and Bayesian LDL mapping of QTL can be implemented as distribution or expectation methods (see Sections 19.3.4 and 19.4.3), with the expectation methods requiring no assumption or inference about the number of QTL alleles and no estimation of allele and haplotype frequencies pertaining to the QTL, but possibly at the expense of slightly reduced power in some situations. While the LDL QTL mapping methods of Farnir *et al.* (2002), Perez-Enciso (2003) and Lou *et al.* (2005) are distribution methods assuming biallelic QTLs in outbred pedigrees, the method of Almasy *et al.* (1999) for a single marker, Meuwissen *et al.* (2002) for multiple linked markers in a half-sib design, and Gao and Hoeschele (2005) for multiple linked markers and complex pedigrees are all expectation methods. Distribution methods require more assumptions and are more highly parameterized, which may inversely affect their robustness, although somewhat more flexible and robust implementations have been proposed for disease genes (Liu *et al.*, 2001) and for QTL (Sillanpaa and Bhattacharjee, 2005). In the LDL expectation methods, IBD probabilities in founders are no longer taken to be 0 but rather are estimated on the basis of similarity among marker haplotypes, either by Monte Carlo simulation (Meuwissen and Goddard, 2000) or deterministically using approximate coalescence theory (Meuwissen and Goddard, 2001). While Meuwissen *et al.* (2002) assume that marker haplotypes are known or inferred with certainty in a half-sib design, Gao and Hoeschele (2005) allow for marker haplotype uncertainty in complex pedigrees. However, the last method does not yet allow for (substantial amounts of) missing marker data and does not yet incorporate marker LD in the pedigree founders. It is a two-step procedure consisting of marker haplotype inference followed by QTL mapping conditional on the haplotype information, which may be suboptimal in the case of missing data relative to methods inferring marker haplotypes and ordered QTL genotypes simultaneously (e.g. Perez-Enciso, 2003), but it is likely computationally more feasible and possibly more robust.

The development of LD and LDL QTL mapping methods is still an active area of research. Systematic comparisons among the different LD and LDL QTL mapping methods in different settings (e.g. population histories) would clearly be beneficial. Most methods map only a single QTL and should be extended to multiple QTLs, while some methods already map several QTLs simultaneously (e.g. Sillanpaa and Bhattacharjee, 2005). LDL QTL mapping methods should also allow for multiple ancestral mutations or genetic heterogeneity (Liu *et al.*, 2001; Morris *et al.*, 2002) allow for multiple ancestral mutations in LD mapping of disease genes). Lastly, existing methods should be extended to multiple traits as in Meuwissen and Goddard (2004).

19.8   CONCLUDING REMARKS

Much progress has been achieved in the development of statistical methods and software for linkage mapping of QTL in outbred populations and recently for LD and joint LD and

linkage (LDL) mapping, in particular with the development of the Bayesian methodology, its MCMC implementations, and with the development of software packages such as SOLAR. This work is, however, not yet completed and likely will not be completed in the near future, or ever (?), in particular with the advances in high-throughput phenotyping technologies including physiological traits, expression profiles or etraits, proteomic profiling, metabolic traits and high-throughput cell-based assays (Williams, 2006). We have begun analyzing these high-dimensional phenotype data with existing methods for linkage, LD and LDL QTL mapping methods with success, but some innovative methods or modifications of existing methods will be required to extract the most information from these data (e.g. Kulp and Jagalur, 2006; Liu *et al.*, 2007). We will capitalize on these techniques in the highly promising fields of genetical genomics (e.g. Jansen and Nap, 2001; Bing and Hoeschele, 2005; Drake *et al.*, 2006). Further improvements and innovations should include the incorporation of epistasis along more traditional lines (Yi *et al.*, 2005) that are limited (not in theory but in practice) to two-locus interactions and via novel approaches providing a more general approach to epistasis without limitation to low order (e.g., Hanlon and Lorenz, 2005), extensions of peeling and descent graph genotype samplers to multiple linked loci (in the presence of missing data), and systematic performance comparisons among methods in different settings (population histories and structures, etc.) via genetic analysis workshops and other venues.

Acknowledgments

REFERENCES

Abecasis, G.R., Cardon, L.R. and Cookson, W.O.C. (2000). A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics* **66**, 279–292.

Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101.

Allison, D.B. (1997). Transmission- disequilibrium tests for quantitative traits. *American Journal of Human Genetics* **60**, 676–690.

Allison, D.B., Heo, M., Kaplan, N. and Martin, E.R. (1999). Sibling-based tests of linkage and association for quantitative traits. *American Journal of Human Genetics* **64**, 1754–1763.

Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics* **62**, 1198–1211.

Almasy, L., Williams, J.T., Dyer, T.D. and Blangero, J. (1999). Quantitative trait locus detection using combined linkage/disequilibrium analysis. *Genetic Epidemiology* **17**(Suppl. 1), S31–S36.

Barnard, G.A. (1949). Statistical inference. *Journal of the Royal Statistical Society, Series B* **11**, 115–139.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Bing, N. and Hoeschele, I. (2005). Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170**, 533–542.

Botstein, D., White, R.L., Skolnick, M.H. and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction length fragment length polymorphisms. *American Journal of Human Genetics* **32**, 314–331.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison Wesley, Reading, MA.

Broman, K.W. and Speed, T.P. (2002). A model selection approach for identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society, Series B* **64**, 641–656.

Cannings, C., Thompson, E.A. and Skolnick, M.H. (1978). Probability functions on complex pedigrees. *Advances in Applied Probability* **10**, 26–61.

Carlin, B.P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society, Series B* **77**, 473–484.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *The American Statistician* **49**, 327.

Chipman, H., Edwards, E.I. and McCulloch, R.E. (2001). The practical implementation of Bayesian model selection. In *Model Selection*, P. Lahiri, ed. Institute of Mathematical Statistics, Beachwood, OH, pp. 65–116.

Churchill, G.A. and Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1997). *Multiple Shrinkage and Subset Selection in Wavelets*, ISDS Discussion Paper 95–37. Institute of Statistics and Decision Sciences, Duke University, Durham, NC.

Coppieters, W., Kvasz, A., Farnir, F., Arranz, J.J., Grisart, B., Mackinnon, M. and Georges, M. (1998). A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sib pedigrees: application to milk production in a granddaughter design. *Genetics* **149**, 1547–1555.

Crow, J.F. and Kimura, M. (1970). *An Introduction to Population Genetic Theory*. Harper & Row, New York.

Davis, S., Schroeder, M., Goldin, L.R. and Weeks, D. (1996). Nonparametric simulation-based statistics for detecting linkage in general pedigrees. *American Journal of Human Genetics* **58**, 867–880.

DeBoer, I.J.M. and Hoeschele, I. (1993). Genetic evaluation methods for populations with dominance and inbreeding. *Theoretical and Applied Genetics* **86**, 245–258.

Dekkers, J.C.M. and Dentine, M.R. (1991). Quantitative genetic variance associated with chromosomal markers in segregating populations. *Theoretical and Applied Genetics* **81**, 212–220.

Demenais, F., Lathrop, M. and Lalouel, J.M. (1986). Robustness and power of the unified model in the analysis of quantitative measurements. *American Journal of Human Genetics* **38**, 228–234.

Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322.

Drake, T.A., Schadt, E.E. and Lusis, A.J. (2006). Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. *Mammalian Genome* **17**, 466–479.

Du, F.-X. and Hoeschele, I. (2000a). Estimation of additive, dominance, and epistatic variance components using finite-locus models implemented with a single-site Gibbs and a descent graph sampler. *Genetical Research* **76**, 187–198.

Du, F.-X. and Hoeschele, I. (2000b). A note on genotype and allele elimination in complex pedigrees with incomplete genotype data. *Genetics* **156**, 2051–2062.

Du, F.-X., Hoeschele, I. and Gage-Lahti, K.M. (2000). Estimation of additive and dominance variance components in finite polygenic models and complex pedigrees. *Genetical Research* **74**, 179–187.

Elsen, J.M., Knott, S.A., Le Roy, P. and Haley, C.S. (1997). Comparison between some approximate likelihood methods for quantitative trait locus detection in progeny test designs. *Theoretical and Applied Genetics* **95**, 236–245.

Elsen, J.M. and Le Roy, P. (1990). In *Proceedings of the Fourth World Congress on Genetics Applied to Livestock Production*, W.G. Hill, R. Thompson and J.A. Wolliams, eds. Vol. XV, Edinburgh, pp. 37–49.

Elston, R.C. and Stewart, J. (1971). A general model for the analysis of pedigree data. *Human Heredity* **21**, 523–542.

Falconer, D.S. and Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*. Longman, London.

Fan, R. and Jung, J. (2003). High-resolution joint linkage disequilibrium and linkage mapping of quantitative trait loci based on sibship data. *Human Heredity* **56**, 166–187.

Farnir, F., Grisart, B., Coppieters, W., Riquet, J., Berzi, P., Cambisano, N., Karim, L., Mni, M., Moisio, S., Simon, P., Wagenaar, D., Vikki, J. and Georges, M. (2002). Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics* **161**, 275–287.

Fernández, S.A. and Fernando, R.L. (2002). Determining peeling order using sparse matrix algorithms. *Journal of Dairy Science* **85**, 1623–1629.

Fernández, S.A., Fernando, R.L., Guldbrandtsen, B., Totir, L.R. and Carriquiry, A.L. (2001). Sampling genotypes in large pedigrees with loops. *Genetics, Selection, Evolution* **33**, 337–368.

Fernández, C. and Steel, M.F.J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* **93**, 359–371.

Fernando, R.L., Stricker, C. and Elston, R.C. (1993). An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. *Theoretical and Applied Genetics* **87**, 89–93.

Fernando, R.L., Stricker, C. and Elston, R.C. (1994). The finite polygenic mixed model: an alternative formulation for the mixed model of inheritance. *Theoretical and Applied Genetics* **88**, 573–580.

Fulker, D.W., Cherny, S.S. and Cardon, L.R. (1995). Multipoint interval mapping of quantitative trait loci using sib pairs. *American Journal of Human Genetics* **56**, 1224–1233.

Fulker, D.W., Cherney, S.S., Sham, P.C. and Hewitt, J.K. (1999). Combining linkage and association sib pair analysis for quantitative traits. *American Journal of Human Genetics* **64**, 259–267.

Gao, G. and Hoeschele, I. (2005). Approximating identity-by-descent matrices using multiple haplotype configurations on pedigrees. *Genetics* **171**, 365–376.

Gao, G. and Hoeschele, I. (2006). A rapid conditional enumeration method for haplotyping in pedigrees. *Genetics, Selection, Evolution* (submitted).

Gao, G., Hoeschele, I., Sorensen, P. and Du, F.-X. (2004). Conditional probability methods for haplotyping in pedigrees. *Genetics* **167**, 2055–2065.

Garcia-Cortes, L.A., Moreno, C., Varona, L. and Altarriba, J. (1995). Estimation of prediction-error variances by resampling. *Journal of Animal Breeding and Genetics* **112**, 176–182.

Garcia-Cortes, L.A. and Sorensen, D. (1996). On a multivariate implementation of the Gibbs sampler. *Genetics, Selection, Evolution* **28**, 121–126.

Garner, C. and Slatkin, M. (2002). Likelihood-based disequilibrium mapping for two-marker haplotype data. *Theoretical Population Biology* **61**, 153–161.

Genin, E., Martinez, M. and Clerget-Darpoux, F. (1995). Posterior probability of linkage and maximal lod score. *Annals of Human Genetics* **59**, 123–132.

George, V.T. and Elston, R.C. (1988). Generalized modulus power transformations. *Communications in Statistics: Theory and Methods* **17**, 2933–2952.

George, E.I. and Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747.

George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

George, E.I. and McCulloch, R.E. (1996). Stochastic search variable selection. In *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson and D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 203–214.

George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

George, A.W., Visscher, P.M. and Haley, C.S. (2000). Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics* **156**, 2081–2092.

Gessler, D.D.G. and Xu, S. (1996). Using the expectation or the distribution of identical- by-descent for mapping quantitative trait loci under the random model. *American Journal of Human Genetics* **59**, 1382–1390.

Geyer, C.J. and Thompson, E.A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90**, 909–920.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (eds) (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Godsill, S.J. (2001). On the relationship between MCMC model uncertainty methods. *Journal of Computational and Graphical Statistics* **10**, 230–248.

Godsill, S.J. (2003). Proposal densities and product space methods. In *Highly Structured Stochastic Systems*, P.J. Green, N.L. Hjort and S. Richardson, eds. Oxford University Press, London, New York, Oxford, pp. 199–203.

Goldgar, G.E. (1990). Multipoint analysis of human quantitative genetic variation. *American Journal of Human Genetics* **47**, 957–967.

Graser, H.-U., Smith, S. and Tier, B. (1987). A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. *Journal of Animal Science* **64**, 1362–1370.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Green, P.J. (2003). Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems*, P.J. Green, N.L. Hjort and S. Richardson, eds. Oxford University Press, London, New York, Oxford, pp. 179–198.

Grignola, F. and Hoeschele, I. (1997). Mapping linked quantitative trait loci via residual maximum likelihood. *Genetics, Selection, Evolution* **29**, 529–544.

Grignola, F.E., Hoeschele, I. and Tier, B. (1996). Residual maximum likelihood to map quantitative trait loci: methodology. *Genetics, Selection, Evolution* **28**, 479–490.

Hanlon, P. and Lorenz, A. (2005). A computational method to detect epistatic effects contributing to a quantitative trait. *Journal of Theoretical Biology* **235**, 350–364.

Heath, S. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics* **61**, 748–760.

Heath, S. (1999). The problems of large human pedigrees: the application of MCMC methods to an island pedigree. In *Invited Lecture at the Highly Structured Stochastic Systems Workshop on Bayesian and MCMC Methods in Gene Mapping*, Lammi, Finland, March 24–27. `http://www.rni.Helsinki.fi/biometry/seminars.html`.

Hoeschele, I. (1988). Statistical techniques for detection of major genes in animal breeding data. *Theoretical and Applied Genetics* **76**, 311–319.

Hoeschele, I. (1999). Mapping complex trait genes in a dairy cattle pedigree for milk production and health and in a human pedigree for cholesterol, using Bayesian and other methods. In *Invited Lecture at the Highly Structured Stochastic Systems Workshop on Bayesian and MCMC Methods in Gene Mapping*, Lammi, Finland, March 24–27. `http://www.rni.Helsinki.fi/biometry/seminars.html`.

Hoeschele, I. and Romano, E.O. (1993). On the use of marker information from granddaughter designs. *Journal of Animal Breeding and Genetics* **110**, 429–429.

Hoeschele, I. and VanRaden, P.M. (1993). Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. *Theoretical and Applied Genetics* **85**, 953–960.

Hoeschele, I., Uimari, P., Grignola, F.E., Zhang, Q. and Gage, K.M. (1997). Advances in statistical methods to map quantitative trait loci in outbred populations. *Genetics* **147**, 1445–1456.

Jansen, R.C., Johnson, D.L. and van Arendonk, J.A.M. (1998). A mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. *Genetics* **148**, 391–399.

Jansen, R.C. and Nap, J.P. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics* **17**, 388–391.

Janss, L.L.G., van Arendonk, J.A.M. and van der Werf, J.H.J. (1995a). Computing approximate monogenic model likelihoods in large pedigrees with loops. *Genetics, Selection, Evolution* **27**, 567–579.

Janss, L.L.G., Thompson, R. and van Arendonk, J.A.M. (1995b). Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. *Theoretical and Applied Genetics* **91**, 1137–1147.

Jensen, C.S. (1997). Blocking Gibbs sampling for inference in large and complex Bayesian networks with applications in Genetics. Ph.D. thesis, Department of Computer Science, Aalborg University, Denmark.

Jensen, C.S. and Kong, A. (1996). Blocking Gibbs sampling for linkage analysis in large pedigrees with loops. Technical Report R-96-2048, Aalborg University, Department of Computer Science, Denmark.

Kerr, R.J. and Kinghorn, B.P. (1996). An efficient algorithm for segregation analysis in large populations. *Journal of Animal Breeding and Genetics* **113**, 457–469.

Knott, S.A., Elsen, J.M. and Haley, C.S. (1994). *Proceedings of the Fifth World Congression Genetics Applied to Livestock Production*, Vol. 21, Department of Animal and Poultry Science, University of Guelph, Guelph.

Knott, S.A. and Haley, C.S. (1992). Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics* **132**, 1211–1222.

Knott, S.A., Haley, C.S. and Thompson, R. (1991). Methods of segregation analysis for animal breeding data: a comparison of power. *Heredity* **68**, 299–311.

Kong, A. (1991). Efficient methods for computing linkage likelihoods of recessive diseases in inbred pedigrees. *Genetic Epidemiology* **8**, 81–103.

Kruglyak, L. and Lander, E.S. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

Kulp, D.C. and Jagalur, M. (2006). Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics* **7**, 125.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhya. Series B* **60**, 65–81.

Lam, J.C., Roeder, K. and Devlin, B. (2000). Haplotype fine-mapping by evolutionary trees. *American Journal of Human Genetics* **66**, 659–673.

Lander, E.S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 2363–2367.

Lange, K.L. (1997). *Mathematical and Statistical Methods for Genetic Analysis*. Springer-Verlag, New York.

Lange, K.L. and Goradia, T.M. (1987). An algorithm for automatic genotype elimination. *American Journal of Human Genetics* **40**, 250–256.

Lange, K.L., Kittle, R.J.A. and Taylor, J.M.G. (1989). Robust statistical modeling using the *t*-distribution. *Journal of the American Statistical Association* **84**, 881–896.

Le Roy, P., Elsen, J.M. and Knott, S.A. (1989). Comparison of four statistical methods for detection of a major gene in a progeny test design. *Genetics, Selection, Evolution* **21**, 341–357.

Lin, S. (1995). A scheme for constructing an irreducible Markov chain for pedigree data. *Biometrics* **51**, 318.

Lin, S. (1996). Multipoint linkage analysis via Metropolis jumping kernels. *Biometrics* **52**, 1417.

Lin, S. and Speed, T.P. (1997). An algorithm for haplotype analysis. *Journal of Computational Biology* **4**, 535–546.

Lin, S., Thompson, E.A. and Wijsman, E. (1993). Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data. *IMA Journal of Mathematics Applied in Medicine and Biology* **10**, 1.

Lin, S., Thompson, E.A. and Wijsman, E. (1994). Finding non- communicating sets for Markov chain Monte Carlo estimates on pedigrees. *American Journal of Human Genetics* **54**, 695.

Liu, B., De La Fuente, A. and Hoeschele, I. (2007). From genetics to gene networks: evaluating approaches for integrative analysis of genetic marker and gene expression data for the purpose of gene network inference. *BMC Genomics* (submitted).

Liu, J.S., Sabatti, C., Teng, J., Keats, B.J.B. and Risch, N. (2001). Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Research* **11**, 1716–1724.

Lou, X.-Y., Casella, G., Todhunter, R.J., Yang, M.C.K. and Wu, R. (2005). A general statistical framework for unifying interval and linkage disequilibrium mapping: toward high-resolution mapping of quantitative traits. *Journal of the American Statistical Association* **100**, 158–171.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edition. Chapman & Hall, New York.

McPeek, M.S. and Strahs, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *American Journal of Human Genetics* **65**, 858–875.

Meuwissen, T.H.E. and Goddard, M.E. (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**, 421–430.

Meuwissen, T.H.E. and Goddard, M.E. (2001). Prediction of identify by descent probabilities from marker haplotypes. *Genetics, Selection, Evolution* **33**, 605–634.

Meuwissen, T.H.E. and Goddard, M.E. (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics, Selection, Evolution* **36**, 261–279.

Meuwissen, T.H.E., Karlsen, A., Lien, S., Olsaker, I. and Goddard, M.E. (2002). Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161**, 373–379.

Meyer, K. and Smith, S.P. (1996). Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. *Genetics, Selection, Evolution* **28**, 23–49.

Morris, A.P., Whittaker, J.C. and Balding, D.J. (2000). Bayesian fine-scale mapping of disease loci, by hidden Markov models. *American Journal of Human Genetics* **67**, 155–169.

Morris, A.P., Whittaker, J.C. and Balding, D.J. (2002). Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *American Journal of Human Genetics* **70**, 686–707.

Morton, N.E. (1955). Sequential tests for the detection of linkage. *Annals of Human Genetics* **7**, 277–317.

Morton, N.E. (1998). Significance levels in complex inheritance. *American Journal of Human Genetics* **62**, 690–697.

Morton, N.E. and MacLean, C.J. (1974). Analysis of family resemblance III: complex segregation analysis of quantitative traits. *American Journal of Human Genetics* **26**, 489–503.

Narita, A. and Sasaki, Y. (2004). Detection of multiple QTL with epistatic effects under a mixed inheritance model in an outbred population. *Genetics, Selection, Evolution* **36**, 415–433.

Neumaier, A. and Groeneveld, E. (1999). Restricted maximum likelihood estimation of covariances in sparse linear models. *Genetics, Selection, Evolution* **30**, 3–26.

O'Connell, J.R. (2000). Zero-recombinant haplotyping: applications to fine mapping using SNPs. *Genetic Epidemiology* **19**(Suppl. 1), S64–S70.

Ott, J. (1989). Computer-simulation methods in human linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 4175–4178.

Patterson, H.D. and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 545.

Perez-Enciso, M. (2003). Fine-mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. *Genetics* **163**, 1497–1510.

Pong-Wong, R., Haley, C.S. and Woolliams, J.A. (1999). Behaviour of the additive finite locus model. *Genetics, Selection, Evolution* **31**, 193–211.

Qian, D. and Beckmann, L. (2002). Minimum-recombinant haplotyping in pedigrees. *American Journal of Human Genetics* **70**, 1434–1445.

Rannala, B. and Reeve, J.P. (2001). High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *American Journal of Human Genetics* **69**, 159–178.

von Rohr, P. and Hoeschele, I. (2002). Robust Bayesian QTL analysis using skewed student-*t* distributions. *Genetics, Selection, Evolution* **34**, 1–22.

Satagopan, J.M. and Yandell, B.S. (1996). Estimating the number of quantitative trait loci via Bayesian model determination In *Contributed Paper, Session on Genetic Analysis of Quantitative Traits and Complex Diseases, Joint Statistical Meetings*, Chicago, IL.

Satagopan, J.M., Yandell, B.S., Newton, M.A. and Osborn, T.C. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144**, 805–816.

Schork, N.Y. (1993). Extended multi-point identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. *American Journal of Human Genetics* **53**, 1306.

Scott, W.D. (1992). *Multivariate Density Estimation*. Wiley, New York.

Service, S.K., Temple, D.W., Freimer, N.B. and Sandkuijl, L.A. (1999). Linkage disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *American Journal of Human Genetics* **64**, 1728–1738.

Sham, P. (1998). *Statistics in Human Genetics*. Wiley, New York.

Sheehan, N. and Thomas, A. (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics* **49**, 163–175.

Sillanpää, M.J. and Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**, 1373–1388.

Sillanpää, M.J. and Arjas, E. (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **151**, 1605–1619.

Sillanpaa, M.J. and Bhattacharjee, M. (2005). Bayesian association-based fine mapping in small chromosomal segments. *Genetics* **169**, 427–439.

Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics* **58**, 1323–1337.

Sobel, E., Lange, K., O'Connell, J.R. and Weeks, D.E. (1996). Haplotyping algorithm. In *IMA Volumes in Mathematics and Its Applications*, *Vol. 81: Genetic Mapping and DNA Sequencing*, T.P. Speed and M.S. Waterman, eds. Springer-Verlag, New York, pp. 89–110.

Soller, M. and Andersson, L. (1998). Genomic approaches to the improvement of disease resistance in farm animals. *Revue Science et Technique* **17**, 329–345.

Southey, B.R. and Fernando, R.L. (1998). In *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, *Vol. 26. Animal Genetics and Breeding Unit*. University of New England, Armidale, pp. 221–224.

Spelman, R.J., Coppieters, W., Karim, L., van Arendonk, J.A.M. and Bovenhuis, H. (1996). Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. *Genetics* **144**, 1799–1808.

Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**, 506–516.

Stephens, D.A. and Fisch, R.D. (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* **54**, 1334–1347.

Stephens, D.A. and Smith, A.F.M. (1993). Bayesian inference in multipoint gene mapping. *Annals of Human Genetics* **57**, 65–82.

Stricker, C. and Fernando, R.L. (1998). In *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, *Vol. 26. Animal Genetics and Breeding Unit*. University of New England, Armidale, pp. 25–32.

Stricker, C., Schelling, M., Du, F.-X., Hoeschele, I., Fernández, S.A. and Fernando, R.L. (2002). A comparison of efficient genotype samplers for complex pedigrees and multiple linked loci In *Proceedings of the 7th World Congress on Genetics Applied to Livestock Production* **32**, 637–644.

Stuart, A. and Ord, J.K. (1987). *Distribution Theory*, *Kendal's Advanced Theory of Statistics, Vol. 1*, 5th edition. Oxford University Press, New York.

Tanner, M.A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd edition. Springer-Verlag, Berlin.

Tapadar, P., Ghosh, S. and Majumder, P.P. (2000). Haplotyping in pedigrees via a genetic algorithm. *Human Heredity* **50**, 43–56.

Terwilliger, J.D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *American Journal of Human Genetics* **56**, 777–787.

Thaller, G. and Hoeschele, I. (2000). Fine-mapping of quantitative trait loci in half-sib families using current recombinations. *Genetical Research* **76**, 87–104.

Thallmann, R.M., Bennett, G.L., Keele, J.W. and Kappes, S.M. (2001a). Efficient computation of genotype probabilities or loci with many alleles: I. Allelic peeling. *Journal of Animal Science* **79**, 26–33.

Thallmann, R.M., Bennett, G.L., Keele, J.W. and Kappes, S.M. (2001b). Efficient computation of genotype probabilities or loci with many alleles: II. Iterative method for large complex pedigrees. *Journal of Animal Science* **79**, 34–44.

Thoday, J.M. (1961). Location of polygenes. *Nature* **191**, 368–370.

Thomas, D.C. and Cortessis, V. (1992). A Gibbs sampling approach to linkage analysis. *Human Heredity* **42**, 63–76.

Thomas, A., Gutin, A., Abkevich, V. and Bansal, A. (2000). Multilocus linkage analysis by blocked Gibbs sampling. *Statistics and Computing* **10**, 259–269.

Thomas, D.C., Richardson, S., Gauderman, J. and Pitkaniemi, J. (1997). A Bayesian approach to multi-point mapping in nuclear families. *Genetic Epidemiology* **14**, 903–908.

Thompson, E.A. (1994). Monte Carlo likelihood in genetic mapping. *Statistical Science* **9**, 355–366.

Thompson, E.A. and Heath, S.C. (1999). Estimation of conditional multilocus gene identity among relatives. In *Statistics in Molecular Biology and Genetics*, *IMS Lecture Notes*, F. Sellier-Moiseiwitsch, ed.Institute of Mathematical Statistics, American Mathematical Society, Providence, RI, pp. 95–113.

Thompson, E.A. and Skolnick, M.H. (1977). In *Proceedings of the International Conference on Quantitative Genetics*, E. Pollack, O. Kempthorne and T.B. Bailey Jr., eds. Iowa State University Press, Ames, IA, pp. 815–818.

Tier, B. and Henshall, J. (2001). A sampling algorithm for segregation analysis. *Genetics, Selection, Evolution* **33**, 587–603.

Uimari, P. and Hoeschele, I. (1997). Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics* **146**, 735–743.

Uimari, P. and Sillanpää, M.J. (2001). Bayesian oligogenic analysis of quantitative and qualitative traits in general pedigrees. *Genetic Epidemiology* **21**, 224–242.

Uimari, P., Zhang, Q., Grignola, F.E., Hoeschele, I. and Thaller, G. (1996). Analysis of QTL workshop I granddaughter design data using least-squares, residual maximum likelihood, and Bayesian methods. *Journal of Quantitative Trait Loci* **7**(2).

Vieland, V.J. (1998). Bayesian linkage analysis, or: how I learned to stop worrying and love the posterior probability of linkage. *American Journal of Human Genetics* **63**, 947–954.

Wang, T., Fernando, R.L., van der Beek, S. and van Arendonk, J.A.M. (1995). Covariance between relatives for a marked quantitative trait locus. *Genetics, Selection, Evolution* **27**, 251–275.

Wang, T., Fernando, R.L., Stricker, C. and Elston, R.C. (1996). An approximation to the likelihood for a pedigree with loops. *Theoretical and Applied Genetics* **93**, 1299–1309.

Wang, C.S., Rutledge, J.J. and Gianola, D. (1993). Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics, Selection, Evolution* **25**, 41–62.

Weir, B.S. (1996). *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Weller, J.I. (1986). Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**, 627–640.

Weller, J.I., Song, J.Z., Heyen, D.W., Lewin, H.A. and Ron, M. (1998). A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150**, 1699–1706.

Wijsman, E. (1987). A deductive method of haplotype analysis in pedigrees. *American Journal of Human Genetics* **41**, 356–373.

Williams, R.W. (2006). Expression genetics and the phenotype revolution. *Mammalian Genome* **17**, 496–502.

Wu, P. and Zeng, Z.-B. (2001). Joint linkage and linkage disequilibrium mapping in natural populations. *Genetics* **157**, 899–909.

Xiong, M. and Guo, S.-W. (1997). Fine-scale mapping based on linkage disequilibrium: theory and applications. *American Journal of Human Genetics* **60**, 1513–1531.

Xu, S. (1995). A comment on the simple regression method for interval mapping. *Genetics* **141**, 1657–1659.

Xu, S. (1996). Computation of the full likelihood function for estimating variance at a quantitative trait locus. *Genetics* **144**, 1951–1960.

Xu, S. (1998a). Further investigation on the regression method of mapping quantitative trait loci. *Heredity* **80**, 364–373.

Xu, S. (1998b). Mapping quantitative trait loci using multiple families of line crosses. *Genetics* **148**, 517–524.

Xu, S. and Atchley, W.R. (1995). A random model approach to interval mapping of quantitative trait loci. *Genetics* **141**, 1189–1197.

Yi, N. (2004). A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics* **167**, 967–975.

Yi, N. and Xu, S. (2001). Bayesian mapping of quantitative trait loci under complicated mating designs. *Genetics* **157**, 1759–1771.

Yi, N., Yandell, B.S., Churchill, G.A., Allison, D.B., Eisen, E.J. and Pomp, D. (2005). Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics* **170**, 1333–1344.

Zhang, Q., Boichard, D., Hoeschele, I., Ernst, C., Eggen, A., Murkve, B., Pfister-Genskow, M., Witte, L.A., Grignola, F.E., Uimari, P., Thaller, G. and Bishop, M.D. (1998). Mapping quantitative trait loci for milk production and health of dairy cattle in a large outbred pedigree. *Genetics* **149**, 1959–1973.

Zhao, L.P., Aragaki, C., Hsu, L. and Quaioit, F. (1998). Mapping of complex traits by single nucleotide polymorphisms. *American Journal of Human Genetics* **63**, 225–240.

Inferences from Mixed Models in Quantitative Genetics

D. Gianola

Department of Animal Sciences, Department of Biostatistics and Medical Informatics, and Department of Dairy Science, University of Wisconsin, Madison, WI USA

Statistical methods that have been applied for inferring genetic values in animal breeding are reviewed. Landmarks include statistical genetic models; best linear unbiased prediction; Henderson's mixed model equations and associated computing techniques; variance and covariance component estimation, with emphasis on likelihood-based methods; Bayesian procedures; methods for categorical responses, longitudinal data and survival analysis. The problem caused by the effects of selection on inferences and semiparametric procedures for massive genomic data are discussed. An inventory of some available computing software is presented. Finally, some areas for future development are discussed.

20.1 INTRODUCTION

Genetic selection programs of livestock or plants aim to maximize the rate of increase of some merit function (e.g. economic value of a cultivar) expected to have a genetic basis. Typically, animals with the highest merit are kept as parents of the subsequent generation and those with the lowest merit are culled. Merit can be a linear or nonlinear function of genetic values for several traits considered to be important from the perspective of producing economic returns or benefits to mankind. The genetic component of merit cannot be observed, so it must be inferred from data on the candidates for selection or on their relatives. Here, there are at least three types of statistical problems: (1) assessing whether the traits intervening in the merit function have a genetic basis; (2) developing reasonably accurate methods for inferring merit ('genetic evaluation'), and (3) deciding what to do with the animals having the 'best' such evaluations, e.g. designing mating plans that are optimal in some sense. Problem 1 is usually known as '*estimation of genetic parameters*'. The second problem is referred to as *estimation* (*prediction*) *of breeding*

values, and is conceptually inextricable from the first; loosely speaking, breeding value is a function of the deviation of the progeny of an individual from the population mean. The third problem will not be dealt with here. What follows is motivated primarily by problems in animal breeding, but a great portion of the theory is applicable to plant improvement as well.

The information available for assessing the genetic basis of traits and for inferring merit consists of records of performance. Examples are records on growth rate, milk yield, and composition; records on diseases such as mastitis in dairy cattle; egg production in layers, litter size in pigs, calving difficulty in cattle, and survival or length of productive life in dairy cows. Some traits are recorded on a 'continuous' scale (e.g. milk yield) and some are discrete, such as counts (litter size) or categorical assignments. Hence, adequate probability modeling must often go beyond a normal distribution, although this assumption may be convenient and even useful. In the last decade, data on molecular markers has become increasingly available, but its use in marker-assisted genetic improvement is still at early stages. Furthermore, the analytical paradigm is changing, in the light of the explosive amount of postgenomic data.

Many observable traits (continuous or discrete) seem to have a polygenic mode of inheritance, and are subject to strong environmental influences. Also, there are sex-limited traits, such as milk production (observed in females only) and scrotal circumference in bulls, which is believed to have a positive genetic correlation with fertility in cows. In dairy cattle, it is more relevant to infer the genetic merit of males, because of the impact these can have on the rate of improvement. For example, as a result of artificial insemination and of widespread availability of frozen semen and embryos, some dairy bulls produce thousands of daughters in several countries, creating, thus, an opportunity for international sire evaluation, albeit at the expense of more complicated statistical modeling and implementation (Schaeffer, 1985).

Animal breeding data sets can be large (e.g. millions of lactation records in dairy cattle breeding), multivariate (several traits may need to be modeled simultaneously), seemingly Gaussian in some instances (e.g. logarithm of concentration of somatic cells in milk, an indication of udder disease), or decidedly nonnormal in others, such as with the discrete traits mentioned above. Data structure can be cross-sectional or longitudinal (e.g. growth curves in broilers), extremely unbalanced, and perhaps exhibiting a pattern of nonrandom missingness. For example, not all first lactation cows produce a second lactation, due to sequential selection for higher production, reproductive failure, or disease. Also, some sires are used more intensively than others, because of perceived differences in genetic value, so there is genetic selection as a consequence of variation in their contribution to offspring born in the following generation.

Given the preceding, it is not surprising that statistics has been so important in animal breeding. Examples of some of the more relevant statistical methods and problems discussed in animal breeding in the last 35 years or so are in Hill (1974; 1980); Henderson (1977); Thompson (1977; 1979; 1982); Dempfle (1982); Gianola *et al.* (1986); Schaeffer and Kennedy (1986); Meyer (1990); Ducrocq (1990); Gianola and Hammond (1990); Sorensen *et al.* (1994); Foulley and Quaas (1994); Bidanel (1998); Tempelman and Firat (1998); Wang (1998); Sorensen and Gianola (2002) and Gianola (2006).

Our objectives are to describe some of the statistical methods for inferring breeding values that have been applied in animal breeding. Some specific historical landmarks are

described in Section 20.2. Section 20.3 discusses several specific problems, and offers conjectures about possible additional developments.

20.2 LANDMARKS

20.2.1 Statistical Genetic Models

The models for quantitative genetic analysis employed in animal breeding consist of the following components: (1) a mathematical function relating observations to location parameters and 'random' effects (Bayesians view all unknowns as being random, in the sense of possessing a subjective uncertainty distribution). The 'random' effects may include genetic components, such as additive genetic values (Falconer and Mackay, 1996), dominance and epistatic deviations (from additivity) and permanent environmental effects. All these contribute to correlations between relatives or between longitudinal records of performance. (2) Genetic and environmental dispersion parameters, such as components of variance and covariance (the latter appear in multivariate models, or in situations where a multivariate structure must be embedded in a model for a single response variate), and (3) assumptions about the form of the joint distribution of the observations and of the random effects (in a Bayesian context, the assumptions apply to the joint distribution of all unknowns and of the data).

Most often, functional forms employed in (1) above have been linear. Although convenient, this is not always a sensible specification. Concerning (3), the most widely used and abused assumption has been that of multivariate normality. This is because it is often postulated that traits are inherited in a multifactorial manner, i.e. that there is a large number of genes acting additively, and that the effects of gene substitutions are infinitesimally small. Molecular information is beginning to suggest that the assumption of many genes acting together on quantitative traits may not be unreasonable, at least in some instances. For example, a study in dairy cattle (Zhang *et al.*, 1998) using genetic markers suggested the presence of 'quantitative trait loci (QTL)' affecting fat percentage in milk in chromosomes 2, 6, 14, 26, 28, and this type of investigation is still at early stages. If alleles act additively and have small effects, their sum produces a normal process rapidly. In this context, it is not always clear how much it is gained using such marker information modelwise, at least from a statistical point of view. Some statistical models for marker-assisted genetic evaluation require knowledge of map distances. This introduces more parameters in the model, and additional sources of uncertainty.

Fisher (1918) gave the foundations of the infinitesimal model, and his seminal paper discussed the implications of Mendelian inheritance at the phenotypic level. He posited:

$$observation = genetic\ value + residual,$$

and gave a precursor of the analysis of variance, by proposing a partition of the genetic variance into additive and dominance components. From these, the expected correlations between different types of relatives follow more or less readily. The additive model, in particular, has been extremely useful and has stood the test of time fairly well. Further, it constituted the statistical genetic point of departure for the development of predictors of breeding value, eventually leading to fairly precise evaluation of dairy sires.

It continues being used, albeit in a more sophisticated, vectorial, manner; this will be discussed subsequently.

Independently of Fisher, Wright (1921), arrived at similar results (at least for the additive part) using a method known as *path coefficients*. The method consists of describing a system of correlations with standardized random effects linear models, where each variate has null mean and unit variance. This procedure, although powerful in the hands of Wright, faded away in animal breeding, except at the classroom level. This is because 'path diagrams' convey visually relations of causality or of covariation assumed in a model. Reasons for the 'fall' of path coefficients include lack of generality, i.e. inability to take into account interactions and nuisance parameters, and unattractiveness from the point of view of computer implementation. However, the method has been revived and expanded in the social sciences, in the context of what are called *structural equation* models (Fox, 1984).

Although Fisher had described how some interactions between alleles at different loci could be taken into account, it was not until Cockerham (1954) and Kempthorne (1954) that the total variance from such interactions could be partitioned into what are called *epistatic* components, assuming a large panmictic population in linkage equilibrium. Kempthorne (1954) used the concept of probability of identity by descent developed by Malécot (1948), and disentangled the epistatic variance into several components, depending on the number of loci involved in the expression of the trait. For example, with two loci, the epistatic genetic variance can be expressed as the sum of 'additive x additive', 'additive x dominance', and 'dominance x dominance' components of variance His development allowed expressing the covariance between traits measured in relatives in a random mating population in terms of genetic components of variance and covariance. Results were used subsequently by Henderson (1985; 1988) for inferring dominance and epistatic genetic effects via best linear unbiased prediction (BLUP), a topic to be discussed later.

Additional extensions of statistical genetic models have accommodated, e.g. maternal effects (Falconer, 1965; Willham, 1963; Koerkhuis and Thompson, 1997), cloning and cytoplasmic inheritance (Kennedy and Schaeffer, 1990). Statistically, the model of Willham is interesting because it embeds a multivariate within an univariate structure. This is because it includes covariances between direct genetic effects (expressed in the individual measured) and maternal genetic effects (expressed if the individual becomes a mother having a measured offspring).

20.2.2 Best Linear Unbiased Prediction (BLUP)

20.2.2.1 General

The problem of 'predicting' or 'estimating' genetic merit in candidates for selection is very important in animal breeding. This semantic distinction has created considerable confusion because, statistically, it is nonsensical to speak about 'estimation' of random effects. On the other hand 'prediction' has a futuristic connotation, whereas in animal breeding one is often interested in ranking candidates (e.g. bulls) that already exist. Further, it is more sensible to think in terms of 'inferring' genetic merit, because the problem goes beyond one of merely producing an estimate of location. Often, one may be interested in obtaining a measure of uncertainty or, perhaps, in computing the probability of ordered events involving candidates for selection.

Lush (1931), using path coefficients, gave formulae for assessing the genetic merit of dairy sires, assuming that means and genetic and environmental components of variance were known. It was found that some regression to the mean, i.e. shrinkage, was needed. Robertson (1955) showed that Lush's statistic can be obtained from a weighted average between 'population' information and data, anticipating a Bayesian interpretation. To illustrate, consider the simple model:

$$y_{ij} = \mu + s_i + e_{ij}, \tag{20.1}$$

where y_{ij} is an observation taken on progeny j of sire i, μ is a constant common to all observations, $s_i \sim N(0, v_s)$ is the transmitting ability (half of the additive genetic value) of sire i, and $e_{ij} \sim N(0, v_e)$ is an independently distributed residual peculiar to individual ij. Suppose that μ and the variance components v_s (variance between sires) and v_e (variance within sires) are known 'population' parameters. Here, there are two pieces of information about sire i: (1) what is known about the population and (2) the average performance $\overline{y}_i$ of its n offspring. It would seem 'natural' combining these two pieces in a weighted average between the population mean, 0, and the progeny group mean deviation $\overline{y}_i - \mu$, with the weights being the 'population precision' $\left(v_s^{-1}\right)$ and the precision brought by the data $\left(\frac{n}{v_e}\right)$, respectively. That is:

$$
\begin{aligned}
\widehat{s}_i &= \left(\frac{1}{v_s} + \frac{n}{v_e}\right)^{-1} \left[\frac{1}{v_s}.0 + \frac{n}{v_e}\left(\overline{y}_i - \mu\right)\right] \\
&= \frac{n}{n + \frac{v_e}{v_s}}\left(\overline{y}_i - \mu\right). \tag{20.2}
\end{aligned}
$$

Likewise, a measure of variance is the reciprocal of the total precision:

$$\widehat{v}_i = \left(\frac{1}{v_s} + \frac{n}{v_e}\right)^{-1} = v_s\left(1 - \frac{v_s}{v_s + \frac{v_e}{n}}\right). \tag{20.3}$$

It can be shown that $\widehat{s}_i$ and $\widehat{v}_i$ are the mean and variance of the conditional distribution $\left[s_i | \mu, v_s, v_e, y_{i1}, y_{i2}, \ldots, y_{in}\right]$, under normality assumptions. This is also a posterior distribution in a Bayesian setting in which all parameters are known without error, this being an unrealistic assumption in practice. At any rate, probabilistic inference about the transmitting ability of the sire is completed using the conditional distribution with mean $\widehat{s}_i$ and variance $\widehat{v}_i$. For example, if one has two genetically unrelated sires with data represented by vectors $\mathbf{y}_1$ and $\mathbf{y}_2$, one may wish to compute the following probability:

$$\Pr\left(s_1 > s_2 | \mathbf{y}_1, \mathbf{y}_2, \mu, v_s, v_e\right) = \Pr\left(z > 0\right),$$

where z is a normally distributed random variable with mean $\widehat{s}_1 - \widehat{s}_2$ and variance:

$$v_z = v_s\left[2 - v_s\left(\frac{1}{v_s + \frac{v_e}{n_1}} + \frac{1}{v_s + \frac{v_e}{n_2}}\right)\right].$$

Note that as $n_i \to \infty$, $v_z \to 0$, illustrating how uncertainty dissipates asymptotically.

20.2.2.2   *The Mixed Model Equations*

It was not until Henderson (e.g. 1950, 1963, 1973, 1984) that the problem of 'prediction' of breeding value could be formulated in a more general framework, by deriving what later was known to be BLUP. Henderson *et al.* (1959), posed the linear (univariate or multivariate) mixed effects model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{20.4}$$

where β is a fixed (over repeated sampling) vector, and $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$ are uncorrelated random vectors; $\mathbf{X}$ and $\mathbf{Z}$ are known incidence matrices, and $\mathbf{G}$ and $\mathbf{R}$ are variance–covariance matrices, which are function of (known) dispersion parameters. The vector of random effects $\mathbf{u}$ can include herd effects, breeding values, permanent environmental deviations common to all records of the same (or of a set) of animals, etc. The joint density of $\mathbf{u}$ and $\mathbf{y}$ is:

$$p(\mathbf{u}, \mathbf{y}|\beta, \mathbf{G}, \mathbf{R}) \propto p(\mathbf{y}|\mathbf{u}, \beta, \mathbf{R}) \cdot p(\mathbf{u}|\mathbf{G})$$

$$\propto \exp\left\{-\frac{1}{2}\left[(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{u})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{u}) + \mathbf{u}'\mathbf{G}^{-1}\mathbf{u}\right]\right\}. \tag{20.5}$$

Maximization of (20.5) with respect to β and $\mathbf{u}$ simultaneously leads to Henderson's mixed model equations (MME):

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \tag{20.6}$$

Henderson thought he was maximizing a likelihood function, so he termed $\widehat{\beta}$ and $\widehat{\mathbf{u}}$ as 'maximum likelihood (ML)' estimators of β and of $\mathbf{u}$, respectively. It turns out that $\widehat{\beta}$ is indeed the ML estimator of β under normality assumptions but, technically, $\mathbf{u}$ cannot be 'estimated', as this vector is random. Today, it is known that the objective function maximized by Henderson is a joint posterior density, in a certain Bayesian setting, or a 'penalized' or 'extended' likelihood in some *ad hoc* sense. However, this 'error' had a happy ending, as Henderson and Searle proved in several subsequent papers that even without the normality assumption, $\widehat{\beta}$ is the generalized least-squares estimator of β and $\widehat{\mathbf{u}}$ is the best linear unbiased predictor of $\mathbf{u}$; the inverse of the coefficient matrix in (20.6) yields the covariance matrices of $\widehat{\beta}$ and of $\widehat{\mathbf{u}} - \mathbf{u}$. This holds both for multivariate and univariate settings. Golberger (1992) derived BLUP independently of Henderson.

When β is known, $\widehat{\mathbf{u}}$ is the best linear predictor and, under normality, $\widehat{\mathbf{u}}$ is the minimum mean squared error predictor, i.e. the best, in the mean squared error sense (Henderson, 1973; Searle, 1974). In a multivariate context, with known β, the predictor gives the 'selection index' evaluation derived by Smith (1936) and Hazel (1943) in less general settings (Henderson, 1963).

Bulmer (1980) pointed out, pertinently, that was unclear whether ranking animals with BLUP would maximize expected genetic progress in a single round of selection, and suggested an alternative predictor. The latter was shown to be equal to BLUP by Gianola and Goffinet (1982); Fernando and Gianola (1986) give a discussion of some of the issues.

Animal breeders often misinterpret the unbiasedness property of BLUP. It cannot be overemphasized that BLUP is unbiased over conceptual repeated sampling from the

distribution $\left[\mathbf{y}, \mathbf{u}|\beta, \mathbf{G}\right]$, but not over $\left[\mathbf{y}|\mathbf{u}, \beta, \mathbf{G}\right]$. The latter is the distribution that 'practitioners' tend to have in mind, i.e. one where $\mathbf{u}$ is a vector of *realized* breeding values. It is easy to show that:

$$E\left(\widehat{\mathbf{u}}|\mathbf{u}\right) = \left(\mathbf{Z}'\mathbf{PZ} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\mathbf{PZu},$$

where $\mathbf{P} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ is the usual projection matrix. This indicates that BLUP gives biased predictions of *specific* breeding values, although the bias vanishes asymptotically (as per-animal information increases). Paradoxically, in the limit, it is no longer possible to predict breeding values uniquely as a degeneracy in rank occurs.

20.2.2.3   *Solving the Mixed Model Equations*

The MME algorithm for calculating BLUE and BLUP has been employed worldwide for genetic evaluation of livestock. The order of the linear system in (20.6) can be in the million of equations, specially for models, univariate or multivariate, where a random additive genetic effect is fitted for each animal with a record of production as well as for animals without records in the genealogy, but that need to be included to account properly for genetic covariances between relatives. Hence, iterative methods must be used for solving the MME (e.g. Schaeffer and Kennedy, 1986; Misztal and Gianola, 1987), although approximations are needed to assess the uncertainty of the predictions. The MME appeared in the statistical theory literature late and somewhat sparingly (e.g. Patterson and Thompson, 1971; Harville, 1977; Wolfinger, 1993; Lee and Nelder, 1996; Nelder, 1996). This is surprising because the MME can be used to advantage in connection with computing algorithms for several methods of variance component estimation in generalized mixed effects linear models (Harville and Mee, 1984; Gilmour *et al.*, 1985; Foulley *et al.*, 1987a).

An obvious difficulty, at least in animal breeding, was inverting $\mathbf{G}$ (unless the matrix has an exploitable pattern, such as block-diagonality) when the order of $\mathbf{u}$ was in the hundreds of thousands, or millions, as in the routine genetic evaluation of dairy cows in the United States. Again, Henderson (1976) made a remarkable breakthrough. Let $\mathbf{u}$ be a vector of breeding values, and suppose that $\mathbf{G} = \mathbf{G}_0 \otimes \mathbf{A}$, where $\mathbf{G}_0$ has order equal to the number of traits (a dozen, say) and $\mathbf{A}$ is a matrix of 'additive genetic relationships' (reflecting the probabilities that related individuals carry identical copies of the same allele). Henderson discovered that $\mathbf{A}^{-1}$ can be written directly from a list of parents of the animals. This enabled using all available relationships in genetic evaluation, which leads to more precise inferences about genetic values and, also, to the possibility of correcting biases due to ignoring many relationships in otherwise naive variance component analyses.

20.2.3   Variance and Covariance Component Estimation

From the preceding, it follows that 'prediction' of breeding values depends nontrivially on knowledge of variance and covariance components. Many methods have been developed, but only a few ones stood the test of time. Hofer (1998) gave a review of essentially all methods of estimation that have been applied so far. Because the data sets in animal breeding can be large, unbalanced, and the models have a sizable number of nuisance location parameters, simple, ANOVA-type methods, seldom work well. Henderson (1953),

in a classical paper, described three methods for unbalanced data. The most general, Method 3, uses quadratic forms based on least squares, and yields unbiased estimators. Harvey (1960; 1970) incorporated it in software for variance and covariance component estimation, and this was used amply in animal breeding. Searle (1968; 1971) and his students clarified Henderson's methods, and presented these in matrix form.

Subsequently, Rao's (1971) and LaMotte's (1973) minimum norm quadratic unbiased estimation and its minimum variance version (under normality), respectively, entered into the picture. These estimators can be formulated in terms of solutions to the MME. However, attaining optimality with these procedures requires knowledge of the true parameters and, also, the sampling distributions of the estimators admit negative values of the variance components with nonnull probability. In fact, their multivariate counterparts can lead to embarrassing estimates of covariance matrices. Hence, animal breeders became seduced by ML, assuming normality, where these problems do not occur, at least in point estimation. Here, papers by Hartley and Rao (1967) and Harville (1977) were influential. Many algorithms for ML estimation can be derived using the MME (Harville, 1977; Henderson, 1984; Harville and Callanan, 1990). It is unclear whether the move toward likelihood-based methods was either a consequence of the availability of something new that could be computed with the MME algorithm (here used iteratively, as ML estimators cannot be written explicitly for most models), or of the appeal of the asymptotic properties of the method.

The well-known bias of the ML estimator of the residual variance led to widespread interest in a method called *restricted* maximum likelihood (REML). The basic ideas are in Anderson and Bancroft (1952) and Thompson (1962). However, Patterson and Thompson (1971) gave a more general description, suitable for a mixed effects model. REML can be viewed as an attempt of accounting for the 'loss of degrees of freedom' incurred in estimating the fixed effects. Patterson and Thompson (1971) noted that maximization of the location invariant part of the likelihood leads to estimating equations that are similar to those in ANOVA, in the balanced setting. Their objective was to reduce bias, but it is pertinent to ask whether this occurs at the expense of precision. Patterson and Thompson (1971) and Harville (1977) argued that no information is lost by using such 'restricted' likelihood, whereas Foulley (1993) seemed unconvinced.

We explore this issue a bit further. Consider the sampling model:

$$\mathbf{y} \sim N\left[\mathbf{X}\beta, \mathbf{V}\left(\theta\right)\right], \tag{20.7}$$

where $\mathbf{V}\left(\theta\right)$ is a dispersion matrix, this being a function of variance and covariance components θ. The 'full' likelihood is:

$$l\left[\beta, \theta\right] \propto \left|\mathbf{V}\left(\theta\right)\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}SSE\right]$$
$$\propto \left|\mathbf{V}\left(\theta\right)\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\widehat{SSE}\right].\exp\left\{-\frac{1}{2}SSB\right\}, \tag{20.8}$$

where:

$$SSE = \left(\mathbf{y} - \mathbf{X}\beta\right)' \mathbf{V}^{-1}\left(\theta\right)\left(\mathbf{y} - \mathbf{X}\beta\right),$$
$$\widehat{SSE} = \left(\mathbf{y} - \mathbf{X}\widehat{\beta}\right)' \mathbf{V}^{-1}\left(\theta\right)\left(\mathbf{y} - \mathbf{X}\widehat{\beta}\right),$$

and

$$SSB = \left(\beta - \hat{\beta}\right)' \left[\mathbf{X}'\mathbf{V}^{-1}(\theta)\mathbf{X}\right] \left(\beta - \hat{\beta}\right),$$

for

$$\widehat{\beta} = \left[\mathbf{X}'\mathbf{V}^{-1}(\theta)\mathbf{X}\right]^{-1} \mathbf{X}'\mathbf{V}^{-1}(\theta)\mathbf{y}.$$

Harville (1974) showed that REML is the mode of the posterior distribution of the variance parameters after integrating the fixed effects (over an improper, uniform, prior) out of the joint posterior distribution, this being proportional to the likelihood function. The probability calculus takes uncertainty about fixed effects into account automatically, at least from a Bayesian perspective. The integrated ('restricted') likelihood is then proportional to:

$$l_I[\theta] \propto |\mathbf{V}(\theta)|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\widehat{SSE}\right] \int_{\Re_\beta} \exp\left\{-\frac{1}{2}SSB\right\} \mathbf{d}\beta$$

$$\propto |\mathbf{V}(\theta)|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\widehat{SSE}\right] \left|\mathbf{X}'\mathbf{V}^{-1}(\theta)\mathbf{X}\right|^{-\frac{1}{2}}. \tag{20.9}$$

This follows because the integral above involves a Gaussian kernel, so it can be expressed in closed form. Maximization of $l_I[\theta]$ with respect to θ gives the REML estimates of the dispersion parameters (Harville, 1974). Ignoring constants that do not depend on the parameters, then:

$$\log l[\beta, \theta] = \log l_I[\theta] - \frac{1}{2}SSB + \frac{1}{2}\log\left|\mathbf{X}'\mathbf{V}^{-1}(\theta)\mathbf{X}\right|^{\frac{1}{2}}. \tag{20.10}$$

The next step is quantifying 'information'. Suppose one adopts Fisher's information measure, which is a natural choice in likelihood inference. Recall that information is the expected value over sampling model (20.7) of minus the second derivatives of the log likelihood with respect to the parameters. To illustrate, consider a simple situation, amenable to analytical treatment. In a fixed effects or regression model with p identifiable location parameters and variance θ, the Fisherian information about θ is $\frac{N}{2\theta^2}$ in the full likelihood, whereas that from the restricted likelihood is $\frac{N-p}{2\theta^2}$. This is not surprising when the restricted likelihood is viewed from a Bayesian perspective, as less information is to be expected in a marginal than in a joint distribution. Lacking evidence to the contrary, similar 'losses of information' ought not be surprising in more complex models

In hierarchical or variance component models, both ML and REML are biased, so it would be unfair to focus the discussion on bias only. In general, whatever favors REML in this sense may be compensated by a loss of precision of this estimator. Simulations of Corbeil and Searle (1976) give ambiguous results. Searle *et al.* (1992) give a discussion of the relative merits of the two methods but, in our opinion, fail to include a strong logical argument favoring REML over ML: its Bayesian interpretation, which indicates clearly how uncertainty about fixed effects (acting as nuisance parameters in this case) is accounted for via integration, as shown earlier.

It must be pointed out that representations of the restricted likelihood alternative to (20.9) can be arrived at. Using the notation employed in connection with mixed effects

model (20.4):

$$l_I [\theta] \propto \int_{\Re_\beta} l [\beta, \theta] \, d\beta \propto \int_{\Re_\beta} p (\mathbf{y}|\beta, \theta) \, d\beta$$

$$\propto \int_{\Re_\beta} \left[\int_{\Re_u} p (\mathbf{y}|\beta, \mathbf{u}, \mathbf{R} (\theta)) \, p (\mathbf{u}|\mathbf{G} (\theta)) \, d\mathbf{u} \right] d\beta.$$

After algebra involving combining quadratic forms within exponential functions, followed by integrating a Gaussian kernel, one obtains:

$$l_I [\theta] \propto |\mathbf{R} (\boldsymbol{\theta})|^{-\frac{1}{2}} |\mathbf{G} (\theta)|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} SSR \right) |\mathbf{C} (\theta)|^{\frac{1}{2}}, \qquad (20.11)$$

where:

$$SSR = \mathbf{y}'\mathbf{R}^{-1} (\theta) \, \mathbf{y} - \widehat{\beta}'\mathbf{X}'\mathbf{R}^{-1} (\theta) \, \mathbf{y} - \widehat{\mathbf{u}}'\mathbf{Z}'\mathbf{R}^{-1} (\theta) \, \mathbf{y},$$

and:

$$\mathbf{C} (\theta)^{-1} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1} (\theta) \, \mathbf{X} & \mathbf{X}'\mathbf{R}^{-1} (\theta) \, \mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1} (\theta) \, \mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1} (\theta) \, \mathbf{Z} + \mathbf{G}^{-1} (\theta) \end{bmatrix}.$$

is the coefficient matrix of the MME. Alternative representations of the restricted likelihood lead to different algorithms for computing REML, each having its advantages and disadvantages.

20.2.4 BLUP and Unknown Dispersion Parameters

BLUP exists only when the dispersion parameters θ are known (at least to proportionality in variance component models), so an important question is what method of estimation of θ should be used when the end-point is predicting breeding values? Conceivably, a method deemed best (in some sense) for estimating variance components may not be the best to use in the context of inferring breeding values. Gianola and Fernando (1986) and Gianola *et al.* (1986) employed a Bayesian idea to answer this question. They argued that using BLUE and BLUP with the unknown (co)variance parameters evaluated at the REML estimates corresponds to an approximate integration of the dispersion parameters out of a certain posterior distribution. This leads to an approximate Bayesian solution to the problem of making inferences about breeding values when genetic and environmental variances are unknown. However, the restricted likelihood must be sharp enough for this to work well.

Harville and Carriquiry (1992) studied the problem in some detail. The approximation was found to be excellent in a data set consisting of milk yields of more than 0.5 million cows sired by about 1000 bulls, but less so in another data set on birth weights of 62 lambs, progeny of 23 sires. The latter data set is not representative of what animal breeders encounter with field records, but it is not atypical of experimental settings. Then, how should one carry finite sample inference about genetic values in the absence of sharp knowledge about dispersion parameters? Regrettably, this problem does not have an elegant frequentist solution, conceptually, and analytical treatment is impossible or, at best, algebraically awkward (e.g. Kackar and Harville, 1981).

The current 'state of the art' in animal breeding data analysis is precisely the REML + BLUP 'tandem'. Interestingly, REML has a likelihood justification, as noted,

but lacks a frequentist one. Conversely, BLUP has a frequentist lineage, but it does not arise from a likelihood formalism. Hence, the 'tandem' does not fall in any sharply defined school of 'classical' inference. Does this 'recombinant' correspond to the highest possible state of statistical fitness? The answer seems to be negative, as illustrated by the study of Harville and Carriquiry (1992). We elaborate further on this below.

20.2.5 Bayesian Procedures

Frequentist and likelihood-based approaches dominated the statistical views in animal breeding during most of the century. To a large extent, this is because many animal breeders had been trained in universities such as Cornell, Edinburgh, Iowa State, and North Carolina State, where the teaching of statistics focused on such approaches. What would have happened if animal breeding had been taught in Chicago (Zellner), London (Lindley), Lyon (Malécot) or Yale (Savage)? However, the field did not remain insensitive to the Bayesian revival taking place in the mid 60's. Papers such as Lindley and Smith (1972) provided a clear link between mixed models and hierarchical Bayesian approaches, and Box and Tiao (1973) presented the technical details. Seemingly, a seminar given at Cornell by Solomon (Henderson, personal communication), instigated Ronningen (1971) to investigate connections between BLUP and Bayesian ideas. This was pursued further by Dempfle (1977).

Consider first, the Bayesian view of BLUP. Suppose, as in a BLUP setting, that the dispersion matrices $\mathbf{G}$ and $\mathbf{R}$ associated with the mixed effects linear model (20.4) are known. In a Bayesian context (see Gianola and Fernando, 1986), if the prior distribution of β is taken to be uniform over $\Re^p$, with p being the order of β, and the prior distribution of the random effects is $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$, with the two location vectors being independent *a priori,* the joint posterior distribution can be shown to be:

$$\begin{bmatrix} \beta \\ \mathbf{u} \end{bmatrix} | \mathbf{y}, \mathbf{G}, \mathbf{R} \sim N\left(\begin{bmatrix} \widehat{\beta} \\ \widehat{\mathbf{u}} \end{bmatrix}, \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} \right). \tag{20.12}$$

Because the joint posterior distribution is Gaussian, so are the marginals or any induced conditional distribution. It also follows that any linear combination of β and $\mathbf{u}$ must have a Gaussian posterior distribution as well. For example, suppose one wishes to infer a vector of merits or of 'aggregate genetic values' (in the sense of Hazel, 1943) $\mathbf{h} = \mathbf{Mu}$, of a set of candidates. Here, $\mathbf{M}$ is a constant matrix reflecting the relative economic importance of traits, and $\mathbf{u}$ is a vector of multitrait genetic values. Then, the posterior distribution of $\mathbf{h}$ has mean vector $\widehat{\mathbf{h}} = \mathbf{M}\widehat{\mathbf{u}}$ and covariance matrix $\mathbf{M}\mathbf{C}_u\mathbf{M}'$, where $\mathbf{C}_u$ is the submatrix pertaining to $\mathbf{u}$ of the inverse in (20.12).

A related problem is that of inferring nonlinear merit. To illustrate, consider a second-order merit function. Suppose now that the aggregate genetic value of a candidate has the form:

$$h = \mathbf{m}'\mathbf{u} + \mathbf{u}'\mathbf{Qu},$$

where $\mathbf{m}'$ is a known row vector and $\mathbf{Q}$ is a known matrix, assumed to be symmetric without loss of generality. The posterior distribution of h does not have a closed form, but it can be estimated by Monte Carlo methods by drawing samples of $\mathbf{u}$ from the normal posterior distribution (20.12) and, thus, obtaining samples of h from the preceding

expression. However, the mean and variance of the posterior distribution of h can be calculated analytically as:

$$E\left(h|\mathbf{y}, \mathbf{G}, \mathbf{R}\right) = \mathbf{m}'\widehat{\mathbf{u}} + \widehat{\mathbf{u}}'\mathbf{Q}\widehat{\mathbf{u}} + tr\left(\mathbf{Q}\mathbf{C}_u\right),$$

and

$$Var\left(h|\mathbf{y}, \mathbf{G}, \mathbf{R}\right) = Var\left(\mathbf{m}'\mathbf{u}\right) + Var\left(\mathbf{u}'\mathbf{Q}\mathbf{u}\right) + 2\mathbf{m}'Cov\left(\mathbf{u}, \mathbf{u}'\mathbf{Q}\mathbf{u}\right)$$

$$= \mathbf{m}'\mathbf{C}_u\mathbf{m} + 2tr\left(\mathbf{Q}\mathbf{C}_u\right)^2 + 4\widehat{\mathbf{u}}'\mathbf{Q}\mathbf{C}_u\mathbf{Q}\widehat{\mathbf{u}} + 2\mathbf{m}'\mathbf{C}_u\mathbf{Q}\widehat{\mathbf{u}}.$$

Bulmer (1980) and Gianola and Fernando (1986) showed that the conditional or posterior mean is an optimum ranking rule when all parameters are known. Contrary to the case of a linear merit function, as seen above, the posterior precision of the candidate or, equivalently, the reliability of its evaluation enters nontrivially when inferring second-order merit. In fact, for some simple forms of the merit function, it can be shown that if two candidates have the same posterior mean (genetic evaluation) one would choose the one with the largest posterior variance. For more complex forms of the merit function, prediction of breeding value needs to proceed almost entirely using Monte Carlo methods.

Gianola and Fernando (1986) suggested the Bayesian approach as a general inferential method for solving a large number of animal breeding problems, linear or nonlinear, even in situations where there is uncertainty about all location and dispersion parameters. The first applications of the paradigm used Gaussian approximations to joint or partially marginalized posteriors, because of the technical difficulties. It was not until the advent of Markov chain Monte Carlo (MCMC) methods, however, that the power and flexibility of the Bayesian approach could be exploited in full. There are many MCMC methods, such as the Metropolis–Hastings algorithm, Gibbs sampling, reversible jump, simulated tempering, coupling from the past, etc. For a review of some of the algorithms, see Robert (1996). Undoubtedly, the most popular one has been the Gibbs sampler, although it can be used only under certain conditions. The basic idea in Gibbs sampling is as follows. Suppose that one wishes to infer a parameter (scalar or vector) λ , breeding values, say, from its posterior distribution. Further, suppose that the statistical model requires specifying additional, 'nuisance', parameters δ, e.g. in a mixed effects linear model these would be the fixed effects and the dispersion components. The joint posterior density is then:

$$p\left(\lambda, \delta|\mathbf{y}\right).$$

Typically, the marginal densities $p\left(\lambda|\mathbf{y}\right)$ and $p\left(\delta|\mathbf{y}\right)$are difficult or impossible to arrive at by analytical means. An alternative is to estimate features of the posterior distribution of breeding values having density $p\left(\lambda|\mathbf{y}\right)$ by sampling methods, and the Gibbs sampler is one of such procedures. Here, one needs to form the fully conditional distributions:

$$\left[\delta|\lambda, \mathbf{y}\right]$$

$$\left[\lambda|\delta, \mathbf{y}\right].$$

Subsequently, a sample is drawn from $\left[\delta|\lambda, \mathbf{y}\right]$, and the value is used to update the nuisance parameters in $\left[\lambda|\delta, \mathbf{y}\right]$; next, draw λ from the updated distribution, use the draws to update $\left[\delta|\lambda, \mathbf{y}\right]$, and repeat the process a large number of times, say m. This creates a Markov chain having the joint posterior $\left[\lambda, \delta|\mathbf{y}\right]$ as equilibrium distribution. At

point *m* (it is said that the sampler has 'converged') any subsequent draw belongs to the joint posterior, the implication being that the λ – component of the draw is an extraction from the marginal posterior distribution of interest. Collecting a reasonably large number of samples, such that inferences have a small Monte Carlo error, one can estimate, say, the posterior mean, median, variance, order statistics, or the marginal posterior density of any breeding value at regions of interest in the space of λ. A key aspect of the Gibbs sampler is that the fully conditional distributions must be recognizable and easy to sample from (Gianola *et al.*, 1994). Otherwise, one needs to resort to other sampling methods, such as Metropolis–Hastings or rejection sampling, to produce the necessary draw.

Early applications of Gibbs sampling in animal breeding were those of Wang *et al.* (1993; 1994a), and many papers using MCMC have been published thereafter. An important development was the introduction of Bayesian measures for assessing uncertainty in response to genetic selection from designed experiments (Sorensen *et al.*, 1994; Wang *et al.*, 1994b), this being a problem in animal breeding where the likelihood-frequentist 'tandem' approach can be regarded only as an approximation, even under normality assumptions. The Bayesian method resides in estimating the posterior distribution of measures of genetic change, with these being functions of the unobservable breeding values. The latter are drawn from their posterior distributions via MCMC and from these samples, one constructs draws from the posterior distribution of, say, response to selection. From the entire collection of draws, the posterior distribution of the unobservable genetic change is estimated. Sorensen *et al.* (2001) proposed a method for monitoring the evolution of additive genetic variance in the course of selection.

20.2.6   Nonlinear, Generalized Linear Models, and Longitudinal Responses

20.2.6.1   *Categorical Data*

Limited-information-dependent variables are pervasive in the analysis of fertility and disease data. Animal breeders did not hesitate (and still do not) to use linear models for such variables, even if this at the expense of causing consternation among statisticians. For example, Thompson (1979) stated:

"I have some unease at using linear models for these dichotomous traits."

and suggested some intuitively appealing approaches to mixed model analysis of binary data.

Gianola and Foulley (1983) addressed inferences about fixed and random effects in generalized mixed linear models for ordered categorical responses, a problem that was also studied by Harville and Mee (1984). The two methods give the same answer for predicting breeding values and estimating fixed effects, and yield BLUP when the data are Gaussian, rather than discrete. Their approach is similar to that used by Henderson in his early derivation of BLUP, although viewed in a Bayesian framework. For categorical data, it is assumed that there is an underlying, unobservable, variate called *liability* that follows a mixed effects linear model. Suppose, for simplicity, that a binary random variable is scored, e.g. presence or absence of mastitis in a dairy cow. If liability is larger than a conceptual threshold, mastitis is observed; the cow is nonmastitic otherwise. Since liability cannot be observed, the residual standard deviation in the underlying scale is taken as

unit of measurement. To simplify, suppose that the underlying distribution of liability is logistic, so the conditional probability that datum i is scored as 'mastitis' is:

$$P_i = \frac{\exp(\mathbf{x}_i'\beta + \mathbf{z}_i'\mathbf{u})}{1 + \exp(\mathbf{x}_i'\beta + \mathbf{z}_i'\mathbf{u})},$$

where $\mathbf{x}_i'$ and $\mathbf{z}_i'$ are the ith rows of $\mathbf{X}$ and $\mathbf{Z}$, respectively. The logic is defined to be:

$$\ln\left(\frac{P_i}{1 - P_i}\right) = \mathbf{x}_i'\beta + \mathbf{z}_i'\mathbf{u}.$$

It can be seen that P_i increases as β or $\mathbf{u}$ increase. As in the linear model, take $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ as prior distribution of the random effects, with an uniform prior adopted for β. If the variance of the random effects is known, the mode of the joint posterior distribution of β and $\mathbf{u}$ can be found iterating with:

$$\begin{bmatrix} \mathbf{X}'\mathbf{W}^{[t]}\mathbf{X} & \mathbf{X}'\mathbf{W}^{[t]}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}^{[t]}\mathbf{X} & \mathbf{Z}'\mathbf{W}^{[t]}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}\begin{bmatrix} \widehat{\beta}^{[t+1]} \\ \widehat{\mathbf{u}}^{[t+1]} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{W}^{[t]}\mathbf{y}^{[t]} \\ \mathbf{Z}'\mathbf{W}^{[t]}\mathbf{y}^{[t]} \end{bmatrix}, \qquad (20.13)$$

where the superscript indicates iterate number, $\mathbf{W} = \{P_i(1 - P_i)\}$ is a diagonal matrix with order equal to the number of observations, and:

$$\mathbf{y}^{[t]} = \mathbf{X}\beta^{[t]} + \mathbf{Z}\mathbf{u}^{[t]} + \left(\mathbf{W}^{[t]}\right)^{-1}\left(\mathbf{y} - \mathbf{P}^{[t]}\right),$$

is a pseudodata vector. If a normal distribution with variance 1, instead of the logistic, is adopted for liability (thus leading to a mixed effects probit model), the estimating equations are as in (20.13), except that $\mathbf{W}$ is slightly more difficult to calculate. Inferences are completed by employing a Gaussian approximation centered at the modal value and with dispersion matrix equal to the inverse of the coefficient matrix in (20.13), evaluated at the modal values.

This method was extended by Foulley *et al.* (1983) to models with Gaussian and categorical responses, by Höschele *et al.* (1986) to multivariate binary responses, and by Foulley *et al.* (1987b) to hierarchical models where a categorical response variables depends on a count having a Poisson conditional distribution. Harville and Mee (1984), Foulley *et al.* (1987a; 1990), Tempelman and Gianola (1996; 1999) and Tempelman and Firat (1998) discussed estimation of dispersion components in such settings. In particular, Harville and Mee (1984) and Foulley *et al.* (1987a) employed a Gaussian approximation with the expectation-maximization algorithm employed for calculating 'quasi-REML' or, in a better terminology, 'quasimarginal ML' estimates. For categorical data, Gilmour *et al.* (1985) described a different procedure, based on quasilikelihood; their variance component estimators and predictors of random effects lack formal justification, but reduce to REML and BLUP with Gaussian responses. Sorensen *et al.* (1995) presented a fully Bayesian solution for ordered polychotomies based on Gibbs sampling. In view of the computing power available today, there is not much justification for continued use of linear models or of approximations to the analysis of categorical data.

20.2.6.2 Linear and Nonlinear Models for Longitudinal Data

In the animal and veterinary sciences, there has been a renewed interest in the analysis of longitudinal records of performance. This is, perhaps, a consequence of more intensive

recording systems (for instance, in dairy cattle production it is possible to monitor instantaneous milk flow) and of better statistical methods for the analysis of longitudinal mixed effects models. In particular, linear random regression models or similar approaches have been applied in animal breeding, where there is a large body of literature in connection with the analysis of 'test-day' yields in dairy cattle. Similar applications have been made in meat producing species.

Briefly, the problem of analyzing longitudinal data can be posed as follows. Envisage a setting where, in a randomly drawn sample, each individual is measured longitudinally. For example, male and female rabbits from several breeds may be weighted at several phases of their development, from near birth to the adult stage. Suppose the objective is to study growth patterns of the two sexes in each of the breeds, while taking into account interindividual variability. Typically, there will be variation in the number of measurements per individual, leading to longitudinal unbalancedness. A hierarchical or multistage model consists of a series of nested functional specifications, together with the associated distributional assumptions. In the context of longitudinal data, at the first stage of the model, a mathematical function is used to describe the expected trajectory of individuals, and a stochastic residual having some distribution reflects the departure of the observations from such trajectory. At the second stage, a submodel is used to describe interindividual variation of parameters of the first-stage specification. A second-stage residual reflects the inability of the sub-model to explain completely the variation of the parameters. Additional stages can be imposed in a Bayesian context to describe uncertainty about the parameters.

At the first stage, it is assumed that the trajectory (body weights of the same animal, e.g.) can be described with the parametric model:

$$\mathbf{y}_i = \mathbf{f}_i\left(\theta_i, \mathbf{t}_i\right) + \varepsilon_i; \quad i = 1, 2, \ldots, M, \tag{20.14}$$

where $\mathbf{y}_i = \left\{y_{ij}\right\}$ $(i = 1, 2, \ldots, M; j = 1, 2, \ldots, n_i)$ is an n_i x 1 vector of records on the trajectory of individual i; $\mathbf{f}_i\left(\theta_i, \mathbf{t}_i\right)$ is its expected trajectory (e.g. expected growth curve) given a vector of animal-specific parameters θ_i, of order r x 1, and $\mathbf{t}_i$ is an n_i x 1 vector of known times of measurement. In (20.14), the n_i x 1 residual vector ε_i represents the inability of the function $\mathbf{f}_i\left(\theta_i, \mathbf{t}_i\right)$ of reproducing the observed body weights $\mathbf{y}_i$ exactly. An observation on individual i at time j is then:

$$y_{ij} = f_{ij}\left(\theta_i, t_{ij}\right) + \varepsilon_{ij}, \tag{20.15}$$

so the parameters θ_i dictate the form of the expected trajectory of individual i. The relationship between observed body weights and the parameters may be linear or nonlinear. In a linear specification, the derivatives of the model with respect to the parameters do not depend on θ_i. The entire vector of records can be represented as:

$$\mathbf{y} = \mathbf{f}\left(\theta, \mathbf{t}\right) + \varepsilon, \tag{20.16}$$

where θ is the Mr x 1 vector of parameters of all individuals, $\mathbf{t}$ contains times of measurement and ε is the $\sum_{i=1}^{M} n_i$ x 1 vector of residuals. In general, it is reasonable to assume that the first-stage residuals are independent between individuals, but some dependence within trajectories may exist. Possible dependencies between individuals,

such as genetic or environmental relatedness, can be introduced in the next stage of the model. Assuming normality of the residuals (sometimes, a thick-tailed distribution, such as Student's-t, may be a more sensible specification), the density of the first-stage distribution is expressible as:

$$\mathbf{y}_i \mid \theta_i, \gamma \sim \mathbf{N}\left[\mathbf{f}_i\left(\theta_i, \mathbf{t}_i\right), \mathbf{R}_i\left(\gamma\right)\right]; \ i = 1, 2, \dots, M, \tag{20.17}$$

with $\mathbf{y}_i$ being conditionally independent of $\mathbf{y}_j$, for all such pairs. In (20.17), $\mathbf{R}_i\left(\gamma\right)$ is an $n_i \ x \ n_i$ first-stage variance–covariance matrix, which depends on γ, a vector of dispersion parameters. For example, if residuals are independently and identically distributed within individuals, then $\mathbf{R}_i\left(\gamma\right) = \mathbf{I}_{n_i}\gamma$, where γ is the variance about the expected trajectory, so γ would be a scalar parameter here. The form of the matrix $\mathbf{R}_i\left(\gamma\right)$ depends on the dispersion assumptions needed.

The second stage is a statement of how individual-specific parameters vary according to explanatory factors, these perhaps representing genetic sources of variation. In order to facilitate implementation, it may be convenient to assume that the second stage of the model is linear on the effects of the explanatory variables. However, at least in theory, there is no reason for precluding a nonlinear specification, particularly if this is dictated by mechanistic considerations. If a linear model is adopted, the structure becomes:

$$\theta_i = \mathbf{X}_i\beta + \mathbf{u}_i + \mathbf{e}_i; \ i = 1, 2, \dots, M. \tag{20.18}$$

Above, the vector β represents the effects of p explanatory variables contained in the $r \ x \ p$ matrix $\mathbf{X}_i$; $\mathbf{u}_i$ are subject-specific effects on each of the r parameters, and $\mathbf{e}_i$ is a vector of second-stage residuals. Similarly to the errors in the first stage, these residuals represent discrepancies between the second-stage explanatory structure $\mathbf{X}_i\beta + \mathbf{u}_i$ and the 'true values' θ_i. In animal breeding applications, e.g. the vector $\mathbf{u}_i$ may be additive genetic effects on trajectory parameters, and these may or may not be identifiable separately from the residual vector $\mathbf{e}_i$, depending on the genetic relationship structure.

The second-stage distributional assumptions pertain to the uncertainty induced by the presence of $\mathbf{e}_i$ in model (20.18), given β and $\mathbf{u}_i$. It is often convenient to postulate that:

$$\theta_i \mid \beta, \mathbf{u}_i, \Sigma_e \sim N\left(\mathbf{X}_i\beta + \mathbf{u}_i, \Sigma_e\right), \tag{20.19}$$

implying that $\mathbf{e}_i \mid \Sigma_e \sim N\left(\mathbf{0}, \Sigma_e\right)$, where the second-stage variance–covariance matrix has the form:

$$\Sigma_e = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & . & . & & \sigma_{1r} \\ \sigma_{21} & \sigma_2^2 & . & . & & \sigma_{2r} \\ . & . & . & . & & . \\ . & . & . & . & \sigma_{r-1,r} \\ \sigma_{r1} & \sigma_{r2} & . & \sigma_{r,r-1} & \sigma_r^2 \end{bmatrix}. \tag{20.20}$$

Here, the diagonals elements are the variances of the second-stage residuals and the off-diagonals are corresponding covariances. In some instances, one may wish to assign a thick-tailed or robust distribution to the residuals, e.g. an $r - variate$ t-distribution. In this situation, one would write $\mathbf{e}_i \mid \nu_e, \Sigma_e \sim t_r\left(\mathbf{0}, \Sigma_e, \nu_e\right)$ to denote a t-distribution of dimension r, having a null mean vector, variance–covariance Σ_e, and degrees of freedom ν_e. It must be noted that in a multivariate t-distribution, $\Sigma_e = \frac{\nu_e}{\nu_e - 2}\mathbf{S}_e$, where $\mathbf{S}_e$ is a scale matrix, so $\nu_e > 2$ is a necessary condition for existence of the variance–covariance matrix

(Zellner, 1971). Often, it is assumed that second-stage residuals are mutually independent across individuals. Then, the joint density of all parameters at the second stage can be expressed as:

$$p\left(\theta_1, \theta_2, \ldots, \theta_M \mid \beta, \mathbf{u}_1, \mathbf{u}_{2,} \ldots, \mathbf{u}_M, \ \Sigma_e\right) = \prod_{i=1}^{M} p\left(\theta_i \mid \beta, \mathbf{u}_i, \ \Sigma_e\right). \qquad (20.21)$$

Compactly, all parameters are expressible as:

$$\theta_{Mr\,x\,1} = \mathbf{X}_{Mr\,x\,p}\beta_{p\,x\,1} + \mathbf{u}_{Mr\,x\,1} + \mathbf{e}_{Mr\,x\,1}.$$

This indicates that the second-stage distribution of all parameters of all individuals is:

$$\theta \mid \beta, \mathbf{u}, \Sigma_e \sim N\left(\mathbf{X}\beta + \mathbf{u}, \mathbf{I} \otimes \Sigma_e\right). \qquad (20.22)$$

An alternative formulation can be obtained by arranging individuals within parameters; here, $\mathbf{X}$ must be redefined accordingly, and the covariance matrix of the process would then be $\Sigma_e \otimes \mathbf{I}$. The choice between the two alternative orderings is entirely a matter of computational convenience.

In a Bayesian model, prior distributions must be assigned to all unknown quantities in the statistical system posited. Thus, priors must be stated for β, $\mathbf{u}$, Σ_e and γ. Let the vector $\mathbf{u}$ represent additive genetic effects on the trajectory parameters, in which case a classical (and convenient) assumption made in quantitative genetics is that:

$$\mathbf{u} \mid \mathbf{G}_0 \sim N\left(\mathbf{0}, \mathbf{A} \otimes \mathbf{G}_0\right), \qquad (20.23)$$

where it is implied that parameters are ordered within individuals. Above, $\mathbf{A}$ is the additive genetic relationship matrix between the M individuals, and $\mathbf{G}_0$ is an $r \ x \ r$ additive genetic variance–covariance matrix between parameters, i.e.:

$$\mathbf{G}_0 = \begin{bmatrix} \sigma_{u_1}^2 & \sigma_{u_{12}} & \cdot & \cdot & \sigma_{u_{1r}} \\ \sigma_{u_{21}} & \sigma_{u_2}^2 & \cdot & \cdot & \sigma_{u_{2r}} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \sigma_{u_{(r-1)r}} \\ \sigma_{u_{r1}} & \sigma_{u_{r2}} & \cdot & \sigma_{u_{r(r-1)}} & \sigma_{u_r}^2 \end{bmatrix}.$$

If $\mathbf{G}_0$ is unknown, a prior distribution must be elicited for this matrix as well. Uncertainty about all unknowns would be in the joint prior density:

$$p\left(\beta, \mathbf{u}, \mathbf{G}_0, \Sigma_e, \gamma\right). \qquad (20.24)$$

After data are combined with the prior by formal use of Bayes theorem, inference can proceed. The preceding prior distribution can be bounded, based on prior knowledge of parameter values or on theoretical considerations. It must be emphasized that an advantage of the Bayesian approach resides in the possibility of incorporating external stochastic information into the analysis.

Prior to the advent of MCMC, approximate methods needed to be used. For example, Gianola and Kachman (1983) and Kachman and Gianola (1984) suggested empirical Bayes and REML-type estimators of breeding values and (co)variance components, respectively, for nonlinear functions (the linear model being a particular case) describing

longitudinal trajectories, e.g. lactation or growth. These are identical to those developed later by Lindstrom and Bates (1990) and Laird (1990), for what has been termed *random regression* models. Rekaya (1997); Rodriguez-Zas (1998) and Chang (1999) described fully Bayesian implementations of nonlinear lactation curves via MCMC (rejection sampling, Metropolis–Hastings) including, in the last case, heavy-tailed distributions. A seemingly different approach has been that of 'covariance' functions (Kirkpatrick and Lovsfold, 1989), where the covariance between records of individuals is viewed as a continuous function of time. Meyer (1998) describes a REML implementation of covariance functions, but it is unclear how this can be extended, within the paradigm employed, to a situation where the trajectory dictates an intrinsically nonlinear model.

20.2.6.3   Survival Analysis

An area that has received increased attention in animal breeding has been survival analysis applied to productive lifespan or herd life (Smith and Allaire, 1986; Ducrocq and Casella, 1996; Sorensen *et al.*, 1998, and Korsgaard *et al.*, 1999). As in the medical sciences, 'proportional hazard' models have been employed in animal genetics as well. Here, the two basic concepts are: (1) the probability that an individual will survive beyond time t, or survival function, and (2) the hazard function, i.e. the 'instantaneous probability' that an individual surviving up to time t, will die right after that. A common feature of survival analysis is the presence of censored observations. For example, it may be known that a certain cow was present in the herd at some time, but that she was sold thereafter to another herd for production purposes, with no information about the date of termination of her career. The density of all observations, given the parameters, has two components: one for censored observations, depending on the parameters through the survival function, and another (for uncensored data) where the parameters enter through the hazard function. In these models, the hazard of an individual can be expressed as the product of a baseline hazard and of a proportional hazard that depends on fixed and random effects. Here, empirical Bayes approaches have been used for inferring breeding values (e.g. Ducrocq and Casella, 1996), although fully Bayesian analyses are technically feasible.

A challenge was how to embed a survival analysis within a multitrait setting. This problem is easy to deal with if a linear censored model is employed (Korsgaard *et al.*, 1999), although at the expense of flexibility and realism of the assumptions. Recently, Damgaard and Korsgaard (2006) proposed a bivariate quantitative genetic model for a Gaussian trait and a time-to-event variate. Their statistical treatment is fully Bayesian.

Likewise, there are situations where a multivariate hazard function may need to be modeled. For example, Guo (1999) studied herd life and lifetime prolificacy in sows. Both traits can be subject to censoring, and a bivariate survival analysis would be needed if one wishes to infer the genetic and environmental correlation between the traits. A more complete discussion is in Guo (2006).

20.2.7   Effects of Selection on Inferences

As noted earlier, animal breeding data seldom arise from a truly random mechanism. Except in designed experiments, the 'history of the selection process' is known incompletely, leading to missing data in some statistical sense. A question of importance is the extent to which inferences are distorted when, e.g. selection and assortative mating are

ignored. Important contributions here have been those of Kempthorne and von Krosigk in Henderson *et al.* (1959), Curnow (1961) and, notably, Henderson (1975). Im *et al.* (1989) discussed the inferential problems from a 'missing data' point of view, and Gianola and Fernando (1986) and Fernando and Gianola (1990) gave the Bayesian perspective.

Kempthorne and von Krosigk in Henderson *et al.* (1959) and Curnow (1961), employing normality assumptions, found that the ML estimator of a parameter has the same form with and without selection, provided that all data used for selection decisions are used in the analysis. Im *et al.* (1989) presented the result in more general form, for any distribution. However, this does not imply that the asymptotic distribution of the ML estimator remains unaffected by selection. In order to arrive at the information matrix under selection, one needs to take expectations under the marginal distribution of the observations under selection, rather than under random sampling. Failing to do this, interestingly, leads to correct point estimation, but inadequate interval inferences, over repeated sampling. Hence, selection is not completely ignorable if one wishes to go beyond obtaining a point estimate. Gianola *et al.* (1989) review some of the issues.

Henderson (1975), assuming that dispersion parameters were known, derived best linear unbiased predictors of breeding value under a specific selection model. Here, upon conceptual repeated sampling, incidence matrices and the relationship matrix must remain constant from replication to replication. Henderson's model holds only when the animals are exchangeable, in the sense that any permutation of items yields the same distribution and relationship matrix. This is unrealistic. At any rate, and within these restrictions, he gave conditions for unbiasedness, which have been largely quoted and followed by the animal breeding community. One of these, e.g. states that if selection is based on linear functions of the unobservable breeding values, some random elements in the model (e.g. herds) must be treated as fixed in order to obtain unbiased predictors of the breeding values. First, it is obvious that if one could observe the breeding values for constructing the linear functions on which selection is based, there would be no point in predicting anything, as the state of nature would be known. Second, this does not describe the type of selection encountered in practice. Interestingly, this selection model has been received rather uncritically by animal breeders, with the notable exception of Thompson (1979). At any rate, Henderson (1975) probably constitutes the best frequentist attempt of tackling unbiased prediction of random effects under selection.

In a Bayesian setting, Gianola and Fernando (1986) showed that if all data are used for constructing the joint posterior distribution of the unknown parameters, selection can be ignored. This was elaborated further by Fernando and Gianola (1990) and extended by Gianola *et al.* (1999). This holds at the level of the marginal posteriors for any unknown quantity, irrespective of whether this is a breeding value, a genetic correlation in a multivariate threshold model or the degrees of freedom of a t-distribution. These results, however, should not be interpreted from a frequentist point of view. For example, under normality assumptions and known dispersion structure, the mean of the posterior distribution of the breeding values under this type of selection is equal to BLUP ignoring selection. However, 'regular' BLUP is biased under 'location variant' selection (Henderson, 1975). Here, we have a situation where the Bayesian solution cannot cure a frequentist malaise.

Unfortunately, there are situations in which selection is not ignorable. For example, it is to be expected that an analysis of carcass traits in beef cattle ignoring concomitant selection for growth rate would lead to incorrect inference. Here, it is essential to

attempt modeling the 'missing data' or selection process or, alternatively, one perhaps should consider using robust methods of inference. For example, if selection (natural or artificial) moves a population toward some intermediate optimum, this must be taken into account somehow. If such selection is according to a Gaussian fitness functions, the distribution after selection remains normal, but parameters change (e.g. Bulmer, 1980). A precise statement of the problem of inference under selection is in Sorensen *et al.* (2001).

20.2.8   Massive Molecular Data: Semiparametric Methods

20.2.8.1   *Motivation*

Massive quantities of genomic data are now available, with potential for enhancing accuracy of prediction of genetic value of, e.g. candidates for selection in animal and plant breeding programs, or for molecular classification of disease status in subjects (Golub *et al.*, 1999). For instance, Wong *et al.* (2004) reported a genetic variation map of the chicken genome containing 2.8 million single-nucleotide polymorphisms (SNPs), and demonstrated how the information can be used for targeting specific genomic regions. Likewise, Hayes *et al.* (2004) found 2507 putative SNPs in the salmon genome that could be valuable for marker-assisted selection in this species.

The use of molecular markers as aids in genetic selection programs has been discussed extensively. Important early papers are Soller and Beckmann (1982) and Fernando and Grossman (1989), with the latter focusing on BLUP of genetic value when marker information is used. Most of the literature on marker-assisted selection deals with the problem of locating one or few QTL using flanking markers. However, in the light of current knowledge about genomics, the widely used single-QTL search approach is naive, since there is evidence of abundant QTLs affecting complex traits, as discussed, e.g. by Dekkers and Hospital (2002). This would support the infinitesimal model of Fisher (1918) as a sensible statistical specification for many quantitative traits, with complications being the accommodation of nonadditivity and of feedbacks (Gianola and Sorensen, 2004). Dekkers and Hospital (2002) observe that existing statistical methods for marker-assisted selection do not deal well with complexity posed by quantitative traits. Some difficulties are: specification of 'statistical significance' thresholds for multiple testing, strong dependence of inferences on model chosen (e.g. number of QTLs fitted, distributional forms), inadequate handling of nonadditivity and ambiguous interpretation of effects in multiple-marker analysis, due to collinearity.

In this section, we discuss how large-scale molecular information, such as those conveyed by SNPs, can be employed for marker-assisted prediction of genetic value for quantitative traits in the sense of, e.g. Meuwissen *et al.* (2001), Gianola *et al.* (2003) and Xu (2003). The focus is on inference of genetic value, rather than detection of QTL. A main challenge is that of positing a functional form-relating phenotypes to SNP genotypes (viewed as thousands of possibly highly colinear covariates), to polygenic additive genetic values, and to other nuisance effects, such as sex or age of an individual, simultaneously. A more detailed presentation is in Gianola *et al.* (2006).

Standard quantitative genetics theory gives a mechanistic basis to the mixed effects linear model, treated either from classical or Bayesian perspectives. Meuwissen *et al.* (2001) and Gianola *et al.* (2003) exploit this connection and suggest highly parametric structures for modeling relationships between phenotypes and effects of hundreds or

thousands of molecular markers. A first concern is the strength of their assumptions (e.g. linearity, multivariate normality, proportion of segregating loci, spatial within-chromosome effects); it is unknown if their procedures are robust. Secondly, colinearity between SNP or marker genotypes is bound to exist, because of the sheer massiveness of molecular data plus cosegregation of alleles. While adverse effects of colinearity can be tempered when marker effects are treated as random variables, statistical redundancy is undesirable.

The genome seems to be much more highly interactive than what standard quantitative genetic models can accommodate. In theory, genetic variance can be partitioned into orthogonal additive, dominance, additive × additive, additive × dominance, dominance × dominance, etc., components, only under highly idealized conditions. These include linkage equilibrium, absence of natural or artificial selection, and no inbreeding or assortative mating. Arguably, these conditions are violated in nature and in breeding programs. Actually, marker-assisted selection exploits existence of linkage disequilibrium, and even chance creates disequilibrium. Further, estimation of nonadditive components of variance is notoriously difficult, even under standard assumptions. Therefore, it is doubtful whether standard quantitative genetic approaches can model fine-structure relationships between genotypes and phenotypes adequately, unless either departures from assumptions have mild effects, or statistical constructs turn out to be more robust that what is expected on theoretical grounds. These considerations suggest that a nonparametric treatment of the data could be valuable. On the other hand, application of the additive genetic model in selective breeding of livestock has produced remarkable dividends, as shown in Dekkers and Hospital (2002). Hence, a combination of nonparametric modeling of effects of molecular variables (e.g. SNPs) with features of the additive polygenic mode of inheritance is appealing.

20.2.8.2   *Kernel Regression on SNP Markers*

Consider a stylized situation in which each of a series of individuals possesses a measurement for some quantitative trait denoted as y, as well as information on a possibly massive number of genomic variables, such as SNP 'genotypes', represented by a vector $\mathbf{x}$. In the main, $\mathbf{x}$ is treated as a continuously valued vector of covariates, even though SNP genotypes are discrete (coding is done via dummy variates). Also, $\mathbf{x}$ could represent gene expression measurements from microarray experiments; here, it would be legitimate to regard this vector as continuous. Although gene expression measurements are typically regarded as response variables, there are contexts in which this type of information could be used in an explanatory role (Mallick *et al.*, 2005).

Let the relationship between y and $\mathbf{x}$ be represented as

$$y_i = g\left(\mathbf{x}_i\right) + e_i; \; i = 1, 2, \ldots, n, \tag{20.25}$$

where:

- y_i is a measurement, such as plant height or body weight, taken on individual i;

- $\mathbf{x}_i$ is a $p \times 1$ vector of dummy SNP or microsatellite covariates observed on i, and $g\left(.\right)$ is some unknown function relating these 'genotypes' to phenotypes. Define $g\left(\mathbf{x}_i\right) = E\left(y_i \mid \mathbf{x}_i\right)$ as the conditional expectation function, i.e. the mean phenotypic

value of an infinite number of individuals, all possessing the p-dimensional geno-type $\mathbf{x}_i$.

- $e_i \sim (0, \sigma^2)$ is a random residual, distributed independently of $\mathbf{x}_i$ and with variance σ^2.

The conditional expectation function is

$$g(\mathbf{x}) = \frac{\int y p(\mathbf{x}, y) \, dy}{p(\mathbf{x})}. \tag{20.26}$$

Following Silverman (1986), consider a nonparametric kernel estimator of the p-dimensional density of the covariates:

$$\widehat{p}(\mathbf{x}) = \frac{1}{n h^p} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right), \tag{20.27}$$

where $K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)$ is a kernel function and h is a window width or smoothing parameter. In (20.27), $\mathbf{x}$ is the value ('focal point') at which the density is evaluated and $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$ is the observed p-dimensional SNP genotype of individual i in the sample. Hence, (20.27) estimates population densities (or frequencies). If $\widehat{p}(\mathbf{x})$ is to behave as a multivariate probability density function, then it must be true that the kernel function is positive and that it integrates to 1. A nonparametric estimator of $g(\mathbf{x})$ is given by

$$\widehat{g}(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) y_i, \tag{20.28}$$

where

$$w_i(\mathbf{x}) = \frac{K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)},$$

is a weight that depends on the kernel function and window width chosen, and on the $\mathbf{x}_i$ (i.e. genotypes) observed in the sample. The linear combination of the observations (20.28) is called the *Nadaraya–Watson estimator of the regression function*. As seen in (20.28), the fitted value at coordinate $\mathbf{x}$ is a weighted average of all data points, with the value of the weight depending on the 'proximity' of $\mathbf{x}_i$ to $\mathbf{x}$ and on the value of the smoothing parameter h. For instance, if the kernel function has the Gaussian form:

$$K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) = \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left[-\frac{1}{2}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)'\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)\right],$$

this has a maximum value of $(2\pi)^{-\frac{p}{2}}$ when $\mathbf{x} = \mathbf{x}_i$, and tails off to 0 as the distance between $\mathbf{x}$ and $\mathbf{x}_i$ increases. The values of $w_i(\mathbf{x})$ decrease more abruptly as $h \to 0$. Observations with $\mathbf{x}_i$ coordinates closer to the focal point $\mathbf{x}$ are weighted more strongly in the computation of the fitted value $\widehat{E}(y \mid \mathbf{x})$.

20.2.8.3   Discrete Kernels

For a biallelic SNP, there are three possible genotypes at each 'locus', as in stylized Mendelian situations. In a standard (parametric) analysis of variance representation, incidence situations (or additive and dominance effects at each of the loci) are described via two dummy binary variables per locus, and all corresponding epistatic interactions can be assessed from effects of cross-products of these variables. This leads to a highly parameterized structure and to formidable model selection problems.

Consider now the nonparametric approach. For an $\mathbf{x}$ vector with p coordinates, its statistical distribution is given by the probabilities of each of the 3^p combinations of binary outcomes. With SNPs, p can be very large (possibly much larger than n), so it is hopeless to estimate the probability distribution of genotypes accurately from observed relative frequencies, and smoothing is required (Silverman, 1986). Kernel estimation extends as follows: for binary covariates the number of disagreements between a focal $\mathbf{x}$ and the observed $\mathbf{x}_i$ in subject i is given by

$$d\left(\mathbf{x}, \mathbf{x}_i\right) = \left(\mathbf{x}_i - \mathbf{x}\right)' \left(\mathbf{x}_i - \mathbf{x}\right),$$

where $d(.)$ takes values between 0 and p. For binary covariates, Silverman (1986) suggests the 'binomial' kernel

$$K\left(\mathbf{x}, \mathbf{x}_i, h\right) = h^{p - d(\mathbf{x}, \mathbf{x}_i)} \left(1 - h\right)^{d(\mathbf{x}, \mathbf{x}_i)},$$

with $\frac{1}{2} \leq h \leq 1$. It follows that the kernel estimate of the probability of observing the focal value $\mathbf{x}$ is

$$\widehat{p}\left(\mathbf{x}\right) = \frac{1}{n} \sum_{i=1}^{n} h^{p - d(\mathbf{x}, \mathbf{x}_i)} \left(1 - h\right)^{d(\mathbf{x}, \mathbf{x}_i)}. \tag{20.29}$$

If $h = 1$, the estimate is just the proportion of cases for which $\mathbf{x}_i = \mathbf{x}$; if $h = \frac{1}{2}$, every focal point gets an estimate equal to $\left(\frac{1}{2}\right)^p$, irrespective of the observed values $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n$.

The nonparametric estimator of the regression function is

$$\widehat{g}(\mathbf{x}) = \frac{\displaystyle\sum_{i=1}^{n} h^{p - d(\mathbf{x}, \mathbf{x}_i)} \left(1 - h\right)^{d(\mathbf{x}, \mathbf{x}_i)} y_i}{\displaystyle\sum_{i=1}^{n} h^{p - d(\mathbf{x}, \mathbf{x}_i)} \left(1 - h\right)^{d(\mathbf{x}, \mathbf{x}_i)}}.$$

20.2.8.4   Semiparametric Kernel Mixed Model

Consider now a situation for which there might be an operational or mechanistic basis for specifying at least part of a model. For instance, suppose y is a measure on some quantitative trait, such as milk production of a cow. Animal breeders have exploited to advantage the infinitesimal model of quantitative genetics (Fisher, 1918). In this section, we combine features of the infinitesimal model with a nonparametric treatment of genomic data, and present semiparametric implementations.

Model (20.25) is expanded as

$$y_i = \mathbf{w}_i' \beta + \mathbf{z}_i' \mathbf{u} + g(\mathbf{x}_i) + e_i; \quad i = 1, 2, \ldots, n, \tag{20.30}$$

where β is a vector of nuisance location effects and $\mathbf{u}$ is a $q \times 1$ vector containing additive genetic effects of q individuals (these effects are assumed to be independent of those of the markers), some of which may lack a phenotypic record; $\mathbf{w}_i'$ and $\mathbf{z}_i'$ are known nonstochastic incidence vectors. As before, $g(\mathbf{x}_i)$ is some unknown function of the SNP data. It will be assumed that $\mathbf{u} \sim N\left(\mathbf{0}, \mathbf{A}\sigma_u^2\right)$, where σ_u^2 is the 'unmarked' additive genetic variance and $\mathbf{A}$ is the additive relationship matrix. Let $\mathbf{e} = \{e_i\}$ be the $n \times 1$ vector of residuals, and assume that $\mathbf{e} \sim N\left(0, \mathbf{I}\sigma_e^2\right)$, where σ_e^2 is the residual variance. Note that the model implies that

$$y_i - g(\mathbf{x}_i) = \mathbf{w}_i'\beta + \mathbf{z}_i'\mathbf{u} + e_i, \tag{20.31}$$

and

$$y_i - \mathbf{w}_i'\beta - \mathbf{z}_i'\mathbf{u} = g(\mathbf{x}_i) + e_i. \tag{20.32}$$

The preceding means that: (1) the offset $y_i - g(\mathbf{x}_i)$ follows a standard mixed effects model, and (2) if β and $\mathbf{u}$ were known, one could use (20.28) to estimate $g(\mathbf{x}_i)$ employing $y_i - \mathbf{w}_i'\beta - \mathbf{z}_i'\mathbf{u}$ as 'observations'.

One possible strategy is to carry out a mixed model analysis. This follows from representation (20.31). First, estimate $g(\mathbf{x}_i)$, for $i = 1, 2, \ldots, n$, via $\widehat{g}(\mathbf{x}_i)$, as in (20.28). Then, carry out a mixed model analysis using the 'corrected' data vector and pseudomodel

$$\mathbf{y}^* = \{y_i - \widehat{g}(\mathbf{x}_i)\} = \mathbf{W}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where $\mathbf{W} = \{\mathbf{w}_i'\}$ and $\mathbf{Z} = \{\mathbf{z}_i'\}$ are incidence matrices of appropriate order. The pseudomodel ignores uncertainty about $g(\mathbf{x})$, since $\widehat{g}(\mathbf{x}_i)$ is treated as if it were the true regression (on SNPs) surface.

Under the standard multivariate normality assumptions of the infinitesimal model, one can estimate the variance components σ_u^2 and σ_e^2 from $\mathbf{y}^*$ via REML, and form empirical best linear unbiased estimators (BLUE) and predictors of β and $\mathbf{u}$, respectively, by solving the Henderson MME

$$\begin{bmatrix} \mathbf{W}'\mathbf{W} & \mathbf{W}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{W} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{W}'\mathbf{y}^* \\ \mathbf{Z}'\mathbf{y}^* \end{bmatrix}. \tag{20.33}$$

The ratio $\frac{\sigma_e^2}{\sigma_u^2}$ is evaluated at REML estimates of the variance components. Solving system (20.33) is a standard problem in animal breeding even for very large q, since $\mathbf{A}^{-1}$ is easy to compute. The two stage procedure could be iterated several times, i.e. use the solutions to (20.33), to obtain a new estimate of $g(\mathbf{x})$ using $y_i - \mathbf{w}_i'\widehat{\beta} - \mathbf{z}_i'\widehat{\mathbf{u}}$ as 'data', then update the pseudodata $\mathbf{y}^*$, etc. A Bayesian approach can be used instead, using the corrected data $y_i - \widehat{g}(\mathbf{x}_i)$ as observations. Under standard assumptions made for the prior and the likelihood, one can draw samples $j = 1, 2, \ldots, m$ from the pseudo posterior distribution $[\beta, \mathbf{u}, \sigma_u^2, \sigma_e^2 | \mathbf{y}^*]$ via Gibbs sampling, and then form 'semiparametric' draws of the total genetic value. Irrespective of whether classical or Bayesian viewpoints are adopted, this approach ignores the error of estimation of $g(\mathbf{x})$, as noted earlier.

20.2.8.5 *Reproducing Kernel Hilbert Spaces Mixed model*

What follows is inspired by developments in Mallick *et al.* (2005) for classification of tumors using microarray data. The underlying theory is outside the scope of the paper.

Only essentials are given here, and foundations are in Wahba (1990; 1999). Using the structure of (20.30), consider the penalized sum of squares

$$SS\big[g(\mathbf{x}), h\big] = \sum_{i=1}^{n} \big[y_i - \mathbf{w}_i'\beta - \mathbf{z}_i'\mathbf{u} - g(\mathbf{x}_i)\big]^2 + h \,\|g(\mathbf{x})\|, \qquad (20.34)$$

where, as before, h is a smoothing parameter (possibly unknown) and $\|g(\mathbf{x})\|$ is some norm or 'stabilizer'. For instance, in smoothing splines, $\|g(\mathbf{x})\|$ is a function of the second derivatives of $g(\mathbf{x})$ integrated between end-points that comprise the data. The second term in (20.34) acts as a penalty: if the unknown function $g(\mathbf{x})$ is rough, in the sense of having slopes that change rapidly, the penalty increases. The main problem here is that of finding the function $g(\mathbf{x})$ that minimizes (20.34). Since $SS\big[g(\mathbf{x}), h\big]$ is a functional on $g(\mathbf{x})$, this is a variational or calculus of variations problem over a space of smooth curves. The minimizer admits the representation

$$g(.) = \alpha_0 + \sum_{j=1}^{n} \alpha_j K\left(., \mathbf{x}_j\right),$$

where $K\,(.,.)$ is called a *reproducing kernel*. A possible choice for the kernel is the single smoothing parameter Gaussian function

$$K_h\left(\mathbf{x}, \mathbf{x}_j\right) = \exp\left[-\frac{\left(\mathbf{x} - \mathbf{x}_j\right)'\left(\mathbf{x} - \mathbf{x}_j\right)}{h}\right].$$

We embed these results into (20.30), leading to the specification

$$y_i = \mathbf{w}_i'\beta + \mathbf{z}_i'\mathbf{u} + \sum_{j=1}^{n} \exp\left[-\frac{\left(\mathbf{x}_i - \mathbf{x}_j\right)'\left(\mathbf{x}_i - \mathbf{x}_j\right)}{h}\right]\alpha_j + e_i, \qquad (20.35)$$

with the intercept parameter α_0 included as part of β. Note that there are as many regressions α_j as there are data points. However, the roughness penalty in the variational problem leads to a reduction in the effective number of parameters in reproducible kernel Hilbert spaces regression (RKHS), as it occurs in smoothing splines.

Define the $1 \times n$ row vector

$$\mathbf{t}_i'\,(h) = \left\{\exp\left[-\frac{\left(\mathbf{x}_i - \mathbf{x}_j\right)'\left(\mathbf{x}_i - \mathbf{x}_j\right)}{h}\right]\right\}, \quad j = 1, 2, .., n;$$

the $n \times 1$ column vector $\alpha = \left\{\alpha_j\right\}$, $j = 1, 2, \ldots, n$; and the $n \times n$ matrix

$$\mathbf{T}\,(h) = \begin{bmatrix} \mathbf{t}_1'\,(h) \\ \mathbf{t}_2'\,(h) \\ . \\ . \\ . \\ \mathbf{t}_n'\,(h) \end{bmatrix}.$$

Then (20.35) can be written in matrix form as

$$\mathbf{y} = \mathbf{W}\beta + \mathbf{Z}\mathbf{u} + \mathbf{T}\,(h)\,\alpha + \mathbf{e}.$$

Suppose, further, that the α_j coefficients are exchangeable according to the distribution $\alpha_j \sim N\left(0, \sigma_\alpha^2\right)$. Hence, for a given smoothing parameter h, we are in the setting of a mixed effects linear model.

Given h, σ_u^2, σ_e^2, and σ_α^2 (at a given h, the three variance components may be estimated by, e.g. REML) one can obtain predictions of the polygenic breeding values $\mathbf{u}$ and of the coefficients α from the solutions to the system

$$\begin{bmatrix} \mathbf{W}'\mathbf{W} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{T}(h) \\ \mathbf{Z}'\mathbf{W} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2} & \mathbf{Z}'\mathbf{T}(h) \\ \mathbf{T}'(h)\mathbf{W} & \mathbf{T}'(h)\mathbf{Z} & \mathbf{T}'(h)\mathbf{T}(h) + \mathbf{I}\frac{\sigma_e^2}{\sigma_\alpha^2} \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{\mathbf{u}} \\ \widehat{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{W}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{T}'(h)\mathbf{y} \end{bmatrix}. \quad (20.36)$$

At least in animal breeding, it is not feasible to have all individuals genotyped for SNPs. On the other hand, the number of animals with phenotypic information available is typically in the order of hundreds of thousands, and genotyping is selective, e.g. young bulls that are candidates for progeny testing in dairy cattle production. Animals lacking molecular data are not a random sample from the population, and ignoring this issue may lead to biased inferences. Unless missingness of molecular data is ignorable, the procedures given below require modeling of the missing data process, which is difficult and may lack robustness. Here, it is assumed that missingness is ignorable, enabling use of likelihood-based or Bayesian procedures as if selection had not taken place. Consider the following two *ad hoc* procedures.

Let the vector of phenotypic data be partitioned as $\mathbf{y} = \left[\mathbf{y}_1' \ \mathbf{y}_2'\right]'$, where $\mathbf{y}_1$ $(n_1 \times 1)$, consists of records of individuals lacking SNP data, whereas $\mathbf{y}_2$ $(n_2 \times 1)$ includes phenotypic data of genotyped individuals. Often, it will be the case that $n_1 > p >> n_2$. We adopt the model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix} \beta + \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} \mathbf{u} + \begin{bmatrix} \mathbf{0} \\ \mathbf{T}_2(h) \end{bmatrix} \alpha + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}. \quad (20.37)$$

For the sake of flexibility, assume that $\mathbf{e}_1 \backsim N\left(\mathbf{0}, \mathbf{I}_{n_1}\sigma_{e_1}^2\right)$ and $\mathbf{e}_2 \backsim N\left(\mathbf{0}, \mathbf{I}_{n_2}\sigma_{e_2}^2\right)$ are mutually independent but heteroscedastic vectors. In short, the key assumption made here is that the random effect α affects $\mathbf{y}_2$ but not $\mathbf{y}_1$, or equivalently, that it gets absorbed into $\mathbf{e}_1$. With this representation, the mixed model equations take the form

$$\begin{bmatrix} \sum_{i=1}^{2} \frac{1}{\sigma_{e_i}^2}\mathbf{W}_i'\mathbf{W}_i & \sum_{i=1}^{2} \frac{1}{\sigma_{e_i}^2}\mathbf{W}_i'\mathbf{Z}_i & \frac{1}{\sigma_{e_2}^2}\mathbf{W}_2'\mathbf{T}_2(h) \\ \sum_{i=1}^{2} \frac{1}{\sigma_{e_i}^2}\mathbf{Z}_i'\mathbf{W}_i & \sum_{i=1}^{2} \frac{1}{\sigma_{e_i}^2}\mathbf{Z}_i'\mathbf{Z}_i + \mathbf{A}^{-1}\frac{1}{\sigma_u^2} & \frac{1}{\sigma_{e_2}^2}\mathbf{Z}_2'\mathbf{T}_2(h) \\ \frac{1}{\sigma_{e_2}^2}\mathbf{T}_2'(h)\mathbf{W}_2 & \mathbf{T}_2'(h)\mathbf{Z} & \frac{1}{\sigma_{e_2}^2}\mathbf{T}_2'(h)\mathbf{T}_2(h) + \mathbf{I}\frac{1}{\sigma_\alpha^2} \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{\mathbf{u}} \\ \widehat{\alpha} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{2} \frac{1}{\sigma_{e_i}^2}\mathbf{W}_i'\mathbf{y}_i \\ \sum_{i=1}^{2} \frac{1}{\sigma_{e_i}^2}\mathbf{Z}_i'\mathbf{y}_i \\ \frac{1}{\sigma_{e_2}^2}\mathbf{T}_2'(h)\mathbf{y}_2 \end{bmatrix}. \quad (20.38)$$

If SNP data are missing completely at random and h, σ_u^2, σ_e^2 and σ_α^2 are treated as known, then $\widehat{\beta}$ is an unbiased estimator of β, and $\widehat{\mathbf{u}}$ and $\widehat{\alpha}$ are unbiased predictors of $\mathbf{u}$ and α,

respectively. They are not 'best', in the sense of having minimum variance or minimum prediction error variance, because the smooth function $g(\mathbf{x})$ of the SNP markers is missing in the model for individuals that are not genotyped (Henderson, 1974).

An alternative consists of writing the bivariate model

$$
\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix} \beta + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{T}_2(h) \end{bmatrix} \alpha + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix},
$$

and then assign to the polygenic component, the multivariate normal distribution

$$
\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{A}\sigma_{u_1}^2 & \mathbf{A}\sigma_{u_{12}} \\ \mathbf{A}\sigma_{u_{12}} & \mathbf{A}\sigma_{u_2}^2 \end{bmatrix} \right).
$$

Here, $\sigma_{u_1}^2$ and $\sigma_{u_2}^2$ are additive genetic variances in individuals without and with molecular information, respectively, and $\sigma_{u_{12}}$ is their additive genetic covariance. Computations would be those appropriate for a two-trait linear model analysis.

Gianola *et al.* (2006) simulated phenotypic and genotypic values for a sample of N unrelated individuals, for each of two situations. The trait was determined by either five biallelic QTL, having additive gene action, or by five pairs of biallelic QTL, having additive by additive gene action. Under nonadditive gene action, the additive genetic variance was null, so all genetic variance was of the additive $\times$ additive type. Heritability (ratio between genetic and phenotypic variance) in both settings was set to 0.5. Genotypes were generated for a total of 100 biallelic markers, including the 'true' QTL; all loci were simulated to be in gametic-phase equilibrium. Since all individuals were unrelated and all loci were in gametic-phase equilibrium, only the QTL genotypes and the trait phenotypes would provide information on the genotypic value. In most real applications, the location of the QTL will not be known, and so many loci that are not QTL will be included in the analysis. An RKHS mixed model was used to predict genotypic values, given phenotypic values and genotypes at the QTL and at all other loci. The model included a fixed intercept α_0 and a random RKHS regression coefficient α_i for each subject; additive effects were omitted from the model, as individuals were unrelated, precluding separation of additive effects from residuals, in the absence of knowledge of variance components. The genetic value g_i of a subject was predicted as

$$
\hat{g}_i = \hat{\alpha}_0 + \mathbf{t}'_i(h)\hat{\boldsymbol{\alpha}},
$$

where $\hat{\boldsymbol{\alpha}}_0$ and $\hat{\alpha}$ were obtained by solving (20.36) for this model, using a Gaussian kernel at varying functions of h. The mean squared error of prediction (MSEP) of genetic value was calculated as

$$
\text{MSEP}(h, \sigma_e^2/\sigma_\alpha^2) = \sum_i (g_i - \hat{g}_i)^2, \tag{20.39}
$$

and a grid search was used to determine the values of h and $\sigma_e^2/\sigma_\alpha^2$ that minimized (20.39). To evaluate the performance of $\hat{g}_i$ as a predictor, another sample of 1000 individuals (prediction) was simulated, including genotypes, genotypic values and phenotypes. This was deemed preferable to doing prediction in the training sample, to reduce dependence between performance and $(h, \sigma_e^2/\sigma_\alpha^2)$, whose values were assessed in the training sample The genotypic values of the subjects in prediction were predicted, given their genotypes, using $\hat{g}_i$. Genotypic values were also predicted using a multiple linear regression (MLR) mixed model with a fixed intercept and random regression coefficients on the linear effects of genotypes.

When gene action was strictly additive, the two methods (each fitting $k = 100$ loci) had the same accuracy, indicating that RKHS performed well even when the parametric assumptions were valid. On the other hand, when inheritance was purely additive $\times$ additive, the parametric MLR was clearly outperformed by RKHS, irrespective of the number of loci fitted. An exception occurred at $k = 100$ and $N = 1000$; here, the two methods were completely inaccurate. However, when N was increased from 1000 to 5000, the accuracy of RKHS jumped to 0.85, whereas MLR remained inaccurate. The accuracy of RKHS decreased (with N held at 1000) when k increased. We attribute this behavior to the use of a Gaussian kernel when, in fact, covariates are discrete.

Our expectation is that the nonparametric function of marker genotypes, $g(\mathbf{x})$, would capture all possible forms of interaction, but without explicit modeling. The procedures should be particularly useful for highly dimensional regression, including the situation in which the number of SNP variables (p) exceeds amply the number of data points (n). Instead of performing a selection of a few 'significant' markers based on some *ad hoc* method, information on all molecular polymorphisms is employed, irrespective of the degree of colinearity. This is because the procedures should be insensitive to difficulties caused by colinearity, given the forms of the estimators, e.g. (20.28). It is assumed that the assignment of genotypes to individuals is unambiguous, i.e. $\mathbf{x}$ gives the genotypes for SNP markers free of error.

20.2.9 Computing Software

Because of the sheer size of animal breeding data sets, much effort has been devoted to making BLUP and REML computationally feasible, even in multivariate models. A comparison of packages is in Misztal (1998); see Hofer (1998) for additional discussion. The most widely used packages for mixed effects linear models are DFREML (Meyer, 1991), DMU (Jensen and Madsen, 1994), MTDFREML (Kriese *et al.*, 1994), VCE (Groeneveld, 1994) and ASREML (Gilmour and Thompson, 1998). Some packages include MCMC implementations (Van Tassel and Van Vleck, 1996; Groeneveld and Garcia Cortes, 1998; Janss, 1998). Janss and de Jong (1999) fitted an univariate mixed effects model with about 1.4 million location effects including about 700 000 additive genetic values (with a relationship matrix, $\mathbf{A}$, of corresponding order) to milk yield of Dutch Cattle, and used Gibbs sampling. They estimated the posterior distributions of heritability in a precise manner. This illustrates a situation where the MCMC machinery allows estimating a complete distribution, whereas deterministic likelihood computations are seemingly not feasible. Some software for nonlinear models, survival analysis and limited dependent variables is available, but it is far from general. An example is the SURVIVAL KIT, for analysis of survival models (Ducrocq and Sölkner, 1998), and the Danish DMU, which now has some finite mixture model capabilities. Animal breeders are increasingly using R, a free software environment for statistical computing and graphics (`http://www.r-project.org`).

20.3 FUTURE DEVELOPMENTS

Animal breeders have taken up new statistical ideas and technology at a fairly rapid pace. Some expected developments and issues are outlined below.

20.3.1   Model Development and Criticism

It is just not sensible to expect that all quantitative traits in animal breeding can be described suitably by a linear model with the 'herd-year + animal + permanent environment + residual' standard specification under Gaussian assumptions. Given the continued growth in computer power and algorithms, there is flexibility for fitting more realistic functional forms and distributions, as well as for challenging models stringently, this being an area that has not received enough attention. One relative measure of model adequacy, at least in a Bayesian framework, consists of assessing the posterior probabilities of each one in a set of competing models. This requires computing Bayes factors or using reversible jump methods to estimate the posterior probability distribution of the models. For example, Strandén and Gianola (1997) found strong evidence against a model with Gaussian errors. A specification assuming a t-distribution for the residuals, was at least 10^5 times more probable than its Gaussian counterpart; this would correspond to a value of about 23 in the scale of the likelihood ratio test. An interesting and comprehensive assessment of model for describing litter size in pigs is given by Sorensen and Waagepetersen (2003).

Analysis of residuals is an important diagnosis tool. For instance, in a multitier hierarchical model for longitudinal data (e.g. milk yield) one may examine the fit of different specifications at the level of the trajectory, at the level of variation of the parameters describing the trajectory and at the level of the different sub-populations embedded in the analysis. Strandén (1996) undertook a Bayesian analysis of residuals of models describing cross-sectional milk yield data. He examined the posterior distributions of residuals, detected outliers, and found how a robust distribution led to a better fit. Similarly, Rodriguez-Zas (1998) used MCMC for criticizing several longitudinal models for describing somatic cell scores in Holsteins, and detected several aberrant observations within individuals. Further, she looked at posterior distributions of Mahalanobis distance measures for multivariate outliers, this being done for detecting individuals whose vector of 'random regression' parameters depart from what the model predicts. This type of analysis extends naturally to discrete data in a Bayesian framework, so this may be in the agenda for further work.

20.3.2   Model Dimensionality

A second domain of interest is related to the danger of making too strong assumptions concerning the dimension of a model. Animal breeders would seem to believe that an analysis with a highly dimensional model is necessarily better, or that it provides a 'gold standard', than a model based on less ambitious assumptions. In the absence of knowledge about the state of nature, there is no reason to expect why a more highly parameterized model should provide the 'best' description of reality (Malécot, 1947). For example, suppose that one has traits $A, B, C, \ldots, Z$, and that a multiple-trait Gaussian model with as many dimensions as there are letters in the alphabet is fitted for predicting breeding values and estimating genetic parameters. This is equivalent to constructing a probability model of the type:

$$\Pr(A) \, x \, \Pr(B \mid A) \ldots x \quad \Pr(Z \mid Y, X, \ldots, B, A).$$

This can be seen as a ladder, but what happens if one or more of the steps are false? There may be good reason for believing that there is approximate Gaussianity at the

margins. Unfortunately, this is not a sufficient condition for defining a jointly Gaussian process, as all conditional distributions must be also Gaussian for this to hold. Strandén and Gianola (1997), e.g. using Bayes factors, found that univariate repeatability models (with either Gaussian or t-distributed errors) were much more plausible than bivariate models for describing first and second lactation milk yield in Ayrshire cattle. Even if one adopts a high-dimensional model, Rekaya *et al.* (1999) illustrate how a parsimonious parameterization of a genetic variance–covariance matrix can lead to dramatically more precise inferences about genetic correlations, relative to a standard multitrait approach. Factor analytic representations for reducing the number of parameters in a model are receiving an increased attention (De los Campos, 2005).

20.3.3 Robustification of Inference

There has been some work in fitting thick-tailed distributions (Strandén and Gianola, 1999; Rodriguez-Zas, 1998; Rodriguez-Zas, *et al.*, 1998; Rosa, 1998), Bayesian nonparametrics (Saama, 1999) and use of splines (White *et al.*, 1999). Concerning heavy-tailed distributions, Strandén (1996) and Strandén and Gianola (1999) described how univariate and multitrait mixed effects linear models could be extended to accommodate t-distributions, to attain a more robust analysis. Strandén and Gianola (1997) found that models with independent and identically distributed univariate or bivariate t-errors were more plausible than their Gaussian counterparts when describing milk yield in cattle. In this study, the posterior distribution of the degrees of freedom concentrated between values of 6 and 10, pointing away from the validity of the Gaussian assumption. On the other hand, Rodriguez-Zas (1998) found that the posterior means of the degrees of freedom ranged between 20 and 24 when using nonlinear 'random regressions' for describing somatic cell scores in Holstein cows. Here, Bayes factors were not decisive against the Gaussian assumption. Rosa (1998) pointed out that the t-models could be extended easily (at least from a Bayesian point of view) to accommodate lack of symmetry in the distribution of random effects. This had been suggested by Fernandez and Steel (1998a; 1998b) for the first tier of a hierarchical model. Rosa *et al.* (1998; 2003) fitted seven distributions (Gaussian; univariate and multivariate t; univariate and multivariate slash; univariate and multivariate contaminated normals) to birth weight data in mice. The Gaussian and the three multivariate robust distributions received the least degree of support. Within the three univariate robust distributions, the slash and the contaminated normal led to models that were between five and six times more probable (*a posteriori*) than the univariate-t. The Gaussian model was about 10^{-26} times less likely than the univariate contaminated normal. This suggests that analyses based on Gaussian assumptions are dangerous for *entire probabilistic inference*. While assuming normality (linearity) may not be cause serious problems from the point of view of point prediction of breeding values, it may be create difficulties when calculating probabilities of ordered events. For example, using Norwegian binary mastitis data, Heringstad *et al.* (1999, unpublished), estimated via MCMC in a Bayesian threshold model, the posterior probability that at least 10 of the sires out of the top 25 (posterior means) would be those with the largest transmitting abilities in the liability scale; the analysis involved 257 young sires and about 12 000 records. This type of probabilistic computation can be used to screen among models, choosing those that give the highest probability that the ranking of true values, corresponds to the rankings from the evaluation, given the data. Clearly, using a Gaussian assumption for $0 - 1$ data would lead to spurious

probabilities. Calculating the probability of correctly ordering random variables that are neither independent not identically distributed is an old problem in animal breeding (Henderson, 1973).

20.3.4   Inference Under Selection

Dealing with data derived from cryptic selection processes remains one of the biggest challenges in animal breeding. Hence, *developing selection models* is an area that should received more attention beyond *ad hoc* simulations conducted to see what happens under narrow conditions. For example, Gianola and Hill (1999, unpublished) derived best linear unbiased prediction under nor-optimal stabilizing selection. Inferences depend on knowledge of the optimum toward which the population is being moved, and of a sharpness parameter matrix. In principle, this parameter can be estimated by comparing data before and after selection. However, if one had the data observed before selection takes place, such selection is ignorable from the point of view of Bayesian or likelihood inference. As mentioned, Sorensen *et al.* (2001) developed an approach that allows inferring the trajectory of additive genetic variance (under an infinitesimal model) in the course of selection. Using simulated data, they showed that the posterior distribution of the additive genetic variance at any generation covered well the true value. Their Bayesian analysis was much more precise than regression of offspring on parents, this being a form of conditional likelihood inference that has been advocated for selected data. An obvious extension consists of studying the dynamics of the genetic correlation between traits, and this is straightforward.

20.3.5   Mixture Models

Finite mixture models, used in biology and in genetics since Pearson (1894),will probably play an increasing role in animal breeding. These models can uncover heterogeneity due to hidden structure or incorrect assumptions. Unknown loci with major effects can create 'bumps' in a phenotypic distribution, and this heterogeneity may be resolved by fitting a mixture, i.e. by calculating conditional probabilities that a datum is drawn from one of the several potential, yet unknown, genotypes. Concealed heterogeneity produces curious phenomena. For instance, the offspring–parent regression depends on the mixing proportions, and the genetic correlation between a 'mixture trait' and a Gaussian character is a function of the mixing proportions and of the ratio of genetic variances between mixture components. Ignoring this can give a misleading interpretation of the genetic structure of a population, and of expected response to selection when applied to a heterogeneous trait.

Many QTL detection procedures are based on ideas from mixture models, and inference about some quantitative genetic effects via finite mixture models may be warranted in practice. For example, consider mastitis, an inflammation of the mammary gland of cows and goats associated with bacterial infection. Genetic variation in susceptibility exists, and selection for increased resistance is feasible. However, recording of mastitis events is not routine in most nations, and milk somatic cell count (SCC) is used as a proxy in genetic evaluation of sires (via mixed effects linear models), because an elevated SCC is associated with mastitis. SCC is both an indicator of mastitis and a measure of response to infection It is not obvious how the SCC information should be treated optimally in genetic evaluation. It is reasonable to expect that SCC observations

taken on healthy and diseased animals display different distributions, which are 'hidden' in the absence of disease recording. Finite mixture models were suggested in this context by Detilleux and Leroy (2000), Ødegård *et al.* (2003; 2005) and Boettcher *et al.* (2005). Although software for running mixture models with random effects (Bayesian perspective) already exists, such as the Danish DMU, it not feasible to carry out a national genetic evaluation, at least at present. An option (perhaps crude) is to use the good old conditional posterior mode (also know as penalized likelihood estimation), dating back to the procedure leading to the discovery of the MME by Henderson. Assuming that the variance of the random effects is known, a stationary point of the density (conditional posterior mode) can be located by iterating with a reweighted system of mixed model equations, where the weights enter via a diagonal matrix containing the conditional probabilities that the records belong to the first component of the mixture, given the current values of the parameters. The additive genetic variance could be inferred using a pseudo-ML or pseudo-REML step, as in Foulley *et al.* (1987a). Formulae for updating the probabilities and the residual variances of the two components of the mixture are simple, but are not presented here. The approach can be made more complex by introducing a mixture for the random effects. Ødegård *et al.* (2005) gave a structure in which the probability of membership is modeling hierarchically, producing a prediction of breeding value for putative resistance to mastitis using SCC data; their computations are not (yet) feasible at the level of national genetic evaluation.

When dealing with counts, e.g. number of episodes of a disease, the number of observed 0's is often larger than what could be expected under some distribution, such as Poisson. In the context of disease, one can think of a population consisting of two components: a 'perfect one' (animals that never get the disease because they are resistant), and a 'liable' one, consisting of individuals that may get the disease. The probability of observing a zero is contributed by two sources. If the liable component is Poisson, then the mixture is called a *zero-inflated Poisson* (*ZIP*) model. More generally, the idea is to construct a hierarchical model where a transform (probit, logit) of the probability of perfection is governed by genetic effects, while a transform of the Poisson parameter is under genetic control as well. The two random effects might be correlated, as in a maternal effects model. Rodrigues-Motta (2006) and Rodrigues-Motta *et al.* (2006) have developed a fully Bayesian analysis via MCMC of a ZIP model, with application to number of mastitis episodes in Norwegian dairy cattle. If the model fits, it would be possible to select animals on the basis of the chance of being truly resistant to the disease, but one cannot rule out lack of exposure.

It is clear that the *statistical use of molecular information* in inferences about genetic values, and in resolving genetic complexity is an area of great importance. Höschele *et al.* (1997) present some developments. **Chapters 26** and **27**, as well as others in this book, deal with this problem.

Acknowledgments

REFERENCES

Anderson, R.L. and Bancroft, T.A. (1952). *Statistical Theory in Research*. McGraw-Hill, New York.

Bidanel, J.P. (1998). Benefits and limits of increasingly sophisticated models for genetic evaluation: the example of pig breeding. *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Vol. 25, Animal Genetics and Breeding Unit, Armidale, pp. 577–584.

Boettcher, P.J., Moroni, P., Pisoni, G. and Gianola, D. (2005). Application of a finite mixture model to somatic cell scores of Italian goats. *Journal of Dairy Science* **88**, 2209–2216.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading.

Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.

Chang, Y.M. (1999). Bayesian analysis of lactation curves in dairy sheep. Master of Science Thesis, University of Wisconsin-Madison.

Cockerham, C.C. (1954). An extension of the concept of partitioning hereditary variance for the analysis of covariances among relatives when epistasis is present. *Genetics* **39**, 859–882.

Corbeil, R.R. and Searle, S.R. (1976). A comparison of variance component estimators. *Biometrics* **32**, 779–791.

Curnow, R.N. (1961). The estimation of repatability and heritability from records subject to culling. *Biometrics* **17**, 553–566.

Damgaard, L.H. and Korsgaard, I.R. (2006). A bivariate quantitative genetic model for a linear Gaussian trait and a survival trait. *Genetics, Selection, Evolution* **38**, 45–64.

Dekkers, J.C.M. and Hospital, F. (2002). The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics* **3**, 22–32.

De los Campos, G. (2005). Structural equation models with applications in quantitative genetics. M. S. Thesis, University of Wisconsin-Madison.

Dempfle, L. (1977). Relation entre BLUP (best linear unbiased prediction) et Estimateurs Bayesiens. *Annales de Génetique et de Séction Animale* **9**, 27–32.

Dempfle, L. (1982). Problems in estimation of breeding values. *Proceedings of the Second World Congress on Genetics Applied to Livestock Production*. Neografis, Madrid, pp. 104–118.

Detilleux, J. and Leroy, P.L. (2000). Application of a mixed normal mixture model to the estimation of mastitis-related parameters. *Journal of Dairy Science* **83**, 2341–2349.

Ducrocq, V. (1990). Estimation of genetic parameters arising in nonlinear models. *Proceedings of the Fourth World Congress on Genetics Applied to Livestock Production*, Vol. XII, Joyce Darling, Penicuik, pp. 419–428.

Ducrocq, V. and Casella, G. (1996). Bayesian analysis of mixed survival models. *Genetics, Selection, Evolution* **28**, 505–529.

Ducrocq, V. and Sölkner, J. (1998). The survival kit: a Fortran package for the analysis of survival data. *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Vol. 22, Animal Genetics and Breeding Unit, Armidale, pp. 51–52.

Falconer, D.S. (1965). Maternal effects and selection response. In *Genetics Today*, S.J. Geerts, ed. Pergamon, Oxford, pp. 763–774.

Falconer, D.S. and Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*. Longman, Essex.

Fernandez, C. and Steel, M.F.J. (1998a). On Bayesian modelling of fat tails and skewness. *Journal of the American Statistical Association* **93**, 359–371.

Fernandez, C. and Steel, M.F.J. (1998b). On the dangers of modelling through continuous distributions: a Bayesian perspective. In *Bayesian Statistics*, Vol. 6, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds. Oxford University Press, Oxford, pp. 1–19.

Fernando, R.L. and Gianola, D. (1986). Optimal properties of the conditional mean as a selection criterion. *Theoretical and Applied Genetics* **72**, 822–825.

Fernando, R.L. and Gianola, D. (1990). Statistical inferences in populations undergoing selection or non-random mating. In *Advances in Statistical Methods for Genetic Improvement of Livestock*, D. Gianola and K. Hammond, eds. Springer-Verlag, Berlin, pp. 437–453.

Fernando, R.L. and Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. *Genetics, Selection, Evolution* **21**, 467–477.

Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Royal Society (Edinburgh) Transactions* **52**, 399–433.

Foulley, J.L. (1993). A simple argument showing how to derive restricted maximum likelihood. *Journal of Dairy Sciences* **76**, 2320–2324.

Foulley, J.L., Gianola, D. and Im, S. (1990). Genetic evaluation for discrete polygenic traits in animal breeding. In *Advances in Statistical Methods for Genetic Improvement of Livestock*, D. Gianola and K. Hammond, eds. Springer-Verlag, Berlin, pp. 361–409.

Foulley, J.L., Gianola, D. and Thompson, R. (1983). Prediction of genetic merit from data on categorical and quantitative variates with an application to calving difficulty, birth weight and pelvic opening. *Genetics, Selection, Evolution* **15**, 407–424.

Foulley, J.L., Im, S., Gianola, D. and Höschele, I. (1987a). Empirical Bayes estimation of parameters for *n* polygenic binary traits. *Genetics, Selection, Evolution* **19**, 197–224.

Foulley, J.L., Gianola, D. and Im, S. (1987b). Genetic evaluation for traits distributed as Poisson-bionomial with reference to reproductive traits. *Theoretical and Applied Genetics* **73**, 870–877.

Foulley, J.L. and Quaas, R.L. (1994). Statistical analysis of heterogeneous variances in Gaussian linear mixed models. *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production*, Vol. 18, University of Guelph, Guelph, pp. 341–348.

Fox, J. (1984). *Linear Statistical Models and Related Methods*. John Wiley & Sons, New York.

Gianola, D. (2006). Statistics in animal breeding: angels and demons In *Proceedings of the Eight World Congress of Genetics Applied to Livestock Production, CD Paper 00-03*, Belo Horizonte: Insituto Prociencia, Brasil.

Gianola, D. and Fernando, R.L. (1986). Bayesian methods in animal breeding theory. *Journal of Animal Science* **63**, 217–244.

Gianola, D., Fernando, R.L., Im, S. and Foulley, J.L. (1989). Likelihood estimation of quantitative genetic parameters when selection occurs: models and problems. *Genome* **31**, 768–777.

Gianola, D., Fernando, R.L. and Stella, A. (2006). Genomic assisted prediction of genetic value with semi-parametric procedures. *Genetics* **173**, 1761–1776.

Gianola, D. and Foulley, J.L. (1983). Sire evaluation for ordered categorical data with a threshold model. *Genetics, Selection, Evolution* **15**, 201–224.

Gianola, D., Foulley, J.L. and Fernando, R.L. (1986). Prediction of breeding values when variances are not Known. *Proceedings of the Third World Congress on Genetics Applied to Livestock Production*, Vol. XII, Agricultural Communications, University of Nebraska, Lincoln, pp. 356–370.

Gianola, D. and Goffinet, B. (1982). Sire evaluation with best linear unbiased predictors. *Biometrics* **38**, 1085–1088.

Gianola, D. and Hammond, K., eds. (1990). *Advances in Statistical Methods for Genetic Improvement of Livestock*, Springer-Verlag, Berlin.

Gianola, D. and Hill, W.G. (1999). Selection for an intermediate optimum and best linear unbiased prediction, (In Preparation).

Gianola, D. and Kachman, S.D. (1983). Prediction of breeding value in situations with nonlinear structure: categorical responses, growth functions and lactation curves. In *34th Annual Meeting, European Association of Animal Production*, Madrid. Summaries, p. 172.

Gianola, D., Perez-Enciso, M. and Toro, M.A. (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics* **163**, 347–365.

Gianola, D., Piles, M.M. and Blasco, A. (1999). Bayesian inference about parameters of a longitudinal trajectory when selection operates on a correlated trait. *Proceedings of International Symposium in Animal Breeding and Genetics*. Universidade Federal de Viçosa, Brasil, pp. 101–132.

Gianola, D., Rodriguez-Zas, S. and Shook, G.E. (1994). The Gibbs sampler in the animal model: a primer. In *Seminaire Modele Animal*, J.L. Foulley and M. Molenat, eds. INRA, Jouy-en-Josas, pp. 47–56.

Gianola, D. and Sorensen, D. (2004). Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics* **176**, 1407–1424.

Gilmour, A.R., Anderson, R.D. and Rae, A.L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**, 593–599.

Gilmour, A.R. and Thompson, R. (1998). Modelling variance parameters in ASREML for repeated measures data. *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Vol. 27, Animal Genetics and Breeding Unit, Armidale, pp. 453–457.

Golberger, A.S. (1992). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association* **57**, 369–375.

Golub, T.R., Slonim, D., Tamayo, P., Huard, C., Gasenbeek, M., Mesiro, J., Coller, H., Loh, M., Downing, J., Caliguri, M. Bloomfield, C. and Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Groeneveld, E. (1994). VCE- A multivariate multimodel REML (Co)variance component estimation package. *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production*, Vol. 22, University of Guelph, Guelph, pp. 47–48.

Groeneveld, E. and Garcia Cortes, L.A. (1998). VCE 4.0: A (Co)variance component package for frequentists and BAYESIANS. *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Vol. 27, Animal Genetics and Breeding Unit, Armidale, pp. 455–456.

Guo, S.F. (1999). Application of survival and censored linear models to the analysis of herdlife and lifetime prolificacy in landrace sows Master of Science Thesis, University of Wisconsin-Madison.

Guo, S.F. (2006). Statistical models for analysis of herd life and prolificacy traits in swine. Ph.D. Thesis, University of Wisconsin-Madison.

Hartley, H.O. and Rao, J.N.K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93–108.

Harvey, W.R. (1960). *Least-squares analysis of data with unequal subclass numbers*. United States Department of Agriculture, Agricultural Research Service, Washington. Bulletin 20-8.

Harvey, W.R. (1970). Estimation of variance and covariance components in the mixed model. *Biometrics* **26**, 485–504.

Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–385.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–340.

Harville, D.A. and Callanan, T.P. (1990). Computational aspects of likelihood-based inference for variance components. In *Advances in Statistical Methods for Genetic Improvement of Livestock*, D. Gianola and K. Hammond, eds. Springer-Verlag, Heidelberg, pp. 136–176.

Harville, D.A. and Carriquiry, A.L. (1992). Classical and Bayesian prediction as applied to an unbalanced mixed linear model. *Biometrics* **48**, 987–1003.

Harville, D.A. and Mee, R.W. (1984). A mixed model procedure for analyzing ordered categorical data. *Biometrics* **40**, 393–408.

Hayes, B., Laerdahl, J., Lien, D., Adzhubei, A. and Høyheim, B. (2004). Large scale discovery of single nucleotide polymorphism (SNP) markers in Atlantic Salmon (Salmo salar), AKVAFORSK, Institute of Aquaculture Research. www.mabit.no/pdf/hayes.pdf.

Hazel, L.N. (1943). The genetic basis for constructing selection indexes. *Genetics* **28**, 476–490.

Henderson, C.R. (1950). Specific and general combining ability. In *Heterosis*, J.W. Gowen, ed. Iowa State College Press, Ames, pp. 352–370.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226–252.

Henderson, C.R. (1963). Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding*, W.D. Hanson and H.F. Robinson, eds. National Academy of Sciences-National Research Council, Washington, DC, pp. 141–163, Publication 992.

Henderson, C.R. (1973). Sire evaluation and genetic trends. *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush*. American Society of Animal Science and the American Dairy Science Association, Champaign, pp. 10–41.

Henderson, C.R. (1974). General flexibility of linear model techniques for sire evaluation. *Journal of Dairy Science* **57**, 963.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–449.

Henderson, C.R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32**, 69–83.

Henderson, C.R. (1977). Prediction of future records. In *Proceedings of the International Conference on Quantitative Genetics*, E. Pollak, O. Kempthorne and T.B. Bailey, eds. The Iowa State University Press, Ames, pp. 615–638.

Henderson, C.R. (1984). *Application of Linear Models in Animal Breeding*. University of Guelph, Guelph.

Henderson, C.R. (1985). Best linear unbiased prediction of non-additive genetic merits in noninbred populations. *Journal of Animal Science* **60**, 111–117.

Henderson, C.R. (1988). Progress in statistical methods applied to quantitative genetics since 1976. In *Proceedings of the Second International Conference on Quantitative Genetics*, B.S. Weir, E.J. Eisen, M.M. Goodman and G. Namkoong, eds. Sinauer, Sunderland, pp. 85–90.

Henderson, C.R., Searle, S.R., Kempthorne, O. and vonKrosigk, C.M. (1959). Estimation of enviromental and genetic trends from records subject to culling. *Biometrics* **15**, 192–218.

Hill, W.G. (1974). Heritabilities: estimation problems and the present state of information. *Proceedings of the First World Congress on Genetics Applied to Livestock Production*, Vol. I, Graficas Orbe, Madrid, pp. 343–351.

Hill, W.G. (1980). Design of quantitative genetic selection experiments. In *Selection Experiments in Laboratory and Domestic Animals*, A. Robertson, ed. Commonwealth Agricultural Bureaux, Slough.

Hofer, A. (1998). Variance component estimation in animal breeding: a review. *Journal of Animal Breeding and Genetics* **115**, 247–265.

Höschele, I., Foulley, J.L., Colleau, J.J. and Gianola, D. (1986). Genetic evaluation for multiple binary responses. *Genetics, Selection, Evolution* **18**, 299–320.

Höschele, I., Uimari, P., Grignola, F.E., Zhang, Q. and Gage, K. (1997). Advances in statistical methods to map quantitative trait loci in outbred populations. *Genetics* **147**, 1445–1457.

Im, S., Fernando, R.L. and Gianola, D. (1989). Likelihood inferences in animal breeding under selection: a missing data theory viewpoint. *Genetics, Selection, Evolution* **21**, 399–414.

Janss, L.L.G. (1998). MaGGic: a package of subroutines for genetic analysis with Gibbs sampling. *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Vol. 27, Animal Genetics and Breeding Unit, Armidale, pp. 459–460.

Janss, L.L.G. and de Jong, J. (1999). MCMC based estimation of variance components in a very large dairy cattle data set. *Computational Cattle Breeding 99*. MTT, Helsinki.

Jensen, J. and Madsen, P. (1994). DMU: a package for the analysis of multivariate mixed models. *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production*, Vol. 22, University of Guelph, Guelph, pp. 45–46.

Kachman, S.D. and Gianola, D. (1984). A Bayesian estimator of variance and covariance components in nonlinear growth models. *Journal of Animal Science* **59**(Suppl. 1), 176.

Kackar, R.N. and Harville, D.A. (1981). unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics A: Theory and Methods* **10**, 1249–1261.

Kempthorne, O. (1954). The correlation between relatives in a random mating population. *Royal Society (London) Proceedings* **B143**, 103–113.

Kennedy, B.W. and Schaeffer, L.R. (1990). Reproductive technology and genetic evaluation. In *Advances in Statistical Methods for Genetic Improvement of Livestock*, D. Gianola and K. Hammond, eds. Springer-Verlag, Heidelberg, pp. 507–532.

Kirkpatrick, M. and Lofsvold, D. (1989). The evolution of growth trajectories and other complex quantittative characters. *Genome* **31**, 778–783.

Koerkhuis, A.N.M. and Thompson, R. (1997). Models to estimate maternal effects for juvenile body weight in broiler chickens. *Genetics, Selection, Evolution* **29**, 225–249.

Korsgaard, I.R., Lund, M.S., Sorensen, D., Gianola, D., Madsen, P. and Jensen, J. (1999). Multivariate Bayesian analysis of Gaussian, right-censored Gaussian, ordered categorical and binary traits in animal breeding. *Computational Cattle Breeding 99*. MTT, Helsinki.

Kriese, L.A., Boldman, K.G., Van Vleck, L.D. and Kachman, S.D. (1994). A flexible set of programs to estimate (co)variances for messy multiple trait animal models using derivative free REML and sparse matrix techniques. *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production*, Vol. 22, University of Guelph, Guelph, pp. 43–44.

Laird, N.M. (1990). Analysis of linear and nonlinear growth models with random parameters. In *Advances in Statistical Methods for Genetic Improvement of Livestock*, D. Gianola and K. Hammond, eds. Springer-Verlag, Heidelberg, pp. 329–343.

LaMotte, L.R. (1973). Quadratic estimation of variance components. *Biometrics* **32**, 793–804.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society* **B58**, 619–678.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society* **B34**, 1–41.

Lindstrom, M.J. and Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673–687.

Lush, J.L. (1931). The number of daughters necessary to prove a sire. *Journal of Dairy Science* **14**, 209–220.

Malécot, G. (1947). Statistical methods and the subjective basis of scientific knowledge. *Genetics, Selection, Evolution* **31**, 269–298.

Malécot, G. (1948). *Les Mathématiques de L'Hérédité*. Masson et Cie, Paris.

Mallick, B.K., Ghosh, D. and Ghosh, M. (2005). Bayesian classification of tumours by using gene expression data. *Journal of the Royal Statistical Society* **B67**, 219–234.

Meyer, K. (1990). Present status of knowledge about statistical procedures and algorithms to estimate variance and covariance components. *Proceedings of the Fourth World Congress on Genetics Applied to Livestock Production*, Vol. XII, Joyce Darling, Penicuik, pp. 407–418.

Meyer, K. (1991). Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. *Genetics, Selection, Evolution* **23**, 67–83.

Meyer, K. (1998). Modeling repeated records: covariance functions and random regression models to analise animal breeding data. *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Vol. 25, Animal Genetics and Breeding Unit, Armidale, pp. 517–520.

Misztal, I. (1998). Comparison of software packages in animal breeding. *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Vol. 22, Animal Genetics and Breeding Unit, Armidale, pp. 3–10.

Misztal, I. and Gianola, D. (1987). Indirect solution of mixed model equations. *Journal of Dairy Science* **70**, 716–723.

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001). Is it possible to predict the total genetic merit under a very dense marker map? *Genetics* **157**, 1819–1829.

Nelder, J.A. (1996). Hierarchical GLMs In *Proceedings of the XVIIIth International Biometrics, Invited Papers*, Amsterdam, The Netherlands, 137–139.

Ødegård, J., Jensen, J., Madsen, P., Gianola, D., Klemetsdal, G. and Heringstad, B. (2003). Mixture models for detection of mastitis in dairy cattle using test-day somatic cell scores: a Bayesian approach via Gibbs sampling. *Journal of Dairy Science* **86**, 3694–3703.

Ødegård, J., Jensen, J., Madsen, P., Gianola, D., Klemetsdal, G. and Heringstad, B. (2005). A Bayesian liability-normal mixture model for analysis of a continuous mastitis-related trait. *Journal of Dairy Science* **88**, 2652–2659.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A* **185**, 71–110.

Rao, C.R. (1971). Estimation of variance and covariance components: MINQUE theory. *Journal of Multivariate Analysis* **1**, 257–275.

Rekaya, R. (1997). Analisis Bayesiano de Datos de Producción en los Dias del Control Para la Selección de Caracteres Lecheros Tesis Doctoral, Universidad Politecnica de Madrid.

Rekaya, R., Weigel, K.A., and Gianola, D. (1999). Bayesian estimation of a structural model for genetic covariances for milk yield in five regions of the USA In *European Association for Animal Production. 50th Annual Meeting*, Zurich, Switzerland, p. 7.

Robert, C. (1996). *Méthodes de Monte Carlo par Châines de Markov*. Economica, Paris.

Robertson, A. (1955). Prediction equations in quantitative genetics. *Biometrics* **11**, 95–98.

Rodrigues-Motta, M. (2006). Zero-inflated Poisson models for quantitative genetic analysis of count data with applications to mastitis in dairy cows. Ph. D. Thesis, University of Wisconsin-Madison.

Rodrigues-Motta, M., Gianola, D., Chang, Y.M. and Heringstad, B. (2006). A zero-inflated Poisson model for genetic analysis of number of mastitis cases in Norwegian red cows In Proceedings of the 8th World Congress of Genetics Applied to Livestock Production, CD Paper 26-05. Belo Horizonte, Instituto Prociencia, Brasil.

Rodriguez-Zas, S.L. (1998). Bayesian analysis of somatic cell score lactation patterns in Holstein cows using nonlinear mixed effects models. Ph.D. Thesis, University of Wisconsin-Madison.

Rodriguez-Zas, S.L., Gianola, D. and Shook, G.E. (1998). Bayesian analysis of nonlinear mixed effects models for somatic cell score lactation patterns in holsteins. *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Vol. 25, Animal Genetics and Breeding Unit, Armidale, pp. 497–500.

Ronningen, K. (1971). Some properties of the selection index derived by Henderson's mixed model method. *Zeitschrift für Tierzuchtung und Züchtungsbiologie* **8**, 186–193.

Rosa, G.J.M. (1998). Análise Bayesiana de Moldeos Lineares Mistos Robustos via Amostrador de Gibbs. Dr. Agr. Thesis, Escola Superior de Agricultura Luiz de Queiroz, Piracicaba: S∼ao Paulo.

Rosa, G.J.M., Gianola, D. and Padovani, C.R. (1998). Bayesian analysis of some robust mixed linear models with an application to birth weight in mice In Sixth Valencia International Meeting on Bayesian Statistics Abstracts Alcossebre, Spain, p. 122.

Rosa, G.J.M., Padovani, C.R. and Gianola, D. (2003). Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. *Biometrical Journal* **45**, 573–590.

Saama, P. (1999). Posterior exploration of Markov chains in a bayesian analysis of discrete finite mixture models. *Computational Cattle Breeding 99*. MTT, Helsinki.

Schaeffer, L.R. (1985). Model for international evaluation of dairy sires. *Livestock Production Science* **12**, 105–115.

Schaeffer, L.R. and Kennedy, B.W. (1986). Computing solutions to mixed model equations. *Proceedings of the Third World Congress on Genetics Applied to Livestock Production*, Vol. XII, Agricultural Communications, University of Nebraska, Lincoln, pp. 382–393.

Searle, S.R. (1968). Another look at Henderson's methods of estimating variance components. *Biometrics* **24**, 749–778.

Searle, S.R. (1971). Topics in variance component estimation. *Biometrics* **27**, 1–76.

Searle, S.R. (1974). Prediction, mixed models and variance components. In *Reliability and Biometry*, F. Proschan and R.J. Serfling, eds. Society for Industrial and Applied Mathematics, Philadelphia.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*. John Wiley & Sons, New York.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Smith, F.H. (1936). A discriminant function for plant selection. *Annals of Eugenics* **7**, 240–250.

Smith, S.P. and Allaire, F.R. (1986). Analysis of failure times measured on dairy cows: theoretical considerations in animal breeding. *Journal of Dairy Science* **69**, 217–227.

Soller, M. and Beckmann, J.S. (1982). Restriction fragment length polymorphisms and genetic improvement. *Proceedings of the 2nd World Congress on Genetics Applied to Livestock Production* **6**, 396–404.

Sorensen, D., Andersen, S. Gianola, D. and Korsgaard I. (1995). Bayesian inference in threshold models using Gibbs sampling. *Genetics, Selection, Evolution* **27**, 229–249.

Sorensen, D.A., Andersen, S., Jensen, J., Wang, C.S. and Gianola, D. (1994). Inferences about genetic parameters using the Gibbs sampler. *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production*, Vol. 18, University of Guelph, Guelph, pp. 321–328.

Sorensen, D.A., Fernando, R.L. and Gianola, D. (2001). Inferring the trajectory of genetic variance in the course of artificial selection. *Genetical Research* **77**, 83–94.

Sorensen, D. and Gianola, D. (2002). *Likelihood, Bayesian and MCMC methods in quantitative genetics*. Springer, New York.

Sorensen, D.A., Gianola, D. and Korsgaard, I.R. (1998). Bayesian mixed effects model analysis of a censored normal distribution with animal breeding applications. *Acta Agriculturae Scandinavica* **A48**, 222–229.

Sorensen, D. and Waagepetersen, R. (2003). Normal linear models with genetically structured residual variance heterogeneity: a case study. *Genetical Research* **82**, 207–222.

Sorensen, D.A., Wang, C.S., Jensen, J. and Gianola, D. (1994). Bayesian analysis of genetic change due to selection using Gibbs sampling. *Genetics, Selection, Evolution* **26**, 333–360.

Strandén, I. (1996). Robust mixed effects linear models with t-distributions and application to dairy cattle breeding. Ph.D. Thesis, University of Wisconsin-Madison.

Strandén, I. and Gianola, D. (1997). Gaussian versus student-t mixed effects linear models for milk yield in Ayrshire cattle. European Association for Animal Production. 48th Annual Meeting, Vienna, Austria, pp. 16.

Strandén, I. and Gianola, D. (1999). Mixed effects linear models with t-distributions for quantitative genetic analysis: a Bayesian approach. *Genetics, Selection, Evolution* **31**, 25–42.

Tempelman, R.J. and Firat, M.Z. (1998). Beyond the linear mixed model: perceived versus real benefit. *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Vol. 25, Animal Genetics and Breeding Unit, Armidale, pp. 605–612.

Tempelman, R.J. and Gianola, D. (1996). A mixed effects model for overdispersed count data in animal breeding. *Biometrics* **52**, 265–279.

Tempelman, R.J. and Gianola, D. (1999). Genetic analysis of fertility in dairy cattle using negative binomial mixed models. *Journal of Dairy Science* **82**, 1834–1847.

Thompson, W.A. (1962). The problem of negative estimates of variance components. *Annals of Mathematical Statistics* **33**, 273–289.

Thompson, R. (1977). Estimation of quantitative genetic parameters, *Proceedings of the International Conference on Quantitative Genetics*, 639–657 E. Pollak, O. Kempthorne and T.B. Bailey, eds. The Iowa State University Press.

Thompson, R. (1979). Sire evaluation. *Biometrics* **35**, 339–353.

Thompson, R. (1982). Methods of estimation of genetic parameters. *Proceedings of the Second World Congress on Genetics Applied to Livestock Production*, Vol. V, Neografis, Madrid, pp. 95–103.

Van Tassel, C.P. and Van Vleck, L.D. (1996). Multiple-trait Gibbs sampler for animal models; flexible programs for Bayesian and likelihood based (co)variance component inference. *Journal of Animal Science* **74**, 2586–2597.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GAVC. In *Advances in Kernel Methods*, B. SCHÖLKOPF, C. BURGES and A. SMOLA, eds. MIT Press, Cambridge, pp. 68–88.

Wang, C.S. (1998). Implementation issues in Bayesian analysis in animal breeding. *Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production*, Vol. 25, Animal Genetics and Breeding Unit, Armidale, pp. 481–488.

Wang, C.S., Rutledge, J.J. and Gianola, D. (1993). Marginal inference about variance components in a mixed linear model using Gibbs sampling. *Genetics, Selection, Evolution* **25**, 41–62.

Wang, C.S., Rutledge, J.J. and Gianola, D. (1994a). Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genetics, Selection, Evolution* **26**, 91–115.

Wang, C.S., Gianola, D., Sorensen, D.A., Jensen, J., Christensen, A. and Rutledge, J.J. (1994b). Response to selection for litter size in Danish landrace pigs: a Bayesian analysis. *Theoretical and Applied Genetics* **88**, 220–230.

White, I.M.S., Thompson, R. and Brotherston, R. (1999). Genetic and environmental smoothing of lactation curves with cubic splines. *Journal of Dairy Science* **82**, 632–638.

Willham, R.L. (1963). The covariance between relatives for characters composed of components contributed by related individuals. *Genetics* **19**, 18–27.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika* **80**, 791–795.

Wong, G., Liu, B., Andersson, L., Groenen, M., Hunt, H.D., Cheng, H.H (2004). A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* **432**, 717–722.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20**, 557–585.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**, 789–801.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons, Inc., New York.

Zhang, Q., Boichard, D., Hoeschele, I., Ernst, C., Eggen, A., Murkve, B., Pfister-Genskow, M., Witte, L.A., Grignola, F.E., Uimari, P., Thaller, G. and Bishop, M.D. (1998). Mapping quantitative trait loci for milk production and health of dairy cattle in a large outbred pedigree. *Genetics* **149**, 1959–1973.

21

Marker-assisted Selection and Introgression

L. Moreau

INRA, UMR de Génétique Végétale, Ferme du Moulon, France

F. Hospital

INRA, Université Paris Sud, Orsay, France

and

J. Whittaker

Department of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK

Good maps of molecular markers now exist for many species. In this chapter, we discuss the statistical methodology that has been developed to facilitate the exploitation of these maps in commercial breeding programs. Methods appropriate both for populations derived from inbred line crosses and outbred populations, particularly dairy cattle populations, are considered, and the possible utility of such methods discussed. There is a consensus that incorporating marker information into breeding programs can increase selection response, but the value of such schemes once marker-typing costs are allowed for is less clear-cut.

21.1 INTRODUCTION

A great deal of effort has been put in recent years to develop ways in which molecular marker information can be exploited in identifying, locating and estimating the effect of loci underlying quantitative traits (*quantitative trait loci* (*QTLs*)): see **Chapter 18** and **Chapter 19** for reviews of this work. One goal of this work is to answer fundamental scientific questions regarding, for instance, the number of loci affecting quantitative traits and the effect of these loci on the trait, as a step toward dissecting the genetic

architecture of complex traits. Another is that the identification of QTLs raises the possibility of exploiting these loci directly in breeding programs, and this is potentially of great economic value. Discussion of the possible uses of molecular marker information to accomplish this sort of goal goes back at least to Neimann-Sorenson and Robertson (1961), but has been given renewed impetus in recent years by rapid advances in molecular biology. There are a number of ways in which marker information might be used for these purposes, but we shall consider two broad areas here.

Firstly, we may have identified a gene with desirable properties in a *donor* population, which we want to move into a *recipient* population; typically, the recipient is a commercial, elite line, while the donor is not, and the donor is therefore inferior at most loci to the recipient. We therefore wish to produce a population in which the desired allele is fixed or very common but the proportion of donor genome is low. This has become known as *introgression*, or when marker information is used, *marker-assisted introgression* (MAI) and is typically accomplished by way of repeated backcrosses.

Secondly, suppose we have a population in which one or, more usually, many QTLs are segregating. We wish to increase the value of some trait in subsequent generations by selecting certain individuals from the population to form the next generation. Traditionally, this has been done solely on the basis of phenotypic information on individuals and their relatives, but incorporating information on the marker genotypes of individuals in *marker-assisted selection* (MAS) has the potential to improve the rate of progress.

The above division is not entirely clear-cut, but seems to have become reasonably widely accepted in the literature and forms a convenient basis for the present chapter. In both cases, different approaches are required for inbred and outbred populations; in each case, we shall begin by considering populations resulting from inbred line crosses before tackling the more complex problems raised by outbred populations.

21.2   MARKER-ASSISTED SELECTION: INBRED LINE CROSSES

21.2.1   Lande and Thompson's Formula

The seminal paper here is Lande and Thompson (1990), and the description below in the main follows their development. The key idea is that crossing two inbred lines generates linkage disequilibrium between markers and QTLs; this is exploited by the QTL mapping methods for inbred lines described in **Chapter 18**. However, Lande and Thompson (1990) pointed out that it is not necessary to map the QTL if our aim is to perform MAS. We need only identify the markers that are correlated with the trait or traits of interest and estimate the correlation between each of these selected markers and the trait; we can then base our selection criteria upon this marker information. For Gaussian traits, this is most simply done by regression of phenotype on marker-type.

More formally, consider an F_2 population derived from a cross between two inbred lines, with each line assumed homozygous for different alleles at all loci under consideration. We label the alleles at the kth QTL in the first line Q_k, and the alleles at the jth marker locus M_j. The corresponding alleles in the second line are labelled q_k and m_j. For each individual in the F_2 population, we know the phenotype y and the number of M_j alleles at the jth marker locus, x_j, where $x_j = 1$ if the individual has genotype $M_j M_j$, $x_j = 0$ if the individual has genotype $M_j m_j$ and $x_j = -1$ if the individual has

genotype $m_j m_j$. Supposing there are l markers, the marker genotype of the ith individual is thus described by $\mathbf{x} = (x_1, x_2, \ldots, x_l)$. To simplify the discussion, we shall make the additional assumption that QTL effects combine additively both within and between loci, i.e. the *genetic value z* is given by

$$z = \sum_{k=1}^{n} a_k g_k, \tag{21.1}$$

where n is the number of QTLs, a_k is the effect of the kth QTL and g_k the number of Q_k alleles at the kth QTL. We shall also assume that y is related to the genetic value by

$$y = z + \varepsilon, \tag{21.2}$$

where $\varepsilon \sim N(0, \sigma_e^2)$.

We wish to identify the individuals that should be selected to produce the next generation. Typically, this is done indirectly by estimating the genetic value of the individuals in the population and choosing individuals to reproduce based on these estimates in such a way as to maximise selection response over the desired period. One obvious approach is to regress phenotype on marker-type, i.e. to use the estimator $\hat{z}_s$, where

$$\hat{z}_s = \sum_{j \in \mathcal{A}} \hat{a}_j x_j \tag{21.3}$$

and $\mathcal{A}$ is the set of markers for which effects $\hat{a}_j$ have been fitted. The predicted values $\hat{z}_s$ are often known as the *molecular scores*. In principle, fitting this model by maximum likelihood rather than least squares might be preferred because, for a finite number of QTLs, the errors in the above model follow a mixture of Gaussians rather than a Gaussian distribution (see **Chapter 18**, for more on mixture models and QTLs), but, in practice, the difference is barely noticeable and least-squares estimation has considerable computational advantages (Martínez and Curnow, 1992; Haley and Knott, 1992). We thus have a standard linear regression prediction problem; the only difficult element of this is the choice of $\mathcal{A}$. The problem of choosing a linear model so as to trade off the variance and bias of the estimator of interest (here $\hat{z}$), which increase and decrease respectively as the number of variables included in the model increases, and thus to minimise prediction error, has been much discussed in the statistical literature (Miller, 1990). Several regression techniques and criteria for declaring markers as significant can be used and might influence MAS efficiency. See below for a discussion on this point.

We could base selection on $\hat{z}_s$ alone, but as $\hat{z}_s$ only explains part of the genetic variation, Lande and Thompson (1990) suggested combining it with y using the usual theory on selection indices for multiple correlated traits (e.g. Bulmer, 1980) to give

$$\hat{z} = b_0 y + b_1 \hat{z}_s, \tag{21.4}$$

with

$$\mathbf{b} = \mathbf{P}^{-1} \mathbf{G}, \tag{21.5}$$

where

$$\mathbf{P} = \begin{bmatrix} \text{var}(y) & \text{cov}(y; \hat{z}_s) \\ \text{cov}(y; \hat{z}_s) & \text{var}(\hat{z}_s) \end{bmatrix} \tag{21.6}$$

and

$$\mathbf{G} = (\text{cov}(z; y), \text{cov}(z; \hat{z}_s))^{\mathrm{T}}$$

$$= (\text{var}(z), \text{cov}(z; \hat{z}_s))^{\mathrm{T}}. \tag{21.7}$$

For this to be of use, we need estimates of $\text{var}(z)$ and $\text{cov}(z; \hat{z}_s)$. Some authors (e.g. Lande and Thompson, 1990; Gimelfarb and Lande, 1994a) have suggested estimating $\text{cov}(z; \hat{z}_s)$ by $\text{cov}(y; \hat{z}_s)$, arguing that all the phenotypic variance explained by the markers is by definition genetic. With this assumption, $\text{cov}(z; \hat{z}_s) = \text{cov}(y; \hat{z}_s) = \text{var}(\hat{z}_s)$. Hence the weighted coefficients are equal to

$$\mathbf{b} = \begin{pmatrix} b_y \\ b_m \end{pmatrix} = \begin{pmatrix} \frac{\text{var}(z) - \text{var}(\hat{z}_s)}{\text{var}(y) - \text{var}(\hat{z}_s)} \\ \frac{\text{var}(y) - \text{var}(z)}{\text{var}(y) - \text{var}(\hat{z}_s)} \end{pmatrix} = \begin{pmatrix} \frac{h^2 - R^2}{1 - R^2} \\ \frac{1 - h^2}{1 - R^2} \end{pmatrix}, \tag{21.8}$$

where $h^2 = \text{var}(z)/\text{var}(y)$ is equal to the trait heritability and R^2 is the percentage of phenotypic variance explained by the markers (i.e. the R^2 of the regression of phenotype on marker-type). Clearly, this assumption cannot be true since marker effects are estimated on the basis of the performances, so $\text{var}(\hat{z}_s)$ is not equal to $\text{var}(z_s)$ and, moreover, markers are selected because they explain a high proportion of the phenotypic variance; therefore, using the same data to select markers and to estimate marker effects clearly leads to $\text{cov}(y; \hat{z}_s)$ overestimating $\text{cov}(z; \hat{z}_s)$ and hence to overestimation of the weight to be placed on molecular score in the selection index. Indeed, this can lead to all weight being placed on the molecular score (Whittaker *et al.*, 1995; Gimelfarb and Lande, 1994a; Hospital *et al.*, 1997) and also, in some cases (if the estimated R^2 is higher than h^2), to put a negative weight on the phenotype.

A number of solutions to this problem have been proposed. Lande and Thompson (1990) suggest that in generations after the first, variable selection be done on the previous generation and model fitting on the current generation so as to give unbiased estimates of regression coefficients. Zhang and Smith (1992; 1993) generate two independent sets of F_2 data from the same population, applying their marker selection procedure to one to give $\mathcal{A}$ and then obtaining unbiased marker effects for this $\mathcal{A}$ from the other set of data. The same set of markers $\mathcal{A}$ is then used in all subsequent generations, so that bias in estimates due to the marker selection procedure is eliminated. However, this is clearly not an efficient use of data: much of the information from the first of these F_2 data sets is wasted. Also, Gimelfarb and Lande (1994a) have shown that MAS is more efficient if the marker selection procedure is repeated every generation.

Cross-validation approaches have been proposed to correct for the model selection bias (Whittaker *et al.*, 1997; Utz *et al.*, 2000). Whittaker *et al.* (1997) developed a cross-validation-based estimate of $\text{cov}(z; \hat{z}_s)$, and showed that this gave performance almost equivalent to using the actual value of $\text{cov}(z; \hat{z}_s)$, despite the estimation of $\text{cov}(z; \hat{z}_s)$ being rather poor in, for example, mean square error terms. This suggests that selection is reasonably robust to errors in the selection index weights. Perhaps this should not be surprising, as the position is similar to selection indices involving information on relatives, where it is known (Sales and Hill, 1976) that selection is reasonably robust to errors in the selection index weights. Finally, note that care must be taken to avoid over fitting the data and so overestimating the total proportion of genetic variance explained by the markers. The key requirement seems to be to always maintain some weight on y in the selection

index to allow discrimination between individuals with identical $\hat{z}_s$ values (Hospital *et al.*, 1997). Such individuals can be quite common, particularly in later generations as many markers become fixed.

Once the genetic values of candidates are predicted using Lande and Thompson's index, the best individuals are chosen and randomly intercrossed in order to produce the next generation. The same whole process is then repeated in the subsequent generations: new individuals are marker-typed and phenotyped, a new selection of markers is performed before predicting the genetic values of candidates. Hence, this strategy of marker-assisted recurrent selection requires at each generation to phenotype all individuals and to genotype them for markers covering the entire genome. Thus, MAS using this strategy is more costly than purely phenotypic selection. This raises the question of the economic interest of MAS compared to phenotypic selection.

21.2.2 Efficiency of Marker-assisted Selection

Much effort has been put into attempting to compare MAS with purely phenotypic selection by theory or by computer simulation. Theoretical assessment is difficult because of the complexity of the process that must be modelled, particularly over a number of generations. The prediction of MAS genetic gain requires the prediction of linkage disequilibrium between markers and QTL. This can be easily done in the initial unselected population, but, in subsequent generations, selection modifies allele frequencies and creates linkage disequilibrium even between unlinked chromosomes, which makes prediction hardly possible for complex genetic models. So deterministic models were used to predict genetic gains in the first generation (Lande and Thompson, 1990; Moreau *et al.*, 1998) or in subsequent ones but using a simplified model with only one marker linked to one QTL (Luo *et al.*, 1997).

Assuming large sample sizes as well as the normality of $\hat{z}$ and $\hat{z}_s$, Lande and Thompson (1990) predicted the relative efficiency (RE) of one cycle of marker-assisted selection compared to conventional phenotypic selection. They showed that the relative efficiency of MAS depends on two factors: the trait heritability, h^2, and the proportion of additive genetic variance explained by the markers, $p = \dfrac{R^2}{h^2}$. Finally, they get

$$RE = \sqrt{\frac{p}{h^2} + \frac{(1-p)^2}{1 - h^2\, p}},\tag{21.9}$$

for selection combining molecular score and phenotype, and

$$RE = \sqrt{\frac{p}{h^2}},\tag{21.10}$$

for selection based on the molecular score alone.

Considering these equations, it can be shown that MAS combining molecular score and phenotype is expected to be more efficient than phenotypic selection and than selection based on the molecular score alone. The advantage of MAS over phenotypic selection increases when the percentage of genetic variance explained by the markers increases and when the trait heritability decreases. Indeed, if the heritability is high, phenotypic selection is already very efficient and markers are of little use. However, if h^2 is low, the power of detection of marker effects is low (unless the sample size used for QTL

detection is large) and p cannot be expected to be very high. Hence, there is an optimum in RE (Hospital *et al.*, 1997; Moreau *et al.*, 1998) that does not show up in Lande and Thompson's equations. Moreover, Lande and Thompson's formulas are only valid for large samples and ignore the overestimation of marker effects when the same sample is used both for marker selection and for marker effect estimation. Moreau *et al.* (1998) derived a finite-population expression for the relative efficiency in the first generation of selection, by assuming that all markers are unlinked, so that the estimated regression coefficients for the markers are uncorrelated. Despite this simplification, their results are reasonably consistent with simulation results using more realistic genetic models and show that the Lande and Thompson formula overstates the relative efficiency of MAS considerably especially because of the overestimation of marker effects. Other studies dealing with the efficiency of MAS were based on stochastic simulations (e.g. Hospital *et al.*, 1997; 2000; Whittaker *et al.*, 1995; 1997; Gimelfarb and Lande, 1994a; 1994b; 1995; Zhang and Smith, 1992; 1993; Edwards and Page, 1994; Bernardo, 2004; Bernardo and Charcosset, 2006; Bernardo *et al.*, 2006). We briefly review several factors affecting MAS efficiency before discussing some alternatives to the Lande and Thompson's index that have been proposed.

21.2.2.1 *Model Selection Process*

Several strategies of model selection can be used to select markers to be included in the model. A common, but in some ways unsatisfactory (e.g. Breiman, 1995), solution is to choose the number of variables to be included using criterion such as Akaike's information criterion (AIC) or Mallow's C_p (Miller, 1990). A scheme based on Mallow's C_p was implemented in Whittaker *et al.* (1995). Other work (e.g. Gimelfarb and Lande, 1994a; Zhang and Smith, 1992; 1993) used a predetermined number of markers, whilst Hospital *et al.* (1997) used a stepwise variable selection scheme based on F tests using pre-specified significance thresholds. The choice of the number of markers or significance thresholds can be problematic, and AIC (or related criteria) are to be preferred on theoretical grounds, though, in practice, differences between the approaches tend to be minor (Whittaker *et al.*, 1995; Hospital *et al.*, 1997). Whatever the method, a compromise must be found between the power of detection and the risk of false positives and of over fitting the predictive model. Several studies found that there is an optimum type I error risk for MAS efficiency. Moreau *et al.* (1998) showed that this optimum is higher when the trait heritability is lower and advocated to use to rather permissive type I error risk (at least 5 % for low trait heritability). This is a much higher risk than the ones usually considered in QTL detection experiments (0.1 %). This was confirmed by simulations by Hospital *et al.* (1997) and Bernardo (2004). The latter even advocated the use of a permissive type I error risk between 0.1 and 0.2, depending on trait heritability.

In QTL detection (see **Chapters 18**, **19**, and **37**), it has been suggested that false discovery rate (FDR, Fernando *et al.*, 2004; Benjamini and Yekutieli, 2005), i.e. the proportion of false association detected over all the detected associations, might be a better parameter to consider instead of the type I error risk. Bernardo (2004) showed that optimal type I error for MAS corresponded to large FDR. So, increasing power is more important than the risk of including false positive in the selection index. Moerkerke *et al.* (2006) presented a method to balance type I and type II error risk in MAS experiments. However, they did not test the efficiency of their method by simulation.

21.2.2.2 *Experimental Design*

The simulation results described above clearly highlight the key effect on MAS efficiency of the sample size used for detection of marker effects. A large sample size increases the power of detection, increases the accuracy of marker-effects estimates and reduces the FDR. At least 100 or more individuals are needed at this step for MAS to be more efficient than phenotypic selection (Moreau *et al.*, 1998; Hospital *et al.*, 1997). Using large sample sizes is far from standard practice in plant breeding where breeders usually prefer to derive a smaller number of progenies per cross but develop a larger number of different crosses to increase the chance of obtaining superior genotypes. For many crop species, genotype by environment interaction effects can be as important as the additive variance (Moreau *et al.*, 2004). To limit the impact of QTL by environment interaction, and to correct for spatial heterogeneities, segregating populations are usually evaluated at several locations with several replicates per location. Increasing the total number of replicates increases the broad sense heritability of the trait and consequently increases the efficiency of conventional selection. In this context, using markers in selection means that experimental design must be reconsidered. Moreau *et al.* (2000) showed that, for a given cost, the optimal design for MAS is to evaluate a population of larger sample size than that for phenotypic selection in a smaller number of trials. Taking marker-genotyping cost into account reduces the relative benefit of MAS compared to phenotypic selection. MAS appears interesting only for traits with low heritability provided the investment is high enough to genotype and phenotype a large sample size.

In a related way, simulations showed that it is not beneficial to use a high marker density. First, this increases the cost of MAS. Second, even without considering cost, it is not necessary to have a high marker density for MAS to be useful in inbred line crosses, presumably because most of the benefits of the marker information arise in the early few generations of selection when linked markers are highly correlated. Gimelfarb and Lande (1995) found that increasing the marker density could even reduce response to selection, though this might be partly dependent upon the variable selection process used.

21.2.3   **Refinements**

So far, we have described the basic MAS procedure for inbred line crosses using Lande and Thompson's index. A number of enhancements or alternative approaches have been developed, some of which we describe now.

21.2.3.1   *Mixed Model Formulation*

Instead of estimating the molecular score in a first step, then combining it with the phenotype in a second step, Moreau *et al.* (1999) proposed to analyse the performances using the following mixed model:

$$y = \mu + z_s + z^* + e, \tag{21.11}$$

where z_s corresponds to the fixed effect of the markers (i.e. the molecular score as defined by Lande and Thompson (1990)), and z^* to the random genetic effect not explained by the markers ($z_s - z$), e being the residual experimental error. As discussed by Gianola *et al.* (2003), z^* and e cannot be distinguished in a case where there is only one record

per individual and no family structure. However, in many crop species, the phenotypic evaluation of an individual is often replicated (using clones or progeny testing), which makes it possible to distinguish between z^* and e. The genetic value z can thus be directly predicted by the sum of the estimated molecular score and the best linear unbiased predictor (BLUP) of the residual genetic variance.

$$\hat{z} = \hat{z}_s + \hat{z^*}. \tag{21.12}$$

This is equivalent to the Lande and Thompson's index except that the genetic value is decomposed into two parts that are jointly estimated and does not require us to explicitly compute the weight that must be given to each genetic term (that depends on h^2 and R^2). Unlike Lande and Thompson's index, this formulation allows one to not only take into account that some individuals may have fewer performances than others but also to include some effects related to the experimental design ('block effects', etc). It is also possible to take into account that error terms of genotypes sown in neighbour plots in a given field trial are not independent and correlations between errors can be modelled (Cullis and Gleeson, 1991). Moreau *et al.* (1999) showed that modelling spatial trend using autoregressive integrated moving average (ARIMA) models corrected the predictions of genetic value for spatial heterogeneity within trial and increased the accuracy of MAS. Smith *et al.* (2002) proposed to extend this to multiple trials.

A further improvement could be to consider marker effects as random instead of fixed and to take into account covariances (when they exist) between polygenic values. Doing this, the above model becomes equivalent to the marker-BLUP model used for outbred populations (Goddard, 1992, see below) but with a very simple population structure. In such a model, the regression coefficients are treated as random effects: as a consequence, estimated genetic values are shrunk back towards zero. This is more attractive than treating the regression coefficients as fixed effects, since it avoids the tendency of fixed effect models to overestimate the effect of large QTL (see Gianola *et al.*, 2003 for a discussion). In their simulations, Zhang and Smith (1993) used Lande and Thompson's index but considered marker effects either as random or as fixed. They showed that considering them as random led to better predictions of breeding value.

After the first generation of selection, a family structure appears in the population that might be used to increase the accuracy of breeding prediction. Lande and Thompson (1990) considered full and half sib families and showed how to incorporate information on relatives into their selection index. In the mixed model described above, a matrix of variance–covariances between relatives can be included to predict residual polygenic value. Taking information on relatives into account increases the efficiency of phenotypic selection and might also increase the efficiency of MAS but probably to a less extent. Indeed, marker effects, especially when treated as random, already take into account similarities between individuals. Zhang and Smith (1992; 1993) used simulation to compare MAS with BLUP and found that combining marker information and BLUP on phenotype in a selection index as described above gave superior performance to selection on phenotypic BLUP alone.

21.2.3.2   *Shrinkage Estimation, an Alternative to Model Selection*

We have commented earlier on the problems involved in selecting $\mathcal{A}$, the subset of markers to be included in the model. An alternative approach, often argued to be superior to subset

selection (e.g. Breiman 1992; 1995), would be to include all variables in the regression model, but to shrink the regression coefficients back towards zero, for instance, by using ridge regression. The usual least-squares estimates would then be replaced by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{21.13}$$

where $\mathbf{I}$ is the identity matrix. This perhaps makes most sense in a Bayesian framework, where the shrinkage is then towards a Gaussian prior with zero mean and variance–covariance matrix $\sigma_p^2\mathbf{I}$, with λ reflecting the strength of our confidence in the prior distribution relative to the amount of information in the data. Whittaker *et al.* (2000) showed that such an approach can give slightly improved performance to subset selection in MAS, increasing mean response slightly and reducing the variance in response substantially. As we would expect, ridge regression is favoured relative to subset selection for dense marker maps and relatively uniform distributions of QTL effect sizes (Whittaker *et al.*, 2000). Note a further advantage of ridge regression: the addition of the $\lambda\mathbf{I}$ term reduces collinearity and prevents the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ from becoming singular or near-singular. Singular $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ often occur after several generations of selection as the population becomes increasingly inbred, so this is potentially important here.

Gianola *et al.* (2003) and Xu (2003), argued that ridge regression might not be optimal especially where the number of markers exceeds the number of individuals. From a Bayesian point of view, ridge regression implies that the regression coefficients are independent and identically distributed and are thus all shrunk towards a common mean (zero). This may be questionable since a marker linked to QTL does not deserve to be shrunk as much as if it was unlinked. Xu (2003) adapted the Bayesian approach of Meuwissen *et al.* (2001), initially developed for outbred populations, to populations derived from inbred crosses. Though he did not test by simulation if this method could increase MAS efficiency compared to using a ridge regression, results in terms of QTL detection seem promising.

21.2.3.3 *Non-Gaussian Traits*

So far, we have assumed that traits have a Gaussian distribution: this is true for many traits of interest, and often transformation of non-Gaussian traits to approximate normality is possible. For QTL mapping, good results have also been obtained using linear regression for non-Gaussian traits (Visscher *et al.*, 1996a), and the same might be expected to hold for MAS. Nevertheless, a method that allows MAS for non-Gaussian traits would be useful. It is clearly possible to develop a generalised linear model (McCullagh and Nelder, 1989) version of the linear models used above to do this, but this does not seem to have been done, possibly because work on non-Gaussian responses has suggested that least-squares-based approaches are often adequate (Visscher *et al.*, 1996a). A related approach was implemented by Lange and Whittaker (2001), who developed a MAS method using quasi-likelihood (Heyde, 1997) and generalised estimating equation (Liang and Zeger, 1986) approaches. See Thomson (2003) for a related GEE approach. The great advantage of such models is that, rather than specifying a probability model for the data as we must with likelihood approaches, we merely need to make suitable first- and second-moment assumptions to obtain consistent parameter estimates, regardless of the true underlying distribution of the data.

21.2.3.4 Selection for Multiple Traits

Often, we wish to improve several traits simultaneously. For Gaussian traits, the above theory is easily extended to consider a vector of phenotypic values $\mathbf{y}$ and a corresponding vector of molecular scores $\hat{\mathbf{z}}_s$ (Lande and Thompson, 1990). Writing

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{yy} & \mathbf{P}_{ys} \\ \mathbf{P}_{ys} & \mathbf{P}_{ss} \end{bmatrix} \tag{21.14}$$

for the phenotypic variance–covariance matrix,

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{yy} & \mathbf{G}_{ys} \\ \mathbf{G}_{ys} & \mathbf{G}_{ss} \end{bmatrix} \tag{21.15}$$

for the corresponding additive genetic variance–covariance matrix, and $\mathbf{d}$ for the vector of relative economic weights of the traits, standard theory gives the weights $(\mathbf{b}_y, \mathbf{b}_s)$ for $\mathbf{y}$ and $\hat{\mathbf{z}}_s$ as

$$\begin{pmatrix} \mathbf{b}_y \\ \mathbf{b}_s \end{pmatrix} = \begin{bmatrix} \mathbf{P}_{yy} & \mathbf{P}_{ys} \\ \mathbf{P}_{ys} & \mathbf{P}_{ss} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{G}_{yy}\mathbf{d} \\ \mathbf{G}_{ys}\mathbf{d} \end{pmatrix}. \tag{21.16}$$

The problems of estimating the required variances and covariances remain, but, assuming that populations are sufficiently large to allow these problems to be ignored, MAS again outperforms selection on phenotype alone (Lande and Thompson, 1990; Xie and Xu, 1998).

There has been surprisingly little work done on multiple traits in more realistic settings. Lange and Whittaker (2001) considered the above scheme and a version of the ridge regression approach used in Whittaker *et al.* (2000) extended to multivariate phenotypes using results from Brown and Zidek (1980). They also tested a method based on generalised estimating equations (Liang and Zeger, 1986; McCullagh and Nelder, 1989) that deals with multivariate non-Gaussian traits. In both cases, results were broadly similar to the univariate case in terms of comparisons between MAS and phenotypic selection. Methods based on multivariate Gaussian data are often adequate, but, for some situations, particularly where responses are highly skewed, the generalised estimating equation approach, which allows for non-normality of the traits, can give improvements compared to the usual approach of marginal transformation of the traits to approximate normality followed by use of the standard Gaussian theory. It seems that the transformation of the traits to approximate normality destroys the correlation structure between the traits, whilst the proper modelling of correlation structure allowed by the estimating equation approach can give considerable improvements in efficiency.

Although it is possible to reduce selection on multiple traits to a univariate analysis of total economic value, the multivariate analysis is in general more efficient. This is easily seen by noting that, for example, the relative weights on the molecular scores are not proportional to the corresponding economic values in (21.16).

21.2.3.5 Unrecorded Phenotypes

Once the regression coefficients in (21.3) have been estimated, it is possible to use z_s as an estimate of genetic value even if no phenotypic information is available for that individual.

The most obvious application is to sex-limited traits such as milk production, but two-stage selection schemes have also received considerable attention (Tanksley *et al.*, 1981; Soller and Beckmann, 1983; Lande and Thompson, 1990; Xie and Xu, 1998). The idea here is to perform selection on immature individuals (seedlings, embryos, or juveniles) based on $\hat{z}_s$, followed by phenotypic selection or MAS of the selected individuals on reaching adulthood. Selection response is of course reduced relative to a single stage of selection on all individuals as adults, but costs are also greatly reduced. Xie and Xu (1998) argued that two-stage MAS schemes are always more effective than one-stage MAS, though they assumed that populations were infinite and that the regression coefficients in (21.3) were known without error. There is clearly considerable cost involved in estimating these regression coefficients. Hospital *et al.* (1997) considered several schemes where an initial cycle of index selection was followed by one or two generations of selection using the regression coefficients estimated in the initial generation. Hospital *et al.* (1997) found that the genetic gains from such schemes were roughly comparable to the gains given by same number of generations of phenotypic selection. Costs would be considerably less than those incurred in using index selection in every generation because the phenotypes of the individuals are only required in the first generation, whilst genotyping is only necessary for the markers in $\mathcal{A}$. There is also an important saving in time over index or phenotypic selection. Hospital *et al.* (1997) suggest that this sort of approach makes MAS viable for medium- or high-heritability traits; indeed, because selection response is more variable under MAS than under phenotypic selection for low heritability, they argue that these may be the situations where MAS is of most interest.

This approach to MAS has been applied experimentally with success on maize (Johnson, 2004). However, it has to be noted that its efficiency relies on the stability of the marker effects estimated at an earlier generation of selection. As selection proceeds, markers get fixed for favourable alleles at some QTL, and recombination events may occur between markers and QTL. Hence, the markers initially selected become less and less predictive of the breeding value of candidates. It then becomes necessary to re-detect new associations after several generations of selection on markers alone as advocated by Hospital *et al.* (1997). Other sources of instability of the marker effects might come from interactions between QTL effects and environmental conditions (that might differ between the time of marker-effects detection and the time of genetic gain evaluation) and also from epistatic interactions effects between QTL. More specifically, in the case of 'less than additive' epistasis (as found in tomato by Eshed and Zamir, 1996), the effects of favourable alleles at QTL are reduced when favourable alleles at other QTL become fixed by selection. This reduces the interest of MAS compared to what could be expected on the basis of marker effects estimated prior to selection. These two phenomena that can be important for complex traits (see Moreau *et al.*, 2004; Blanc *et al.*, 2006 for examples on maize grain yield) are often ignored in simulations studies that usually assume a rather simple genetic model where there is no interaction between QTL and the environmental conditions and no (Hospital *et al.*, 1997) or limited interactions between QTL (Bernardo, 2004; Bernardo *et al.*, 2006). This leads to an overestimation of the efficiency of selection on markers alone. Indeed, non-additive effects were often found to be responsible for the lack of success of some MAS selection experiments (Moreau *et al.*, 2004). These non-additive effects have a lower impact when marker effects are re-estimated at each generation since new predictions take into account changes in the environmental conditions and the genetic background. In this case, the simple additive model, even if it is false, behaves quite well.

Podlich *et al.* (2004) carried out simulations of MAS efficiency using a genetic model where QTL were assumed to be involved in a complex network of epistatic interactions also depending on environmental factors. They clearly showed that, when non-additive effects are important, the best strategy (called *mapping as you go*) is to continually revise estimates of marker effects.

The assumption of additivity within and between QTL in Lande and Thompson's index, though convenient, is not essential. Dominance can be included in the model of prediction as well as digenetic interactions between markers or interactions with environmental covariates. The main limitation is that these non-additive effects are hardly detectable and estimable. Using cycles of MAS without re-estimating marker effects calls for a better knowledge of the genetic architecture of quantitative traits and efficient ways of estimating non-additive effects to better anticipate the results of such a selection.

21.2.3.6 *Optimisation of the Efficiency of MAS over Several Generations*

Several simulation results draw the conclusion that, even in a simple additive case, when effects are re-estimated regularly, MAS combining phenotype and markers gives improved response over phenotypic selection in early generations but the advantage declines with time, and phenotypic selection may overtake MAS in later generations. As already mentioned, part of the explanation comes from the recombination events that accumulate between markers and QTL, reducing linkage disequilibrium. However, Hospital *et al.* (1997) showed that the long-term superiority of phenotypic selection over MAS was due to the higher level of fixation of unfavourable alleles at QTL of small effect with MAS than with phenotypic selection. The high efficiency of MAS for fixing favourable alleles at major QTL results in involuntarily fixing unfavourable alleles at other QTL of minor effects by 'hitchhiking'. The truncation selection of candidates based on Lande and Thompson's index and followed by random mating of the selected individuals optimises the genetic response in the next generation but not necessarily in the subsequent ones. In a specific case, where all the QTL are known, Hospital *et al.* (2000), proposed an empirical strategy of selection that consists in giving equal weights to all the QTL and taking into account complementarity between selected individuals to avoid losing favourable alleles. They showed that this was more efficient than selecting on the molecular score only.

Other authors (van Berloo and Stam; 1998; 2001; Charmet *et al.*, 1999; Bernardo *et al.*, 2006) suggested the use of complementarities between individuals for QTL not only to select the best individuals but also to choose the best couples between selected individuals that are more likely to produce superior genotypes in the next generation. van Berloo and Stam (1998; 2001) and Charmet *et al.* (1999) used this strategy to predict the values of the best inbred lines that can be derived from each couple. Bernardo *et al.* (2006) simulated four generations of MAS and compared different intensities of selection and different strategies of intercrossing including the one based on the expected probability of obtaining superior genotypes in the next generation. They found that this strategy was only slightly better than random mating of the selected genotypes; their major conclusion was that the intensity of selection was the most important factor. Compared to other studies on MAS optimisation, they did not consider that QTL were known without error but based their selection on QTL detected with a high type I error risk. In such a case, the estimation of the expected progeny variances were certainly very poor, which might explain the lack of efficiency of this approach. There is clearly a need, especially in plant

breeding, to move from the recurrent selection of populations as defined in Lande and Thompson's paper to a 'genotype building' strategy that aims to assemble as quickly as possible favourable alleles to create superior varieties.

21.2.3.7 *Broadening Genetic Variability*

As previously mentioned, MAS does not fit with usual breeding practices. Each year, breeders derive tens of populations of small sample sizes from different crosses to cover a large genetic diversity and increase the probability of finding superior genotypes. Conversely, MAS leads to high investment in a few specific populations. Moreover, as marker-QTL associations are specific to each population, the detection and estimation of marker effects must be re-run in each population with no benefit from previous results. To overcome this, one solution is to perform QTL detection and MAS jointly on different biparental populations. Among all the multi-parental designs that can be considered, connected biparental populations (populations that have one parental line in common) are particularly interesting. Connections between populations increase the power and the accuracy of QTL detection compared to single-population analyses, they enable one to estimate simultaneously the different parental alleles that segregate and to identify the most favourable ones for selection and in some cases they allow one to test for epistasis interaction effects between the QTL and the genetic background (Rebai and Goffinet, 2000; Jannink and Jansen, 2001; Blanc *et al.*, 2006). Such designs were found by simulations to increase the efficiency of MAS compared to single–biparental population (Blanc *et al.*, 2006).

In breeding programs, however, populations are often not directly connected, but can be connected to each other to a certain extent by using (as parents) related lines that are both derived from an ancestral elite inbred line. Even if the number of inbred lines involved in such a pedigree can be large, the actual number of different alleles that segregate can be much lower. For those complex pedigree structures, fixed models considering one parameter per marker and per parental line are no longer suitable. Random models derived from those developed for outbred populations have recently been proposed to identify QTL in complex plant breeding pedigrees (Crepieux *et al.*, 2004). Other strategies have been proposed to reduce the number of parameters to be estimated, such as using marker haplotypes (Jansen *et al.*, 2003) or phenotypic values (Jannink and Wu, 2003) to infer the actual number of different alleles that need to be estimated. In such a multi-allelic context, models recently proposed are very often adapted from those used for association-mapping studies and those initially developed for outbred populations.

21.3   MARKER-ASSISTED SELECTION: OUTBRED POPULATIONS

Inbred populations have the great advantage that we can assume initial complete association between QTL and markers in the founding lines. This is not true of outbred populations, and so a number of additional complications must be faced when trying to implement MAS schemes in outbred populations. Most of the work on outbred populations has been motivated by possible application to animal breeding, and we shall concentrate on that context, but related problems arise in other areas, such as tree breeding (e.g.

Strauss *et al.*, 1992). Two main approaches have been suggested. In the first, MAS is performed within families: this is the simplest approach as we need then only ascertain the preferred marker types in the founding individuals and select accordingly. Selection on a population level across families, although more difficult, has the potential to give greater increases in genetic response and can be done by incorporating marker information into the standard BLUP procedure (see **Chapter 20**, for more detail on BLUP). In both cases, it is usual to assume that the QTL has been mapped, or at least that the markers flanking the QTL have been identified. We discuss these approaches in turn, starting with the marker-assisted BLUP approach.

21.3.1 MAS via BLUP

Fernando and Grossman (1989) showed how information on a single marker believed to be linked to a QTL could be incorporated into the standard BLUP estimate of expected breeding value. This method was then extended to deal with multiple markers by Goddard (1992). We shall concentrate on the variant of this approach described in Meuwissen and Goddard (1996). To simplify notation, we describe the model for a single marked QTL, but this can easily be extended to any number of QTLs linked to markers.

Suppose $\mathbf{y}$ is a vector of phenotypic records for the trait in question, $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{u}$ is a vector of random polygenic effects, $\mathbf{e}$ a vector of random environmental effects, $\mathbf{q}$ a vector of random QTL effects and $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{Q}$ are incidence matrices. Then we fit the mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{Z}\mathbf{Q}\mathbf{q} + \mathbf{e}. \tag{21.17}$$

The variance matrices of the random effects are written as

$$\mathrm{var}(\mathbf{e}) = V_e\mathbf{I},$$
$$V_q\mathbf{G} = \mathrm{var}(\mathbf{q}),$$
$$V_u\mathbf{A} = \mathrm{var}(\mathbf{u}), \tag{21.18}$$

where the scalars V_e, V_u and V_q are the environmental variance, variance of polygenic effects and variance of the QTL effect for a single gamete respectively. To simplify the exposition, we treat the variance components V_e, V_u and V_q as known, perhaps, from the same QTL mapping experiment that produced the estimate of QTL location. Alternatively, they could be estimated using REML, as described in **Chapter 20**.

Note that the incidence matrices $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{Q}$ link fixed effects to records, animals to records and QTL effects to records respectively. Thus, every row of $\mathbf{Q}$ has two elements equal to 1, with all other elements equal to 0. Of course, QTL alleles cannot be observed directly, but must be inferred from marker information. We assume that all alleles of founding individuals are distinct; thus, the number of QTL alleles in the base population is twice the number of base animals. In the next generation, inference on marker haplotype is used to follow the transmission of QTL alleles, with $\mathbf{Q}$ linking the transmitted QTL allele to the progeny's phenotype. Where there is uncertainty about which QTL allele has been transmitted, either because a recombination has occurred between the markers flanking the QTL or because marker haplotypes were uninformative, a new QTL allele is formed. This allele is linked to the progeny phenotype through the $\mathbf{Q}$ matrix and the effect of the new allele is linked to its parent's alleles by assuming that the expected value

of the new QTL effect is the mean of the parental QTL effects. The following example, based on Meuwissen and Goddard (1996), is instructive.

Consider two individuals in the base generation and write the effects of their QTL alleles as (q_1, q_2) and (q_3, q_4), respectively. Suppose the individuals are mated to produce a third individual and that this individual receives the marker haplotype corresponding to the QTL allele with effect q_1 from one parent, but that a recombination between the flanking markers makes the QTL allele transmitted from the other parent uncertain. A new allele is therefore formed with a value of q_5, where

$$E(q_5) = 0.5(q_3 + q_4),$$

$$\text{var}(q_5) = V_q,$$

and

$$\text{cov}(q_5; q_4) = \text{cov}(0.5q_4; q_4) = 0.5V_q,$$

$$\text{cov}(q_5; q_3) = \text{cov}(0.5q_3; q_3) = 0.5V_q, \tag{21.19}$$

since the effects of founding alleles are assumed independent. This gives

$$\mathbf{Qq} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \end{bmatrix} \tag{21.20}$$

and

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0.5 & 0.5 & 1 \end{bmatrix}. \tag{21.21}$$

Thus, this procedure produces a 'pedigree' of QTL alleles, so that $\mathbf{G}$ has the structure of a numerator relationship matrix. The inverses $\mathbf{A}^{-1}$ and $\mathbf{G}^{-1}$ can thus be obtained using the results in Henderson (1976).

Putting $\lambda = V_e/V_u$ and $\alpha = V_e/V_q$, we can therefore get estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$ by solving the usual mixed model equations (see **Chapter 20**):

$$\begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{Z} & \mathbf{X}^{\mathrm{T}}\mathbf{ZQ} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \mathbf{A}^{-1}\lambda & \mathbf{Z}^{\mathrm{T}}\mathbf{ZQ} \\ \mathbf{Q}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{X} & \mathbf{Q}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z} & \mathbf{Q}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{ZQ} + \mathbf{G}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{q}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{y} \\ \mathbf{X}^{\mathrm{T}}\mathbf{y} \\ \mathbf{Q}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{y} \end{bmatrix}. \tag{21.22}$$

21.3.2 Comments

The scheme described above does not attempt to use information about QTL location to make probabilistic statements about the most likely source of QTL alleles where a recombination has prevented certain identification of the QTL allele transmitted. We have also ignored the possibility of double recombination between flanking markers. The model could easily be extended to incorporate these phenomena (e.g. van Arendonk *et al.*, 1994), but this would add considerably to the computational complexity and seems unlikely to

make much difference in practice unless markers are very widely spaced, and then MAS will probably not be effective in any case. In addition, Meuwissen and Goddard (1999) comment that when the flanking markers are closely spaced so that the probability of double recombination is small, attempting to incorporate double recombinants leads to a set of mixed model equations, which is almost singular and therefore difficult to solve.

A further drawback of this BLUP approach is that the random effects **q** are unlikely to have Gaussian distributions; although the derivation of BLUP estimates does not depend on normality, some of the desirable properties of BLUP estimators do. However, Goddard (1992) argues that departures from normality should not decrease the accuracy with which breeding values are estimated.

As we have already stated, this approach extends easily to multiple QTLs; indeed, Goddard (1992) explicitly assumed multiple QTLs. However, analyses quickly become very computationally expensive as the number of marked QTLs or the number of animals in the pedigree increase, since for each animal we have an equation for the polygenic effect plus two equations for each marked QTL. This has motivated a number of attempts to produce methods in which some of these equations can be eliminated. For example, a reduced animal model can be used to eliminate equations for animals which are not parents, QTL effects for non-genotyped individuals can be absorbed into the polygenic effect, or marked QTL effects can be combined with the polygenic effect to give total genetic merit (Bink *et al.*, 1998; Cantet and Smith, 1991; Saito and Iwaisaki, 1997). The most general and effective approach to reducing the number of equations is due to Hoeschele (1993); however, because of its generality, this approach is not straightforward to use. Several authors have therefore developed simpler, though less effective, methods for use in particular circumstances. Examples include Meuwissen and Goddard's (1999) approach for use where many of the animals in the pedigree have not been marker-typed, for example in commercial dairy cattle populations where the size of the population (more than 1 million cows) makes exhaustive genotyping impracticable and Saito *et al.*'s (1998) consideration of carcass traits, where phenotypic information is only available on slaughtered animals, which therefore never become parents.

Simulation studies indicate that these MAS schemes can be substantially more effective than conventional (phenotypic) selection, particularly where selection occurs before phenotypic values for the trait are recorded (an example would be fertility) or for carcass traits (Meuwissen and Goddard, 1996). The increased selection response given by MAS is particularly marked for low- heritability traits, and, as we would expect, when marked QTLs contribute a large proportion of genetic variance. Most studies have assumed a single marked QTL, but Spelman and Bovenhuis (1998b) found similar results where two marked QTLs were considered. They noted that the benefits of MAS were most apparent in the early generations of selection when the two QTLs were on the same chromosome, explaining this result in terms of the covariance between the two QTLs. They also found, again as we would expect, that decreasing the interval between the two markers flanking the QTL increased the efficiency of MAS by increasing the accuracy with which QTL effects can be estimated.

Clearly, the initial QTL mapping stage is crucial to the success of MAS programs of this type. This is an area of concern because the precision of estimates from QTL mapping programs is typically rather low. In particular, confidence intervals for QTL location tend to be wide. Choosing the wrong pair of flanking markers can lead to serious losses of genetic gain, compared to the use of the correct interval. It is therefore tempting

to use widely spaced markers to increase the chance that the selected interval contains the true QTL location, but Spelman and van Arendonk (1997) show that this can be counterproductive by comparing 20- and 5-cM intervals: the extra accuracy gained by using a 5-cM marker interval when that interval does contain the QTL outweighs the increased chance that the 20-cM interval contains the QTL. Of course, this must be strongly dependent on assumptions about the precision of the initial estimation of QTL location.

Moreover, estimates of QTL effect are both variable and biased upwards because of the stringent significance thresholds used in these experiments. The effect of errors in the estimates of QTL location and the variance due to the QTL, i.e. the variance component V_q, has been considered by Spelman and van Arendonk (1997). They found that overestimation of V_q, in their case assuming that V_q was equal to 10 or 15 % of the phenotypic variance when the actual value was 5 %, had little effect in early generations of the MAS program but did give a noticeable decline in long-term response, relative to use of the correct value, because the overestimation of V_q leads to too much weight being applied to QTL genotype and so to a loss of polygenic variation, which becomes important in later generations of the selection program. This is consistent with the above discussion on 'hitchhiking'. Severe overestimation of V_q, for example, selection on a non-existent QTL because of type I error, gives inferior performance relative to purely phenotypic BLUP schemes, as we might expect. Spelman and van Arendonk (1997) comment that the loss in response can be reduced if V_q is re-estimated during the selection program.

However, even when the QTL effect is assumed to be known without error, several authors showed that rapid fixation of favourable allele at QTL resulted in a lower response on the polygenic value so that after several generations MAS becomes less efficient than purely phenotypic BLUP. This effect, often referred to as the Gibson effect (Gibson, 1994), is similar to the lack of long-term efficiency of Lande and Thompson's index found by simulation in inbred populations. Using an approach based on the optimal control theory, Dekkers and van Arendonk (1998) showed that the weight given to the QTL in early generations should be decreased to maintain long-term efficiency of MAS. This approach was later extended to multiple QTL (Dekkers *et al.*, 2002; Chakraborty *et al.*, 2002). Results showed that the best strategy is to select equally on all the QTL, regardless of the percentage of variance they explain. However, this strategy of optimisation is computationally demanding and requires the prediction of the selection response over the time horizon. It relies on some strong assumptions, such as an infinite sample size and normal distribution of the polygenic genetic value, and cannot usually be applied in the case of finite populations or to large number of QTL selected simultaneously. Another concern in outbred populations is that selecting on markers may increase the rate of inbreeding in the population and thus affect the long-term efficiency of selection. Several strategies were proposed to also control the contributions of selected individuals (Manfredi *et al.*, 1998; Villanueva *et al.*, 2002; Meuwissen and Sonessen, 2004).

21.3.3 Within-family MAS

The BLUP method described above is attractive conceptually, but is complex and computationally demanding. For dairy cattle populations, family sizes are sufficiently large, and therefore some of the benefits of MAS may be achieved by using simpler alternatives based on selection within families. Two such schemes have been suggested, termed *top-down* and *bottom-up* approaches. Both suppose that a QTL has been identified

in a previous QTL mapping program; the basic idea is to select individuals with good alleles at this locus for progeny testing. In the absence of marker information, this would be done using expected breeding value calculated using BLUP.

The top-down scheme (Kashi *et al.*, 1990) is based on the granddaughter design. Let the current generation of elite sires be generation 0. Now consider the progeny-tested sons of a given generation 0 sire. The progeny-tested sons are grouped according to the QTL allele inherited from the generation 0 sire, determined from their marker haplotype, and if a significant difference exists between the alleles, then it is deemed that the generation 0 sire is heterozygous for the QTL. Typically, marker haplotype here means a single marker assumed to be tightly linked to the QTL of interest, so this grouping is straightforward, although it is not obvious how to choose the significance threshold to be used. Now, the generation 1 sires have already been progeny tested, so MAS cannot be applied to these individuals. Instead, the marker information is used to select which of the sons of the progeny-tested sons (grandsons of the generation 0 sire) are progeny tested. Where sires have received the favoured marker haplotype, grandsons with this haplotype are selected; where sires have received the unfavoured haplotype, grandsons with the sires maternal haplotype are selected. Note that this assumes the following ordering on marker haplotypes:

$$\text{favoured grandsire} > \text{dam population} > \text{unfavoured grandsire.} \tag{21.23}$$

This is reasonable if there are only two QTL haplotypes, so that the favoured QTL haplotype has value a_Q, the unfavoured haplotype value a_q and the dam population has mean $a_q + P(Q)(a_Q - a_q)$, but becomes more doubtful with larger numbers of QTL haplotypes.

A similar scheme could be used to identify maternal grandsires heterozygous for the QTL. Grandsons would then be progeny tested only if they had inherited the favoured allele from the paternal grandsire and the favoured allele from the maternal grandsire.

The bottom-up (Mackinnon and Georges, 1997) approach is fundamentally similar but is based on the daughter design. Daughters of a generation 0 sire are used to test for an effect of the paternal marker inherited on phenotype. If this test is significant, a QTL is judged to be segregating in the generation 0 sire. Sons who inherit the favoured allele from this sire are progeny tested; sons who inherit the unfavoured paternal allele are not. Note that dams in general have too few offspring to allow identification of dams heterozygous for the QTL, so selection cannot be applied to maternally inherited alleles using this scheme. However, it may be possible to use this scheme for paternally inherited alleles and the top-down scheme for maternally inherited alleles.

Spelman and Garrick (1998) showed by simulation experiments based on the New Zealand breeding scheme for Holstein–Friesian cows that these within-family schemes could increase genetic gain by 1–5 %, and argued that these relatively modest figures would be sufficient to make MAS economically viable assuming current genotyping costs. They found the bottom-up selection to be more effective than the top-down selection, with the highest gains being obtained when the bottom-up selection was used for paternal alleles and the top-down selection was used for maternal alleles. Interestingly, both Spelman and Garrick (1998) and Mackinnon and Georges (1997) found that the greatest genetic gains were obtained from the bottom-up scheme when the threshold at which we declare a grandsire to be heterozygous is rather low. This is presumably because the cost of a false positive (falsely declaring a sire to be heterozygous) is low compared to the cost of

assuming a sire to be homozygous when a QTL is in fact segregating in his descendants. Again, this is consistent with the above discussion on 'hitchhiking'.

21.4   MARKER-ASSISTED INTROGRESSION

Different plant or animal lines have different genes of commercial value. The previous section on MAS is concerned with crossbreeding, where existing lines are crossed to create a hybrid variety superior to the parental lines. We now consider the situation where one of the parental lines is broadly superior to the other, but the inferior line contains a small number of genes of interest, or, in the extreme case, a single valuable gene. Our objective is to transfer these genes into the other, superior, parental line whilst leaving the rest of that line's genome unchanged. Accordingly, we speak of moving genes from the *donor* line to the *recipient* line. An example would be the transfer of a gene for increased litter size (Rothschild *et al.*, 1996) from the slow-growing and fat Meishan pig into commercial lines. Where markers are used, this process has become known as *Marker-Assisted Introgression* (*MAI*). Typically, repeated backcrosses to the recipient line are made, with markers used to identify the allele which is to be introgressed and to select against the remainder of the donor genome, followed by intercrossing or selfing of the final backcross population to generate individuals homozygous for the introgressed allele. Markers can thus be used in two ways (Hospital and Charcosset, 1997). In *foreground selection*, markers are used to detect the presence of the introgressed gene where direct phenotypic ascertainment would be impossible or too costly (Tanksley, 1983; Melchinger, 1990). In *background selection* (Hillel *et al.*, 1990; 1993), markers are used to accelerate recovery of the recipient genome.

Of course, all this assumes that the genes to be introgressed actually exist and will exhibit the anticipated effect in the recipient background. This is far from certain when the introgressed gene is a QTL, given the low power of QTL mapping. However, we ignore this problem here, assuming that an introgression program will only be initiated if the existence of a QTL has been 'confirmed' in some way.

21.4.1   Inbred Line Crosses

Here, we can assume that the parental lines are initially homozygous for different alleles, so that there is no ambiguity about the parental origin of marker alleles. Early work in this area assumed that the introgressed gene can be detected with certainty, i.e. there is a marker synonymous with the introgressed gene. Foreground selection is then straightforward and we can concentrate on background selection, specifically on how efficient MAI is in recovering the recipient genome, typically measured in terms of *genomic proportion*, the relative proportion of an individual's genome that originates from the recipient line.

A considerable amount is known about the speed at which the recipient genome is recovered in the absence of selection (Stam and Zeven, 1981; Hospital *et al.*, 1992; Hill, 1993; Visscher *et al.*, 1996b). Dealing with selection on markers is more difficult, but some results have been obtained and a number of simulation studies have been performed (Hospital *et al.*, 1992; Visscher *et al.*, 1996b; Frisch *et al.*, 1999b; Hospital, 2001; 2002). The general conclusion is that, depending on selection intensity, selecting

on markers recovers the recipient genome approximately two generations faster than if marker information was not used. More optimistic forecasts of the speed of recovery using markers have been made (Hillel *et al.*, 1990; Markel *et al.*, 1997), but these papers ignore some important aspects of the selection process (Hospital *et al.*, 1992; Visscher, 1999). For instance, they assume that linked chromosomal segments segregate independently; they therefore underestimate the variance in genomic proportion during backcrossing and overestimate the speed at which the recipient genome is recovered (Visscher, 1999).

After selection on markers is performed, the genotype of the selected individuals at non-marker loci is never completely known, unless very dense marker maps are used, which is very costly. Hence, it is better to use few markers, but optimise their positions, not only to increase the *mean* genomic proportion but also to control the variance of this proportion among individuals. The variance in genomic composition of backcross progeny sharing the same genotype at selected molecular markers at the end of the programme was used as a criterion for the number of individuals that should be genotyped (Visscher, 1996; Servin, 2005), based on earlier developments from Hill (1993). Using a different computation for the reduction of the variance in the donor genome proportion due to selection, Frisch and Melchinger (2005) adapted standard normal distribution selection theory to MAI.

We now consider the case where introgressed genes cannot be detected with certainty; indeed, we may only have an estimate of the gene's location, for instance from a QTL mapping program. Background selection on the non-QTL-bearing chromosomes can proceed as above, but foreground selection must now be based on the genotype at markers close to the estimated QTL location. The question then is how to optimise the number and location of markers used in foreground selection, in order to best control the presence of a gene of uncertain location. Denote by M and Q respectively the events that a given individual has the desired donor marker genotype and QTL genotypes. Consider a chromosome of length L; if we knew that the location of a QTL was $\rho \in (0, L)$ and that the left and right flanking markers of the QTL were at locations ρ_L and ρ_R respectively, then writing $r(\cdot, \cdot)$ for the recombination fraction between two points, we have that in the tth generation of backcrossing

$$P(Q|M, \rho) = \frac{[1 - r(\rho_L, \rho)]^t [1 - r(\rho, \rho_R)]^t}{[1 - r(\rho_L, \rho_R)]^t}. \tag{21.24}$$

Therefore, if we have obtained a density $P(\rho)$ for the location of the QTL from, for example, a previous QTL mapping experiment,

$$P(Q|M) = \int_0^L \frac{[1 - r(\rho_L, \rho)]^t [1 - r(\rho, \rho_R)]^t}{[1 - r(\rho_L, \rho_R)]^t} P(\rho) \, d\rho. \tag{21.25}$$

There is a slight problem here in that many QTL mapping approaches do not provide $P(\rho)$. In fact, as most of the QTL mapping methodology works within the frequentist framework, it is not strictly possible to consider ρ as having a density. *Ad hoc* approaches have been suggested, however. Both Visscher *et al.* (1996b), who did this calculation for two markers, and Hospital and Charcosset (1997), who extended the approach to any number of markers, suppose that $P(\rho)$ is Gaussian, centred on the true QTL location and with variance obtained, for example, from the confidence interval for QTL location given by the QTL mapping experiment. Hospital and Charcosset (1997) go on to maximise $P(Q|M)$

with respect to marker location. They show that even where considerable uncertainty about the QTL location exists, a small number of markers (2−4) gives values of $P(Q|M)$ close to 1 at $t = 3$, when foreground selection alone is used on the QTL-carrying chromosome. They also showed that the population sizes needed to give at least one individual with the desired marker genotype with high probability are reasonable.

We also wish to perform background selection on the donor gene-carrying chromosome (whether the target donor gene to introgress is a known gene or a QTL). This raises the interesting question of where to switch from selection for donor marker alleles (at marker loci that control the donor target, herein called *central markers*) to selection for recipient marker alleles (at marker loci that flank the target region on each side, herein called *flanking marker*).

For simplicity, let us say that an individual is 'R1' if it has donor genotype at the central markers but recipient genotype at only one of the flanking markers (and donor genotype at the flanking markers on the other side), and 'R2' if it has donor genotypes at the central markers and recipient genotypes at *both* of the flanking markers (i.e. on both sides). Genotypes R1 and R2 are sometimes called *single* and *double recombinants*. Note that here the recombinations leading to 'R2' do not necessarily take place at the same generation, which gives space for optimisation. Intuitively, for close flanking markers, R2 genotypes are highly unlikely to be obtained in one single generation (BC_1) so that at least two BC generations should be performed, with selection for a R1 genotype in BC_1 (with recombination on one side of the target), and an R2 genotype in BC_2 with a recombination on the other side (Young and Tanksley, 1989). The underlying mathematics has been worked out recently. Hospital and Charcosset (1997) were the first to formalise this problem, together with the optimisation of central marker positioning in the case of QTL introgression. The above approach can then be applied to optimise marker location, though this requires some compromise between background and foreground selection efficiency. The solution derived by Hospital and Charcosset (1997) was used by Frisch *et al.* (1999a) with numerical applications in the context of 'next-generation' optimisation (population size at generation $BC_{(n+1)}$ is optimised given that the genotype selected at generation BC_n is known). However, Hospital (2001) showed that a better optimisation is obtained when considering all the planned generations simultaneously because the optimal population size at each BC generation depends on the total duration of the breeding scheme. Optimising population sizes over several successive generations requires some numerical calculations. Hospital and Decoux (2002) designed a computer program (*popmin*) that performs these calculations easily.

One can run the *popmin* program to investigate any particular situation. However, the general conclusions are as follows. First, planning to perform a total of more than two BC generations is in general recommended. Second, it is often preferable to genotype more individuals in advanced BC generations than in early BC generations (e.g. for a BC scheme lasting two generations, genotype *more* individuals in BC_2 than in BC_1, *not* the reverse). The latter result is counter intuitive. A hand-waving explanation is that, because at each generation an average 50 % of donor genes return to recipient type by chance due to normal backcross process, it is better to wait until chance events have occurred, before screening favourable genotypes in the remains. This reduces the average number of genotyping over the entire BC scheme. Hence, a typical marker-assisted backcross scheme should involve three to four BC generations in most cases, unless rapid success is sought for particular reasons. Planning to perform three or more BC generations has two main

advantages: first, it permits a more drastic reduction of linkage drag while reducing the genotyping effort. Second, it increases the probability of success (obtaining a R2 genotype) in advanced BC generations. Several studies of the optimisation of such programs have been performed using simulations (Frisch *et al.*, 1999b; Frisch and Melchinger, 2001; van Berloo *et al.*, 2001; Hospital, 2002; Ribaut *et al.*, 2002; Stam, 2003).

Hospital (2001) computed the mean and variance of the length of the intact donor segment around the target gene, for R2 individuals, in any BC generation. This gives the efficacy of background selection for the reduction of linkage drag. The numerical results indicate that the expected length of donor segment on each side of the target gene is approximately half the distance between the target and the flanking marker in BC_1. This length does not decrease much in more advanced BC generations, except for long distances (markers 20 cM from the target or more). Then, using very close markers is the only way to reduce linkage drag substantially. And for short marker distances, optimisation of population sizes as described above becomes very important.

One may also wish to introgress multiple target genes at a time. Hospital and Charcosset (1997) considered the possibility of introgressing multiple QTL. Frisch and Melchinger (2001) studied the case of two known genes. The above discussion extends easily, the main complication being that large population sizes may be required to give a reasonable probability of having at least one individual with the desired marker genotype, and so the intensity of background selection that is possible is rather limited. Hospital and Charcosset (1997) instead suggest a 'pyramidal design' where QTLs are first introgressed one at a time into the recipient in a number of simultaneous but separate backcross programs to produce lines fixed for one of the desired QTL, and these lines are then intercrossed to accumulate QTLs in the same genotype. They suggest that this design will allow more intense background selection than attempting to introgress the same number of QTLs in a single large backcross program. The 'pyramiding' design is described later in this chapter.

21.4.2 Outbred Populations

Where the parental lines are not completely inbred, a number of complications arise. Firstly, the same marker alleles can exist in both donor and recipient lines, so that markers identify neither the introgressed allele nor the desired background genotype with certainty. Ignoring this problem and simply using the methods described above for inbred lines can lead to a dramatic reduction in the efficiency of introgression (van Heelsum *et al.*, 1997a). Results can be improved by properly allowing for the uncertainty about the presence of the introgressed allele, or founding individuals can be selected in such a way as to ensure that markers are fully informative, i.e. that no marker allele found in the founding donor individuals is found in the founding recipient individuals and vice versa (van Heelsum *et al.*, 1997b). We might also expect this problem to become less acute as the density of marker maps increases, leading to more informative marker haplotypes.

Considering the efficiency of MAI is also more complicated in outbred populations. By definition, outbred populations vary for the traits of interest, and therefore phenotypic selection can be performed on the population. The value of an MAI program must therefore be judged relative not to the initial population, but to the population which would have resulted had phenotypic selection been performed instead of the MAI program. The MAI program produces a population, which is superior at the introgressed locus, but inferior elsewhere. This genetic lag arises owing to a number of reasons. Firstly, the selection to ensure that individuals have the appropriate marker alleles at the introgressed

locus means that fewer individuals are available for use in the phenotypic selection programme; secondly, older individuals tend to be used in the backcross program; finally, it is likely that the recipient population had considerably higher average genetic merit than the donor population (Gama *et al.*, 1992; Visscher and Haley, 1999).

Visscher and Haley (1999) investigated the efficiency of background selection using markers and by selection on phenotype alone in introgression experiments and showed that marker-based selection was valuable where the initial breed difference was large, in terms of the within-breed phenotypic variance. If the initial breed difference is small, then phenotypic selection gives lower genetic lag than selection on markers. This makes intuitive sense because markers explain only between-breed variation and so selection on markers is most effective when this is large relative to the within-breed variance. Wall *et al.* (2005) included the prediction of linkage drag and recipient individuals' genomic contributions to the carrier chromosome in the prediction of genetic lag, which turned out to be even higher than that predicted by Gama *et al.* (1992) or Visscher and Haley (1999). This can be reduced by using larger offspring size, if permitted by the species. Else, one should rely on reproductive technology. More generally, marker-assisted introgression is less easy in animals than it is in plants, in particular because of reduced offspring per parent (Koudande *et al.*, 1999; van der Waaij and van Arendonk, 2000). Genetic lag relative to a continuously selected elite population is between 1 and 3 generations, suggesting that for introgression programmes to be viable the introgressed allele must have a value of 1 to 3 generations of genetic gain. This implies that MAI is best suited to the introgression of loci of large effect.

21.5   MARKER-ASSISTED GENE PYRAMIDING

MAI as described above is limited to introgression of very few genes at a time because, when backcrossing for multiple genes, very few individuals are available that carry all the targets after the foreground selection step, leaving little intensity for a possible background selection. In the MAS strategies described earlier in this chapter, possibly more genes were monitored, but the genetic background was not controlled. Moreover, most MAS strategies were studied in the context of classical recurrent selection with random mating of the selected individuals. Here, we want to briefly review some methods that bridge the gap between MAS and MAI. These are sometimes called *genotype building strategies* because favourable alleles are seen as 'building blocks' which one wants to assemble as quickly as possible to create superior varieties. Here, we assume that an ideal genotype (ideotype) has been previously defined at a collection of loci (either known loci, or quantitative trait loci). It is assumed that gene effects are well estimated, and sustainable, so that the selection is only at the molecular level and simply consists in screening the products of meioses (recombination) taking place in successive generations in order to obtain the ideotype as fast as possible.

When several favourable genes are originally hosted by two different parents, the simplest strategy involves production of an F_2, F_3, or (if possible) recombinant inbred lines (RIL) or doubled-haploid (DH) population. Then, screen the population based on molecular markers for individuals homozygous at the requested loci (van Berloo and Stam, 1998). If all the genes cannot be fixed in a single step of selection, it is necessary to cross

again selected individuals with incomplete, but complementary, sets of homozygous loci (Charmet *et al.*, 1999). However, such strategies are limited to small numbers of target loci because the population size necessary to fix the target genes increases exponentially with the number of loci: for example, in a RIL population, the frequency of homozygote is $(1/2)^k$ for k unlinked target loci, i.e. less than 1/1000 individuals for 10 target loci.

For even more loci, recurrent selection should be used. Hospital *et al.* (2000) studied a selection of 50 QTL flanked by marker pairs. The breeding scheme involved selection and random mating of the selected individuals for several generations. The best strategy was to select at each generation a set of individuals that is complementary for their genotypes at the QTL. With this strategy, selection of 3 to 5 individuals among a total of 200 for 10 generations increased the frequency of favourable alleles at the 50 QTL up to 100 % when markers are located exactly on the QTL, but only to 92 % when marker-QTL distance is 5 cM because of recombination taking place between the markers and the QTL. The authors conclude that the only way to accelerate the response to selection, so that favourable QTL alleles are fixed before marker-QTL linkage disequilibrium vanishes, is to replace random mating by pair-wise mating of individuals based on their genotypes.

When the target genes are originally hosted by multiple parents, one can perform a marker-assisted *gene pyramiding* scheme, involving several initial crosses between the parents. For example, if four genes are present in four different lines (L1–L4), one can combine the four genes into a single line in a two-step procedure, crossing e.g. L1 to L2 and L3 to L4, then crossing the hybrids (L1 × L2) to (L3 × L4). This has been applied with some success, for example, in rice (Steele *et al.*, 2006; Ashikari and Matsuoka, 2006). However, if we extend this problem to more loci, possibly linked on the same chromosome, then it is not trivial to decide in which order the pair-wise crosses must be performed, and there is currently no theory available to solve this problem. Yet, a first step in the optimisation of gene pyramiding schemes was provided by Servin *et al.* (2004) as follows.

Assuming that individuals can be selected and mated according to their genotype, the best procedure corresponds to an optimal succession of crosses over several generations (*pedigree*). Assuming that a collection of parents P_i is available, such that each P_i is homozygous for a given target gene G_i, there are several ways to cross those parents to get finally the ideal genotype I combining all target genes. Actually, for n genes, the total number of possible pedigrees is as follows:

$$\mathcal{A}(n) = \prod_{k=2}^{n}(2k-3) = (2n-3)(2n-5)\cdots 1, \tag{21.26}$$

which increases very quickly with n: 105 pedigrees for $n = 5$, 135 for $n = 8$, 3.4×10^7 for $n = 10$ and so on.

Servin *et al.* (2004) provide an algorithm that generates all possible pedigrees. For each pedigree, they compute the probability to obtain the desired genotype from the known recombination fractions between the target loci. Then they deduce the number of individuals (population sizes) that should be genotyped over successive generations until the desired genotype is obtained. The different pedigrees can then be compared on the basis of the population sizes they require and on their total duration (in number of generations) to find the best gene pyramiding scheme. As an example, the optimal gene pyramiding schemes for eight target genes are compared with the reference genotype selection method

with random mating of Hospital *et al.* (2000) described above. The results indicate that optimal gene pyramiding methods can combine the eight targets in three generations less, or with far fewer genotyped individuals than the reference method. An important practical conclusion is that intermating of selected individuals at each generation should be performed in two steps: rather than directly crossing selected genotypes (e.g. $G_1 \times G_2$), it is better to perform a 'double-cross' hybrid $(G_1 \times X) \times (G_2 \times X)$ to avoid the large sample sizes requested by double recombination (Servin *et al.*, 2004), see Figure 21.1.

There is now a growing awareness, at least in the plant community, that exotic genetic resources may hide genes for agriculturally important traits of larger effects than those already segregating in the commercial populations, which was indeed shown in some plants (Eshed *et al.*, 1996). On the basis of this fact, there is a need to use molecular markers to 'unlock the genetic potential from the wild' (Tanksley and McCouch 1997). This will require expensive community resources (Gur and Zamir, 2004; Fernie *et al.*, 2006). At the breeding level, it is likely that use of these resources may be best achieved by the development of numerous 'introgression lines libraries' as has already started in many species (e.g. Liu *et al.*, 2006; Tian *et al.*, 2006; Canady *et al.*, 2005) combined with new selection methods able to 'pyramid' rapidly all detected QTL into new improved genetic material.

21.6 DISCUSSION

All the MAS approaches discussed here exploit the correlation, or in genetic terms linkage disequilibrium or allelic association, between marker loci and QTL. It is quite clear that one of the limits of MAS is that the genes underlying QTL are usually unknown. Recent advances in positional cloning led to the identification of some of those QTL and one might reasonably assume that the huge efforts in sequencing and functional genetics will result in the identification of many other QTL in the next few years. Once QTL are known, it will be possible to define markers totally linked to the causal polymorphism and use them in selection. This will remove the risk of recombination between markers and QTL and increase the efficiency of MAS. However, as highlighted by Bernardo and Charcosset (2006), even if some of the QTL were known, one would still need to test and estimate the allelic effects within the population of interest. Knowing the genes removes some of the statistical hazards related to model selection, but does not remove the risk of model overfitting, especially if the number of known QTL is high compared to the number of individuals.

It has been suggested (Smith, 1967; Smith and Smith, 1993) that, if marker maps are sufficiently dense, markers in nearly complete association with QTL could be identified, even in outbreeding populations, without knowing the genes underlying the QTL. This is clearly closely related to the association-mapping approaches currently popular in human genetics; further details can be found in **Chapter 19** and **Chapter 36**, but see also Risch and Merikangas (1996). In humans, very large numbers of markers are needed: current sets of $300-500\,K$ have been estimated to capture around $80\,\%$ of common genetic variation (Barrett and Cardon, 2006; Pe'er *et al.*, 2006, **Chapter 37**). New technologies allowing marker sets of this size to be genotyped simultaneously at relatively low cost are currently revolutionizing association studies in humans. These technologies also open exciting
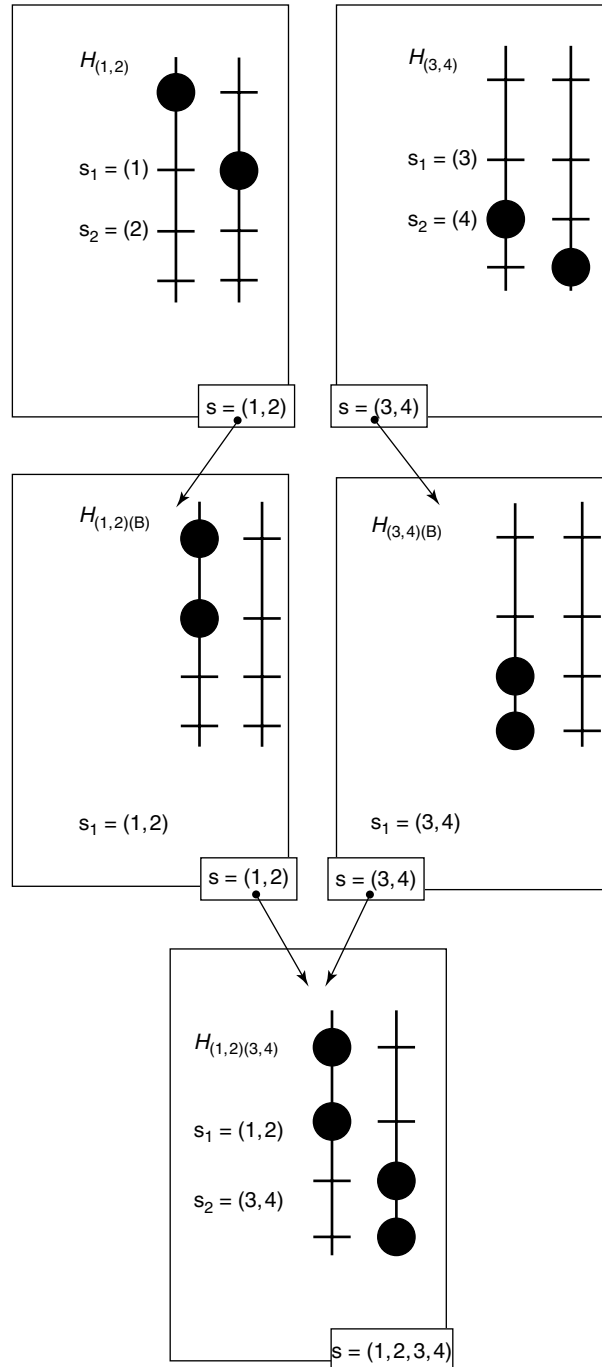
Figure 21.1 The two-step hybridisation procedure for obtaining an intermediate genotype carrying favourable alleles at four loci (1, 2, 3, 4) from two parents carrying favourable alleles at loci (1, 2) and (3, 4). The first step is performed by crossing each parent with a *blank* genotype (not represented here). The resulting offspring, carrying the corresponding target genes in coupling phase on one of their chromosomes, are then mated to obtain the desired genotype, $H(1, 2)(3, 4)$. [Reproduced by permission of the Genetics Society of America.]

new prospects for MAS, especially as the forces generating allelic association, notably hybridisation and selection, are more prevalent in animal or plant populations than in humans, and so it should be easier to capture the relevant genetic variation. New methods of QTL detection have already been proposed using dense maps. Meuwissen *et al.* (2002) proposed a method of fine mapping combining classical QTL detection method and the use of high marker density in animal populations. When detecting QTL in a population that assembles a large genetic diversity, it is important to take the structure of the population into account. Contrary to human populations, plant and animal populations have often been strongly selected and experienced several bottlenecks so drift could generate associations between marker haplotypes and traits. Admixture can be high and may result, if not correctly accounted for, in false associations between marker polymorphism and traits of interest.

It is reasonable to assume that when the marker map is dense enough, all QTL are in linkage disequilibrium with markers at the population level, so the prediction of the polygenic value of individuals can be based on the whole marker data set. In such a case, the number of markers is likely to exceed the number of individuals. The most promising current approach in such a context is the implementation of Bayesian approaches via Markov Chain Monte Carlo. This has the advantage of treating QTL effects as random and so shrinking back towards a prior distribution, providing natural ways of dealing with missing marker or phenotypic information and allowing genetic values to result from integration over different models, so that the model selection problems described above are avoided (see a discussion on that point in Gianola *et al.* (2003)). Some work has been done for outbred populations (Bink *et al.*, 1997; 1998) mainly using Gibbs samplers, and later extended to inbred populations (Xu, 2003). Other related work includes the use of reversible jump Metropolis–Hastings samplers for QTL mapping in inbred lines (e.g. Sillanpää and Arjas, 1998). Meuwissen *et al.* (2001) compared several approaches, including least-squares estimates of marker effects, BLUP models and a Bayesian approach, and showed that the Bayesian approach implemented using Markov Chain Monte Carlo provided good predictions of the genetic values. However, in contrast to least-squares or BLUP approaches that can be easily implemented, Bayesian approaches can be extremely computationally demanding, and monitoring convergence in the complex parameter spaces required is not easy. Ter Braak *et al.* (2005) highlighted that inadequate priors in Xu (2003) yielded to improper posterior distribution, slow convergence and thus incorrect results.

A key advantage of these Bayesian approaches is in removing the need to choose a subset of markers associated with the trait on which to apply selection. As noted earlier, this model selection step is problematic statistically; moreover, for high throughput genotyping methods, there is no or little reduction in cost by reducing the number of markers used in this way. Even if the cost per marker data point is low, the genotyping cost per individual remains high. So the economic interest of such approaches needs to be addressed.

For all these reasons, methods relying on family level associations will certainly remain important. Even here, increased density of marker maps is of course helpful, for instance, in increasing the informativeness of marker haplotypes. Those marker haplotypes could be a way to make the link between different results obtained in association studies or in QTL detection experiments. In many species, a large number of QTL results have now been published for a given trait using different genetic backgrounds, different environments,

etc. The crucial point in the next few years will be to integrate all this information to better understand the genetic architecture of quantitative traits. This would help us in anticipating results of MAS better and thus in defining new efficient strategies of selection. Dense marker techniques are certainly useful tools but raise new questions. So, the development of theory for MAS and MAI will still remain an active area. It is striking to note the convergence between approaches developed recently for inbred and outbred populations. Even if some specific issues remain, this convergence might facilitate the development of general MAS methods that would benefit to both animal and plant breeders.

The work summarised in this chapter is almost unanimous in concluding that appropriately designed MAS programs can increase selection response compared to purely phenotypic selection, but it is less clear that MAS is economically viable once the additional costs of marker genotyping are allowed for. This will change as genotyping costs are reduced, but, at present, the most attractive applications of MAS are in situations in which the use of MAS allows a corresponding decrease in phenotyping costs, for example, by reduction of the generation interval, or for introgression of, or selection for, a single major gene. In the latter case, the efficiency of MAS is critically dependent on the accuracy of the mapping programme, which identified the QTL. In particular, it is essential that the identified QTL is genuine and not a false positive. In many cases, this will require an initial positive result to be confirmed in a second, independent study (Spelman and Bovenhuis, 1998a).

Acknowledgments

REFERENCES

Ashikari, M. and Matsuoka, M. (2006). Identification, isolation and pyramiding of quantitative trait loci for rice breeding. *Trends in Plant Science* **11**, 344–350.

van Arendonk, J.A.M., Tier, B. and Kinghorn, B.P. (1994). Use of multiple genetic markers in prediction of breeding values. *Genetics* **137**, 319–329.

Barrett, J.C. and Cardon, L.R. (2006). Evaluating coverage of genome-wide association studies. *Nature Genetics* **38**, 659–662.

Benjamini, Y. and Yekutieli, D. (2005). Quantitative trait loci analysis using the false discovery rate. *Genetics* **171**, 783–790.

van Berloo, R., Aalbers, H., Werkman, A. and Niks, R.E. (2001). Resistance QTL confirmed through development of QTL-NILs for barley leaf rust resistance. *Molecular Breeding* **8**, 187–195.

van Berloo, R. and Stam, P. (1998). Marker assisted selection in autogamous RIL populations: a simulation study. *Theoretical and Applied Genetics* **96**, 147–154.

van Berloo, R. and Stam, P. (2001). Simultaneous marker-assisted selection for multiple traits in autogamous crops. *Theoretical and Applied Genetics* **102**, 1107–1112.

Bernardo, R. (2004). What proportion of declared QTL in plants are false? *Theoretical and Applied Genetics* **109**, 419–424.

Bernardo, R. and Charcosset, A. (2006). Usefulness of gene information in marker-assisted recurrent selection: a simulation appraisal. *Crop Science* **46**, 614–621.

Bernardo, R., Moreau, L. and Charcosset, A. (2006). Optimising the number and fitness of selected individuals in recurrent selection. *Crop Science* **46**, 1972–1980.

Bink, M.C.A.M., van Arendonk, J.A.M. and Quaas, R.L. (1997). Breeding value estimation with incomplete marker data. *Genetics, Selection, Evolution* **30**, 45–58.

Bink, M.C.A.M., Quaas, R.L. and van Arendonk, J.A.M. (1998). Bayesian estimation of dispersion parameters with a reduced animal model including polygenic and QTL effects. *Genetics, Selection, Evolution* **30**, 103–125.

Blanc, G., Charcosset, A., Mangin, B., Gallais, A. and Moreau, L. (2006). Connected populations for detecting QTL and testing for epistasis, an application in maize. *Theoretical and Applied Genetics* **113**, 206–224.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* **87**, 738–754.

Breiman, L. (1995). Better subset selection using the nonnegative garrote. *Technometrics* **37**, 373–384.

Brown, P.J. and Zidek, J.V. (1980). Adaptive multivariate ridge regression. *Annals of Statistics* **8**, 64–74.

Bulmer, M. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.

Canady, M.A., Meglic, V. and Chetelat, R.T. (2005). A library of solanum lycopersicoides introgression lines in cultivated tomato. *Genome* **48**, 685–697.

Cantet, R.J.C. and Smith, C. (1991). Reduced animal model for marker assisted selection using best linear unbiased prediction. *Genetics, Selection, Evolution* **23**, 221–233.

Chakraborty, R., Moreau, L. and Dekkers, J.C.M. (2002). A method to optimize selection on multiple identified quantitative trait loci. *Genetics, Selection, Evolution* **34**, 145–170.

Charmet, G., Robert, N., Perretant, M.R., Gay, G., Sourdille, P., Groos, C., Bernard, S. and Bernard, M. (1999). Marker-assisted recurrent selection for cumulating additive and interactive QTLs in recombinant inbred lines. *Theoretical and Applied Genetics* **99**, 1143–1148.

Crepieux, S., Lebreton, C., Servin, B. and Charmet, G. (2004). Quantitative trait loci (QTL) detection in multicross inbred designs: recovering QTL identical-by-descent status information from marker data. *Genetics* **168**, 1737–1749.

Cullis, B.R. and Gleeson, A.C. (1991). Spatial analysis of field experiments-an extension to two dimensions. *Biometrics* **47**, 1449–1460.

Dekkers, J.C.M. and van Arendonk, J.A.M. (1998). Optimizing selection for quantitative traits with information on an identified locus in outbred populations. *Genetical Research* **71**, 257–275.

Dekkers, J.C.M., Chakraborty, R., Moreau, L. and Settar, P. (2002). Optimal selection on multiple identified quantitative trait loci. *Genetics Selection Evolution* **34**, 171–192.

Edwards, M.D. and Page, N.J. (1994). Evaluation of marker-assisted selection through computer simulation. *Theoretical and Applied Genetics* **88**, 376–382.

Eshed, Y., Gera, G. and Zamir, D. (1996). A genome-wide search for wild-species alleles that increase horticultural yield of processing tomatoes. *Theoretical and Applied Genetics* **93**, 877–886.

Eshed, Y. and Zamir, D. (1996). Less-than-additive epistatic interactions of quantitative trait loci in Tomato. *Genetics* **143**, 1807–1817.

Fernando, R.L. and Grossman, M. (1989). Marker-assisted selection using best linear unbiased prediction. *Genetics, Selection, Evolution* **21**, 467–477.

Fernando, R.L., Nettleton, D., Southey, B.R., Dekkers, J.C.M., Rothschild, M.F. and Soller, M. (2004). Controlling the proportion of false positives in multiple dependants tests. *Genetics* **166**, 611–619.

Fernie, A.R., Tadmor, Y. and Zamir, D. (2006). Natural genetic variation for improving crop quality. *Current Opinion in Plant Biology* **9**, 196–202.

Frisch, M., Bohn, M. and Melchinger, A.E. (1999a). Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. *Crop Science* **39**, 967–975.

Frisch, M., Bohn, M. and Melchinger, A.E. (1999b). Comparison of selection strategies for marker assisted backcrossing of a gene. *Crop Science* **39**, 1295–1301.

Frisch, M. and Melchinger, A.E. (2001). Marker-assisted backcrossing for simultaneous introgression of two genes. *Crop Science* **41**, 1716–1725.

Frisch, M. and Melchinger, A.E. (2005). Selection theory for marker-assisted backcrossing. *Genetics* **170**, 909–917.

Gama, L.T., Smith, C. and Gibson, J.P. (1992). Transgene effects, introgression strategies and testing schemes in pigs. *Animal Production* **54**, 427–440.

Gianola, D., Perez-Enciso, M. and Toro, M.A. (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics* **163**(1), 347–365.

Gibson, J.P. (1994). Short-term gain at the expense of long-term response with selection of identified loci. *Proceedings of the 5th world Congress Congress of Genetics Applied to Livestock Production*, Vol. 21. University of Guelph, Guelph, pp. 201–204.

Gimelfarb, A. and Lande, R. (1994a). Simulation of marker-assisted selection in hybrid populations. *Genetical Research* **63**, 39–47.

Gimelfarb, A. and Lande, R. (1994b). Simulation of marker assisted selection for non-additive traits. *Genetical Research* **64**, 127–136.

Gimelfarb, A. and Lande, R. (1995). Marker assisted selection and marker QTL associations in hybrid populations. *Theoretical and Applied Genetics* **91**, 522–528.

Goddard, M.E. (1992). A mixed model for analysis of data on multiple genetic markers. *Theoretical and Applied Genetics* **83**, 877–886.

Gur, A. and Zamir, D. (2004). Unused natural variation can lift yield barriers in plant breeding. *PLOS Biology* **2**, 1610–1615.

Haley, C.S. and Knott, S.A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

van Heelsum, A.M., Haley, C.S. and Visscher, P.M. (1997a). Marker-assisted introgression using non-unique marker alleles I: selection on the presence of linked marker alleles. *Animal Genetics* **28**, 181–187.

van Heelsum, A.M., Haley, C.S. and Visscher, P.M. (1997b). Marker-assisted introgression using non-unique marker alleles II: selection on probability of presence of the introgressed allele. *Animal Genetics* **28**, 188–194.

Henderson, C.R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in the prediction of breeding values. *Biometrics* **32**, 69–83.

Heyde, C.C. (1997). *Quasi-Likelihood and its Application*. Springer-Verlag, New York.

Hill, W.G. (1993). Variation in genetic composition in backcrossing programs. *The Journal of Heredity* **84**, 212–213.

Hillel, J.T., Schaap, T., Haberfield, A., Jeffreys, A.J., Plotzky, Y., Cahener, A. and Lavi, U. (1990). DNA fingerprint applied to gene introgression breeding programs. *Genetics* **124**, 783–789.

Hillel, J., Verrinder, A.M., Gibbens, R., Etches, R.J. and Shaver, D.M. (1993). Strategies for the rapid introgression of a specific gene modification into a commercial poultry flock from a single carrier. *Poultry Science* **72**, 1197–1211.

Hoeschele, I. (1993). Elimination of quantitative trait loci equations in an animal model incorporating genetic marker data. *Journal of Dairy Science* **76**, 1693–1713.

Hospital, F. (2001). Size of donor chromosome segments around introgressed loci and reduction of linkage drag in marker-assisted backcross programs. *Genetics* **158**, 1363–1379.

Hospital, F. (2002). Marker-assisted backcross breeding: a case study in genotype building theory. In *Quantitative Genetics, Genomics and Plant Breeding*, M.S. Kang, ed. CAB International, New York, Oxon.

Hospital, F. and Charcosset, A. (1997). Marker-assisted introgression of quantitative trait loci. *Genetics* **147**, 1469–1485.

Hospital, F., Chevalet, C. and Mulsant, P. (1992). Using marker in gene introgression breeding programs. *Genetics* **132**, 1199–1210.

Hospital, F. and Decoux, G. (2002). Popmin: a program for the numerical optimization of population sizes in marker-assisted backcross programs. *The Journal of Heredity* **93**, 383–384.

Hospital, F., Goldringer, I. and Openshaw, S. (2000). Efficient marker-based recurrent selection for multiple quantitative loci. *Genetical Research* **75**, 357–368.

Hospital, F., Moreau, L., Lacoudre, F., Charcosset, A. and Gallais, A. (1997). More on the efficiency of marker-assisted selection. *Theoretical and Applied Genetics* **95**, 1181–1189.

Jannink, J.L. and Jansen, R.C. (2001). Mapping epistatic quantitative trait loci with one dimensional genome searches. *Genetics* **157**, 445–454.

Jannink, J.L. and Wu, X.L. (2003). Estimating allelic number and identity in state of QTLs in interconnected families. *Genetical Research* **81**, 133–144.

Jansen, R.C., Jannink, J.L. and Beavis, W.D. (2003). Mapping quantitative trait loci in plant breeding populations: use of parental haplotype sharing. *Crop Science* **43**, 829–834.

Johnson, R. (2004). Marker-assisted selection. In *Plant Breeding Reviews. Part I: Long-Term Selection: Maize*, J. Janick, ed. John Wiley & Sons, Hoboken, NJ, pp. 293–309.

Kashi, Y., Hallerman, E. and Soller, M. (1990). Marker assisted selection of candidate bulls for progeny testing programmes. *Animal Production* **51**, 63–74.

Koudande, O.D., Thomson, P.C. and Van Arendonk, J.A.M. (1999). A model for population growth of laboratory animals subjected to marker-assisted introgression: how many animals do we need? *Heredity* **82**, 16–24.

Lande, R. and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**, 743–756.

Lange, C. and Whittaker, J.C. (2001). Mapping Quantitative Trait Loci using generalized estimating equations. *Genetics* **159**, 1325–1337.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Liu, S.B., Zhou, R.G., Dong, Y.C., Li, P. and Jia, J.Z. (2006). Development, utilization of introgression lines using a synthetic wheat as donor. *Theoretical and Applied Genetics* **112**, 1360–1373.

Luo, Z.W., Thompson, R. and Woolliams, J.A. (1997). A population genetics model of marker-assisted selection. *Genetics* **146**(3), 1173–1183.

Mackinnon, M.J. and Georges, M. (1997). A bottom up approach to marker assisted selection. *Livestock Production Science* **54**, 227–248.

Manfredi, E., Barbieri, M., Fournet, F. and Elsen, J.M. (1998). A dynamic deterministic model to evaluate breeding strategies under mixed inheritance. *Genetics Selection Evolution* **30**, 127–148.

Markel, P., Shu, P., Ebeling, C., Carlson, G.A., Nagle, D.L., Smutko, J.S. and Moore, K.J. (1997). Theoretical and empirical issues for marker-assisted breeding of congenic mouse strains. *Nature Genetics* **17**, 280–284.

Martínez, O. and Curnow, R.N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480–488.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Melchinger, A.E. (1990). Use of molecular markers in breeding for oligogenic disease resistance. *Plant Breeding* **104**, 1–19.

Meuwissen, T.H.E. and Goddard, M.E. (1996). The use of marker haplotypes in animal breeding schemes. *Genetics, Selection, Evolution* **28**, 161–176.

Meuwissen, T.H.E. and Goddard, M.E. (1999). Marker assisted estimation of breeding values when marker information is missing on many animals. *Genetics, Selection, Evolution* **31**, 375–394.

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Meuwissen, T.H.E., Karlsen, A., Lien, S., Olsaker, I. and Goddard, M.E. (2002). Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161**, 373–379.

Meuwissen, T.H.E. and Sonesson, A.K. (2004). Genotype-assisted contribution selection to maximize selection response over a specified time period. *Genetical Research* **84**, 109–116.

Miller, A.J. (1990). *Subset Selection in Regression*. Chapman and Hall, London.

Moerkerke, B., Goetghebeur, E., De Riek, J. and Roldan-Ruiz, I. (2006). Significance and impotence: towards a balanced view of the null and the alternative hypotheses in marker selection for plant breeding. *Journal of the Royal Statistical Society, Series A* **169**, 61–79.

Moreau, L., Charcosset, A. and Gallais, A. (2000). Economic efficiency of one cycle of marker-assisted selection. *Crop Science* **40**, 329–337.

Moreau, L., Charcosset, A. and Gallais, A. (2004). Use of trial clustering to study QTL*environment effects for grain yield and related traits in maize. *Theoretical and Applied Genetics* **110**, 92–105.

Moreau, L., Charcosset, A., Hospital, F. and Gallais, A. (1998). Marker-assisted selection efficiency in populations of finite size. *Genetics* **148**, 1353–1365.

Moreau, L., Monod, H., Charcosset, A. and Gallais, A. (1999). Marker-assisted selection efficiency in populations of finite size. *Theoretical and Applied Genetics* **98**, 234–242.

Neimann-Sorenson, A. and Robertson, A. (1961). The association between blood groups and several production characteristics in three Danish cattle breeds. *Acta Agriculturae Scandinavica* **11**, 163–196.

Pe'er, I., de Bakker, P.I.W., Maller, J., Yelensky, R., Altshuler, D. and Daly, M.J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics* **38**, 663–667.

Podlich, D.W., Winkler, C.R. and Cooper, M. (2004). Mapping as you go: an effective approach for marker-assisted selection of complex traits. *Crop Science* **44**, 1560–1571.

Rebai, A. and Goffinet, B. (2000). More about quantitative trait locus mapping with diallel design. *Genetical Research* **75**, 243–247.

Ribaut, J.-M., Jiang, C. and Hoisington, D. (2002). Simulation experiments on efficiencies of gene introgression by backcrossing. *Crop Science* **42**, 557–565.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Rothschild, M.F., Jacobson, C., Vaske, D.A., Tuggle, C.K., Wang, L., Short, T.H., Eckardt, G.R., Sasaki, S., Vincent, A., McLaren, D.G., Southwood, O., Van der Steen, H., Mileham, A. and Plastow, G. (1996). The estrogen receptor locus is associated with a major gene influencing litter-size in pigs. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 201–205.

Saito, S. and Iwaisaki, H. (1997). A reduced animal model approach to predicting total additive genetic merit for marker-assisted selection. *Genetics, Selection, Evolution* **29**, 25–34.

Saito, S., Matsuda, H. and Iwaisaki, H. (1998). Best linear prediction of additive genetic merit using a combined-merit sire and dam model for marker-assisted selection. *Genes and Genetic Systems* **73**, 65–69.

Sales, J. and Hill, W.G. (1976). Effect of sampling errors on efficiency of selection indices. 1. Use of information from relatives for single trait improvement. *Animal Production* **22**, 1–17.

Servin, B. (2005). Using markers to reduce the variation in the genomic composition in marker-assisted backcrossing. *Genetical Research* **85**, 151–157.

Servin, B., Martin, O.C., Mezard, M. and Hospital, F. (2004). Toward a theory of marker-assisted gene pyramiding. *Genetics* **168**, 513–523.

Sillanpää, M.J. and Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line data. *Genetics* **148**, 1373–1388.

Smith, C. (1967). Improvement in metric traits through specific genetic loci. *Animal Production* **9**, 349–358.

Smith, A., Cullis, B., Luckett, D., Hollamby, G. and Thompson, R. (2002). Exploring variety-environment data using random effects AMMI models with adjustments for spatial field trend: Part2: applications. In *Quantitative Genetics, Genomics and Plant Breeding*, M.S. Kang, ed. CABI Publishing, Oxon, pp. 337–351.

Smith, C. and Smith, D.B. (1993). *Animal Breeding Abstracts* **61**, 197–204.

Soller, M. and Beckmann, J.S. (1983). Genetic polymorphism in varietal identification and genetic improvement. *Theoretical and Applied Genetics* **67**, 25–33.

Spelman, R. and van Arendonk, J.A.M. (1997). Effect of inaccurate parameter estimates on genetic response to marker-assisted selection in an outbred population. *Journal of Dairy Science* **80**, 3399–3410.

Spelman, R. and Bovenhuis, H. (1998a). Moving from QTL experimental results to the utilization of QTL in breeding programmes. *Animal Genetics* **29**, 77–84.

Spelman, R. and Bovenhuis, H. (1998b). Genetic response from marker assisted selection in an outbred population for differing marker bracket sizes and with two identified quantitative trait loci. *Genetics* **148**, 1389–1396.

Spelman, R.J. and Garrick, D.J. (1998). Genetic and economic responses for within-family marker assisted selection in dairy cattle breeding schemes. *Journal of Dairy Science* **81**, 2942–2950.

Stam, P. (2003). Marker-assisted introgression: speed at any cost? In *EUCARPIA Leafy Vegetables 2003*, Th.J.L. van Hintum, A. Lebeda, D. Pink and J.W. Schut, eds. EUCARPIA, Valencia, pp. 117–124.

Stam, P. and Zeven, A.C. (1981). The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. *Euphytica* **30**, 227–238.

Steele, K.A., Price, A.H., Shashidhar, H.E. and Witcombe, J.R. (2006). Marker-assisted selection to introgress rice QTLs controlling root traits into an Indian upland rice variety. *Theoretical and Applied Genetics* **112**, 208–221.

Strauss, S.H., Lande, R. and Namkoong, G. (1992). Limitations of molecular-marker-aided selection in forest tree breeding. *Canadian Journal of Forest Research* **22**, 1050–1061.

Tanksley, S.D. (1983). Molecular markers in plant breeding. *Plant Molecular Biology Reporter* **1**, 3–8.

Tanksley, S.D. and McCouch, S.R. (1997). Seed banks and molecular maps: unlocking the genetic potential from the wild. *Science* **277**, 1063–1066.

Tanksley, S.D., Medino-Filho, D.H. and Rick, C.M. (1981). The effect of isozyme selection on metric characters in an interspecific backcross of tomato: basis of an early screening procedure. *Theoretical and Applied Genetics* **60**, 291–296.

Ter Braak, C.J.F., Boer, M.P. and Bink, M.C.A.M. (2005). Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. *Genetics* **170**(3), 1435–1438.

Tian, F., Li, D.J., Fu, Q., Zhu, Z.F., Fu, Y.C., Wang, X.K. and Sun, C.Q. (2006). Construction of introgression lines carrying wild rice (Oryza rufipogon Griff.) segments in cultivated rice (Oryza sativa L.) background and characterization of introgressed segments associated with yield-related traits. *Theoretical and Applied Genetics* **112**, 570–580.

Thomson, P.C. (2003). A generalized estimating equations approach to quantitative trait locus detection of non-normal traits. *Genetics, Selection, Evolution* **35**, 257–280.

Utz, H.F., Melchinger, A.E. and Schon, C.C. (2000). Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics* **154**, 1839–1849.

Villanueva, B., Pong-Wong, R. and Woolliams, J.A. (2002). Marker-assisted selection with optimised contributions of the candidates to selection. *Genetics Selection Evolution* **34**, 679–703.

Visscher, P.M. (1996). Proportion of the variation in genetic composition in backcrossing programs explained by genetic markers. *The Journal of Heredity* **87**(2), 136–138.

Visscher, P.M. (1999). Speed congenics: accelerated genome recovery using genetic markers. *Genetical Research* **74**, 81–85.

Visscher, P.M. and Haley, C.S. (1999). On the efficiency of marker-assisted introgression. *Animal Science* **68**, 59–68.

Visscher, P.M., Haley, C.S. and Knott, S.A. (1996a). Mapping QTLs for binary traits in backcross and F_2 populations. *Genetical Research* **68**, 55–63.

Visscher, P.M., Haley, C.S. and Thompson, R. (1996b). Marker-assisted introgression in backcross breeding programs. *Genetics* **144**, 1923–1932.

van der Waaij, E.H. and van Arendonk, J.A.M. (2000). Introgression of genes responsible for disease resistance in a cattle population selected for production: genetic and economic consequences. *Animal Science* **70**, 207–220.

Wall, E., Visscher, P.M., Hospital, F. and Woolliams, J.A. (2005). Genomic contributions in livestock gene introgression programmes Genet. *Genetics, Selection, Evolution* **37**, 291–313.

Whittaker, J.C., Curnow, R.N., Haley, C.S. and Thompson, R. (1995). Using marker-maps in marker-assisted selection. *Genetical Research* **66**, 255–265.

Whittaker, J.C., Haley, C.S. and Thompson, R. (1997). Optimal weighting of information in MAS. *Genetical Research* **69**, 137–144.

Whittaker, J.C., Thompson, R. and Denham, M.C. (2000). Marker-assisted selection using ridge regression. *Genetical Research* **75**, 249–252.

Xie, C. and Xu, S. (1998). Efficiency of multistage marker-assisted selection in the improvement of multiple quantitative traits. *Heredity* **80**, 489–498.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**, 789–801.

Young, N.D. and Tanksley, S.D. (1989). RFLP analysis of the size of chromosomal segments retained around the *tm-2* locus of tomato during backcross breeding. *Theoretical and Applied Genetics* **77**, 353–359.

Zhang, W. and Smith, C. (1992). Computer simulation of marker-assisted selection utilizing linkage disequilibrium. *Theoretical and Applied Genetics* **83**, 813–820.

Zhang, W. and Smith, C. (1993). Simulation of marker-assisted selection utilizing linkage disequilibrium: the effects of several additional factors. *Theoretical and Applied Genetics* **86**, 492–496.

Part 5

Population Genetics

Mathematical Models in Population Genetics

C. Neuhauser

Department of Ecology, Evolution and Behavior, University of Minnesota, Saint Paul, MN, USA

Throughout the history of population genetics, mathematical models have played an important role in elucidating the effects of mutation and selection on the genetic diversity of organisms. Mathematical models provided the theoretical foundation of neo-Darwinism; sophisticated mathematical tools aided Kimura in establishing the neutral molecular theory. Mathematical models in population genetics today are crucial in the development of statistical tools for analyzing molecular data.

This chapter emphasizes models for selection, but also includes the discussion of neutral models. After a brief history of the role of selection in evolution, basic mathematical models are introduced together with the diffusion approximation. A discussion of coalescent theory follows, with primary focus on selection. A short discussion on how to detect selection concludes the chapter.

22.1 A BRIEF HISTORY OF THE ROLE OF SELECTION

> This preservation of favourable variations and the rejection of injurious variations, I call Natural Selection. Variations neither useful nor injurious would not be affected by natural selection, and would be left a fluctuating element, as perhaps we see in the species called polymorphic. (Darwin, 1985)

Charles Darwin was the first to formulate the concept of natural selection and to apply it to evolution and adaptation. Darwin developed his theory of evolution without a knowledge of the source of variation. Today we know that the hereditary information of most organisms is encoded in deoxyribonucleic acid (DNA) and that variation is caused by mutations; the definition of natural selection remains the same, namely the differential reproductive success of different genotypes.

When Darwin first proposed his theory of natural selection, he believed that genetic diversity was primarily driven by natural selection and that variations accumulated

gradually over time, as expressed in the quote *Natura non facit saltum* ('Nature makes no leaps'). The concept of gradual evolution met immediately with criticism since the fossil record had not yielded transitional forms at that time that would indicate a gradual change from one species to another, as would be expected under gradual evolution.

Shortly after Darwin's proposal of the nature of evolution, an Austrian monk, Gregor Mendel, carried out experiments on peas in 1865 and discovered the basic rules of inheritance. The importance of Mendel's discoveries was only realized in 1900 when the plant breeders de Vries, Correns, and Tschermak independently obtained plant breeding data that could be interpreted by Mendel's rules of inheritance. The cause of variation, namely mutations, was first described by Hugo de Vries. Based on plant breeding experiments, de Vries concluded that mutations caused drastic, nongradual changes.

The relative importance of mutations versus selection as the driving forces of evolution was a matter of dispute. Proponents of the Darwinian Theory asserted that evolution proceeds by small steps, namely selection operating on small variations; whereas proponents of the Mendelian theory believed that evolution proceeds by large leaps caused by mutations.

This controversy continued until the early 1930s when Fisher (1930), Haldane (1932), and Wright (1931) synthesized the Mendelian theory of inheritance and the Darwinian theory of evolution; this synthesis, called the *neo-Darwinism* or *synthetic theory of evolution*, formed the foundation of the modern theory of population genetics. It emphasizes the importance of natural selection acting on variations caused by mutations in the course of evolution.

When protein sequences became available in the 1960s, it soon became clear that genetic diversity was far greater than had been expected. This prompted Kimura (1968a), and King and Jukes (1969) to question the importance of natural selection as the driving force of evolution. Instead, they proposed that most variation was selectively neutral. This started a heated debate between the proponents of the neutral theory and the selectionists. In subsequent work (Kimura, 1968b; 1977; 1979; 1983; Kimura and Ohta, 1973) developed this idea much further; the controversy has not been resolved. A principal conclusion of the neutral theory is that genetic diversity is largely caused by random genetic drift, implying that the genetic diversity seen in populations is a transient phenomenon: Mutations are introduced at random and they either go to fixation or are lost solely due to stochastic forces.

22.2   MUTATION, RANDOM GENETIC DRIFT, AND SELECTION

Mutant alleles that have little effect on the phenotype of the organism may remain in the population until they either become fixed or lost due to stochastic forces. Other alleles are maintained in or quickly eliminated from a population by selective forces. Mathematical models that are based on the laws of inheritance can illuminate the role and relative importance of stochastic and selective forces. An excellent reference for classical mathematical population genetics is Ewens (2004); many of the models discussed here (and much more) can be found in his book.

The simplest mathematical models track allele frequencies in a randomly mating, monoecious population. Changes in allele frequencies are caused by mutation, random

genetic drift, and selection. To understand their effects, we discuss each factor separately at first.

22.2.1 Mutation

Consider a locus with two alleles A_1 and A_2, and let $p(n)$ and $q(n) = 1 - p(n)$ be the frequencies of A_1 and A_2, respectively, at generation n. We assume that generations are nonoverlapping, i.e. in each generation the entire population undergoes random mating; furthermore, A_1 mutates to A_2 at rate u and A_2 to A_1 at rate v, and an allele can mutate at most once per generation. In an infinitely large population, the dynamics can be described by a deterministic equation. In the absence of selective forces and under random mating, the gene frequency of A_1 in the next generation is obtained in the following way: An A_1 allele in generation $n + 1$ could have been either an A_1 allele in generation n that did not mutate (with probability $1 - u$), or an A_2 allele that mutated from A_2 to A_1 (with probability v). Therefore, the gene frequency of A_1 in generation $n + 1$ is given by

$$p(n + 1) = (1 - u)p(n) + v(1 - p(n)). \qquad (22.1)$$

This can be solved in terms of the initial gene frequency of A_1, $p(0)$, and one finds

$$p(n) = \frac{v}{u + v} + \left(p(0) - \frac{v}{u + v} \right) (1 - u - v)^n. \qquad (22.2)$$

The mutation probabilities u and v are typically quite small (of order 10^{-6} or 10^{-5}). In the long run (i.e., when n is very large), the term $(1 - u - v)^n$ approaches zero. For instance, assume $u = v = 10^{-6}$; then $(1 - u - v)^n$ is equal to 0.1353 when $n = 10^6$ and equal to 2.061×10^{-9} when $n = 10^7$. This indicates that changes caused by mutations alone might be quite slow, a realization that prompted population geneticists in the 1930s (Wright, 1931; Haldane, 1932) to suggest that, though mutations are the source of variation, their role in evolution might be limited – see Nei (1987) for a discussion.

Eventually – i.e. in the limit as n tends to infinity – the right-hand side of (22.2) approaches the value $v/(u + v)$. This value is called an *equilibrium*. The equilibrium is also characterized by $p(n + 1) - p(n) = 0$, which means that the gene frequencies do not change over time. If we denote the equilibrium frequency of A_1 by $\hat{p}$, set $p(n) = p(n + 1) = \hat{p}$ in (22.1), and solve for $\hat{p}$, we find (as before)

$$\hat{p} = \frac{v}{u + v}.$$

If $\hat{q} = 1 - \hat{p}$ denotes the equilibrium frequency of A_2, then $\hat{q} = u/(u + v)$. Provided u and v are both positive, we conclude that mutation allows for the maintenance of the two alleles in the population.

22.2.2 Random Genetic Drift

The assumption of an infinite population allowed us to use a deterministic formulation when we investigated the effects of mutation. In a finite population, the random sampling of gametes alone causes changes in gene frequencies. This process is known as *random genetic drift*. To investigate the consequences of random genetic drift, we again look at a single locus with two alleles, A_1 and A_2.

Assume a randomly mating diploid population of size N (or, which is the same in this case, a haploid population of size $2N$) with nonoverlapping generations. Each generation, $2N$ gametes are sampled at random from the parent generation. If $Y(n)$ denotes the number of gametes of type A_1 at generation n, then, in the absence of mutation and selection, the number of A_1 alleles at time $n+1$ is given by the binomial distribution. Namely, the probability that there are j gametes of type A_1 at generation $n+1$, given that there were i gametes of type A_1 at generation n, is

$$P(Y(n+1) = j \mid Y(n) = i) = \binom{2N}{j} p^j (1-p)^{2N-j}, \tag{22.3}$$

where $p = i/2N$. This model is known as the *Wright–Fisher model*. If we follow a population that evolves according to the model defined in (22.3), we will observe that its behavior is quite unpredictable due to the stochastic nature of the model. In particular, this means that if we follow different populations, all of equal size and each following the dynamics described in (22.3), they will follow different trajectories over time (see Figure 22.1). This discrete-time stochastic process is an example of a Markov chain; we will discuss Markov chains in more detail below. An elementary reference on Markov processes is, for instance, Karlin and Taylor (1975).
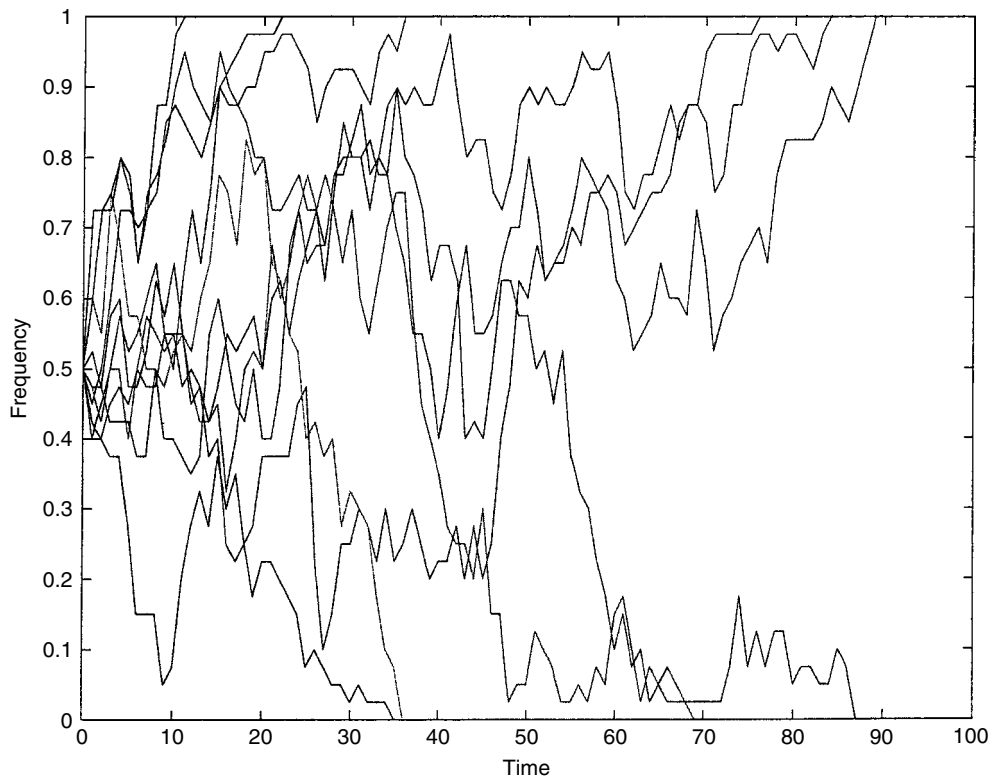
Figure 22.1 Ten trajectories for the neutral Wright–Fisher model when the population size $N = 20$ and the initial gene frequency of A_1 is 0.5.

Since there are no mutations in the model, eventually one of the two alleles will be lost (and the other one fixed), as can be seen in Figure 22.1. The larger the population size, the longer this process of fixation takes. Regardless of the population size, if initially the frequency of A_1 alleles is π_1, then one can show that the probability of fixation of allele A_1 is π_1 (and consequently, the probability of fixation of A_2 is $1 - \pi_1$). This can be understood intuitively, following an argument by Ewens (1972). Namely, after a long enough time, all individuals in the population must have descended from just one of the individuals present at generation 0. The probability that this common ancestor was of type A_1 is equal to the relative frequency of A_1 alleles at generation 0, which is π_1.

22.2.3 Selection

We now turn to the third component, selection. Selection can act on different parts of the life history of an organism; differential fecundity and viability are just two examples. The simplest model of viability selection assumes that selection affects survival between the zygote and adult stage of a diploid organism in a randomly mating population of infinite size, in which generations do not overlap. It is assumed that each genotype has a fixed, specified fitness. Since the population size is infinite, changes in allele frequencies can be described by deterministic equations. In the case of one gene with two alleles, A_1 and A_2, there are three genotypes: A_1A_1, A_1A_2, and A_2A_2. We denote their respective fitnesses by w_{11}, w_{12}, and w_{22}. In the case of viability selection, the fitness w_{ij} reflects the relative survival chances of zygotes of genotype A_iA_j. If the population is in Hardy–Weinberg equilibrium and the frequencies of A_1 and A_2 in generation n are $p(n)$ and $q(n) = 1 - p(n)$, respectively, then, ignoring mutations, the gene frequencies at the next generation are

$$p(n + 1) = \frac{p(n)[p(n)w_{11} + q(n)w_{12}]}{\bar{w}}$$

and

$$q(n + 1) = \frac{q(n)[p(n)w_{12} + q(n)w_{22}]}{\bar{w}},$$

where $\bar{w}$, the average fitness, is chosen so that $p(n + 1) + q(n + 1) = 1$, i.e.

$$\bar{w} = p^2(n)w_{11} + 2p(n)q(t)w_{12} + q^2(n)w_{22}.$$

The predictions of this model are straightforward. If $w_{11} > w_{12} > w_{22}$ (or $w_{11} < w_{12} < w_{22}$), the case of directional selection, $A_1(A_2)$ becomes fixed in the population. If $w_{11}, w_{22} < w_{12}$, the case of overdominance, a stable polymorphism results. If $w_{12} < w_{11}, w_{22}$, the case of underdominance, the polymorphism is unstable and depending on the initial gene frequencies, either A_1 or A_2 becomes fixed.

22.2.4 The Wright–Fisher Model

The Wright–Fisher model is the basic model for reproduction in a finite population that can utilize several mutation models and selection schemes and is at the heart of many models that describe how gene frequencies evolve in the presence of random drift, mutation, and selection (the neutral version of this model was introduced in (22.3)).

The model for a diploid population is defined as follows. Generations are nonoverlapping and the population size N is held constant. As before, we consider a single locus with

two alleles, A_1 and A_2, with genotypes A_1A_1, A_1A_2, and A_2A_2 and respective fitnesses w_{11}, w_{12}, and w_{22}. Furthermore, we assume that the population is randomly mating and in Hardy–Weinberg equilibrium. Suppose there are i genes of type A_1 and $2N - i$ genes of type A_2. Then, assuming selection affects survival between the zygote and adult stage as before, and denoting by $p(n) = i/2N$ the gene frequency of A_1 at generation n, the gene frequency of A_1 after selection becomes

$$\phi_1(n) = \frac{p(n)[p(n)w_{11} + (1 - p(n))w_{12}]}{\bar{w}(n)},$$

where $\bar{w}(n) = p(n)^2 w_{11} + 2p(n)(1 - p(n))w_{12} + (1 - p(n))^2 w_{22}$ is the average fitness. If mutation follows selection, then, assuming symmetric mutation with probability u, the gene frequency of A_1 after mutation becomes

$$\psi_1(n) = \phi_1(n)(1 - u) + (1 - \phi_1(n))u.$$

The N individuals of the next generation are formed by sampling $2N$ independent gametes from the pool according to a binomial resampling scheme. That is, if ψ_1 is the frequency of A_1 after selection and mutation, then the probability that there are j genes of type A_1 in the following generation is

$$\binom{2N}{j} \psi_1^j (1 - \psi_1)^{2N-j}. \tag{22.4}$$

The gene frequency at generation $n + 1$ is then $j/2N$.

This is another example of a discrete-time Markov chain; expression (22.4) is called the *transition probability* of this chain since it describes how the chain evolves from generation to generation. An important feature of this process is that its future depends only on the present state; this property makes this stochastic process a Markov process. The transition probabilities do not depend explicitly on time, and the chain is therefore called *time-homogeneous*. Even for Markov chains with such simple-looking transition probabilities as in (22.4), it is difficult to compute quantities of biological interest exactly, such as the expected time until fixation. For large populations, however, it is often possible to approximate such a chain by a diffusion process. This method was first used by Fisher (1922) and Wright (1931), and later greatly extended by Kimura (1964). (The mathematically rigorous treatment of diffusion processes is due to Kolmogorov, 1931.)

22.3   THE DIFFUSION APPROXIMATION

A diffusion process is a continuous-time stochastic process that tracks a quantity that changes continuously in time and whose future depends only on the present state (the precise definition is somewhat more technical but not needed in the following). An excellent introduction to diffusion processes at an elementary level is Chapter 15 in Karlin and Taylor (1981).

The idea behind using a diffusion process as an approximation of genetic models for large but finite populations is that for many finite-size models, when viewed on a suitable time scale, in the limit as the population size tends to infinity, the change in relative gene

frequencies is continuous and results in a well-defined process. The limiting process is typically easier to study than the original process.

We denote the diffusion process by $\{X(t) : t \geq 0\}$. We think of t as representing time and refer to $X(t)$ as the state at time t; for instance, $X(t)$ could be the relative gene frequency of a particular allele at time t. A diffusion process is characterized by two quantities, the mean and the variance of the infinitesimal displacement, called *drift* and *diffusion*, respectively. The displacement during the time interval $(t, t + h)$ is denoted by $\Delta_h X(t) = X(t + h) - X(t)$. Then the drift parameter is defined as

$$a(x, t) = \lim_{h \to 0} \frac{1}{h} E[\Delta_h X(t) \mid X(t) = x].$$

The diffusion parameter is defined as

$$b(x, t) = \lim_{h \to 0} \frac{1}{h} E[(\Delta_h X(t))^2 \mid X(t) = x].$$

The meaning of these quantities is as follows: For small h, $a(x, t)h$ is approximately the mean of the displacement $\Delta_h X(t)$ during the time interval $(t, t + h)$ since

$$E[\Delta_h X(t) \mid X(t) = x] = a(x, t)h + o(h).$$

The quantity $b(x, t)h$ is approximately the variance of the displacement $\Delta_h X(t)$ during the interval $(t, t + h)$ for small h since

$$\text{var}[\Delta_h X(t) \mid X(t) = x] = E[(\Delta_h X(t))^2 \mid X(t) = x] - \left(E[\Delta_h X(t) \mid X(t) = x] \right)^2$$
$$= b(x, t)h - (a(x, t)h)^2 + o(h) = b(x, t)h + o(h).$$

In all of our examples, the infinitesimal drift and diffusion parameters will not depend on t; this is so since the underlying Markov chains will be time homogeneous. In such cases, we can simply write $a(x)$ and $b(x)$ instead of $a(x, t)$ and $b(x, t)$.

As a first example, we consider the neutral Wright–Fisher model with symmetric mutation for a randomly mating diploid population of size N (i.e. $2N$ gametes). We assume a one-locus model with two alleles, A_1 and A_2, with mutation probability u, and denote by $Y(n)$ the number of A_1 gametes at generation n. The transition probabilities are given by

$$P(Y(n + 1) = j \mid Y(n) = i) = \binom{2N}{j} \psi_1^j (1 - \psi_1)^{2N-j},$$

where

$$\psi_1 = \frac{i}{2N}(1 - u) + \frac{2N - i}{2N} u.$$

To compute the drift and diffusion parameter, we define the scaled process

$$X_N(t) = \frac{Y(\lfloor 2Nt \rfloor)}{2N}, \qquad t \geq 0,$$

where $\lfloor 2Nt \rfloor$ is the largest integer less than or equal to $2Nt$. To find the infinitesimal drift parameter, we compute

$$2NE\left[X_N\left(t + \frac{1}{2N}\right) - X_N(t) \,\middle|\, X_N(t) = \frac{i}{2N} \right]$$

$$= E\left[Y(\lfloor 2Nt \rfloor + 1) - i \mid Y(\lfloor 2Nt \rfloor) = i\right]$$

$$= 2N\psi_1 - i = i(1 - u) + (2N - i)u - i = 2Nu\left(1 - \frac{i}{N}\right).$$

We set $h = 1/2N$, implying that we measure time in units of $2N$ generations, and let N tend to infinity (or h tends to 0). To do this, we also need to scale the mutation parameter, namely, we assume that $\lim_{N\to\infty} 4Nu = \theta$. We then find, with $x = i/2N$,

$$a(x) = \lim_{h\to 0} \frac{1}{h} E\left[X_N(t + h) - X_N(t) \mid X_N(t) = x\right] = \frac{\theta}{2}(1 - 2x).$$

To find the infinitesimal diffusion parameter, we compute

$$2NE\left[\left(X_N\left(t + \frac{1}{2N}\right) - X_N(t)\right)^2 \middle| X_N(t) = \frac{i}{2N}\right]$$

$$= \frac{1}{2N} E\left[(Y(\lfloor 2Nt \rfloor) - i)^2 \mid Y(\lfloor 2Nt \rfloor) = i\right]$$

$$= \frac{1}{2N} 2N\psi_1(1 - \psi_1),$$

since $2N\psi_1(1 - \psi_1)$ is the variance of a binomial distribution with parameters ψ_1 and $2N$. With $x = i/2N$ and $\lim_{N\to\infty} 4Nu = \theta$, we see that $\psi_1 \to x$ as $N \to \infty$. Hence the infinitesimal diffusion parameter is

$$b(x) = x(1 - x).$$

To obtain nontrivial limits of the drift and diffusion parameters, we needed to assume that $\lim_{N\to\infty} 4Nu$ exists (we denoted the limit by θ). Of course, the mutation probability per gene per generation, u, does not depend on the population size but rather is a fixed number. We therefore cannot expect that this limit exists in reality (though we can stipulate that it exists in a mathematical model). How, then, should we interpret this limit? First, the diffusion limit is an *approximation* to the real model when the population size is fixed but finite. Second, we mentioned earlier that the mutation probability is typically quite small, namely of the order of 10^{-5} or 10^{-6}. We should therefore interpret the existence of the limit $\lim_{N\to\infty} 4Nu$ as a guide to when the approximation might be good; namely, we expect the approximation to be good when the population size is of the order of the reciprocal of the mutation probability.

In the literature, one typically finds $4Nu = \theta$ instead of $\lim_{N\to\infty} 4Nu = \theta$. Both are to be interpreted in the same way. In the following, we will adopt the convention of writing $4Nu = \theta$ instead of $\lim_{N\to\infty} 4Nu = \theta$.

Below, we will need the diffusion limit of a haploid Wright–Fisher model with mutations and selection. It is a one-locus model with two alleles A_1 and A_2 for a haploid population of size N. Mutations occur at birth with probability u; i.e. the offspring of an individual is of the same type with probability $1 - u$ and of the other type with probability u. Furthermore, allele A_2 has a selective advantage with selection parameter s. Let $Y_1(n)$ denote the number of individuals of type A_1 at generation n. Then the transition

probabilities are given by

$$P[Y_1(n+1) = j \mid Y_1(n) = i] = \binom{N}{j} \psi_1^j (1 - \psi_1)^{N-j}, \qquad (22.5)$$

where

$$\psi_1 = \frac{p(1-u) + (1-p)(1+s)u}{p + (1-p)(1+s)}, \qquad p = \frac{i}{N}.$$

If we set $\theta = 2Nu$ and $\sigma = 2Ns$, measure time in units of N generations, and let N tend to infinity, then

$$a(x) = -\frac{\sigma}{2}x(1-x) + \frac{\theta}{2}(1-2x), \qquad (22.6)$$

$$b(x) = x(1-x). \qquad (22.7)$$

(Recall that for a haploid population of size N, there are only N gametes; so instead of the factor $2N$, we only have a factor of N in the above scaling.)

The selection parameter s is scaled in the same way as the mutation parameter u. This has an important implication: The diffusion limit for a model with selection can only be carried out provided the selection intensity is of the order of the reciprocal of the population size. Since the diffusion limit is only a good approximation for large populations, the diffusion limit is only useful for weak selection.

What is the advantage of using the diffusion limit? As mentioned earlier, it is rarely possible to obtain exact results of biological interest for the original Markov chain. Using the diffusion limit allows one to take advantage of a number of analytical tools that facilitate the computation of various quantities. We provide two such applications, namely the computation of quantities associated with fixation, and the Kolmogorov forward equation, which can be used to find the stationary distribution.

22.3.1 Fixation

The diffusion approximation can be used to compute certain functionals associated with the process, such as the probability of fixation or the mean time to fixation.

Suppose $X(t)$ denotes the gene frequency at time t. We define $T(y)$ as the (random) time until $X(t)$ reaches y. For $0 \le x_1 < x < x_2 \le 1$, we define the probability that the process reaches x_2 before x_1 when starting at x,

$$u(x) = P[T(x_2) < T(x_1) \mid X(0) = x],$$

and the average time until the process reaches either x_1 or x_2,

$$v(x) = E[\min(T(x_1), T(x_2)) \mid X(0) = x].$$

One can show that, for $0 \le x_1 < x < x_2 \le 1$, $u(x)$ satisfies the differential equation

$$0 = a(x)\frac{du}{dx} + \frac{1}{2}b(x)\frac{d^2u}{dx^2}, \qquad \text{with } u(x_1) = 0 \text{ and } u(x_2) = 1,$$

and $v(x)$ satisfies

$$-1 = a(x)\frac{dv}{dx} + \frac{1}{2}b(x)\frac{d^2v}{dx^2}, \qquad \text{with } v(x_1) = v(x_2) = 0;$$

see Karlin and Taylor (1981) for more detail.

We will apply this to the diffusion approximation of the neutral Wright–Fisher model of genetic drift considered in (22.3), which has no mutation and which is called the *random drift model*; in this case, $a(x) = 0$ since u (and hence θ) is equal to 0, and $b(x) = x(1 - x)$. In the random drift model, eventually, one of the two alleles will be lost. We assume that the initial frequency of one of the alleles is x. By solving the respective differential equations, one can show that the probability of fixation of a particular allele is equal to its initial frequency, and that the mean time until fixation in the diffusion limit is $v(x) = -2[x \ln x + (1 - x) \ln(1 - x)]$, $0 < x < 1$. For instance, when $x = 0.5$, $v(x)$ is equal to 1.39; since time is measured in units of $2N$ generations, this means that it takes approximately $2.78N$ generations until fixation when the population size N is large.

22.3.2 The Kolmogorov Forward Equation

One can show that in the diffusion limit, the conditional probability density that the gene frequency is x at time t given that it was p at time 0, denoted by $\chi(p, x; t)$, satisfies the Kolmogorov forward equation or Fokker–Planck equation

$$\frac{\partial \chi(p, x; t)}{\partial t} = \frac{1}{2} \frac{\partial^2}{\partial x^2}[b(x, t)\chi(p, x; t)] - \frac{\partial}{\partial x}[a(x, t)\chi(p, x; t)].$$

This equation is analytically much more tractable and is the starting point for many investigations of quantities of biological interest. In the time-homogeneous case, we simply replace $a(x, t)$ and $b(x, t)$ by $a(x)$ and $b(x)$.

Solving the Fokker–Planck equation for a given initial gene frequency allows one to study how gene frequencies change over time. A particular important case is when the distribution of gene frequencies reaches an equilibrium. In this case, we expect that as t tends to infinity the limit of the conditional probability density $\chi(p, x; t)$ converges, and that the limit does not depend on the initial state p. We denote this limit by $\rho(x)$ (if it exists), which is then called a *stationary density*. In the time-homogeneous case, $\rho(x)$ would then satisfy

$$0 = \frac{1}{2} \frac{d^2}{dx^2}[b(x)\rho(x)] - \frac{d}{dx}[a(x)\rho(x)]. \tag{22.8}$$

The haploid Wright–Fisher model with mutation and selection introduced earlier has an equilibrium. Using $a(x)$ and $b(x)$ in (22.6) and (22.7), we can compute the stationary density by integrating (22.8). This yields

$$\rho(x) = K x^{\theta-1}(1 - x)^{\theta-1} e^{-\sigma x}, \qquad 0 \le x \le 1, \tag{22.9}$$

where K is a normalizing constant so that $\int_0^1 \rho(x)\, dx = 1$. The density in (22.9) is a special case of Wright's formula (Wright, 1949) and was derived by Kimura (1956) using the diffusion approximation method.

22.3.3 Random Genetic Drift Versus Mutation and Selection

As emphasized above, when a population is finite, the random sampling of gametes introduces a stochastic component into the model. However, the importance of this stochastic component, which we called *genetic drift*, depends on the strength of mutation
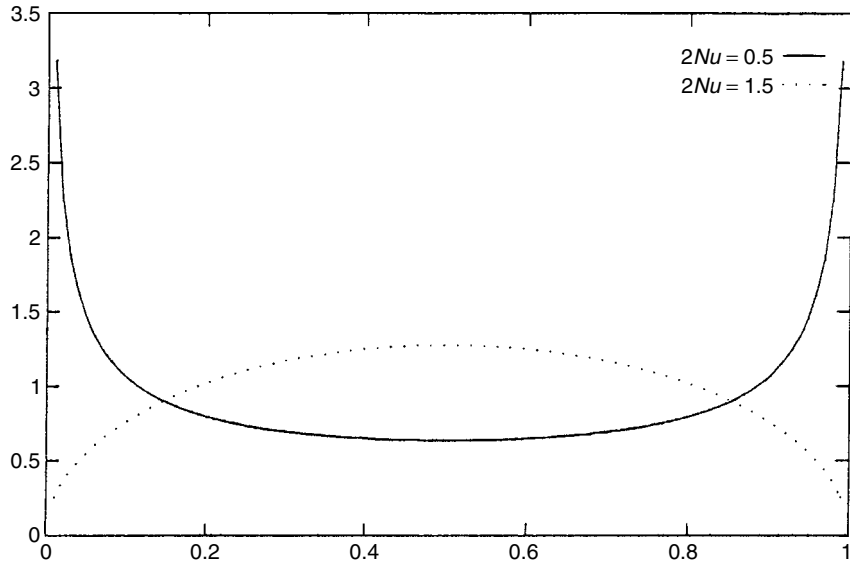
Figure 22.2 The stationary density when $2Nu = 0.5$ and 1.5.

and selection relative to the population size. In general, genetic drift is the more important component in smaller populations.

Let us compare mutation and genetic drift in a population of size N, which evolves according to the one-locus, two-allele haploid Wright–Fisher model with symmetric mutation probability u per gene per generation defined in (22.5). We assume no selection. The stationary density is given by (22.9) when we set $2Nu = \theta$ and $\sigma = 0$. The graph of this function is U-shaped when $2Nu \leq 1$ and unimodal with a peak at $\frac{1}{2}$ when $2Nu \geq 1$ (see Figure 22.2). We see that if $Nu \ll 1$, the population is typically fixed for one or the other allele, meaning that genetic drift is the dominant force; whereas, if $Nu \gg 1$, both alleles will be simultaneously present, meaning that mutational forces are dominant.

To compare random genetic drift versus selection, we again assume the one-locus two-allele Wright–Fisher model with symmetric mutation defined in (22.5), as above, but now we allow allele A_2 to have a selective advantage $1 + s$ over allele A_1. The stationary density is given by (22.9) with $\theta = 2Nu$ and $\sigma = 2Ns$. One can show that if $Ns \ll 1$, selection does not have much of an effect, and the population behaves almost as neutral; whereas if $Ns \gg 1$, selection strongly biases the distribution of alleles towards the favored type, implying that selective forces are dominant (see Figure 22.3).

22.4 THE INFINITE ALLELE MODEL

The models considered so far have a simple genetic structure: One locus with two alleles and mutation changing one allele into the other type and back. This is not a realistic assumption since a gene consists of a large number of nucleotides and thus a mutation occurring at one nucleotide site will likely not result in a type already present in the population, but rather in a novel allele. This prompted Kimura and Crow (1964) to
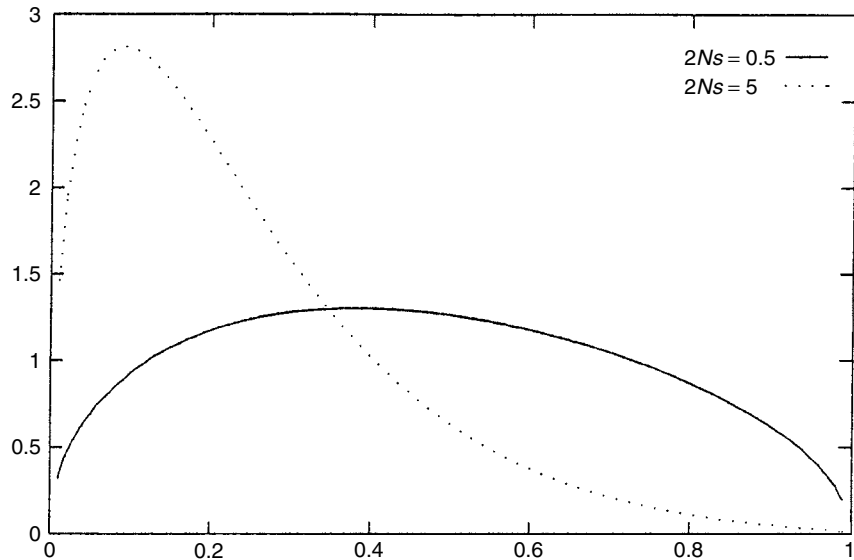
Figure 22.3   The stationary density when $2Nu = 1.5$, $2Ns = 0.5$ and 5.

introduce a new model, the model of infinitely many alleles, or (for short) the infinite allele model.

22.4.1   The Infinite Allele Model with Mutation

The population consists of N diploid individuals and evolves according to a one-locus, neutral Wright–Fisher model with nonoverlapping generations. Mutations occur with probability u per gene per generation, but now every mutation results in a new allele not previously seen in the population.

It is interesting to note that this model was introduced before Kimura proposed his neutral theory; in fact, Kimura and Crow point out that '[i]t is not the purpose of this article to discuss the plausibility of such a system'. Their goal was rather to determine the number of alleles that can be maintained in a population under the extreme case that each mutation would generate a novel allele. This would then provide an upper bound for situations in which mutations could also result in alleles that are already present in the population.

A frequently used measure of genetic diversity is the probability of sampling two alleles of the same type. In the infinite allele model, each allele arises only once, and, therefore, genes in a homozygous individual are identical by descent. If we denote by $F(m)$ the probability that two gametes are identical by descent at generation m, then

$$F(m + 1) = \left[\frac{1}{2N} + \left(1 - \frac{1}{2N} \right) F(m) \right] (1 - u)^2.$$

In equilibrium, $F(m)$ does not depend on m. Denoting the equilibrium by $\hat{F}$, we find

$$\hat{F} = \frac{(1 - u)^2}{2N - (1 - u)^2 (2N - 1)}.$$

The infinite allele model is best studied in the diffusion limit when time is measured in units of $2N$ generations. We set $\theta = 4Nu$ and let N tend to infinity. We find

$$\hat{F} = \frac{1}{\theta + 1}.$$

We conclude from this that in equilibrium the frequency of homozygotes is a decreasing function of mutation pressure.

Since new alleles are constantly introduced into the population and other alleles get lost due to random drift, the actual alleles present are changing over time – even in equilibrium. That is, in equilibrium, the distribution of a particular gene is not stable, but rather the number of alleles attains a steady state. In addition, if we look at the most common allele (whose type will change over time), the second most common allele, and so on, we would find that their frequencies are stable in equilibrium.

22.4.2 Ewens's Sampling Formula

Ewens (1972) investigated the equilibrium properties of samples taken from a population that evolves according to the infinite allele model. He defined the allelic partition of a sample: Denote by a_i the number of alleles present exactly i times in a sample. The vector $(a_1, a_2, \ldots)$ then denotes the allelic partition. If the sample size is n, then $a_{n+1} = a_{n+2} = \cdots = 0$. The number of different alleles in a sample of size n, K_n, is then

$$K_n = \sum_{i=1}^{n} a_i,$$

and the sample size $n = \sum_{i=1}^{n} i a_i$. Ewens obtained the distribution of the allelic partition of a sample in equilibrium in the diffusion limit,

$$P_\theta(a_1, a_2, \ldots, a_n) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!},$$

where $\theta = 4Nu$ and $\theta_{(n)} = \theta(\theta + 1) \cdots (\theta + n - 1)$. (Karlin and McGregor, 1972, gave a formal derivation of this formula.)

An interesting conclusion of this formula is that under neutrality, alleles are not equally likely in equilibrium, but rather the partition is quite lopsided, with a few common alleles and all others relatively uncommon.

Ewens's sampling formula allows one to find the probability distribution of the number of alleles in a sample of size n. Let K_n denote the random variable that counts the number of alleles in a sample of size n in equilibrium; then

$$E[K]_n = \sum_{j=1}^{n} \frac{\theta}{\theta + j - 1}.$$

For large n, this is asymptotically $\theta \ln n$. Furthermore, the variance of K_n for n large is asymptotically $\theta \ln n$ as well.

22.4.3 The Infinite Allele Model with Selection and Mutation

Ethier and Kurtz (1987; 1994) gave a general version of the infinite allele model with selection. Again, the population is diploid and of size N ($2N$ gametes). To distinguish

alleles from each other, we assign each allele a number chosen at random from the interval $(0, 1)$. (This also ensures that each mutation results in a novel allele.) Denote the type of the ith gene by x_i. In each generation, $2N$ gametes are chosen in pairs so that the probability that the ith and the jth gene are selected is

$$P(i, j) = \frac{w_N(x_i, x_j)}{\sum_{1 \leq l, m \leq 2N} w_N(x_l, x_m)},$$

where

$$w_N(x, y) = 1 + \frac{1}{2N} \sigma(x, y),$$

and $\sigma(x, y)$ is a symmetric function of x and y. In the next step, one of the two genes is chosen at random and subjected to mutation; it remains the same type with probability $1 - u$ and mutates to a novel type with probability u, chosen at random from $(0,1)$. This gene is then one of the $2N$ gametes in the next generation.

Ethier and Kurtz (1994) showed that the stationary density of the infinite allele model with selection described above is absolutely continuous with respect to the stationary density of the neutral infinite allele model. Joyce (1995) used this result to show that, for large sample sizes, the allele counts and the total number of alleles under this selection scheme are nearly the same as under neutrality. This convergence, however, is rather slow, and, for small sample sizes, selection has a substantial effect (Li, 1977).

22.5   OTHER MODELS OF MUTATION AND SELECTION

We include three more frequently used models which include mutation and/or selection. The last model mentioned here is a model with overlapping generations.

22.5.1   The Infinitely Many Sites Model

This model was introduced by Watterson (1975) as a model to approximate DNA sequences (see also Ethier and Griffiths, 1987; Griffiths, 1989). Each individual is described by a string of infinitely many, completely linked sites. The mutation probability per site is assumed to be very small so that the total number of mutations per individual per generation is finite and one can assume that no backmutations occur. It is enough to keep track of the segregating sites since none of the other sites carries any information. Each individual is thus represented by just a finite string of sites. The assumption of complete linkage means that there is no recombination between different strings. This assumption is typically made for mitochondrial DNA, which is haploid and maternally inherited.

22.5.2   Frequency-dependent Selection

The fitness of a genotype depends on its frequency (and possibly on the frequencies of other genotypes). An example of this type of selection is gametophytic self-incompatibility, found in many flowering plants. A simple model for a haploid population with minority-advantage frequency-dependent selection is as follows (Wright and Dobzhansky, 1946; Clarke, 1976; Takahata and Nei, 1990): We assume one locus

with infinitely many alleles. The fitness of allele A_k is $1 - ax_k$, where x_k is the frequency of A_k in the population and a is a positive constant. If $F = \sum_{k=1}^{\infty} x_k^2$, then the allelic type A_k contributes a fraction $x_k/(1 - aF)$ to the next generation. Offspring are of the same type with probability $1 - u$ and of a novel type with probability u. If $\theta = 2Nu$ and $\alpha = 2Na$, then in the diffusion scaling, the drift term is $a(x) = -[\alpha x(x - F) + \theta x]/2$ and the diffusion term is $b(x) = x(1 - x)$. Frequency-dependent selection of this form is a powerful mechanism for the maintenance of a polymorphism.

22.5.3 Overlapping Generations

The Wright–Fisher model assumes that generations are discrete (or nonoverlapping). To model overlapping generations, we need to allow individuals to reproduce asynchronously. Moran (1958; 1962) introduced a haploid model in which reproduction occurs continuously in time: Each individual produces an offspring at an exponential rate depending on its fitness. The offspring then undergoes mutation according to a specified mutation process. To keep the population size constant, the offspring then chooses one of the individuals in the population at random and replaces it. In this way, one can define continuous-time analogs for each discrete-time model. It turns out that the corresponding models have the same diffusion limits under suitable scaling of the parameters in the respective processes.

There are many other models that address various aspects of population dynamics. For instance, there is a large amount of literature on population genetic models that take geographic substructure into account (see **Chapter 28**).

22.6 COALESCENT THEORY

22.6.1 The Neutral Coalescent

Looking at the genealogical relationships of a sample of genes is a powerful way to study population dynamics and to infer population parameters, such as the mutation parameter. This method was introduced by Kingman (1982a; 1982b) for neutral models (see **Chapter 25**). A typical genealogy is shown in Figure 22.4. When going back in time, the ancestral lines coalesce until only one line is left; this is the line of the most recent common ancestor.

An important feature of this process is that the family tree and the mutation process can be treated separately under neutrality. We discuss this for a haploid population of size N. Under neutrality, each individual in generation n chooses one parent at random from the previous generation. The ancestral lines of individuals who choose the same parent coalesce.

The dynamics of this genealogical process is not difficult to derive. We will again take advantage of the diffusion approximation, which will allow us to obtain exact results in the limit as the population size goes to infinity. Under neutrality, to simulate the family tree of a sample of size n, we follow the ancestral line of each individual back in time. Suppose now that there are $j \leq n$ ancestral lines left at some time in the past. Each line will choose one ancestor at random from the population. If the population size is
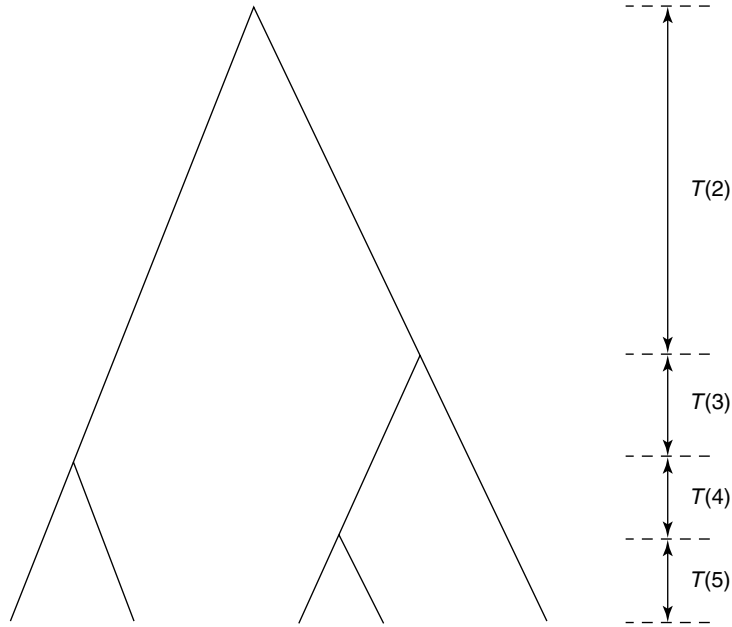
Figure 22.4   The genealogical relationship of a sample of genes.

N, then the probability that the j genes will have no common ancestors in the previous generation is

$$\prod_{i=1}^{j-1}\left(1-\frac{i}{N}\right) = 1 - \frac{\binom{j}{2}}{N} + O\left(\frac{1}{N^2}\right).$$

To prepare for the diffusion limit, we will measure time in units of N generations. If $T(j)$ denotes the time between the coalescing events where the size of the genealogy goes from j to $j-1$ measured in units of N generations. Then for any $t > 0$,

$$P(T(j) > t) = \left(\prod_{i=1}^{j-1}\left(1-\frac{i}{N}\right)\right)^{Nt} \longrightarrow \exp\left[-\binom{j}{2}t\right],$$

as N tends to infinity. We find that in the diffusion limit only pairwise coalescence events occur, and that $T(j)$ is exponentially distributed with parameter $\binom{j}{2}$. The resulting process is called *the coalescent* (Figure 22.4).

An important quantity is the time it takes until a sample of size n is traced back to its most recent common ancestor (MRCA). If we denote this (random) time by $T_{\mathrm{MRCA}}(n)$, then

$$T_{\mathrm{MRCA}}(n) = T(n) + T(n-1) + \cdots + T(2).$$

Since $T(j)$ is exponentially distributed with parameter $\binom{j}{2}$, the expected value of $T(j)$ is $1/\binom{j}{2}$, and hence,

$$E[T_{\mathrm{MRCA}}](n) = \sum_{j=2}^{n} E[T(j)] = \sum_{j=2}^{n} \frac{2}{j(j-1)}$$

$$= 2\sum_{j=2}^{n} \left(\frac{1}{j-1} - \frac{1}{j} \right) = 2\left(1 - \frac{1}{n} \right).$$

From

$$E[T(2)] = 1 \quad \text{and} \quad \lim_{n\to\infty} E[T_{\mathrm{MRCA}}(n)] = 2,$$

we conclude that, on average, the amount of time it takes to reach the MRCA for a very large sample is only about twice that for a sample of size 2.

Using independence, we can compute the variance of $T_{\mathrm{MRCA}}(n)$. We find

$$\mathrm{var}(T_{\mathrm{MRCA}}(n)) = \sum_{j=2}^{n} \mathrm{var}(T(j)) = \sum_{j=2}^{n} \left(\frac{2}{j(j-1)} \right)^2$$

$$= 8\sum_{j=1}^{n-1} \frac{1}{j^2} - 4\left(1 - \frac{1}{n} \right)\left(3 + \frac{1}{n} \right)$$

$$\longrightarrow \frac{8\pi^2}{6} - 12 \approx 1.16 \qquad \text{as } n \to \infty.$$

Since $\mathrm{var}(T(2)) = 1$, we see that the coalescence time $T(2)$ has by far the biggest contribution to the variance.

The mutation process is superimposed on the coalescent. Using the same scaling as in the previous section, namely $\theta = 2Nu$, and measuring time in units of N generations, the mutation process can be described by a Poisson process that puts down mutation events independently on all branches at rate $\theta/2$. It is straightforward to include other mutation schemes in the model. For instance, the coalescent for the neutral infinite allele model has the same structure but we stipulate that at mutation events a novel allele is created.

To simulate a sample of size n, we thus first simulate the genealogical tree, then put mutations on the tree, and finally assign a type to the MRCA and run the process down the tree to obtain a sample of size n. For further details on the neutral coalescent theory, see **Chapter 25**.

It is important to realize that simulating samples of a given size is different from making inferences of population parameters based on a given sample. The coalescent is used for statistical inference; this is described in **Chapter 26**.

22.6.2 The Ancestral Selection Graph

The argument above relied crucially on the assumption that individuals choose their offspring *at random* in the previous generation. This no longer holds under selection. Initial attempts to include selection in the genealogical approach assumed that the gene frequencies were known at all times in the past (Hudson and Kaplan, 1988; Hudson

et al., 1988); this is covered in detail in **Chapter 25**. This approach is successful if the population is in equilibrium and the stationary distribution is known. It turns out, however, that this assumption is not needed when selection is weak. The ancestral selection graph (Neuhauser and Krone, 1997; Krone and Neuhauser, 1997) provides a framework that allows one to study genealogies of samples under selection without knowing the gene frequencies at all times in the past.

We will show how to obtain genealogies for samples under selection using the haploid model for a population of size N, which was defined in (22.5). We follow one locus with two alleles, A_1 and A_2, and assume symmetric mutation with mutation probability u per gene per generation. One of the alleles (A_2) has a selective advantage s.

The genealogy under selection is embedded in a graph, called the *ancestral selection graph*, which has both coalescing and branching events (see Figure 22.5). The coalescing events have the same interpretation as in the neutral case, namely, two ancestral lines join together at the time of their common ancestor. The branching events result in additional ancestral lines. Since the number of actual ancestral lines cannot increase, they constitute potential ancestral lines, reflecting the fact that individuals with a higher fitness have more offspring.

Figure 22.6 illustrates ancestral lines and explains why branching events occur in the presence of selection. If 3 has a selective advantage, then $2'$ and $3'$ have a common ancestor, namely 3; whereas if 3 does not have a selective advantage, then 2 is the ancestor of $2'$ and 3 is the ancestor of $3'$. Since, when constructing genealogies, the gene frequencies in the past are not known, both possibilities need to be carried back until all ancestral lines coalesce and one ancestor, called the *ultimate ancestor* (UA), results. Knowing the type of the UA then allows one to extract the genealogy from the ancestral selection graph. This is illustrated in Figure 22.7, where type 2 has a selective advantage over type 1. We stipulate that a selectively advantageous type on an incoming branch displaces the type on the continuing branch. We see from Figure 22.7 that the MRCA is not necessarily the UA.
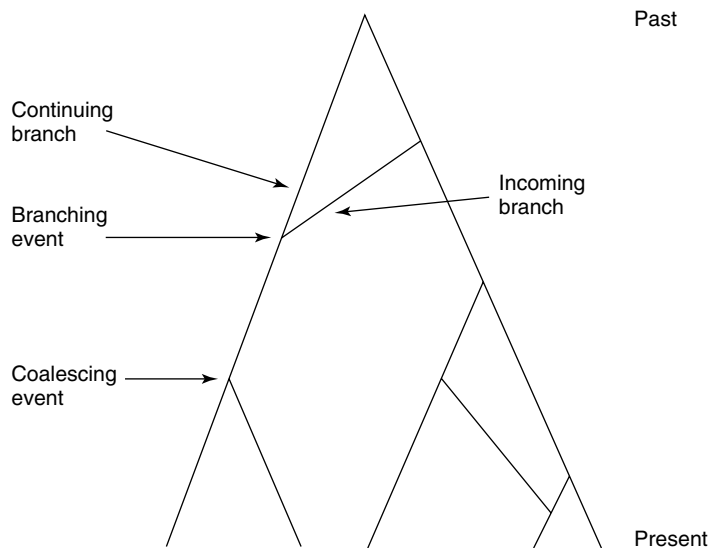
Figure 22.5   The ancestral selection graph with its coalescing and branching structure.
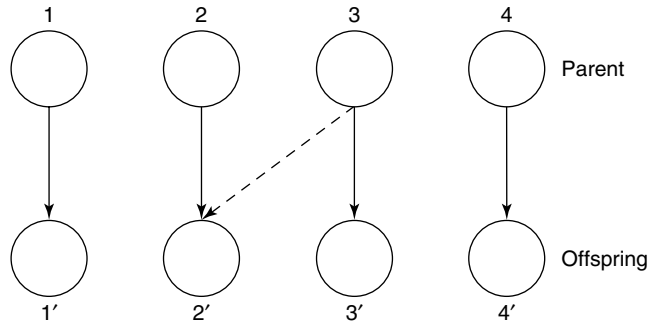
Figure 22.6 Branching events come from differential reproductive success.
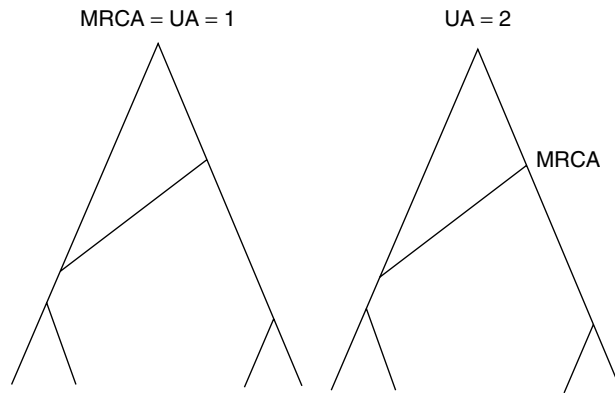
Figure 22.7 Extracting the embedded genealogy from the ancestral selection graph.

The ancestral selection graph can be defined for a large class of models (Neuhauser and Krone, 1997; Neuhauser, 1999), but only in the diffusion limit do we obtain a graph that has a simple structure. The diffusion limit requires that selection is weak, i.e. s is of the order of the reciprocal of the population size (more precisely, $\lim_{N \to \infty} 2Ns = \sigma$). In this limit, the ancestral selection graph for the model defined in (22.5) has the following dynamics. If there are j branches, then

$$\text{coalescing: } j \to j - 1 \quad \text{at rate } \binom{j}{2};$$

$$\text{branching: } j \to j + 1 \quad \text{at rate } \frac{\sigma}{2} j.$$

It follows that coalescing events occur at the same rate as in the neutral coalescent, but, in addition, branching events occur.

The ancestral selection graph allows one to separate the mutation process from the genealogical process as in the neutral case provided the mutation parameter scales with the reciprocal of the population size (i.e, $\lim_{N \to \infty} 2Nu = \theta$). The mutation process is superimposed on the graph: Mutation events occur according to a Poisson process with rate $\theta/2$ independently along each branch of the ancestral selection graph.

The ancestral selection graph described here is for the simplest case of a one-locus two-allele model with symmetric mutation. It can be extended to many other models of selection, including diploid models (Neuhauser and Krone, 1997) and frequency dependent selection (Neuhauser, 1999). The graphs are somewhat more complicated in their branching structure, but have straightforward and intuitive interpretations.

Straightforward simulations of this process are computationally intensive. Recent advances (Slade, 2000a; 2000b; Stephens and Donnelly, 2003; Barton *et al.*, 2004) have led to algorithms that are much less computationally intensive. These advances rely on modeling genealogies conditioned on allele frequencies in the sample.

22.6.3 Varying Population Size

So far, we have always assumed that the population size is constant. There is ample evidence, for instance, that the human population underwent a large expansion in the past. The literature on changing populations sizes, and in particular on bottlenecks, is extensive (Nei *et al.*, 1977; Watterson, 1984; 1989; Slatkin and Hudson, 1991; Rogers and Harpending, 1992; Marjoram and Donnelly, 1994; random environment models are discussed in Donnelly, 1986). It is not difficult to incorporate a changing population size into the coalescent. This was done for the neutral coalescent in Griffiths and Tavaré (1994) and extended to the selection case in Neuhauser and Krone (1997).

To derive the coalescent with varying population size, we begin with the neutral Wright–Fisher model for a haploid population whose size at time 0 (the present) is equal to $N = N(0)$. Denote by $T_N(2)$ the coalescence time of two genes. Then

$$P(T_N(2) > \tau) = \prod_{j=1}^{\tau} \left(1 - \frac{1}{N(j)} \right),$$

where $N(j)$ is the population size j generations in the past. Measuring time in units of N generations, we find, in the limit as N tends to infinity,

$$\lim_{N \to \infty} P(T_N(2) > \lfloor Nt \rfloor) = \lim_{N \to \infty} \prod_{j=1}^{\lfloor Nt \rfloor} \left(1 - \frac{1}{N(j)} \right)$$

$$= \exp \left[-\int_0^t \lambda(u) \; \mathrm{d}u \right],$$

where $\lambda(u)$, the coalescent intensity function, is defined as follows: For $N(j)$ large,

$$\sum_{j=1}^{\lfloor Nt \rfloor} \log \left(1 - \frac{1}{N(j)} \right) \approx - \sum_{j=1}^{\lfloor Nt \rfloor} \frac{1}{N(j)}.$$

If we define $f_N(x) = N(j)/N$ for $x = j/N$, then

$$\lim_{N \to \infty} \sum_{j=1}^{\lfloor Nt \rfloor} \frac{1}{N(j)} = \lim_{N \to \infty} \sum_{x=1/N}^{t} \frac{1}{f_N(x)} \frac{1}{N} = \int_0^t \lambda(u) \; \mathrm{d}u.$$

That is, $\lambda(u) = 1/f(u)$ with $f(u) = \lim_{N \to \infty} f_N(u)$. If we define

$$\Lambda(t) = \int_0^t \lambda(u) \, du,$$

then, as $N \to \infty$, $T_N(2)/N$ converges in distribution to a random variable $T(2)$ with

$$P(T(2) > t) = \exp[-\Lambda(t)].$$

We can now define the coalescent with varying population size. We conclude from the above calculation that at the rescaled time t in the past, each pair of branches coalesces at rate $\lambda(t)$, where $\lambda(t)$ is the coalescent intensity function defined above. If there are j branches present at time t in the past, then a coalescing event occurs at rate $\binom{j}{2}\lambda(t)$. Note that if the population size is constant, i.e. $N(j) = N$ for all $j \geq 0$, then $\lambda(t) = 1$ for all $t \geq 0$, and we obtain the neutral coalescent for fixed population size. A population that expanded rapidly leads to a star-shaped genealogy.

Mutation and selection are not affected by varying population sizes. As in the case of fixed population size, mutation events occur at rate $\theta/2$ along each branch, and branching events occur at rate $j\sigma/2$ if j branches are present.

22.7 DETECTING SELECTION

Explaining molecular differences between individuals and determining the causes of molecular evolution remain a challenging task. Darwin emphasized the role of natural selection as the driving force of evolution, whereas Kimura championed the neutral theory. The availability of molecular data makes it possible to study genetic diversity directly, which should help to resolve the neutrality–selectionist controversy.

A number of statistical tests have been devised to detect selection. This section focuses on tests that are based on coalescent theory, or can at least be understood using coalescent theory. **Chapter 12** discusses statistical methods for phylogenetic analysis of protein coding DNA sequences; this method relies on the ratio of synonymous versus nonsynonymous substitutions.

The null hypothesis for statistical tests of selection typically assumes that the population is neutral. If the observed data deviate too much from what is expected under neutrality, the hypothesis of neutrality is rejected. However, this does not tell one what type of selection acted on the population or if the deviation from the neutral expectation resulted from selection or other forces – for instance, changes in population size, temporally varying environments, or other factors.

One of the first tests was Watterson's (1978) homozygosity test of neutrality. This is based on Ewens's sampling formula and tests whether the observed homozygosity in the sample agrees with that predicted by Ewens's sampling formula. Neither population size nor the mutation parameter needs to be known to apply this test since sample size and the number of distinct alleles in the sample are sufficient to compute the probability of a particular allelic composition given the number of alleles observed in the sample. This test, however, is not very powerful (Gillespie, 1991).

Watterson's test works for allozyme data, which were the prevalent data available then. As DNA sequence data became widespread, there was a need for statistical tests based

on this type of data. Hudson *et al.* (1988) developed a test, the HKA test that compares regions of the genome of two species. This test was developed to detect polymorphism that is maintained by balancing selection.

Tajima (1989) developed a test for DNA sequence data that compares the number of segregating sites and the average number of pairwise nucleotide differences. Under neutrality, the expected number of segregating sites S in a sample of size n is given by (Watterson, 1975)

$$E[S] = \theta \sum_{k=1}^{n-1} \frac{1}{k},$$ (22.10)

where $\theta = 4Nu$, N is the (diploid) population size and u is the mutation probability per gene per generation. The expected number of pairwise nucleotide differences K is given by (Tajima, 1983)

$$E[K] = \theta,$$ (22.11)

where $\theta = 4Nu$ as above.

Selection affects these two quantities differently. The number of segregating sites ignores allele frequencies but the presence of deleterious alleles, which occur in low frequencies, affects this number strongly. The number of pairwise nucleotide differences, on the other hand, is primarily affected by allele frequencies and this is less sensitive to deleterious alleles, which occur at low frequencies.

Both (22.10) and (22.11) suggest a method for estimating θ under neutrality. If the neutrality assumption is violated, however, the two methods should yield different estimates of θ, as explained above. The difference is measured by

$$d = \hat{K} - \frac{\hat{S}}{c_n},$$

where $\hat{K}$ and $\hat{S}$ are the observed pairwise nucleotide differences and the observed number of segregating sites, respectively, and $c_n = \sum_{k=1}^{n-1} k^{-1}$. Tajima (1989) computed the variance of d and derived an estimate of var(d), $\hat{V}(d)$, based on the observed number of segregating sites. He then proposed the test statistic

$$D = \frac{d}{\sqrt{\hat{V}(d)}}.$$

Though Tajima did not base his test on the coalescent, properties of both S and K can be computed using the coalescent. Namely, under the infinite sites model, the number of segregating sites, S, is equal to the total number of mutations in the genealogy of the sampled genes. The total branch length of the genealogical tree is $T_{\text{tot}} = \sum_{j=2}^{n} jT(j)$, where $T(j)$ is exponentially distributed with mean $1/\binom{j}{2}$. Mutations occur independently along branches according to a Poisson process with rate $\theta/2$. Hence,

$$E[S] = \frac{\theta}{2} \sum_{j=2}^{n} j \frac{1}{\binom{j}{2}} = \theta \sum_{j=2}^{n} \frac{1}{j-1} = \theta \sum_{k=1}^{n-1} \frac{1}{k},$$

as in (22.10). The pairwise nucleotide differences can also be derived using the coalescent. Since a site is affected by a mutation at most once under the infinite site model, a pairwise difference occurs if a mutation occurs on either lineage to their common ancestor. Hence,

$$E[K] = \frac{\theta}{2} 2E[T(2)] = \theta,$$

as in (22.11).

A test that is directly based on the coalescent is the test by Fu and Li (1993). They divide the coalescent into external and internal branches (an external branch is a segment that ends in a tip on the genealogy). External branches correspond to younger parts of the tree, internal branches to older parts of the tree. In the case of purifying selection, they expect an excess number of mutations in the external branches since deleterious alleles occur at low frequencies, whereas in the case of balancing selection they expect the opposite. In essence, their test is based on the assumption that branch lengths in the genealogy are affected by selection.

All tests introduced above are based on the assumption that selection will have an effect on the proposed test statistics. Since it is typically not possible to compute the statistical properties of the proposed test statistics under alternatives, a frequent approach is to use simulations to assess the power of proposed tests. This was done, for instance, in Simonsen *et al.* (1995) for Tajima's test statistic and the test proposed by Fu and Li (1993). They simulated alternative hypotheses, such as selective sweeps, population bottlenecks, and population subdivision. In general, they found that Tajima's test was more powerful.

In Golding (1997) and Neuhauser and Krone (1997) the effects of selection on branch lengths were investigated. They found that the shape of the genealogy (i.e. both its topology and the coalescence times) under purifying selection differs only little from that expected under neutrality. Przeworski *et al.* (1999) confirmed their results. Slade (2000a) investigated the effects of selection on branch lengths using the enhanced algorithm of constructing graphs conditioned on the sample. He found as well that the effect of selection on branch lengths is weak. Barton and Etheridge (2004) investigated the effects of both purifying and balancing selection and found that branch lengths are only significantly altered under selection if selection is very strong and deleterious mutations are common in the case of purifying selection or mutation is weak in the case of balancing selection. Overdominant (balancing) selection changes the genealogy (Kaplan *et al.*, 1988). Thus, when using the shape of the genealogy to construct a statistical test for selection, the power of the test depends rather crucially on the alternative.

Acknowledgments

REFERENCES

Barton, N.H. and Etheridge, A.M. (2004). The effect of selection on genealogies. *Genetics* **166**, 1115–1131.

Barton, N.H., Etheridge, A.M. and Sturm, A.K. (2004). Coalescence in a random background. *Annals of Applied Probability* **14**, 754–785.

Clarke, B. (1976). *Genetic Aspects of Host – Parasite Relationships*, A.E.R. Taylor and R.M. Muller, eds. Blackwell, Oxford, pp. 87–103.

Darwin, C. (1985). *The Origin of Species*. Penguin, Harmondsworth, First published in 1859.

Donnelly, P. (1986). A genealogical approach to variable-population size models in population genetics. *Journal of Applied Probability* **23**, 283–296.

Ethier, S.N. and Griffiths, R.C. (1987). The infinitely-many sites model as a measure valued diffusion. *Annals of Probability* **15**, 515–545.

Ethier, S.N. and Kurtz, T.G. (1987). The infinitely-many alleles model with selection as a measure-valued diffusion. In *Stochastic Methods in Biology*, *Lecture Notes in Biomathematics, Vol. 70*, M. Kimura, G. Kallianpur and T. Hida, eds. Springer-Verlag, Berlin, pp. 72–86.

Ethier, S.N. and Kurtz, T.G. (1994). Convergence to Fleming-Viot processes in the weak atomic topology. *Stochastic Processes and their Applications* **54**, 1–27.

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Ewens, W.J. (2004). *Mathematical Population Genetics: Theoretical Introduction*, Vol. 1, 2nd Revised edition, Springer-Verlag.

Fisher, R.A. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh* **42**, 321–341.

Fisher, R.A. (1930). *The Genetical Theory of Natural Selection*, 1st edition. Clarendon, Oxford.

Fu, Y.-X. and Li, W.-H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Gillespie, J.H. (1991). *The Causes of Molecular Evolution*. Oxford University Press, Oxford.

Golding, G.B. (1997). The effect of purifying selection on genealogies. In *Progress in Population Genetics and Human Evolution*, *IMA Volumes in Mathematics and Its Applications, Vol. 87*, P. Donnelly and S. Tavaré, eds. Springer-Verlag, New York, pp. 271–285.

Griffiths, R.C. (1989). Genealogical-tree probabilities in the infinitely-many sites model. *Journal of Mathematical Biology* **27**, 667–680.

Griffiths, R.C. and Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London Series B* **334**, 403–410.

Haldane, J.B.S. (1932). *The Causes of Evolution*. Longmans, Green, London.

Hudson, R.R. and Kaplan, N.L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840.

Hudson, R.R., Kreitman, M. and Aguadé, M. (1988). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

Joyce, P. (1995). Robustness of the Ewens sampling formula. *Journal of Applied Probability* **32**, 609–622.

Kaplan, N.L., Darden, T. and Hudson, R.R. (1988). The coalescent process in models with selection. *Genetics* **120**, 819–829.

Karlin, S. and McGregor, J. (1972). Addendum to a paper by W. Ewens. *Theoretical Population Biology* **3**, 113–116.

Karlin, S. and Taylor, H.M. (1975). *A First Course in Stochastic Processes*. Academic Press, San Diego, CA.

Karlin, S. and Taylor, H.M. (1981). *A Second Course in Stochastic Processes*. Academic Press, San Diego, CA.

Kimura, M. (1956). Stochasting processes in population genetics. Ph.D. thesis, University of Wis-consin, Madison.

Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability* **1**, 177–232.

Kimura, M. (1968a). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

Kimura, M. (1968b). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research* **11**, 247–269.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276.

Kimura, M. (1979). Model of effectively neutral mutations in which selective constraint is incorporated. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 3440–3444.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kimura, M. and Crow, J.F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

Kimura, M. and Ohta, T. (1973). Mutation and evolution at the molecular level. *Genetics (Supplement)* **73**, 19–35.

King, J.L. and Jukes, T.H. (1969). Non-Darwinian evolution: random fixation of selectively neutral mutations. *Science* **164**, 788–798.

Kingman, J.F.C. (1982a). The coalescent. *Stochastic Processes and their Applications* **13**, 235–248.

Kingman, J.F.C. (1982b). On the genealogy of large populations. *Journal of Applied Probability* **19A**, 27–43.

Kolmogorov, A. (1931). Über die analytische methoden inder Wahrscheinlichkeit-srechnung. *Mathematische Annalen* **104**, 415–458.

Krone, S.M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology* **51**, 210–237.

Li, W.-H. (1977). Maintenance of genetic variability under mutation and selection pressures in a finite population. *Proceedings of the National Academic of Sciences of the United States of America* **74**, 2509–2513.

Marjoram, P. and Donnelly, P. (1994). Pairwise comparison of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics* **136**, 673–683.

Moran, P.A.P. (1958). Random processes in genetics. *Proceedings of the Cambridge Philosophical Society* **54**, 60–71.

Moran, P.A.P. (1962). *The Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., Maruyama, T. and Chakraborty, R. (1977). The bottleneck effect and genetic variability in populations. *Evolution* **29**, 1–10.

Neuhauser, C. (1999). The ancestral selection graph and gene genealogy under frequency-dependent selection. *Theoretical Population Biology* **56**, 203–214.

Neuhauser, C. and Krone, S.M. (1997). The genealogy of samples in models with selection. *Genetics* **145**, 519–534.

Przeworski, M., Charlesworth, B. and Wall, J.D. (1999). Genealogies and weak purifying selection. *Molecular Biology and Evolution* **16**, 246–252.

Rogers, A. and Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* **9**, 552–569.

Simonsen, K.L., Churchill, G.A. and Aquadro, C.F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.

Slade, P.F. (2000a). Simulation of selected genealogies. *Theoretical Population Biology* **57**, 35–49.

Slade, P.F. (2000b). Most recent common ancestor probability distributions in gene genealogies under selection. *Theoretical Population Biology* **58**, 291–305.

Stephens, M. and Donnelly, P. (2003). Ancestral inference in population genetics models with selection. *Australian and New Zealand Journal of Statistics* **45**, 395–430.

Slatkin, M. and Hudson, R.R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Takahata, N. and Nei, M. (1990). Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* **124**, 967–978.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276.

Watterson, G.A. (1978). The homozygosity test of neutrality. *Genetics* **88**, 405–417.

Watterson, G.A. (1984). Allele frequencies after a bottleneck. *Theoretical Population Biology* **26**, 387–407.

Watterson, G.A. (1989). The neutral allele model with bottlenecks. In *Mathematical Evolutionary Theory*, M.W. Feldman, ed. Princeton University Press, Princeton, NJ, pp. 26–40.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1949). In *Genetics, Paleontology and Evolution*, G.L. Jepson, G.G. Simpson and E. Mayr, eds. Princeton University Press, Princeton, NJ, pp. 365–389.

Wright, S. and Dobzhansky, T. (1946). Genetics of natural populations. XII. Experimental reproduction of some of the changes caused by natural selection in certain populations of *Drosophila pseudoobscura*. *Genetics* **31**, 125–145.

Inference, Simulation and Enumeration of Genealogies

C. Cannings

Division of Genomic Medicine, University of Sheffield, Sheffield, UK

and

A. Thomas

Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

If we confine our attention to species, such as humans, with two sexes and no selfing, a genealogy is simply a set of individuals and the specification of two binary relations, *mother of* and *father of*. Of course, each individual has at most one mother and one father specified and the set must observe the restrictions imposed by the temporal aspects. This can be most conveniently specified as a list of individual – father – mother triplets with either or both of the latter two fields allowed to be null. Genealogies lend themselves very naturally to graphical representation and anyone involved in genealogical work or plant and animal genetics will likely have drawn a pedigree at some stage to keep track of relationships. These relationships follow a logic that allows us to represent and quantify correlations between pairs or sets of individuals, or their genes, and thus lets us develop algebras of relationship that we discuss here. We also consider two representations of genealogies as graphs that are of primary importance in genetics, discuss their properties and significance and consider the correspondence between them.

23.1 GENEALOGIES AS GRAPHS

Perhaps the neatest way of representing a genealogy is as a *marriage node graph* (Figure 23.1). This is a directed graph $\mathbf{G} = \mathbf{G}(\mathbf{V}, \mathbf{E})$, where the set of nodes $\mathbf{V}$ is partitioned into three subsets: $\mathbf{M}$ the marriages, $\mathbf{f}$ the females and $\mathbf{m}$ the males. The union of $\mathbf{f}$ and $\mathbf{m}$ constitutes $\mathbf{I}$, the set of individuals. The edges, $\mathbf{E}$, which are directed, are similarly partitioned into three subsets, $\mathbf{fM}$ with each element from a female to a marriage, $\mathbf{mM}$ with each element from a male to a marriage and $\mathbf{R}$ with each element
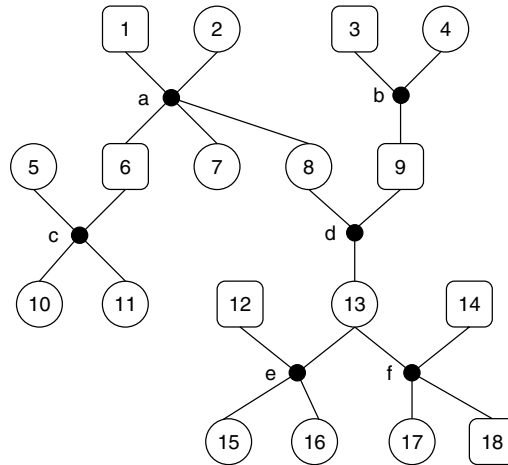
Figure 23.1   A marriage node graph with 18 individuals linked by six marriage nodes. Males are conventionally represented by rectangles, and females by circles. Edges are directed from top to bottom.

from a marriage to an individual, either a male or a female. The restrictions required to make this directed graph a valid genealogy (see Figure 23.1) is simply that (1) there are no cycles, (2) a marriage node may have at most one input edge in **fM** and at most one in **mM**, and (3) an individual node has at most one input in **R**. This representation is well suited for use when drawing a genealogy, particularly when it is complex with many loops. When edges are drawn as angled brackets, instead of straight lines, we see that the result is a picture in the more familiar traditional format, as shown in Figure 23.2.

The second representation is the *moral graph*. This is an undirected graph with a vertex for each individual, edges connecting individuals to their parents, and their parents to each other, so that it is the union of the triangles connecting offspring – father – mother triplets. Although not usually as clear as the marriage node graph when drawn, the structure of the moral graph is important in determining the computational complexity of probability calculations made on genealogies. The name comes from the graphical modelling field where, initially, a directed graph is defined with connections to a dependent, or *daughter*, variable from the variables on which it depends: the *parent* variables. This graph is then *moralised* by connecting, or *marrying*, the parents and discarding the information on direction (Lauritzen and Spiegelhalter, 1988). Although for probability calculations on pedigrees the vertices represent variables for properties of individuals rather than the individuals themselves, the parent – offspring analogy used generally in graphical modelling, in effect, becomes literal for pedigree analysis. An example is given in Figure 23.3.

23.2   RELATIONSHIPS

23.2.1   The Algebra of Pairwise Relationships

It was pointed out earlier that for a genealogy we need only two binary relationships, *mother of* and *father of*, and everything else should follow from that. In practice of
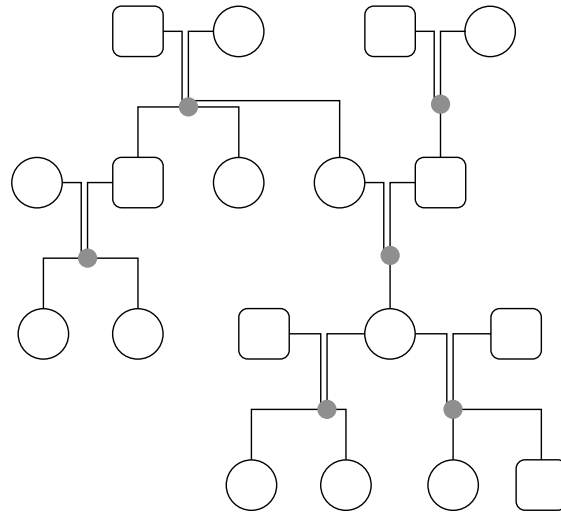
Figure 23.2 The marriage node graph in Figure 23.1 rendered in the more traditional format. As pedigrees get more looped and complex, this format quickly becomes unreadable, and the format in Figure 23.1 is clearer.
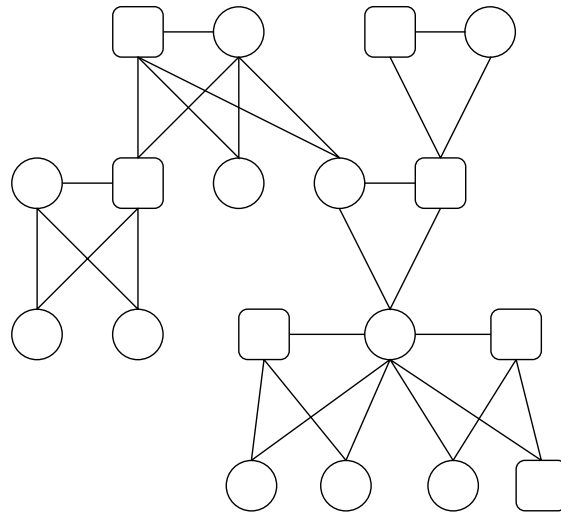
Figure 23.3 The moral graph for the pedigree shown in Figure 23.1.

course there are many other relationships which are defined, though these vary greatly from society to society. The study of relationships, or kinship, is of major import in anthropology, since many of the obligations in societies of all types are based on kinship, though the match between terminology and kinship as represented by the genealogy is not always precise. However, the focus there is often on providing an explanation of why certain individuals are considered to have a particular kinship descriptor.

Here, we briefly comment on the logic of relationships as defined by relative positions in the directed graph of the genealogy. Following Atkins (1974) and Cannings and Thompson (1981), suppose that we denote an edge from a marriage A to an individual B by ARB (R for reproduction) and an edge from an individual B to a marriage A by BMA (M for marriage) where the direction is important. The relationship between any two individuals can then be represented by a string of Rs and Ms, together with their reverses $\overline{R}$ and $\overline{M}$, if the directed graph is a tree, or by a set of such strings if there are loops within the genealogy. We abstract the string of Rs and Ms to define a type of relationship.

Ambiguity can arise by virtue of the possibility of retracing edges in the graph. As a simple example, following Cannings and Thompson (1981), consider a small genealogy, where A, B ,C and D are individuals, and a and b are marriages as in Figure 23.4. Now consider the relationship defined with respect to individual A by $\overline{RMMR}$. Following the edges as defined we can go from A to a, and then from a to D. At this point there is a choice: from D to a or from D to b. If the choice is D to a, then a further choice will take us to A or B. The possible paths are therefore AaDaA, AaDaB and AaDbC, so that the relationship defined is self, sib or half-sib.

It is possible to use this simple system to solve issues of relationship. For example, as discussed by Kendall (1971), there is a well-known riddle regarding relationships: *brothers and sisters have I none but this man's father is my father's son*. It is then necessary to differentiate between males and females. Thus, we split R into S (reproduction producing a son) and D (reproduction producing a daughter), and M into H, the husband of a marriage, and W, the wife of a marriage. Thus, this man's father is $\overline{SH}$, while my father's son is $\overline{RH}HS$, so the relationship between me and this man is summarised by equating these two expressions. We have $\overline{SH} = \overline{RH}HS$. The relationship between the man and me is thus $\overline{SHSHHR}$, so we have two possible relationships (as derived by Kendall) (1) the man is the grandson of my father by a different woman to my mother (I may be male or female), (2) the relationship reduces to $\overline{SHSR}$ so that, since I have no siblings, I must be the man's father.

There are other systems of representation for genealogies used within the anthropological literature, but the focus is usually different from the biological. The interested reader is referred to White and Harary (2001), and references therein, for details of the *P-Systems* approach and other graph systems used in anthropology.
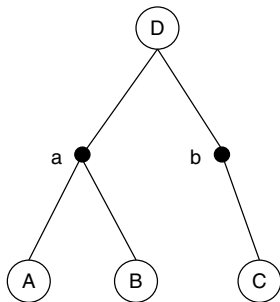
Figure 23.4   A simple pedigree connecting two sibs and a cousin to their common ancestor.

23.2.2 Measures of Genetic Relationship

Here, we discuss measuring and representing the genetic relationship between pairs or sets of individuals, or the genes within an individual. We shall assume throughout that each individual is diploid, i.e. has two genes at each locus (we discuss only the autosomal case, sex-linked genes are treated similarly).

Each individual then has two copies of the gene at any locus of interest. We shall consider sets of copies of the gene in question taken in some specified way (e.g. one taken randomly from each of the individuals A, B and C). However, in accordance with common practice, we shall refer to a set of genes, rather than a set of copies of a gene.

23.2.2.1 Identity by Descent

A set of genes at some locus are said to be *identical by descent* (IBD) if they are all copies of some single ancestral gene, having been passed down by reproduction from that ancestral gene. We shall abbreviate identical by descent by IBD, when appropriate.

Note that identity by descent is an equivalence relation (reflexive, symmetric and transitive), so the set of genes are partitioned into subsets such that genes in the same subset are IBD.

Our measures of relationship are all based on this notion and defined in terms of the probability that the set of genes (at some locus) in question have some specific pattern of identity by descent. This notion was developed by the early work of Cotterman (1940) and Malécot (1948).

23.2.2.2 Pairs of Individuals

In the preceding section, we have referred to the genes *at some locus*, since the definition of identity by descent is referring to a specific set of genes at a specific locus. We now move on to defining and calculating probabilities of various quantities relating to IBD. Here, we can drop the *at some locus* requirement since the probabilities are identical for each locus with similar inheritance (i.e. autosomal or sex-linked). The probabilities are essentially properties of the genealogies.

Definition. The probability that the two distinct genes of an individual A (i.e. that derived from the individual's mother and that from the father) are identical by descent is called the *inbreeding coefficient*, and is denoted by $\mathbf{f}(A)$.

Definition. The probability that a gene randomly selected from individual A and a gene randomly selected from individual B are identical by descent is called the *kinship coefficient* of A and B, and is denoted by $\mathbf{K}(AB)$.

It follows immediately that, for A = B, we have

$$\mathbf{K}(AA) = \frac{1}{2} + \mathbf{f}(A). \tag{23.1}$$

Definition. If individuals A and B are related in some specific way, REL say, then we refer to their kinship coefficient as the *kinship coefficient* of the relationship REL and write this as $\mathbf{K}(REL)$.

Note that if we say two individuals are related in some specific way then it is assumed that the original ancestors of the individuals in the defining genealogy are unrelated.

For example, for a pair of first cousins, which we denote by FC, who share two grandparents, it is assumed that these two individuals are unrelated. We shall consider the pair-wise relationships U = unrelated; S = siblings; FC = first cousins; UN = uncle – niece; DFC = double first cousins. Additionally, we have the pair-wise relationship **I** of an individual with himself.

A fundamental identity allows one to develop a calculus of kinship coefficients deriving that for one relationship in terms of others. Suppose we have a pair of individuals A and B, where B is not an ancestor of A, whose parents are P(A) (father of A), M(A) (mother of A), and similarly P(B) and M(B). Then just using the rules of Mendel one can deduce that

$$\mathbf{K}(AB) = \frac{\mathbf{K}(AP(B)) + \mathbf{K}(AM(B))}{2}. \tag{23.2}$$

The restriction that B is not an ancestor of A is required in order to ensure that there is only one route from the parents of A to individual B. A simple counterexample (see e.g. Lange, 2002) is provided by the case of a three-generation sib mating scheme, where A is in the third generation and B in the second. In order to use recursion in this case one should use the parents of A (which of course include B) rather than the parents of B.

In a similar manner, provided B is not an ancestor of A, nor A of B,

$$\mathbf{K}(AB) = \frac{\mathbf{K}(P(A)P(B)) + \mathbf{K}(P(A)M(B)) + \mathbf{K}(M(A)P(B)) + \mathbf{K}(M(A)M(B))}{4}. \tag{23.3}$$

Hence, for example

$$\mathbf{K}(S) = \frac{\mathbf{K}(I) + \mathbf{K}(U)}{2}$$

$$\mathbf{K}(FC) = \frac{3\mathbf{K}(U) + \mathbf{K}(S)}{4}$$

$$\mathbf{K}(DFC) = \frac{\mathbf{K}(U) + \mathbf{K}(S)}{2}$$

$$\mathbf{K}(UN) = \frac{\mathbf{K}(U) + \mathbf{K}(S)}{2}. \tag{23.4}$$

We may wish to consider coefficients of kinship for a set **W** of individuals $I_i; i = 1, k$, who may not all be different. Then, provided I_1 is not an ancestor of any of the other I_is, we can write

$$\mathbf{K}(W) = \frac{\mathbf{K}(WM1) + \mathbf{K}(WF1)}{2}, \tag{23.5}$$

where **WM1** is formed from **W** by removing the first individual and replacing with his/her mother, and **WF1** similarly replacing with the father.

23.2.3 Identity States for Two Individuals

As pointed out earlier, the reason that the upward recursion of coefficients of kinship breaks down is because it (potentially) ignores a route back to the earlier ancestors. This problem, and many others, can be dealt with by using more complete specification of the states of sets of individuals.

The simplest non-trivial example of the set of identity states occurs for two individuals. We label the genes of the first individual as 1 and 2 and those of the second individual as 3 and 4. We then need to identify those genes that are identical by descent, which is done by

considering the distinct partitions of the four genes into equivalence (by identity) classes. There are 15 partitions (1234), (123, 4)∗(124, 3), (134, 2)∗(234,1), (12, 34), (13, 24)∗ (14, 23), (12, 3, 4), (13, 2, 4)∗(14, 2, 3)∗(23,1, 4)∗(24,1, 3), (34,1, 2), (1, 2, 3, 4), where the ∗ indicates that the partitions are equivalent if we do not require to differentiate between the genes within an individuals. Corresponding to each partition, there is what is termed an *identity state*. Thus, the partition (12, 34) is written as state (1, 1, 2, 2) in the obvious way. For ease of reference, we number the nine states obtained in the latter case as per Table 23.1. Figure 23.5 shows the usual diagrammatic representation of those states.

Example. Suppose that two individuals (one male and one female) whose genes at a particular locus have no identity by descent, i.e their identity state is (1, 2, 3, 4), have two offspring. Now each offspring receives a gene from each parent so that in terms of the identity of the genes by descent each will be one of (1, 3), (1, 4), (2, 3), (2, 4) with probabilities 1/4, independently of the other. Collecting up and reducing to their equivalent identity states, we have that with probabilities 1/4, 1/2 and 1/4 the states will be (1, 2, 1, 2), (1, 2, 1, 3) or (1, 2, 3, 4) respectively. Note that the values in the final states do not relate directly to those values in the parents' states.

23.2.4 More Than Two Individuals

For multiple individuals, Cockerham (1971) generalised to higher-order identities on up to four individuals, while Thompson (1974) used the methods of group theory to discuss the case for arbitrary numbers of individuals, giving examples with up to five.

Table 23.1 Canonical expressions for identity states.

Number	State
1	(1,1,1,1)
2	(1,1,1,2)
3	(1,2,1,1)
4	(1,1,2,2)
5	(1,2,1,2)
6	(1,1,2,3)
7	(1,2,3,3)
8	(1,2,1,3)
9	(1,2,3,4)
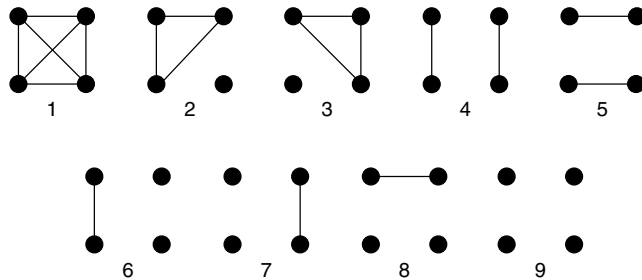
Figure 23.5 Diagram of the identity states of two individuals. The two left-most circles correspond to the genes of the first individual.

Perhaps the neatest treatment is that of Karigl (1981), who generalised the notion of kinship to apply to arbitrary sets of individuals. Thus, suppose that we have sets of individuals $\mathbf{S_1}, \mathbf{S_2} \ldots \mathbf{S}_k$ possibly overlapping. Then, define $\mathbf{K}(\mathbf{S_1}, \mathbf{S_2}, \ldots, \mathbf{S}_k)$ as the probability that a set of genes taken independently one from each of the individuals of set $\mathbf{S}_i$ are identical by descent for every i. Note that

$$\mathbf{K}(\mathbf{S_1}, \mathbf{S_2}, \ldots, \mathbf{S}_k) = \mathbf{K}(\mathbf{P_1}(\mathbf{S_1}), \mathbf{P_2}(\mathbf{S_2}), \ldots, \mathbf{P}_k(\mathbf{S}_k)), \qquad (23.6)$$

where $\mathbf{P}_i(\mathbf{S}_i)$ is a permutation of $\mathbf{S}_i$, and also

$$\mathbf{K}(\mathbf{S_1}, \mathbf{S_2}, \ldots, \mathbf{S}_k) = \mathbf{K}\left(\mathbf{S}_{\rho(1)}, \mathbf{S}_{\rho(2)}, \ldots, \mathbf{S}_{\rho(k)}\right), \qquad (23.7)$$

where $(\rho(\mathbf{1}), \rho(\mathbf{2}), \ldots, \rho(k))$ is a permutation of $\{1, 2, \ldots, k\}$.

Karigl gave the expressions for small numbers of individuals, up to four, which we repeat here in our notation. We shall write $\mathbf{K}(\mathbf{AB}, \mathbf{CD})$ for the case where $k = 2$, $\mathbf{S_1} = \{\mathbf{A}, \mathbf{B}\}$ and $\mathbf{S_2} = \{\mathbf{C}, \mathbf{D}\}$.

We have already defined $\mathbf{K}(\mathbf{AA})$ and $\mathbf{K}(\mathbf{AB})$, and given expressions in terms of the parents of A, given A is not an ancestor of B.

We have, where A is not an ancestor of B, C or D,

$$\mathbf{K}(\mathbf{ABC}) = \frac{\mathbf{K}(\mathbf{P(A)BC}) + \mathbf{K}(\mathbf{M(A)BC})}{2}$$

$$\mathbf{K}(\mathbf{AAB}) = \frac{\mathbf{K}(\mathbf{AB}) + \mathbf{K}(\mathbf{M(A)P(A)B})}{2}$$

$$\mathbf{K}(\mathbf{AAA}) = \frac{1 + 3\mathbf{K}(\mathbf{M(A)P(A)})}{4}$$

$$\mathbf{K}(\mathbf{ABCD}) = \frac{\mathbf{K}(\mathbf{P(A)BCD}) + \mathbf{K}(\mathbf{M(A)BCD})}{2}$$

$$\mathbf{K}(\mathbf{AABC}) = \frac{\mathbf{K}(\mathbf{ABC}) + \mathbf{K}(\mathbf{P(A)M(A)BC})}{2}$$

$$\mathbf{K}(\mathbf{AAAB}) = \frac{\mathbf{K}(\mathbf{AB}) + 3\mathbf{K}(\mathbf{P(A)M(A)B})}{4}$$

$$\mathbf{K}(\mathbf{AAAA}) = \frac{1 + 7\mathbf{K}(\mathbf{P(A)M(A)})}{8}$$

$$\mathbf{K}(\mathbf{AB}, \mathbf{CD}) = \frac{\mathbf{K}(\mathbf{P(A)B}, \mathbf{CD}) + \mathbf{K}(\mathbf{M(A)B}, \mathbf{CD})}{2}$$

$$\mathbf{K}(\mathbf{AA}, \mathbf{BC}) = \frac{\mathbf{K}(\mathbf{P(A)M(A)}, \mathbf{BC}) + \mathbf{BC})}{2}$$

$$\mathbf{K}(\mathbf{AB}, \mathbf{AC}) = \frac{2\mathbf{K}(\mathbf{ABC}) + \mathbf{K}(\mathbf{P(A)B}, \mathbf{M(A)C}) + \mathbf{K}(\mathbf{M(A)B}, \mathbf{F(A)C})}{4}$$

$$\mathbf{K}(\mathbf{AA}, \mathbf{AB}) = \frac{\mathbf{K}(\mathbf{AB}) + \mathbf{K}(\mathbf{P(A)M(A)B})}{2}$$

$$\mathbf{K}(\mathbf{AA}, \mathbf{AA}) = \frac{1 + 3\mathbf{K}(\mathbf{P(A)M(A)})}{4}. \qquad (23.8)$$

In fact, it is relatively easy to express a general recurrence relationship for any $\mathbf{K}(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_k)$. Supposing that A is not an ancestor of any other individual under consideration, and $\mathbf{S}_i = \mathbf{A}^{r_i} \cup \mathbf{T}_i$, where $\mathbf{T}_i$ contains no As. Then

$$\mathbf{K}(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_k) = \frac{1}{2^R} \sum_j \prod_{l=1}^{k} \mathbf{K}(\mathbf{m}^{j_l} \mathbf{f}^{r_l - j_l} \mathbf{T}_l), \tag{23.9}$$

$\mathbf{R}$ being the total number of As, and $\mathbf{j} = (j_1, j_2, \ldots j_k), 1 \leq j_i \leq r_i$ For neatness of expression, $\mathbf{M}(\mathbf{A})$ has been replaced by m, and $\mathbf{P}(\mathbf{A})$ by f. We then replace each expression $m^{j_i} f^{r_i - j_i}$ by m if $j_i = r_1 = 1$, by mm if $j_i = r_i > 1$, by f if $j_i = 0, r_i = 1$, by ff if $j_i = 0, r_i > 1$ and every other expression by mf. We then collect up within each expression for $\mathbf{K}()$ all the terms that have a single m and merge the corresponding $\mathbf{T}_i$s. Similar procedure is adopted for terms with mm, f, ff and fm. Finally, we replace all the mm and ff terms by A, and collect up; note that this last step does not imply that a term with mm is equal to one with A, but rather that the two terms with mm and ff together equal 2 times one with A.

As an example, consider

$$\begin{aligned} 16 \times \mathbf{K}(\mathbf{AAB}, \mathbf{AAC}) = {} & \mathbf{K}(\mathbf{mmB}, \mathbf{mmC}) + 2\mathbf{K}(\mathbf{mmB}, \mathbf{mfC}) + \mathbf{K}(\mathbf{mmB}, \mathbf{ffC}) \\ & + 2\mathbf{K}(\mathbf{mfB}, \mathbf{mmC}) + 4\mathbf{K}(\mathbf{mfB}, \mathbf{mfC}) + 2\mathbf{K}(\mathbf{mfB}, \mathbf{ffC}) \\ & + \mathbf{K}(\mathbf{ffB}, \mathbf{mmC}) + 2\mathbf{K}(\mathbf{ffB}, \mathbf{mfC}) + \mathbf{K}(\mathbf{ffB}, \mathbf{ffC}), \end{aligned} \tag{23.10}$$

leading to

$$\mathbf{K}(\mathbf{AAB}, \mathbf{AAC}) = \frac{\mathbf{K}(\mathbf{AB}, \mathbf{AC}) + \mathbf{K}(\mathbf{AB}, \mathbf{mfC}) + \mathbf{K}(\mathbf{mfB}, \mathbf{AC}) + \mathbf{K}(\mathbf{mfB}, \mathbf{mfC})}{4}. \tag{23.11}$$

Karigl (1981) gives expressions for the nine coefficients of identity in terms of the above. The main benefit of the extended coefficients is that they readily allow the recursive derivation of the coefficients of kinship for individuals on a genealogy. Note that each K-coefficient is expressed in terms of K-coefficients with at most the same number of entries. This therefore allows the use of these to calculate the coefficients given above recursively each time taking A as the most recent individual, thus ensuring that the condition on ancestry is met. Working up through the genealogy finally gives us expressions involving the founders, which will have known values. Karigl (1981) gives a more substantial example for a pedigree of 19 individuals with several loops, and made available programmes to carry out the calculations.

23.2.5 Example: Two Siblings Given Parental States

One often needs to calculate the probabilities of the identity states of a pair of siblings, particularly as part of genetic counselling when the first sibling has some particular genetic disorder and the second is a future offspring about whom one wishes to make a probability statement regarding his/her genetic state. All the information is captured in the probabilities of the gene identity states. We could use Karigl's expressions or proceed directly. Suppose that the parental pair have probabilities $\Pi = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7, \pi_8, \pi_9)$, where π_i is the probability of the ith

state. Then, if a pair of offspring have probabilities Π^*, we can write $\Pi_* = \mathbf{A}\Pi$ where

$$\mathbf{A} = \begin{pmatrix} 1 & 1/4 & 1/4 & 0 & 1/8 & 0 & 0 & 1/16 & 0 \\ 0 & 1/4 & 1/4 & 0 & 1/4 & 0 & 0 & 1/8 & 0 \\ 0 & 1/4 & 1/4 & 0 & 1/4 & 0 & 0 & 1/8 & 0 \\ 0 & 0 & 0 & 0 & 1/8 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1 & 1/4 & 1/2 & 1/2 & 3/16 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/16 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/16 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 3/8 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 \end{pmatrix}, \qquad (23.12)$$

the columns of $\mathbf{A}$ adding to unity, the ninth column corresponding to the example above.

In passing, observe that the matrix above is of upper block diagonal form with block sizes 1, 4, 3 and 1, which correspond to the number of distinct identity states with 1, 2, 3 and 4 distinct genes (by descent). Clearly the process of reproduction can only maintain or reduce the number of distinct genes.

The above matrix can be used as the transition matrix of the Markov chain corresponding to the process of repeated sib mating, see Figure 23.9 (though with columns adding to unity rather than the more usual rows), which is a scheme sometimes used in plant and animal breeding. It is possible in this case, and in more complex examples, to calculate π through time via the eigenvectors and eigenvalues. Fisher (1949) discusses this case in considerable detail, taking the seven identity states, which are relevant when there is no need to differentiate between states 2 and 3, or between 6 and 7. It is then straightforward to calculate the seven eigenvalues 1, 1/4, $(1 \pm \sqrt{5})/4$, 1/2, $-1/8$, 1/4) corresponding to the blocks and to express exactly the value of π. The dominant eigenvalue is $(1 + \sqrt{5})/4$ and indicates the rate of approach to homozygosity.

Consideration of two loci has been addressed by Cockerham and Weir (1968), Weir and Cockerham (1969) and Dennison (1975) among others, and for two and three loci by Thompson (1988). These in the main do not add much to our understanding of the graphical aspects of the genealogies, adding to the complexity but not the conceptual framework, and so are not discussed further here.

23.3   THE IDENTITY PROCESS ALONG A CHROMOSOME

23.3.1   Theory of Junctions

In reality, it is chromosomes that are passed from parent to offspring through the genealogy, so ideally we wish to examine the nature of the identity by descent process along the chromosomes of individuals. This was treated by Fisher (1949) through his theory of junctions, which we discuss in more detail below. This has become more relevant as density of information on the human, and other species, genomes has increased dramatically. Given a set of n individuals, there are $2n$ copies of each chromosome, so the identity state for some particular locus is specified in the obvious way described above.

The description of the identity state of the chromosome as a whole is then given by specifying the identity state at each point along the chromosome; i.e. by specifying a set of intervals within each of which there is some specific identity state. The description of the probabilities associated with the possible realisations is through a random process along the chromosome. This is a continuous time Markov chain (time here denoting distance along the chromosome from one end), with the states corresponding to the identity states.

In Fisher's theory of junctions, a *junction* is a point on a chromosome at which, due to a recombination event, the genetic material on the two sides of the junction has descended by different routes from the ancestors. If the ancestral origin of these two tracts is different, then the junction is termed *external*, and if it is the same, then the junction is termed *internal*. The occurrence of junctions is assumed to be driven by a Poisson process at each segregation, and thus overall by a Poisson process. Whenever a junction occurs, it can then be treated like a mutation and its persistence or loss is governed by the appropriate process for that single locus.

Fisher (1949; 1954; 1959), Bennett (1953) and Gale (1964) investigated the expected number of distinct chromosomal regions (i.e. each separated by a recombination) for a variety of systems of inbreeding: repeated selfing, repeated sib mating, repeated parent – offspring mating and others. Various authors, such as Franklin (1977), Guo (1995), have considered the proportion of the genome which is homozygous by descent in inbred individuals.

23.3.2   Random Walks

In practice, it is usually necessary to concentrate on genealogies of some specific, and relatively simple, structure in order to make much progress. For example, Donnelly (1983) demonstrated that for a pair of first cousins the identity state process along a chromosome was isomorphic to a random walk on the vertices of a three-cube. The transition rates being twice the recombination rate (assuming these to be identical in males and females), six of the vertices corresponding to identity states with zero identity, and two, neighbouring, vertices corresponding to having identity state (1, 2, 1, 3).

Cannings (2003) discusses a general method of deriving the transitions matrix for the continuous time (position along the chromosome) process over the set of identity states. Given the transition matrix for some set of chromosomes at generation n, one can derive that for some new set derived in some specified (Mendelian) manner from these incorporating both the segregation and the recombination process. One can then use the transition matrix to deduce, or estimate numerically, any required variable associated with that particular set of chromosomes.

23.3.3   Other Methods

Bickeboller and Thompson (1996a; 1996b) provided approximations for an arbitrary number of half-sibs, using the Poisson clumping heuristic. Stefanov (2000; 2002; 2004) has provided exact methods for a variety of relationships (grandparent, full- and half-sibs and great grandparental) and Stefanov and Ball (2005) have refined the method in the context of half-sibs.

23.4   STATE SPACE ENUMERATION

23.4.1   Applying the *Peeling* Method

We now proceed to an enumeration problem on genealogies. We suppose that each individual in the genealogy is to be labelled with a genotype and we then enumerate the number of possible such labellings subject to Mendel's first law.

Camp *et al.* (1994) demonstrated how one can use the peeling algorithm to count the number of assignments of genotypes to an arbitrary genealogy that are consistent with Mendel's first law. The rates of growth can be illustrated most clearly in regular mating patterns. At the stage of the peeling algorithm that incorporates the information at a particular nuclear family, the other sections of the pedigree are said to be *below* if connected to the marriage through an offspring and *above* if connected by a parent. For zero-loop pedigrees, these sections are disjoint, although they are not in general. We introduce two functions on these parts, essentially the R-functions of Cannings *et al.* (1978),

$$N_X^U(i) = \text{number of possible states in genealogy and above } X \text{ where } X \text{ has genotype } i$$
(23.13)

and

$$N_X^L(i) = \text{number of possible states in genealogy below } X \text{ given } X \text{ has genotype } i.$$
(23.14)

Consider the nuclear family shown in Figure 23.6, where it is assumed that there are further individuals joined at A, B, C and D who are not shown. Then, supposing that there
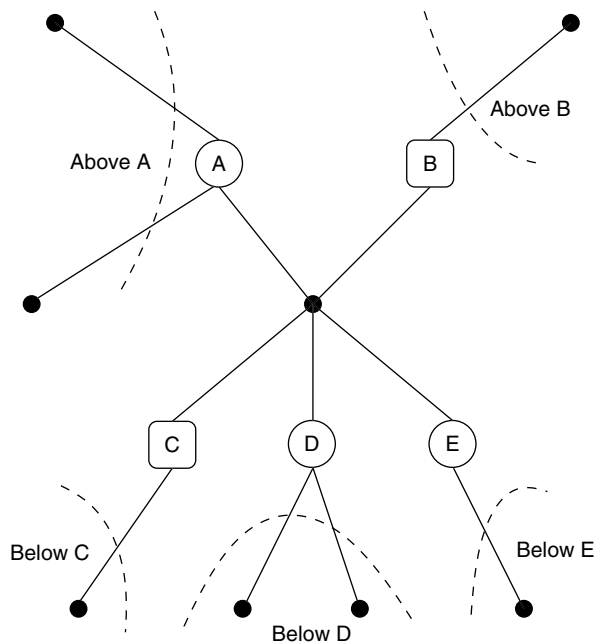
Figure 23.6   The sections of a pedigree above and below a marriage node and attached individuals.

are only two alleles a and b, and indexing the genotypes $aa = 1$, $ab = 2$ and $bb = 2$, then we can write

$$
\begin{aligned}
N_A^L(1) = {} & N_B^U(1)N_C^L(1)N_D^L(1)N_E^L(1) \\
& + N_B^U(2)N_C^L(1,2)N_D^L(1,2)N_E^L(1,2) \\
& + N_B^U(3)N_C^L(2)N_D^L(2)N_E^L(2),
\end{aligned} \tag{23.15}
$$

where

$$
N_I^V(\mathbf{S}) = \Sigma_{k \in \mathbf{S}} N_I^V(k). \tag{23.16}
$$

This is easily seen! For example, A has genotype 1 and B has genotype 2, i.e. the mating is $aa \times ab$ each of the offspring is either aa or ab. Similarly

$$
\begin{aligned}
N_A^L(2) = {} & N_B^U(1)N_C^L(1,2)N_D^L(1,2)N_E^L(1,2) \\
& + N_B^U(2)N_C^L(1,2,3)N_D^L(1,2,3)N_E^L(1,2,3) \\
& + N_B^U(3)N_C^L(2,3)N_D^L(2,3)N_E^L(2,3),
\end{aligned} \tag{23.17}
$$

and $N_A^L(3)$ is obtained from $N_A^L(1)$ by replacing 1 by 3 everywhere.

We also have

$$
\begin{aligned}
N_C^U(1) = {} & N_A^U(1)N_B^U(1)N_D^L(1)N_E^L(1) + N_A^U(1)N_B^U(2)N_D^L(1,2)N_E^L(1,2) \\
& + N_A^U(2)N_B^U(1)N_D^L(1,2)N_E^L(1,2) + N_A^U(2)N_B^U(2)N_D^L(1,2,3)N_E^L(1,2,3)
\end{aligned} \tag{23.18}
$$

and

$$
\begin{aligned}
N_C^U(2) = {} & N_A^U(1)N_B^U(2)N_D^L(1,2)N_E^L(1,2) + N_A^U(2)N_B^U(1)N_D^L(1,2)N_E^L(1,2) \\
& + N_A^U(2)N_B^U(2)N_D^L(1,2,3)N_E^L(1,2,3) + N_A^U(1)N_B^U(3)N_D^L(2)N_E^L(2) \\
& + N_A^U(3)N_B^U(1)N_D^L(2)N_E^L(2).
\end{aligned} \tag{23.19}
$$

Using expressions of the above form, we can easily peel through any genealogy until one has all the information, on the number of possibilities, concentrated on one of the individuals. This then yields the total required.

23.4.2 Recursions

In order to understand the relationship between the *shape* of a genealogy and the number of possible genotype assignments, we examine some regular genealogies. These are made by adding successive nuclear families in a prescribed manner.

Our first example is illustrated in Figure 23.7, where nuclear families with two offspring are joined as shown (it is assumed that there is nothing attached to the Bs and Cs). Now, we write $\mathbf{N}_A^t = (N_{A_t}^L(1), N_{A_t}^L(2))$ (there is no need to keep track of $N_{A_t}^L(3)$ separately since it equals $N_{A_t}^L(1)$). Collecting up the necessary terms, we easily obtain

$$
\mathbf{N}_A^{t+1} = \mathbf{M}\mathbf{N}_A^t, \tag{23.20}
$$
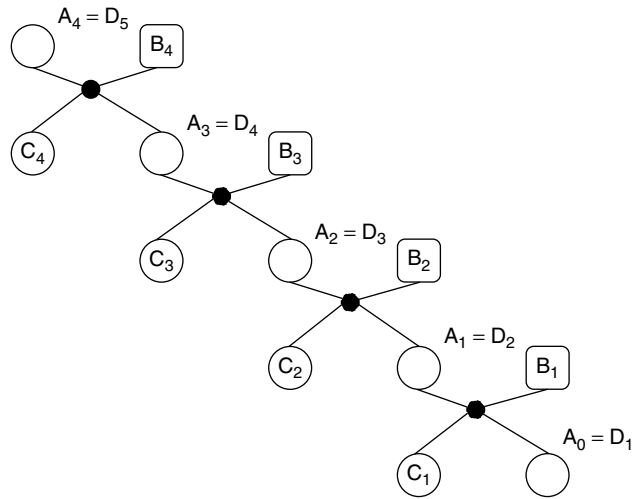
Figure 23.7   A descending line of regular nuclear families.

where

$$\mathbf{M} = \begin{bmatrix} 3 & 3 \\ 10 & 7 \end{bmatrix}. \tag{23.21}$$

Thus, the dominant eigenvalue is $\lambda = 5 + \sqrt{(34)} \approx 10.83$, so the number of assignments grows approximately at rate 2.21 per individual.

Our second example is illustrated in Figure 23.8 and leads to

$$\mathbf{N}_A^{t+1} = \mathbf{M}\mathbf{N}_A^{t}, \tag{23.22}$$

where

$$\mathbf{M} = \begin{bmatrix} 2 & 4 \\ 8 & 9 \end{bmatrix} \tag{23.23}$$

with a dominant eigenvalue of $\lambda = (11 + \sqrt{(177)})/2 \approx 12.15$ giving a growth rate of 2.30 per individual.

Camp *et al.* (1999) prove that the rate 2.30 per individual is the maximum possible among genealogies constructed by adding a new nuclear family with two offspring at each stage, no matter how these are joined on. This information could be of importance in the context of making calculations on a genealogy by, for example MCMC, where knowledge of the size of the sample space could be useful.
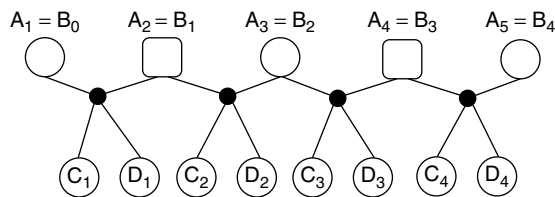
Figure 23.8   A marriage chain.

23.4.3 More Complex Linear Systems

Figure 23.9 shows a repeated sib mating system sometimes used in animal genetics. Here, we link our basic nuclear unit to the existing genealogy at two individuals, as shown, and because the system is growing linearly the recursion is also linear. Here, we switch to the case where there is an arbitrary number, l, of alleles. We work with the seven permutationally distinct states for a pair of unordered sibs

$$(11, 11), (11, 12), (11, 22), (12, 12), (12, 13), (11, 23), (12, 34) \qquad (23.24)$$

(the identity states), noting that these represent respectively

$$l, 4_l C_2, 2_l C_2, {}_l C_2, 6_l C_3, 6_l C_3, 6_l C_4, \qquad (23.25)$$

underlying states.

Now, we can again write

$$\mathbf{N}_S^{t+1} = \mathbf{M} \mathbf{N}_S^t, \qquad (23.26)$$
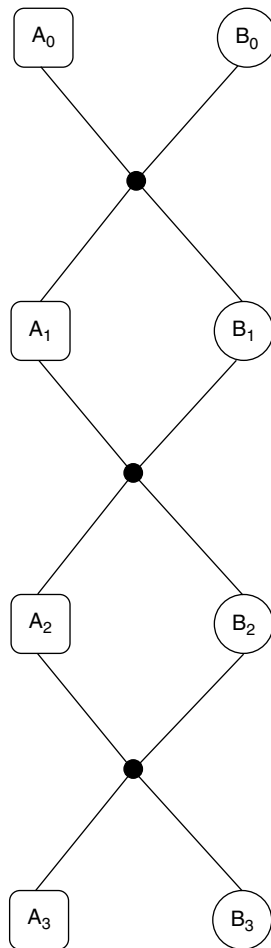
Figure 23.9 Repeated sib mating.

where

$$\mathbf{M} = \begin{bmatrix} 1 & 2(l-1) & 0 & (l-1) & 2_{(l-1)}C_2 & 0 & 0 \\ 0 & 2 & 0 & 1 & 2(l-2) & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 4 & 2 & 1 & 6(l-1) & 4(l-2) & 4_{(l-2)}C_2 \\ 0 & 0 & 0 & 0 & 6 & 2 & 2(l-3) \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}. \tag{23.27}$$

Note that the matrix is block diagonal, essentially since one can only move from a state with the same or fewer distinct elements. Further, the eigenvalues of $\mathbf{M}$ are independent of l. For $l = 2$, the dominant eigenvalue $\lambda \approx 3.7785$ and for $l \geq 3$ $\lambda \approx 6.60$. Thus, since two individuals are added at each stage the respective rates are ≈ 1.944 and 2.569 respectively.

23.4.4   A Non-linear System

As discussed briefly earlier the addition of units in such a way that the number of individuals increases linearly leads to a linear equation for the number of configurations. Figure 23.10 shows a simple non-linear case. The basic unit is a mating with a single offspring A. At each stage, each of the founder members (those who have no parents) are replaced by one of the basic units. The recursive equations here have a bilinear form, rather than linear, so the methods as above are not applicable. Application of the theory of sub-multiplicative sequences, however, gives a rate of growth (for a system with two alleles) of 1.89.

23.5   MARRIAGE NODE GRAPHS

23.5.1   Drawing Marriage Node Graphs

Since, as mentioned earlier, marriage node graphs are a good format for producing drawings of pedigrees, it is worth considering methods for positioning the vertices on a page so as to produce a clear visualisation. Several approaches have been used, for
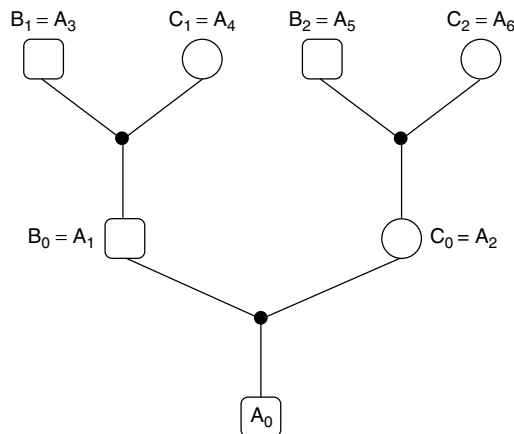
Figure 23.10   A pedigree growing in a non-linear fashion.

instance, Thomas (1994) described using simulated annealing to position vertices on a lattice so as to minimise the total line length, or the amount of ink on a page. A class of methods that works well for general graphs is that of force-directed methods. In these schemes, a quantity representing forces acting between vertices is calculated and minimised, usually using the Newton–Raphson iteration. For example, Fruchterman and Reingold (1991) minimise the quantity

$$\sum_{i=1}^{n} \sum_{j \sim i} d_{i,j}^2 + \alpha \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{d_{i,j}}, \qquad (23.28)$$

where $d_{i,j}$ is the absolute value of the distance in the plane between the ith and jth vertices. The first sum is over connected pairs and represents an attractive force between adjacent vertices. The second sum is a repulsive force pushing apart all pairs.

An attractive feature of this approach is that it lends itself very well to animated graphics with which we can see the vertices move to their optimal positions, and with a few mouse clicks and drags even intervene in the process. Seeing how vertices move from an arbitrary starting point to their final positions is often as informative as the static final picture itself. The first term of 23.28 requires $O(e)$ time to compute, where e is the number of edges, and since pedigrees are relatively sparse graphs this is a quick calculation. However, the second term requires $O(n^2)$ time. This can be greatly improved by ignoring the small repulsions between distant vertices so that the function to minimise becomes

$$\sum_{i=1}^{n} \sum_{j \sim i} d_{i,j}^2 + \alpha \sum_{i=1}^{n} \sum_{j \neq i: |d_{i,j} < \gamma|} \frac{1}{d_{i,j}}, \qquad (23.29)$$

since we can sort vertices into bins depending on their current position and only look for near neighbours in adjacent bins. Since the vertices near any particular vertex tend to be those connected to it, the whole of 23.29 can be computed in time approaching $O(e)$, greatly speeding up calculations. Unfortunately now, the Newton–Raphson scheme does not converge, but oscillates between solutions as the distances between some vertices fluctuate around γ. While this is not a problem when trying to produce a static picture, it leads to an unpleasant flickering effect in animated graphics. We can overcome this by replacing the repulsion term with a function that goes to zero smoothly, i.e. with zero derivative, at a finite value γ. The following modified target function achieves this

$$\sum_{i=1}^{n} \sum_{j \sim i} d_{i,j}^2 + \alpha \sum_{i=1}^{n} \sum_{j \neq i: |d_{i,j} < \gamma|} \frac{(\gamma - d_{i,j})^2}{d_{i,j}}. \qquad (23.30)$$

While the above works well for general graphs, for pedigrees and other directed acyclic graphs, it is nice to have ancestors shown above their descendants. Adding extra terms penalizing the difference between the vertical distance connecting two adjacent vertices and an ideal value θ achieves this, giving the final target function:

$$\sum_{i=1}^{n} \sum_{j \sim i} d_{i,j}^2 + \alpha \sum_{i=1}^{n} \sum_{j \neq i: |d_{i,j} < \gamma|} \frac{(\gamma - d_{i,j})^2}{d_{i,j}}$$

$$+ \beta \left\{ \sum_{i} \sum_{j: j \rightsquigarrow i} (y_i - y_j - \theta)^2 + \sum_{i} \sum_{j: i \rightsquigarrow j} (y_i - y_j + \theta)^2 \right\}, \qquad (23.31)$$

where $i \rightsquigarrow j$ indicates a directed edge from i to j. Something similar can also be achieved by giving each vertex a generation number and fixing the vertical coordinate accordingly. However, for looped pedigrees, this is not a well-defined value as multiple paths connecting an ancestor to an inbred descendant may be of different lengths. Moreover, the repulsive force does not allow vertices fixed on the same horizontal line to cross as it becomes infinite if two vertices coincide exactly.

An animated programme for drawing pedigrees, and also one for general graphs, are available at Alun Thomas's Internet site. These allow the user to interactively change the γ, β and θ parameters and to move and fix specific vertices using mouse controls. A screen shot of the programme is given in Figure 23.11. The pedigree shown is that of the population of Tristan da Cunha as collected in 1962 (Roberts, 1971). Although this contains only 273 people, the multiple, complex relationships between them make it challenging for any pedigree drawing method.

23.5.2 Zero-loop Pedigrees

A *zero-loop* pedigree is one in which there are no loops caused by inbreeding, multiple mating or relative exchange in marriage and corresponds to the case when the marriage
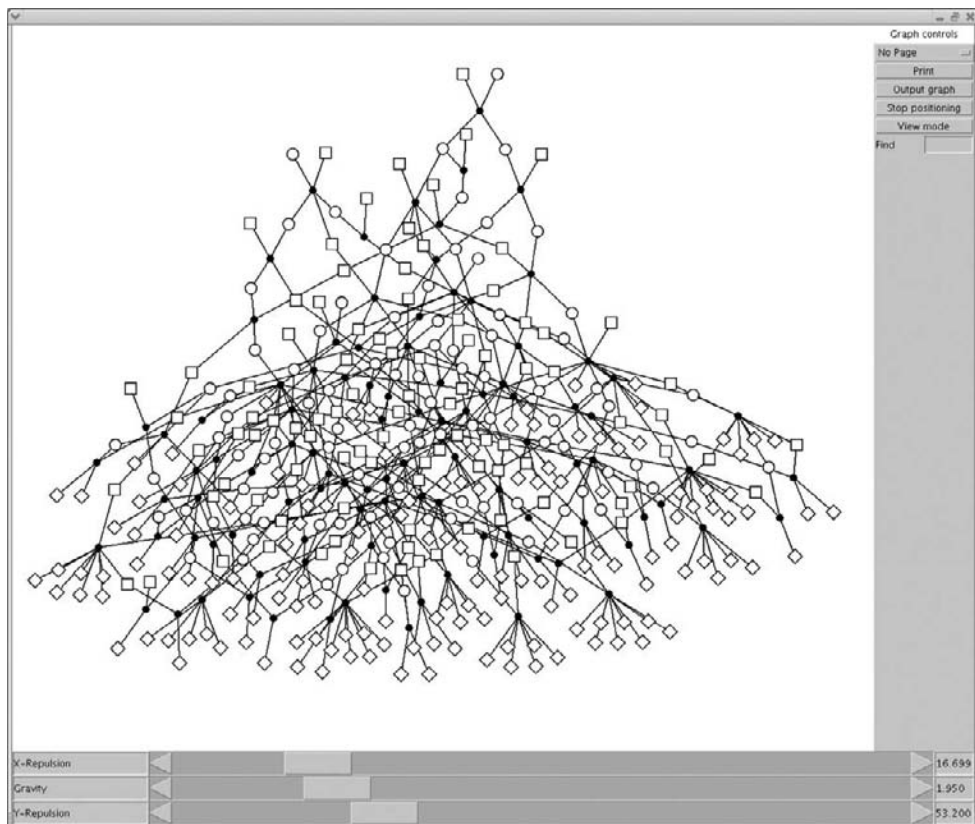
Figure 23.11   A screen shot of the graphical user interface for the ViewPed programme showing a marriage node graph of the pedigree of Tristan da Cunha collected in the 1960s.

node graph is a tree or a collection of unconnected trees: a *forest*. In genetic studies that focus on relatively small pedigrees involving perhaps three or four generations of individuals ascertained from a large, out-bred population, zero-loop pedigrees are the norm. In many cases, when looped pedigrees are found they are either discarded or the loops are avoided because the additional computational complexity that they introduce makes it difficult or impossible to analyse the pedigree. Zero-loop pedigrees are, therefore, of particular interest in genetics.

As discussed below and elsewhere in this book (see **Chapter 24**), probability calculations on pedigrees were the first applications of what became graphical modelling methods. An interesting and non-standard application of graphical modelling can answer the question: *How many ways can the same tree structure be interpreted as a marriage node graph?*. While this may seem an odd question, enumeration is intimately tied up with simulation and the answer will give us an efficient way to simulate zero-loop pedigrees.

Given a specific tree, we can assign vertices roles as either individuals or marriages in only two ways: starting with any vertex designate it an individual, its neighbours as marriages, their neighbours as individuals and so on. To make the other assignment, simply change each vertex's role. Thus, by dealing with each possibility in turn we can assume that we have the structure of the tree and know which vertices are individuals and which are marriages and all that remains is to count in how many ways we can assign individuals as parents or offspring of the marriages. In effect, we have to count the number of ways in which we can assign direction to the edges while following the rules that allow us to interpret the graph as a marriage node graph, which are as follows:

- A marriage node is included only if both parents are specified and there is at least one offspring from the marriage. This is equivalent to insisting that the first element of a triplet is non-null, and that either both or neither of the second and third elements are null.
- An individual must be listed as descended from no more than one marriage node. Individuals whose parents' marriage is not listed are the founders of the pedigree.

For any edge (i, j), let $y_{i,j}$ be 1 if it connects an individual down to its own marriage and 0 if it connects it to its parents' marriage, and for each vertex i let $C_i = \{y_{k,j} : k = i \text{ or } j = i\}$, i.e. the set of direction indicators for the edges that are incident to vertex i. Letting $I[\ldots]$ be an indicator function taking the value 1 when the condition inside the braces is true and 0 when it is false, an allocation of directions $y = \{y_{i,j}\}$ makes a proper marriage node graph if and only if

$$\prod_{\text{individual } i} I\left[\sum_{y \in C_i}(1 - y) \leq 1\right] \prod_{\text{marriage } j} I\left[\sum_{y \in C_j} y = 2\right] I\left[\sum_{y \in C_j}(1 - y) \geq 1\right] = 1.$$

(23.32)

This product defines a Markov graph on the indicator variables and so it can be manipulated using the usual forward – backward methods of graphical modelling. For example, the forward *collect evidence* step computes the total number of allocations of y that result in a valid marriage node graph. If we follow this with a backward simulation
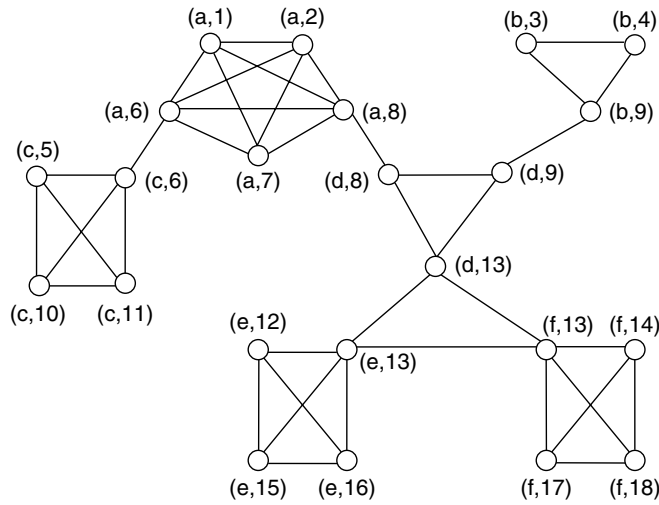
Figure 23.12 The Markov graph for the direction indicators for the edges of the tree structure of the marriage node graph shown in Figure 23.1. The numbers and letters correspond to the individual and marriage vertices of the marriage node graph. Each edge is represented by a letter – number pair reflecting the bipartite structure of the original tree.

step (Dawid, 1992), we sample one of these allocations from a distribution that gives equal probability to each possible allocation.

Figure 23.12 shows the Markov graph for the edge direction indicators corresponding to the tree structure of the marriage node graph in Figure 23.1. You might observe that this graph is *triangulated*, i.e. it contains no cycles of length 4 or more that are un-chorded. Triangulated graphs are amenable to graphical modelling methods, and non-triangulated graphs first have to be made triangulated by inserting additional edges. This triangulation process consists of adding sufficient new edges – *fill-ins* – across loops to break them up into three-cycles. In order to minimise the computations needed to deal with the graphical model, we would like to find the triangulation that makes the largest clique in the graph as small as possible. This is, however, in the general case a significant problem which Markov graphs that are already triangulated avoid.

That the graph in Figure 23.12 is triangulated is not an accident, but is due to the way in which it was constructed. It is a defining property of triangulated graphs that they can be represented as the *intersection graph* of a family of subtrees of a tree (Golumbic, 1980). That is, each vertex of the triangulated graph corresponds to a contiguous subtree of some tree and vertices are connected if and only if the corresponding subtrees intersect. In this case, the subtrees simply comprise the edges of the original tree including the vertices at each end, so that they intersect only when edges have a common end point. The consequence of this is that we can read off the computational requirements for performing the graphical modelling operations described above without having to first triangulate the graph. In particular, we know that the largest clique in the graph will be equal in size to the highest degree of any vertex in the original tree, and so the computational time and storage needed for enumeration and simulation grows as $2^{\max_i |C_i|}$.

Prüfer's constructive bijection (Prüfer, 1918) from the set of trees of size n onto a set of n^{n-2} vectors $(x_1, \ldots x_{n-2})$, where $x_i \in \{1, 2, \ldots n\}$ can be used to generate trees

uniformly at random. Thomas and Cannings (2004) modified this slightly to generate random trees with n individuals and m marriages such that each marriage node has a degree of at least 3, as required for the above conditions. This also allows controlling the numbers of offspring per marriage and marriage per individual according to specific distributions.

23.6 MORAL GRAPHS

23.6.1 Significance for Computation

While a moral graph for a genealogy can be defined without reference to genetic variables, it is also the Markov graph corresponding to the joint distribution of genotypes at a genetic locus for all the individuals in a pedigree. These probabilities are functions of the way that genes arrive at the top of the pedigree, the way they are passed from generation to generation and the way they are expressed as phenotypes of observed individuals. If $g = \{g_1, \ldots g_n\}$ are the genotypes of pedigree members, and $x = \{x_1, \ldots x_n\}$ is the set of observed phenotypes

$$P(x|g) = \prod_{i \in F} \pi(g_i) \prod_{i \in \bar{F}} \tau(g_i | g_{f_i}, g_{m_i}) \prod_{i=1}^{n} \rho(x_i | g_i), \qquad (23.33)$$

where F are the founders of the pedigree and $\pi(g_i)$ is the frequency of genotype g_i in the population from which the founder is drawn, $\tau(g_i | g_{f_i}, g_{m_i})$ encodes the probabilities for transmission of genes from parents f_i and m_i to offspring i and $\rho(x_i | g_i)$ is the probability of the individual's phenotype given the individual's genotype, usually called the *penetrance* function, Each of the three products in this equation assumes conditional independences. The first assumes that founders are independently sampled from some population. The factorisation of the overall transmission probability into the second product is a consequence of the *offspring conditionally independent given genotypes of parents, or OCIGGOP* property (Thompson, 1986) for Mendelian inheritance. The third product assumes that the phenotype is expressed independent of any shared non-genetic factors. The only terms involving multiple genotypes are in the second product and the factors involve the triplets specifying the pedigree. Hence, the moralised graph for the Bayesian network that 23.33 defines is the same as that we have already defined for the pedigree. Probabilities can be calculated on genealogies using standard graphical modelling techniques, although it should be noted that much of this methodology was commonplace in genetics due to the work of Elston and Stewart (1971) and Cannings *et al.* (1978) well in advance of general applications of graphical modelling (Lauritzen and Spiegelhalter, 1988).

The structure of the moral graph is vital in organizing computations efficiently. As Figures 23.13 and 23.14 shows, moral graphs are not necessarily triangulated graphs and, in general, it is a significant problem to find a triangulation of the moral graph with a sufficiently small maximum clique size to make computations feasible, although simulated annealing and heuristic searches such as the *greedy algorithm* have proved to be effective for this (Thomas, 1986). However, the moral graph is not a general graph and has substantial structure.
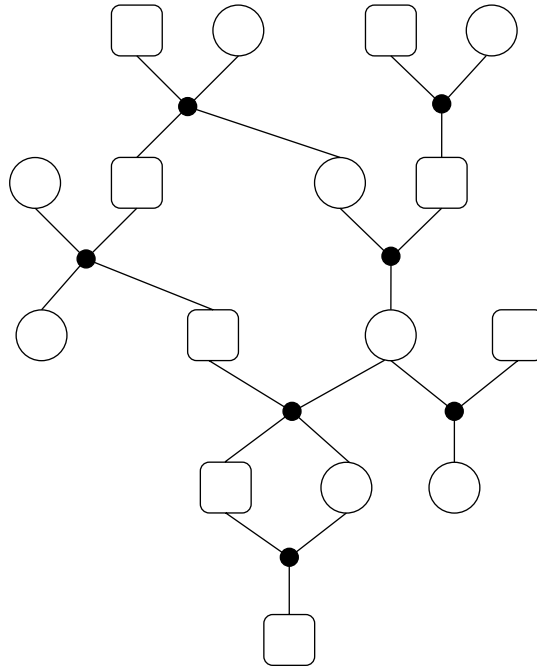
Figure 23.13 A marriage node graph with loops.

23.6.2 Derivation from Marriage Node Graphs

Clearly, starting with a marriage node graph, we can read off the offspring – parent triplets and hence construct the corresponding moral graph. The converse of this is not true since the moral graph is undirected so that different pedigrees can have the same moral graph. However, it is also possible to define the moral graph directly from the marriage node graph as an intersection graph of subtrees as follows. For each individual, define a subtree consisting of his/her vertex, the edge up to but not including their parents' marriage vertex, the edges down to and including each of their own marriages and finally a portion, say a half, of the edge down from each of their own marriages to each of their offspring. This is illustrated in Figure 23.15.

The subtree for any individual intersects with the subtree for a spouse at their marriage node and the upper portion of the edges down from it. Parent and offspring subtrees intersect on the upper portion of the edge down from the connecting marriage node. There are no other ways for subtrees to intersect; hence, the intersection graph is the moral graph.

An immediate consequence of this is that when a pedigree is zero-loop the marriage node graph is a tree and so, from the characterisation of triangulated graphs used above, the moral graph must be triangulated. Thus, for the zero-loop pedigrees used in most genetic studies, no search for an optimal triangulation is necessary. For all but trivial pedigrees, we know that the graph contains a cliques of three vertices corresponding to the offspring – parent triplet cliques. If we can show that there are no larger cliques, then from standard graphical modelling methods we know that the resources required to compute probabilities on a zero-loop pedigree will be, at worst, proportional to tk^3, where
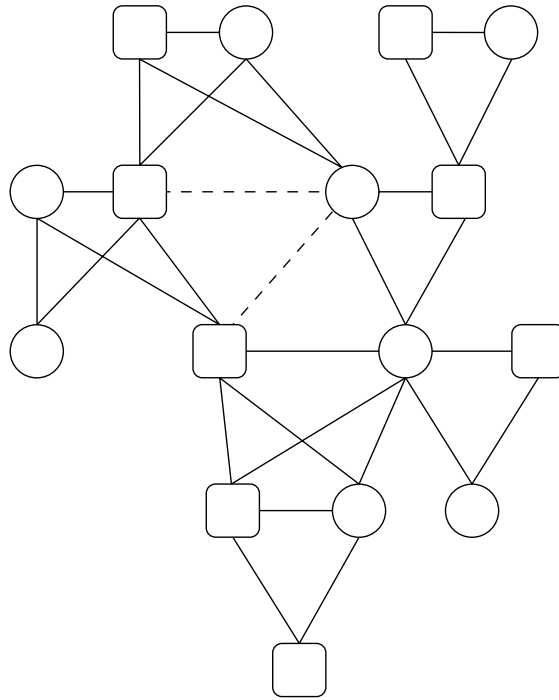
Figure 23.14   The moral graph corresponding to the looped pedigree shown in Figure 23.13. The dotted lines are fill-ins needed to make the graph triangulated.

t is the number of triplets and k is the number of possible genotypes. We can do this using a colouring argument since, for any graph, we know that the size of the largest clique must be no larger than the smallest number of colours that are needed to colour each vertex so that no two adjacent vertices have the same colour (Golumbic, 1980). In fact, it is a property of triangulated graphs that their largest clique size and the smallest possible colouring number must always be equal.

This can be achieved directly from the intersection graph. Starting with an arbitrary marriage node, colour the offspring subtrees all colour 1 and the parent subtrees colours 2 and 3. Working away from the starting marriage node to adjacent ones in a recursive fashion, because the marriage node graph is a tree, at each stage only one of the subtrees adjacent to the marriage node is coloured. If this is an offspring subtree, colour the other offspring subtrees the same colour and allocate the parents the two remaining colours in some order. If it is a parent subtree, allocate the other parent one of the two remaining colours and all offspring subtree the final colour. In this way, each subtree is coloured so that any intersecting subtrees are of different colours. Hence, by giving the vertices of the moral graph the same colour at its corresponding subtree, we can find a three colouring of the graph. Therefore, we can conclude that probability computations on genetic loci in zero-loop pedigrees require at worst tk^3 resources. If fact, we can usually do better than this and make computations using resources of order $t4k^2$ since for any pair of parental genotypes under simple Mendelian inheritance at most four genotypes are possible for any offspring. This model precludes any probability of germ line mutation.
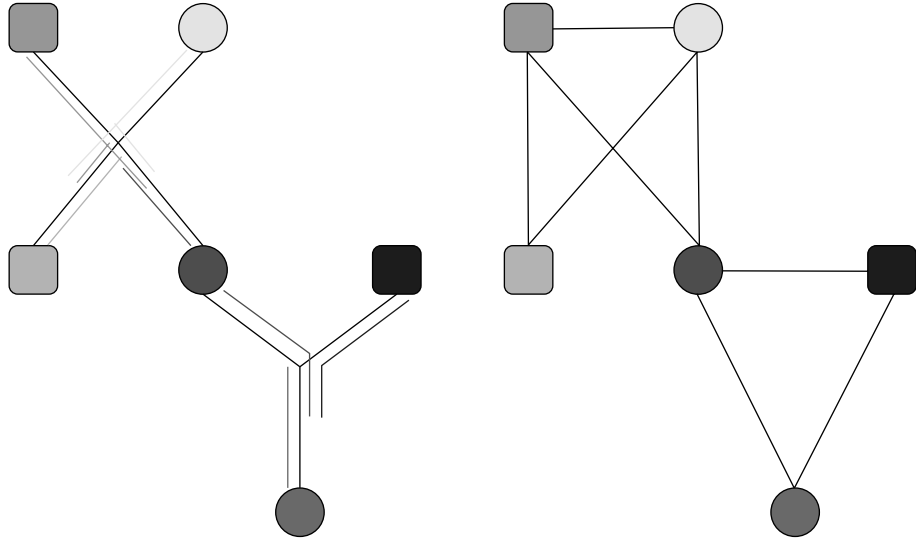
Figure 23.15 A small marriage node graph showing the intersecting subtrees from which a moral graph can be constructed. The colours of the subtrees in the marriage node graph correspond to the individual vertices in the moral graph. Subtrees that intersect in the marriage node graph correspond to connected vertices in the moral graph.

23.6.3   Four Colourability and Triangulation

While it seems obvious that moral graphs of zero-loop pedigrees should be three colourable, it is perhaps surprising at first glance that the moral graph for any pedigree, however large, complex or inbred, can be four coloured. However, this is also straight-forward to show. To do this, first notice that if we make the sub graph of the moral graph consisting of males and the edges between them, then this must be a forest, i.e. a tree or collection of trees. Similarly, the sub graph of females and connecting edges must also be a forest. In effect, we have constructed the ancestral graphs of Y-chromosomes and mitochondria in females respectively. These forests can each be coloured in two colours by taking each founder and giving it one colour, its offspring the second colour, their offspring the first colour again and so on. Since the moral graph is the union of these two forests and a set of edges connecting males to females, these two colourings of the sub graphs can be combined to give a four colouring of the moral graph. That four colours are also necessary in general is illustrated in Figure 23.16, which includes a sib mating that results in a four-clique and which therefore requires four colours. It should also be noted that genealogies are not necessarily planar so that this result is quite separate from the classical four-colour theorem. As a counter example, consider three males each mated with the same three females. The moral graph for this will contain the complete bipartite graph on two sets of three vertices, which cannot be contained in a planar graph.

The consequences of this general result are not as straightforward as in the case of a zero-loop pedigree because a general moral graph needs to be triangulated by adding fill-ins before the computational steps can be read from it. The addition of these fill-ins in general requires connecting vertices of the same colour, thus destroying the colourability
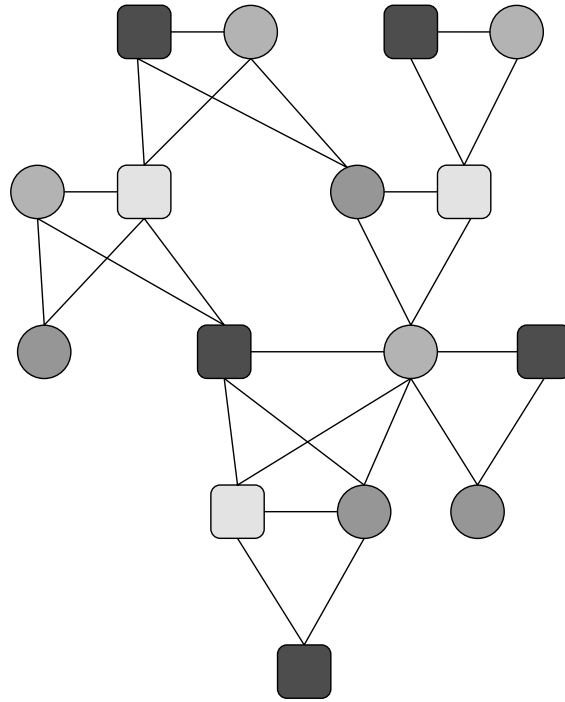
Figure 23.16 A four colouring of the moral graph shown in Figure 23.14. Any genealogy can be four coloured using the same scheme.

and the constraint we have on the largest clique size. However, the colouring can help with heuristic rules for preferring some fill-ins to others. For example, fill-ins connecting males to females never destroy the colourability. Since any loop in the graph must contains at least one individual of each sex, we can always fill it in so that the largest cycles in the graph are four cycles. Unfortunately, except for some simple cases, this leaves an optimisation problem on the same scale as the original.

REFERENCES

Atkins, J.R. (1974). *Grafik: A Multipurpose Kinship Metalanguage*. Mouton and Co.

Bennett, J.H. (1953). Junctions in inbreeding. *Genetica* **26**, 392–406.

Bickeboller, H. and Thompson, E.A. (1996a). Distribution of genome shared IBD by half-sibs: approximation by the Poisson clumping heuristic. *Theoretical Population Biology* **50**, 66–90.

Bickeboller, H. and Thompson, E.A. (1996b). The probability distribution of the amount of an individual's genome surviving to the following generation. *Genetics* **143**, 1043–1049.

Camp, N., Cannings, C. and Sheehan, N. (1994). The number of genotypic assignments on a genealogy. 1: the method and simple examples. *IMA Journal of Mathematics Applied in Medicine and Biology* **11**, 95–106.

Camp, N., Cannings, C. and Sheehan, N. (1999). The number of genotypic assignments on a genealogy 2. general linear systems. *IMA Journal of Mathematics Applied in Medicine and Biology* **16**, 213–236.

Cannings, C. (2003). The identity by descent process along the chromosome. *Human Heredity* **56**, 126–130.

Cannings, C. and Thompson, E.A. (1981). *Genealogical and Genetic Structure*. Cambridge University Press.

Cannings, C., Thompson, E.A. and Skolnick, M.H. (1978). Probability functions on complex pedigrees. *Annals of Applied Probability* **10**, 26–61.

Cockerham, C.C. (1971). Higher order probability functions of identity of alleles by descent. *Genetics* **69**, 235–246.

Cockerham, C.C. and Weir, B.S. (1968). Sib-mating with two linked loci. *Genetics* **60**, 629–640.

Cotterman, C.W. (1940). A calculus for statistico genetics. Ph.D. Thesis, Ohio State University.

Dawid, A.P. (1992). Applications of a general propogation algorithm for probabilistic expert systems. *Statistics and Computing* **2**, 25–36.

Dennison, C. (1975). Probability and genetic relationship: two loci. *Annals of Human Genetics* **39**, 89–104.

Donnelly, K.P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology* **23**, 34–63.

Elston, R.C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523–542.

Fisher, R.A. (1949). *The Theory of Inbreeding*. Oliver and Boyd, Edinburgh.

Fisher, R.A. (1954). A fuller theory of 'junctions' in inbreeding. *Heredity* **8**, 187–197.

Fisher, R.A. (1959). An algebraically exact examination of junction formation and transmission in parent-offspring inbreeding. *Heredity* **13**, 179–186.

Franklin, I.R. (1977). The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. *Theoretical Population Biology* **11**, 60–80.

Fruchterman, T. and Reingold, E. (1991). Graph drawing by force-directed placement. *Software Practice and Experience* **21**, 1129–1164.

Gale, J.C. (1964). Some applications of the theory of junctions. *Biometrics* **20**, 85–117.

Golumbic, M.C. (1980). *Algorithmic Graph Theory and Perfect Graphs*. Academic Press.

Guo, S.-W. (1995). Proportion of genome shared identical by descent by relatives: concept, computation and applications. *American Journal of Human Genetics* **56**, 1468–1476.

Karigl, G. (1981). A recursive algorithm for the calculation of identity coefficients. *Annals of Human Genetics* **45**, 299–305.

Kendall, D.G. (1971). The algebra of genealogy. *Mathematical Spectrum* **4**, 7–8.

Lange, K. (2002). *Mathematical and Statistical Methods for Genetic Analysis*. Springer.

Lauritzen, S.L. and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems. *Journal of the Royal Statistical Society, Series B* **50**, 157–224.

Malécot, G. (1948). *Les Mathematiques de l'Heredite*. Masson et Cie.

Prüfer, H. (1918). Neuer beweis eines satzes uber permutationen. *Archiv fur Mathematik und Physik* **27**, 142–144.

Roberts, D.F. (1971). The demography of Tristan da Cunha. *Population Studies* **25**, 469–475.

Stefanov, V.T. (2000). Distribution of genome shared IBD by two individuals in grandparent relationship. *Genetics* **156**, 1403–1410.

Stefanov, V.T. (2002). Statistics on continuous IBD data: exact distribution evaluation for a pair of full(half)-sibs and a pair of a (great-)grandchild with a (great-)grandparent. *BMC Genetics* **3**, 7.

Stefanov, V.T. (2004). Distribution of the amount of genetic material from a chromosome surviving to the following generation. *Journal of Applied Probability* **41**, 345–354.

Stefanov, V.T. and Ball, F. (2005). Evaluation of identity-by-descent probabilities for half-sibs on continuous genome. *Mathematical Biosciences* **196**, 215–225.

Thomas, A. (1986). Optimal computation of probability functions for pedigree analysis. *IMA Journal of Mathematics Applied in Medicine and Biology* **3**, 167–178.

Thomas, A. (1994). Linkage analysis on complex pedigrees by simulation. *IMA Journal of Mathematics Applied in Medicine and Biology* **11**, 79–93.

Thomas, A. and Cannings, C. (2004). Simulating realistic zero loop pedigrees using a bipartite Prüfer code and graphical modelling. *Mathematical Medicine and Biology* **21**, 335–345.

Thompson, E.A. (1974). Gene identities and multiple relationships. *Biometrics* **30**, 667–680.

Thompson, E.A. (1986). *Pedigree Analysis in Human Genetics*. The Johns Hopkins University Press, Baltimore, MD.

Thompson, E.A. (1988). Two-locus and three-locus gene identity by descent in pedigrees. *IMA Journal of Mathematics Applied in Medicine and Biology* **5**, 261–281.

Weir, B.S. and Cockerham, C.C. (1969). Pedigree mating with two linked loci. *Genetics* **61**, 923–940.

White, D.R. and Harary F. (2001). P-systems: a structural model for kinship studies. *Connections* **24**, 22–33.

Graphical Models in Genetics

S.L. Lauritzen

Department of Statistics, University of Oxford, Oxford, UK

and

N.A. Sheehan

Department of Health Sciences and Genetics, University of Leicester, Leicester, UK

In this chapter, graphical models are introduced and used as a natural way to formulate and address problems in genetics and related areas. Local computational algorithms on graphical models are presented and their relationship with the traditional peeling algorithms discussed. The potential of graphical model representations is explored and illustrated using examples in linkage and association analysis, pedigree uncertainty, forensic identification, and causal inference from observational data.

24.1 INTRODUCTION

Graphs appear in a number of different contexts in genetics to convey information, e.g. about population development and evolution, and relationships between genes and individuals. This chapter focuses on the role of probabilistic graphical models (Lauritzen, 1996) within genetics, and in particular, on aspects of genetics which involve *pedigree analysis*, i.e. the analysis of genetic information among related individuals.

Probabilistic graphical models have their origin in genetics, in path analysis (Wright, 1921; 1923; 1934), which explicitly studies the propagation of hereditary properties through a family tree. They form a natural general framework to express and manipulate a number of important aspects of statistical genetics, e.g. computational algorithms such as 'peeling' (Elston and Stewart, 1971; Cannings *et al.*, 1978; Lander and Green, 1987), but have applications beyond that; e.g. in forensic genetics where complex issues of identification can be naturally expressed in terms of graphical models, in genetic epidemiology where the notion of Mendelian instruments helps to identify causal effects of genes, and in the study of regulatory networks, where graphs are naturally suited

to represent information about the interaction of genes. The latter area is developing particularly rapidly at the time of writing.

In this chapter, we describe basic elements of graphical models, especially Bayesian networks and their use for representing genetic information in pedigrees. We give a relatively detailed description of local computation algorithms in graphs and their use in a number of contexts, in particular, forensic applications, quantitative trait locus (QTL) and linkage analysis, and the handling of pedigree uncertainty. We outline basic elements of causal inference in graphical models and Mendelian randomization, and finally touch upon recent developments in using graphical models for genome-wide association studies and for identifying regulatory networks and patterns of associations in gene expression data.

24.2 BAYESIAN NETWORKS AND OTHER GRAPHICAL MODELS

24.2.1 Graph Terminology

We shall consider a *graph* $\mathcal{G} = (V, E)$ to consist of a finite set V of *vertices* or *nodes* and a set E of *edges* or *links* representing relationships between the nodes. Edges can be either *directed* with arrows indicating the direction of the link, or *undirected*. If there is an undirected edge between a node a and a node b, we say that a and b are *neighbours*, and we say that a is a *parent* of b and b is a *child* of a, if there is a directed edge from a to b. In contrast with the biological interpretation of these terms, a node in a graph can have more than two parents (e.g. node 7 in Figure 24.1).

A *trail* in a graph is defined to be a sequence of edges, each having a node in common with both preceding and succeeding edges. A *path* is a trail with no edges violating the direction of the trail. If all edges of the trail are undirected, the path is *undirected*, and the path is *directed* if all edges are directed. If there is a path from node a to node b (i.e. we can arrive at b by following arrows from a), we say that a is a (graph) *ancestor* of b and b is a (graph) *descendant* of a. A trail beginning and ending with the same node is a *cycle* or *loop*. If all the edges of a graph are directed, it is a directed graph, and if it has no directed cycles, it is a *directed acyclic graph* or DAG. An example of a DAG is displayed in Figure 24.1.

A graph is *connected* if there is a trail between any pair of nodes. A connected graph with no cycles is a *tree*. In this paper, unless otherwise stated, it will be assumed that all graphs are connected.

24.2.2 Conditional Independence

Graphical models (Lauritzen, 1996) exploit graphs to express assumptions of conditional independence to simplify specification, modelling, and analysis of high dimensional

Figure 24.1 A directed acyclic graph. [Reproduced with permission from Lauritzen, S. L. (2000). Causal inference from graphical models. In Barndorff-Nielsen, O. E., Cox, D. R. and Kluppelberg, C., editors, Complex Stochastic Systems, chapter 2, pages 63–107. Chapman & Hall.]
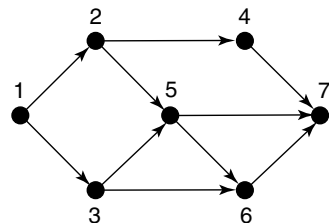
problems (Pearl, 1988; Lauritzen and Spiegelhalter, 1988; Cowell *et al.*, 1999). The conditional independence assumptions essentially split the problem into smaller manageable components.

Random variables X and Y are *conditionally independent* given the random variable Z if the conditional distribution of X given Y and Z is the same as the conditional distribution of X given Z alone, i.e.

$$\mathcal{L}(X \mid Y, Z) = \mathcal{L}(X \mid Z), \tag{24.1}$$

and we then write $X \perp\!\!\!\perp Y \mid Z$. In other words, if $X \perp\!\!\!\perp Y \mid Z$, the value of Y cannot be used to improve the prediction of X once Z is known.

Conditional independence can equivalently be expressed in terms of factorization of the corresponding probability density or probability mass function as

$$X \perp\!\!\!\perp Y \mid Z \quad \Longleftrightarrow \quad f(x, y, z) f(z) = f(x, z) f(y, z) \tag{24.2}$$

$$\Longleftrightarrow \quad \exists a, b : f(x, y, z) = a(x, z) b(y, z). \tag{24.3}$$

24.2.3 Elements of Bayesian Networks

A *Bayesian network* is a DAG with node set V, where the nodes represent random variables, $X = (X_v)_{v \in V}$, having some joint probability distribution function of the form:

$$f(x) = \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}(v)}), \tag{24.4}$$

with $\mathrm{pa}(v)$ denoting the set of parent nodes of the node v and $x_A = (x_v)_{v \in A}$ for any subset $A \subseteq V$. It then holds that any node, given the values at its parents, is conditionally independent of all nodes which are not descendants. This is known as the *directed local Markov property*. The local Markov property for the DAG in Figure 24.1 yields, e.g. $4 \perp\!\!\!\perp \{1, 3, 5, 6\} \mid 2$, $5 \perp\!\!\!\perp \{1, 4\} \mid \{2, 3\}$, and $3 \perp\!\!\!\perp \{2, 4\} \mid 1$.

Further independencies can be deduced from the *global directed Markov property* which gives a complete description of independence relationships associated with a Bayesian network. In fact, the factorization (24.4) is equivalent to either of the local or global directed Markov properties; see (Section 3.2.2) Lauritzen (1996) for details. Note that through (24.4), the joint distribution of a Bayesian network is completely specified from the associated DAG and the conditional distributions of each node, given its parents.

24.2.4 Object-oriented Specification of Bayesian Networks

Bayesian networks involving pedigrees are composed of repeated structures each of identical composition, making these amenable to object-oriented specification. An *object-oriented* Bayesian network (OOBN) (Koller and Pfeffer, 1997) is based on a DAG as above, but each node in the DAG can itself be an OOBN. Simple nodes inside an OOBN can be *internal*, *input*, or *output nodes*. Directed links from an OOBN A to another OOBN B identify output nodes of A with input nodes of B. Each OOBN is typically an *instance* of a *class* of identical OOBNs, thus representing repeated patterns in an efficient manner. In the next section, we shall show examples of such representations of pedigree information. We refer to Dawid *et al.* (2007) for a detailed description of the use of OOBNs in networks representing problems in forensic genetics.

24.3   REPRESENTATION OF PEDIGREE INFORMATION

A pedigree is a group of individuals together with a full specification of all the relationships between them (Thompson, 1986). Individuals without parents in the pedigree are *founders*, are unrelated by definition, and can either be recent or belong to some baseline ancestral generation of interest. Pedigree members with mutual offspring in the pedigree are *spouses* and every spouse pairing is a *marriage*.

24.3.1   Graphs for Pedigrees

A standard diagrammatic representation of a pedigree is shown in Figure 24.2.

A pedigree can also be expressed as a directed graph (Lange and Elston, 1975), the simplest of which is depicted in Figure 24.3(a), where the nodes denote pedigree members, and the arcs connect individuals to their offsprings. A natural extension leads to the *marriage node* graph of Figure 24.3(b) (Thomas, 1985) which has two kinds of node, individual and marriage nodes, and two kinds of arc, connecting an individual to his marriages and connecting a marriage to the offspring of that marriage, respectively (Lange and Elston, 1975; Cannings *et al.*, 1978). Directions on the arcs can be omitted since direction is always *down* from parents to offspring via the relevant marriage node. Consequently, since an individual cannot be his own biological ancestor or descendant, pedigree graphs are always DAGs as there are no directed cycles. Undirected cycles or *loops* can arise, however. These include inbreeding loops, marriage rings, exchange loops, multiple marriage loops, and all kinds of interconnecting combinations of the above (Cannings *et al.*, 1978), where the presence of loops can depend on the particular graphical representation. For example, the loop $14 - 10 - 13 - 9 - 14$ formed by siblings 13 and 14 in the graph of Figure 24.3(a) does not feature in the marriage node graph representation of Figure 24.3(b), but the marriage loop connecting individuals 6, 10, 14, 15, 12, 7, 11 remains.
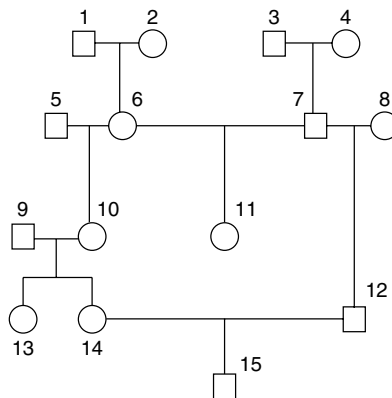
Figure 24.2   A standard representation of a simple pedigree of 14 individuals. As is consistent with common usage, females are represented by circles and males by squares. Individuals 1, 2, 3, and 4 are the baseline founders, while 5, 8, and 9 are recent founders who have married in. Individuals 11, 13, and 15 are *finals* in that they have no marriages.
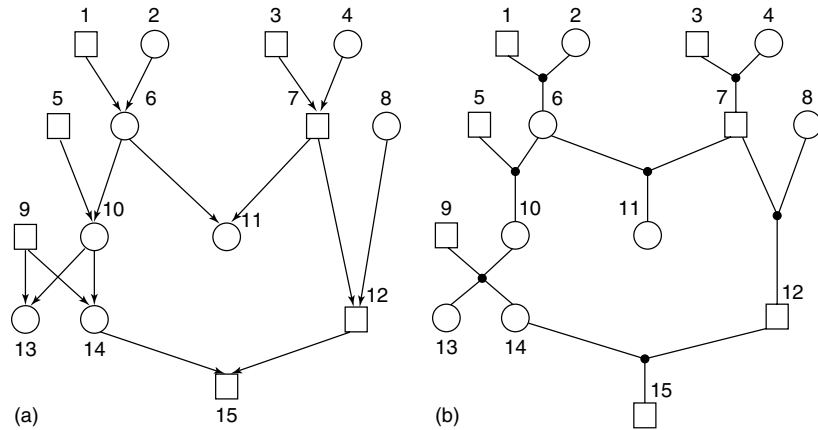
Figure 24.3 The pedigree of Figure 24.2 drawn (a) with nodes representing individuals and directed edges (arcs) connecting individuals to their offsprings and (b) as a marriage node graph with edge directions omitted.

24.3.2 Pedigrees and Bayesian Networks

The heredity of a trait between individuals in a pedigree has a natural expression as a Bayesian network with the graph nodes now representing random variables for which a joint probability distribution satisfying the factorisation in (24.4) can be defined. There are several ways of designing a Bayesian network for a pedigree and these various representations have different properties.

24.3.2.1 Segregation Network

Using the pedigree of Figure 24.2 and a single-locus, discrete genetic trait as an example, we begin with the most direct and complete representation–the *segregation network* (Lauritzen and Sheehan, 2003).

For each individual $i = 1, \ldots, m$, there are two nodes representing the trait genes inherited from his father and mother, respectively. The underlying random variables can assume any of the a allelic types in the relevant genetic system. Following common usage (Thompson, 2001), we will use 0 to label maternal inheritance and 1 for paternal inheritance. Thus the node labelled i^1 is identified with the random variable L_{i^1} assigning the allelic type of the gene inherited by the individual i from his father. For each nonfounder, arcs are directed from the two genes in the father to the paternal gene in the individual and the individual's maternal gene is likewise a (graph) child of the two genes in his mother. An additional node representing the *meiosis* or *segregation* indicator (Thompson, 1994; Sobel and Lange, 1996) is added as a parent to each nonfounder gene node. This is a binary node assuming the values 1 and 0 according to whether the inherited gene is a copy of the paternal or maternal gene in the corresponding parent. In this way, the allelic type of each nonfounder gene is a deterministic function of its (graph) parents. Specifically,

$$L_{i^1} = f(l_{p_i^1}, l_{p_i^0}, S_{p_i,i}) = \begin{cases} l_{p_i^1} & \text{if } S_{p_i,i} = 1 \\ l_{p_i^0} & \text{if } S_{p_i,i} = 0, \end{cases} \tag{24.5}$$
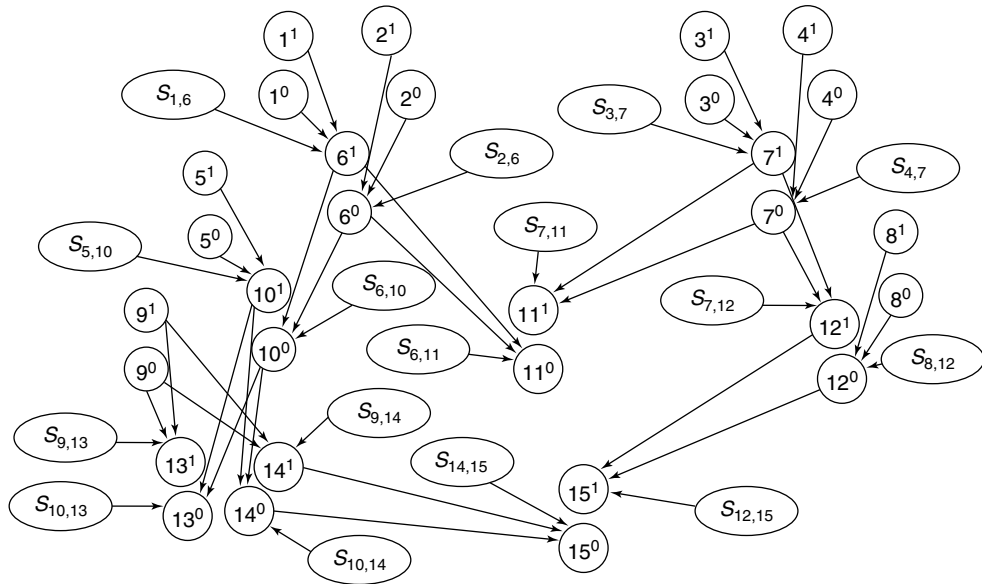
Figure 24.4 The segregation network for the pedigree of Figure 24.2.

for the paternally inherited gene and

$$L_{i^0} = f(l_{m_i^1}, l_{m_i^0}, S_{m_i,i}) = \begin{cases} l_{m_i^1} & \text{if } S_{m_i,i} = 1 \\ l_{m_i^0} & \text{if } S_{m_i,i} = 0, \end{cases} \tag{24.6}$$

for the maternally inherited gene, where m_i and p_i are the labels of the mother and father of the individual i, and $S_{m_i,i}$ and $S_{p_i,i}$ are binary random variables assigning indicators for the segregations to i from the mother and father, respectively. The resulting graph is shown in Figure 24.4.

The laws of inheritance are encoded by letting the segregation indicators be independent with probabilities

$$P(S_{p_i,i} = 1) = \sigma_1, \text{ and } P(S_{m_i,i} = 1) = \sigma_0, \tag{24.7}$$

where $\sigma_1 = \sigma_0 = 1/2$ if inheritance is Mendelian. The assumption of random union of gametes, and hence Hardy–Weinberg proportions for founder genotypes, is implied as the graph clearly indicates that founder genes are independent of each other and of the segregation indicators.

The segregation network can be given a simple object-oriented specification using a master network representing each individual in the pedigree as an OOBN, the class *founder* representing founders of the pedigree, and *schild* representing children as in Figure 24.5. Each instance of the *founder* class is itself an OOBN having nodes that represent the allelic types of the founder, chosen at random from the population and the associated genotype, since this may possibly be observed, see (a) in Figure 24.6. The OOBNs of the class *schild* in Figure 24.6(b) represent the transmission of genetic information from parent to child through yet another OOBN of the class *meiosis* displayed in Figure 24.6(c), which directly describes the segregation of alleles from
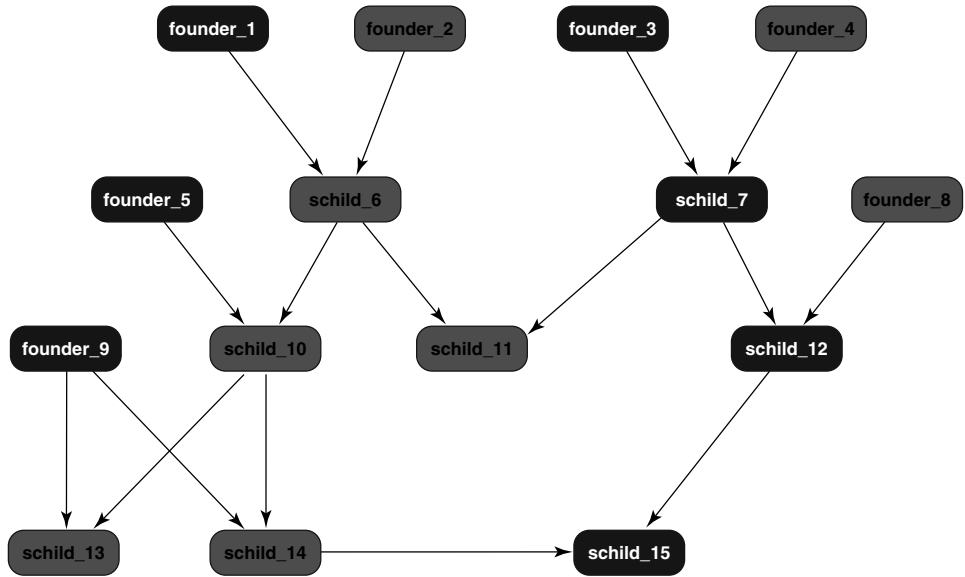
Figure 24.5   The master OOBN for the pedigree of Figure 24.2. There are two OOBN classes: *founder* and *schild*.
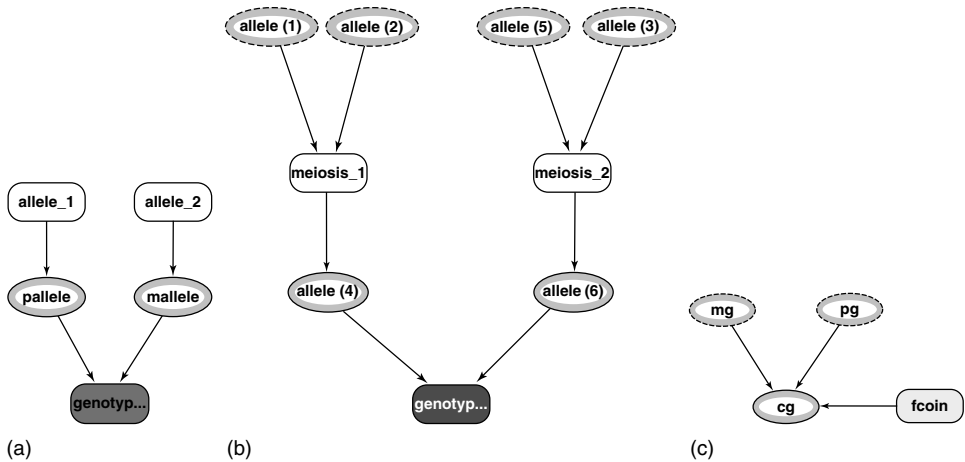
Figure 24.6   (a) The *founder* OOBN has two output nodes determined from the allele class, the latter representing a random allele from the relevant population; (b) the class *schild* has parental alleles as input nodes and represents the transmission of alleles to children via the class *meiosis*; (c) the OOBN of class *fcoin* in the *meiosis* class represents the segregation indicator.

parent to child using the OOBN of the class *fcoin* as the segregation indicator. The latter contains only a single output node with two equally probable values, but defining it as a class object makes it simple to express that this is repeated throughout the pedigree.
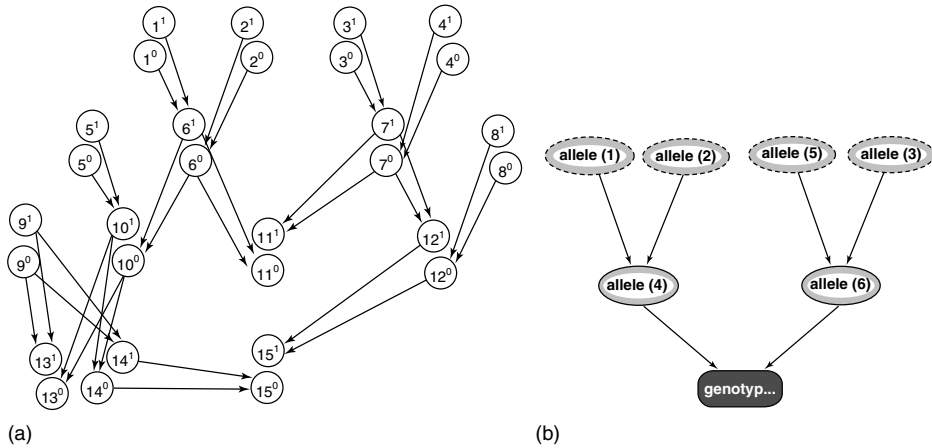
Figure 24.7 (a) Allele network representation of the pedigree; (b) the corresponding modification of the *schild* object to *achild*.

24.3.2.2 Allele Network

In some contexts, the full details of the segregation network may not be required. A convenient reduction is provided by removing the segregation indicators and associated arcs. Figure 24.7(a) shows the corresponding single-locus network for the pedigree of Figure 24.2. The conditional probability distribution of the allelic type of the paternally inherited gene given the (graph) parents, for example, are easily derived from (24.5), (24.6), and (24.7) to be

$$P(L_{i^1} = l \mid l_{p_{i^1}}, l_{p_{i^0}}) = \begin{cases} \sigma_1 & \text{if } l = l_{p_{i^1}} \\ 1 - \sigma_1 & \text{if } l = l_{p_{i^0}} \end{cases}.$$

The maternally inherited gene is handled analogously. We call this the *allele network* (Lauritzen and Sheehan, 2003), but note that it also features in Jensen (1997) and Thompson and Heath (2000). In the object-oriented representation, the corresponding reduction is made by replacing the OOBN *schild* with a simpler object, *achild*, which represents the transmission of alleles probabilistically rather than using the explicit process of meiosis, see Figure 24.7(b).

24.3.2.3 Genotype Network

The visually most parsimonious standard representation, although not necessarily the most useful, is the *genotype network* and uses the graph of Figure 24.3(a) as the underlying DAG with the nodes now representing the genotypes of the individuals rather than the individuals themselves. This representation features in Heath (2003) and Spiegelhalter (1990), is the 'genotype representation' of Jensen (1997), and is the 'genotype network' of Lauritzen and Sheehan (2003). We emphasize that the assumption of Mendelian inheritance is necessary for this representation to be valid, see Lauritzen and Sheehan (2003) for further details.

24.3.2.4 Adding Phenotypic Information

Each of the above networks specifies the inheritance relationships without referring to the observational situation in any given context. Although the genotypes may be identifiable from the phenotypes in many cases, they are often not identifiable and only partial information is available in some situations. To accommodate such data, an extra node can be added for each individual for whom phenotypic information is available and, depending on the purpose of the analysis (e.g. genetic counselling where a future child might be of interest), possibly for some unobserved individuals as well. We let Y_i denote the variable associated with the phenotype of the individual i. In the allele and segregation networks, the node carrying the phenotype Y_i has the two alleles of the individual i as parents, whereas in the genotype network, Y_i has the genotype G_i as its only parent. The conditional distribution of the phenotype Y_i, given its (graph) parents, is the *penetrance distribution* and may take the form of a deterministic relationship (e.g. $Y_i = G_i$) or a more complicated function. If the genotype is itself observable, i.e. $Y_i = G_i$, then we can omit this extra node in the genotype network. For the most parsimonious representation (the genotype network), the local Markov property of the network augmented with phenotypic information is ensured by the phenotype Y_i of any individual being conditionally independent of other variables in the network, given the genotype G_i of that individual. There are some traits for which this conditional independence assumption is clearly not reasonable. A woman's risk of pre-eclampsia is higher for a first pregnancy than a subsequent pregnancy, for example. Genetic imprinting, by which paternal and maternal alleles have differential influence on the phenotype also violates this assumption. In the latter case, one could simply define a network in terms of *ordered* genotypes for which the Markov property would then hold. Alternatively, the augmented allele and segregation networks both contain sufficiently detailed information for the Markov property to hold for more complex genotype–phenotype relationships.

24.4 PEELING AND RELATED ALGORITHMS

Almost every problem associated with pedigree analysis or other complex genetic applications involves a difficult computation. This could be the computation of a likelihood, the probability of an individual having a specific allele, genotype or haplotype, or some other characteristic of the system under investigation. Superficially, such computations seem too complex to be feasible at all and indeed many are not. However, there are a number of related computational algorithms that exploit the local structure of the system. These algorithms yield drastic reductions in the computational complexity. In genetic applications, such computation is typically referred to as *peeling* (Elston and Stewart, 1971; Cannings *et al.*, 1978; Lander and Green, 1987). See also Thompson (2001; 2000), Heath (2003), and Lauritzen and Sheehan (2003) for further discussion.

The peeling algorithms are special cases, or variants, of general algorithms for so-called local computation on graphs (Cowell *et al.*, 1999). In this section, we describe and explain a general algorithm and how it can be applied in this context. The algorithm can be seen as having two phases. During the first phase, a suitable computational structure is established. In the second phase, the computations themselves are executed. The first phase is sometimes referred to as *compilation*, the latter as *propagation*.

24.4.1 Compilation

The compilation process involves the collection of groups of variables into *cliques* so that computations can be performed locally, i.e. only involving functions of sets of variables belonging to the same clique. At the next stage, these cliques are organized in a tree structure, the *junction tree*, which is used to coordinate the local computations in a consistent way to yield the desired correct global result. Finally the numbers to be used in the calculations are associated with the relevant location in the junction tree. The various steps of the compilation process are described in further detail below.

24.4.1.1 From Bayesian Network to Undirected Graph

The local computation algorithms are based on undirected graphs. The first step, therefore, is to transform the Bayesian network structure into an undirected graph. This is done by removing the directions from the existing edges and adding further undirected edges between all pairs of graph parents with a common (graph) child node. The latter process is referred to as *moralising* the graph, i.e. by 'marrying' the (graph) parents. In the resulting *moral graph*, all sets of the form $\{v\} \cup \mathrm{pa}(v)$ are *complete* in the graph, meaning that all pairs of elements are connected with edges. The factorization (24.4) can therefore be written as

$$f(x) = \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}(v)}) = \prod_{C \in \mathcal{C}} \phi_C(x_C), \tag{24.8}$$

where $\mathcal{C}$ denotes the set of *cliques* of the moral graph, i.e. the maximal complete subsets of nodes, and the functions ϕ are the *potentials*. To obtain this factorization, we just collect factors $f(x_v \mid x_{\mathrm{pa}(v)})$ with $\{v\} \cup \mathrm{pa}(v)$ in the clique C, so that the potential ϕ_C is a product of these factors. Since $\{v\} \cup \mathrm{pa}(v)$ is complete in the moral graph, this can always be done. Heath (2003) uses the term *dependency graph* for the moral graph.

Figure 24.8(a) shows a Bayesian allele network corresponding to a modification of the pedigree in Figure 24.2, where the phenotypes of individuals 11 and 15 have been explicitly represented. The corresponding moral graph is displayed in Figure 24.8(b).

24.4.1.2 Triangulation

Computational difficulties associated with pedigree analysis are related to the cycles of the moral graph rather than to the loops of the pedigree graph. The next step of the compilation process is addressing this problem through *triangulation* of the graph by adding *fill-in* edges to the moral graph until all cycles involving more than three nodes have *chords*. When this has been done, and only then, the cliques of the resulting graph can be arranged in a junction tree, see details below.

A triangulation of the graph in Figure 24.8(b) is displayed in Figure 24.9, where six fill-in edges have been added. Such a triangulation is most often found by using an ordering for node *elimination*; when a node is eliminated, fill-in edges are added between any pairs of the node's neighbours, which are not already connected by an edge. The node is then removed together with all its neighbours. The notion of an elimination ordering is identical to what is known as a *peeling sequence*, and the term *peeling* refers to the
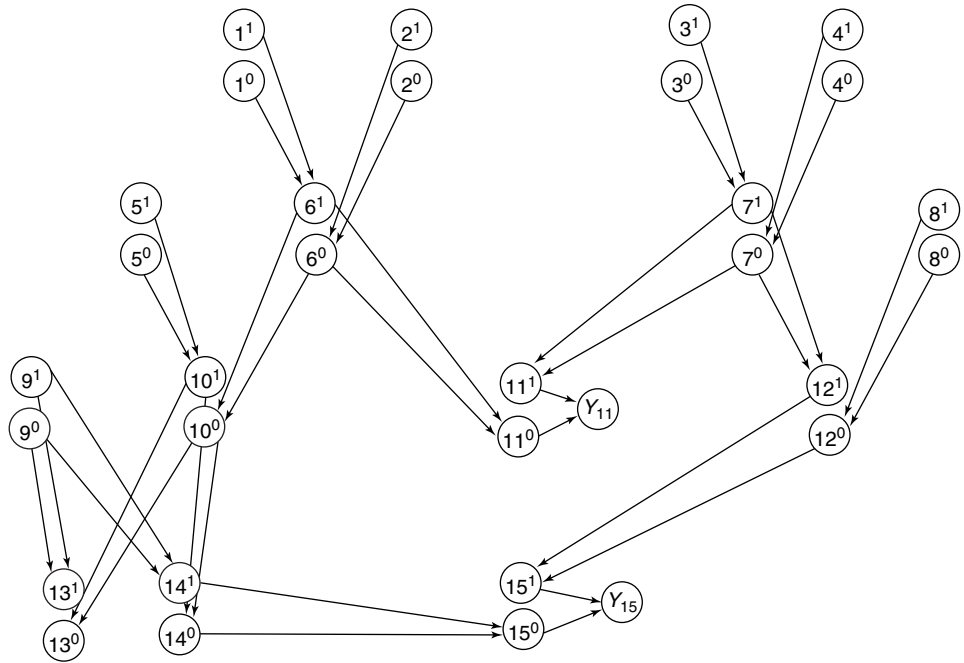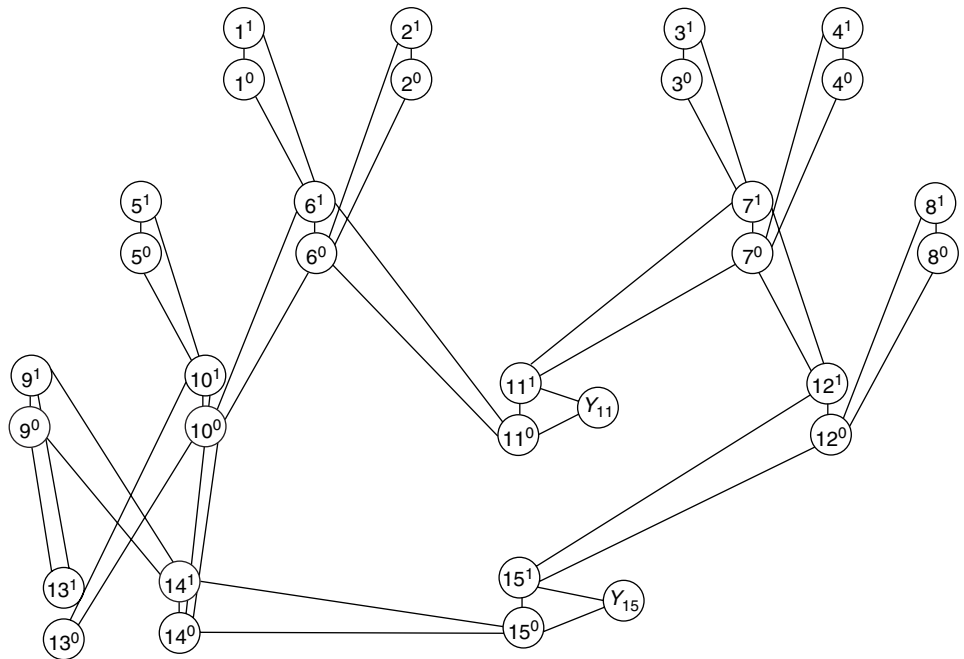
Figure 24.8 (a) Bayesian allele network with phenotypic information on two individuals and (b) its associated moral graph.
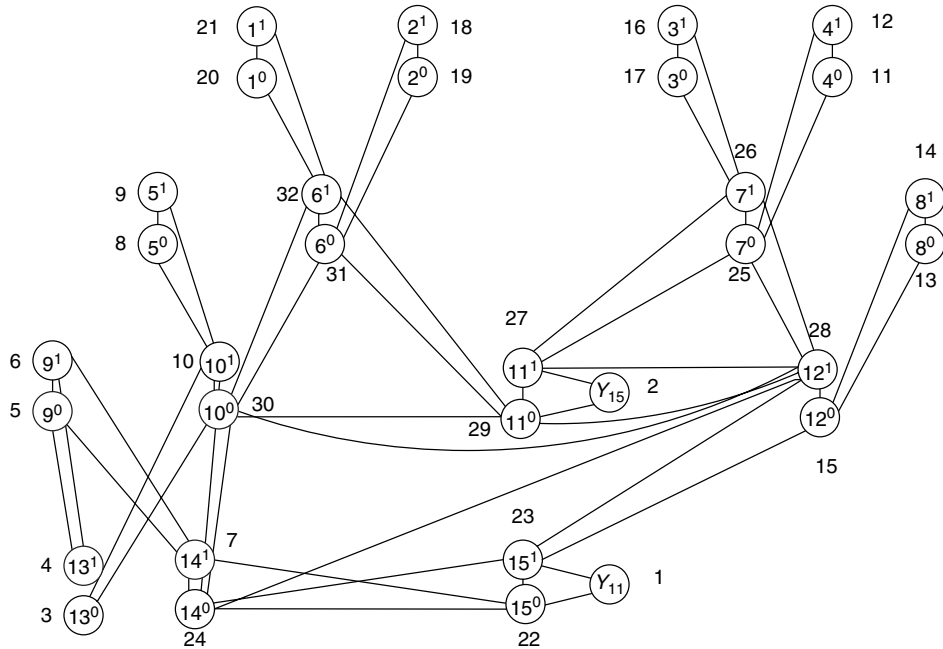
Figure 24.9   A triangulated graph for the Bayesian allele network of Figure 24.8 with phenotypic information represented for two individuals. The numbers $1, \ldots, 32$ indicate the corresponding node-elimination ordering.

elimination process, where one node is 'peeled off' at a time. The factorization (24.8) clearly implies a similar factorization

$$f(x) = \prod_{C \in \mathcal{C}} \phi_C(x_C), \qquad (24.9)$$

where $\mathcal{C}$ now denotes the set of cliques in the triangulated graph, since cliques in the moral graph are complete in any graph with more edges.

Figure 24.9 also displays the elimination order used to produce the given triangulation. A triangulation is not unique and the goal is to generate *cliques* (maximal sets of pairwise connected nodes) which are as small as possible. Optimizing this step is known to be NP (nondeterministic polynomial) complete (Yannakakis, 1981), but Jensen (2002) has implemented an algorithm which, in most cases, runs at reasonable computational speed and is guaranteed to return an optimal triangulation. This is based on the work of (Shoiket and Geiger, 1997; Berry *et al.*, 2000; Bouchitté and Todinca, 2001).

The triangulation step is crucial, as this determines the computational complexity, and thus whether exact computations are at all feasible or whether approximate methods such as Markov chain Monte Carlo (MCMC) will be required.

24.4.1.3   Constructing the Junction Tree

Once the graph has been triangulated, the cliques can easily be identified and connected in what is known as a *junction tree*. This is a tree having the set $\mathcal{C}$ of cliques of a triangulated

graph as nodes, and satisfying the further property that

$$C \cap D \subseteq E \text{ for all } C, D, E \in \mathcal{C} \text{ with } E \text{ between } C \text{ and } D, \qquad (24.10)$$

where E is between C and D if it lies on the unique path from C to D. A junction tree for the triangulated graph in Figure 24.9 is displayed in Figure 24.10.

24.4.1.4 Loading the Junction Tree

The next step is to identify the *potentials* ϕ_C in the factorization (24.9). This is done by collecting factors of the form $f(x_v \mid x_{\mathrm{pa}(v)})$ in (24.8) into cliques which contain both v and $\mathrm{pa}(v)$. For each node v, at least one such clique exists and we choose one of them, say C, and *assign* the node v to C. If $V(C)$ denotes the set of nodes which are assigned to C, we let $\phi_C(x_C) \equiv 1$ for $V(C) = \emptyset$, otherwise,

$$\phi_C(x_C) = \prod_{v \in V(C)} f(x_v \mid x_{\mathrm{pa}(v)}),$$
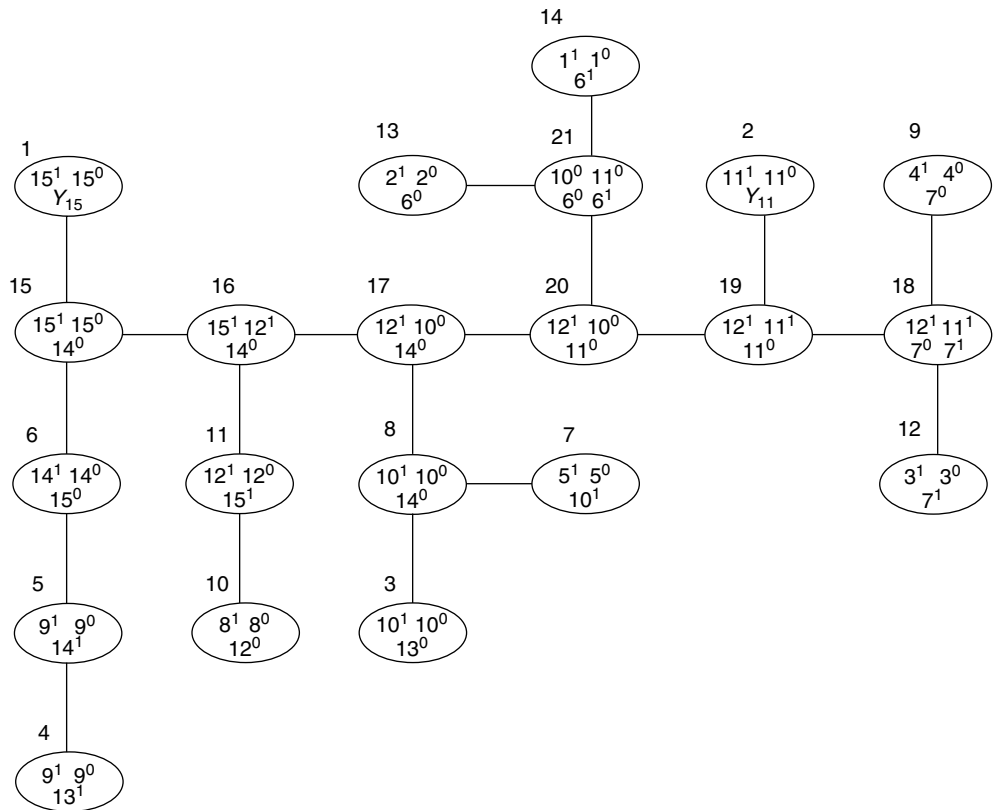
Figure 24.10   Junction tree for the triangulation of the Bayesian allele network shown in Figure 24.9.

whereby (24.9) is clearly satisfied with the joint distribution expressible as a product of potentials over the cliques. This concludes the general part of the compilation process.

24.4.1.5 Incorporating Observations

The compilation process described above has not yet taken account of the data available for the analysis in question. The representation (24.9) gives the joint probability of an arbitrary configuration of variables in the network. However, we want the probability of configurations which are consistent with the observations. This can be obtained if, for all v where $X_v = x_v^*$ is observed, we find a clique C with $v \in C$ and modify the potential there to ϕ_C^* by changing appropriate values to zero. More precisely, we let

$$\phi_C^*(x_C) = \begin{cases} \phi_C(x_C) & \text{if } x_v = x_v^* \\ 0 & \text{otherwise.} \end{cases} \tag{24.11}$$

This then implies that $\prod_C \phi_C^*(x_C)$ is equal to the joint probability of an arbitrary configuration x which is consistent with the observations. The process of forming ϕ^* from ϕ is often referred to as *entering evidence*. If we denote the set of observed nodes with E, we have

$$f(x \mid x_E^*) = \frac{\prod_{C \in \mathcal{C}} \phi_C^*(x_C)}{Z(x_E^*)}, \tag{24.12}$$

where the normalizing constant $Z(x_E^*)$ is the probability of the observations, obtained by summing over all configurations which are consistent with the observations:

$$f(x_E^*) = Z(x_E^*) = \sum_{x:x_E=x_E^*} \prod_{C \in \mathcal{C}} \phi_C(x_C) = \sum_x \prod_{C \in \mathcal{C}} \phi_C^*(x_C). \tag{24.13}$$

This also yields the *likelihood* when comparing different models.

24.4.2 Propagation

In the second part of the algorithm, often referred to as *propagation of evidence*, the actual computations with numbers are made, and the probabilities of interest are calculated. In particular, the sum in (24.13) must be calculated with more sophisticated techniques than brute force, since the number of terms in the sum grows exponentially with the number of nodes in the network.

There are several variants of the general algorithm of which we describe the HUGIN procedure (Jensen *et al.*, 1990), which represents a refinement of the algorithm of Lauritzen and Spiegelhalter (1988). Another variant, known as the *Shafer–Shenoy procedure* (Shenoy and Shafer, 1990), is closer to what is known as *peeling*, but it includes the more general variant used in Thompson (1981) to derive gene probabilities for all individuals in the pedigree.

24.4.2.1 The HUGIN Procedure

With every branch of the junction tree between neighbours C and D, we associate a *separator* $S = C \cap D$. The algorithm used in the software HUGIN (Andersen *et al.*, 1989)

makes specific use of the separators by storing a single potential ψ_S along every branch of the junction tree. Initially, all these separator potentials are identically set to be equal to unity, so the factorization (24.12) implies that

$$f(x \mid x_E^*) \propto \frac{\prod_{C \in \mathcal{C}} \psi_C(x_C)}{\prod_{S \in \mathcal{S}} \psi_S(x_S)}, \tag{24.14}$$

where $\mathcal{S}$ is the set of separators and, initially, $\psi_C = \phi_C^*$ after evidence has been entered.

When a message is sent from C to D via the separator $S = C \cap D$, the following operations are performed:

$$\psi_C^{\downarrow S}(x_C) = \sum_{y_{C \setminus S}} \psi_C(x_S, y_{C \setminus D})$$

$$\tilde{\psi}_D(x_D) = \psi_D(x_D) \frac{\psi_C^{\downarrow S}(x_S)}{\psi_S(x_S)}$$

$$\tilde{\psi}_S(x_S) = \psi_C^{\downarrow S}(x_S),$$

i.e. first the *S-marginal* $\psi_C^{\downarrow S}$ of ψ_C is calculated by summing out over all variables not in S, then the clique potential ψ_D is modified by multiplication with the 'likelihood ratio' $\psi_C^{\downarrow S}/\psi_S$, and finally the separator potential ψ_S is replaced with $\psi_C^{\downarrow S}$. The potential from the clique which sends the message is unmodified, i.e. $\tilde{\psi}_C = \psi_C$. Since we have

$$\frac{\tilde{\psi}_C(x_C) \tilde{\psi}_D(x_D)}{\tilde{\psi}_S(x_S)} = \frac{\psi_C(x_C) \left(\psi_D(x_D) \frac{\psi_C^{\downarrow S}(x_S)}{\psi_S(x_S)} \right)}{\psi_C^{\downarrow S}(x_C)} = \frac{\psi_C(x_C) \psi_D(x_D)}{\psi_S(x_S)},$$

the factorization (24.14) remains valid at all times during the computational procedure.

Messages are now sent between neighbours in the tree according to a specific *schedule*. An efficient message passing schedule allows a clique to send exactly one message to each of its neighbours only after it has already received messages from all its other neighbours. Such a message passing schedule can be implemented via a local control. Alternatively, one can use a global control by first choosing a root R, then making an inward pass through the junction tree, known as COLLECTEVIDENCE, by which messages are sent from the leaves inwards towards R, and subsequently making an outward pass, DISTRIBUTEEVIDENCE, which sends messages in the reverse direction from the root towards the leaves. The first of these phases is illustrated in Figure 24.11.

When exactly two messages have been sent along every branch of the junction tree in an efficient schedule, it holds that

$$f(x_A \mid x_E^*) = \psi_A(x_A)/Z(x_E^*), \quad \text{for all } A \in \mathcal{C} \cup \mathcal{S}. \tag{24.15}$$

The marginal probability and normaliser Z can therefore be found as

$$f(x_E^*) = Z(x_E^*) = \sum_{x_S} \psi_S(x_S),$$
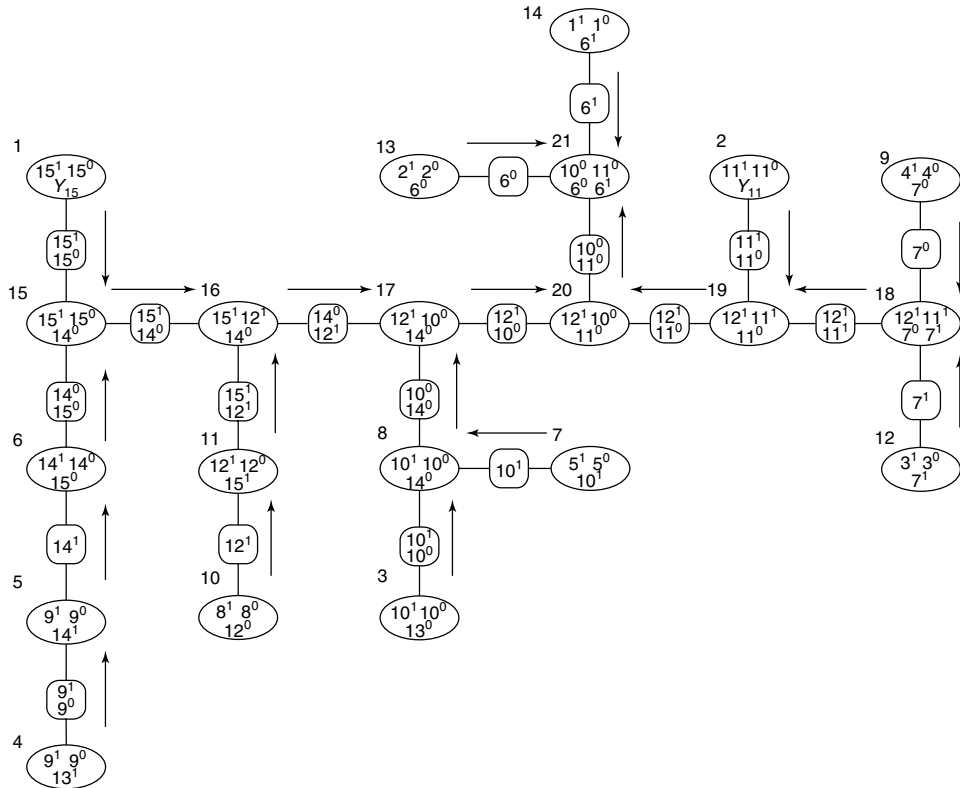
Figure 24.11 The first of the two computational phases in the HUGIN procedure. During COLLECTEVIDENCE, messages are sent towards the root C_{21}.

from any of the separator potentials ψ_S. In particular, the separator with the smallest associated state space can be chosen to calculate this sum.

24.4.3 Random and Other Propagation Schemes

Some generalizations of the message passing schemes described above use different definitions of the marginalisation operation $\downarrow$ and the multiplication in the basic factorization and message computation, but otherwise work in essentially the same fashion (Shenoy and Shafer, 1990; Lauritzen and Jensen, 1997). For example, replacing summation with maximization in (24.13) still yields a valid propagation scheme, known as *max-propagation*. Then, after COLLECTEVIDENCE, the (max) normalisation constant Z is the probability of the most probable configuration of all variables in the network, and this configuration will be identified after DISTRIBUTEEVIDENCE (Dawid, 1992). Since the relation (24.14) remains invariant in the HUGIN procedure (also under max-propagation), one can easily switch between propagation modes.

Another important generalisation is the *random propagate* algorithm described by Dawid (1992). This begins with COLLECTEVIDENCE to a root R using sum-marginalisation, but in the reverse step, a Monte Carlo sample is drawn as follows. After COLLECTEVIDENCE, the potential ψ_R is proportional to the conditional probability distribution of the variables

in the root clique, given the evidence, cf (24.15). Hence, a random configuration $\check{x}_R$ can readily be sampled according to this distribution. The root clique now passes this configuration on to each of its neighbours C as $\check{x}_{R \cap C} = \check{x}_S$, where $S = C \cap R$ is the separator between C and R. After this has been done, each of the neighbouring cliques C chooses a random configuration $\check{x}_{C \setminus S}$ of the remaining variables according to a probability distribution which is proportional to $\psi_C(x_{C \setminus S}, \check{x}_S)$. When the neighbouring cliques have sampled their configurations in this way, they in turn pass on the chosen configuration to their neighbours, and so on. When the sampling stops at the leaves of the junction tree, a configuration $\check{x}$ has been correctly generated from the conditional distribution $f(x \mid x_E^*)$, given the evidence. This procedure is the general version of what (Thompson, 2000, page 95) describes as a variation of the Baum (1972) algorithm, and forms an essential step in many Monte Carlo based computational schemes which are relevant for genetic analyses. In particular, any sampling scheme which carries out a block update on several variables jointly and conditionally on the values of the remaining variables in the network makes use of the *random propagate* algorithm.

24.4.4 Computational Shortcuts

Computational issues have been considered by geneticists for a long time. As a result, a number of shortcuts have been developed which speed up computations beyond the efficiency intrinsic to the local computation algorithms themselves. These shortcuts are all associated with pre-processing before the compilation and propagation steps, and have the purpose of eventually leading to a reduction of the total size of the state spaces associated with the cliques of the final junction tree. Several of these pre-processing steps are, for example, described in Sheehan (2000), Fishelson and Geiger (2002), and Heath (2003). We refer to Lauritzen and Sheehan (2003) for details of these procedures known as, e.g. *trimming, forcing, excluding, allele recoding,* and *delayed triangulation*.

24.5 PEDIGREE ANALYSIS AND BEYOND

We will now use Bayesian network representations for some specific problems involving pedigree analysis.

24.5.1 Single-point Linkage Analysis

Consider a diallelic dichotomous disease segregating through a population with alleles D and d, and affected and normal phenotypes. Typically, we will have some observed phenotypes for the disease and some individuals will be typed at the marker locus. As in Kong (1991), we assume a recessive model for disease with complete penetrance whereby dd homozygotes are always affected and are never normal while both other genotypes are always normal and never affected. Assume that allele frequencies for both loci are known, segregation is Mendelian, founder genotype frequencies are in Hardy–Weinberg Proportions, and the founder population is in linkage equilibrium. The only unknown quantities are the unobserved genotypes and phenotypes, and r, the recombination fraction between the two loci.

To construct a graphical model for the single-point linkage problem, it suffices to focus on a nuclear family comprising a father 1, mother 2, and their offspring 3.

This construction is then replicated for all parent–child triplets in the pedigree, e.g. using an object-oriented method of specification. We will use the segregation network representation as we need to explicitly refer to phase information for linkage. Beginning with the marker locus–the 'α locus'–for each parent $i = 1, 2$, we create two nodes, $i^{1\alpha}$ and $i^{0\alpha}$, for the paternal and maternal genes of the individual with values drawn from the marker allele frequency distribution (e.g. multinomial). Note that this random assignment immediately deals with the fact that phase is unknown in the parents and we have to integrate it out by summing over all possibilities. We can assume that segregation from both parents is Mendelian at this 'first' locus, i.e.

$$P(S^{\alpha}_{1,3} = 1) = P(S^{\alpha}_{2,3} = 0) = 1/2, \tag{24.16}$$

although this is not necessary. As in Section 24.3, the paternally inherited allele of the offspring, $3^{1\alpha}$, is a (graph) child of both alleles in the father, $1^{1\alpha}$ and $1^{0\alpha}$, and of $S^{\alpha}_{1,3}$. Likewise, the maternally inherited allele of 3 is a (graph) child of both genes of 2 and $S^{\alpha}_{2,3}$. The genotype node is a (graph) child of both genes of individual nodes.

Labelling the disease locus as δ, the graph is extended by adding two nodes $i^{1\delta}$ and $i^{0\delta}$ for each parent, 1 and 2, exactly as above. The unobserved disease genotype, G^{δ}_i, is a child node of the corresponding gene nodes and a (graph) parent of the observable phenotype, Y^{δ}_i with link specified by the penetrance function. In this case, the penetrance probabilities are either 0 or 1. For the offspring, 3, we have gene nodes and a segregation indicator exactly as for the marker locus with the difference now being that we must take account of the linkage between the loci. In particular, the value of the segregation indicators $S^{\delta}_{1,3}$ and $S^{\alpha}_{1,3}$ are dependent via the recombination fraction, r. This dependence can be modelled with an undirected link between the corresponding nodes. Formally this would lead to a chain graph representation (Lauritzen, 1996) rather than a DAG. However, for the sake of exposition, we use here the equivalent non-symmetric description through the conditional distribution of $S^{\delta}_{1,3}$, given $S^{\alpha}_{1,3}$, specifically:

$$S^{\delta}_{1,3} \sim \begin{cases} Ber(1-r) & \text{if } S^{\alpha}_{1,3} = 1 \\ Ber(r) & \text{if } S^{\alpha}_{1,3} = 0, \end{cases} \tag{24.17}$$

and similarly for $S^{\delta}_{2,3}$. To complete the graph in Figure 24.12, we now add nodes G^{δ}_3 and Y^{δ}_3 for the offspring's unobserved genotype and phenotype with links defined exactly as above.

Note that this is a full specification of the model similar to that described elsewhere (Kong, 1991; Jensen and Kong, 1999; Thomas *et al.*, 2000). It is important to note that further derivation of the relevant joint and marginal distributions is not necessary as these are a direct result of the induced factorisation (24.4). The graph in Figure 24.12 provides a clear visual representation of the model.

24.5.2 QTL Mapping

Sheehan *et al.* (2002) extend the linkage scenario described above to the problem of detecting a QTL from possibly incomplete marker data for a simple example involving two flanking loci. Two markers are considered with known map positions and it is hypothesised that there is a diallelic QTL somewhere between the two. The trait of interest is any trait measured on a continuum with an associated polygenic effect unlinked to the QTL.
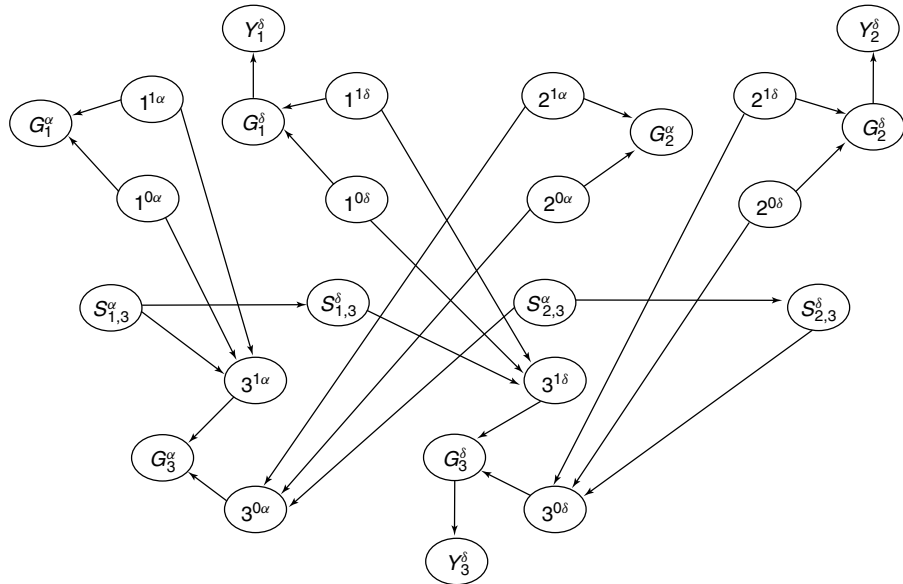
Figure 24.12 The segregation network for two linked loci on individuals 1 (father), 2 (mother), and 3 (offspring). Note that the information on linkage is contained in the directed edge between the segregation indicators $S_{1,3}^{\alpha}$ and $S_{1,3}^{\delta}$ and similarly between $S_{2,3}^{\alpha}$ and $S_{2,3}^{\delta}$.

Marker data are available on a *half-sib* design, common in animal breeding applications, comprising several families each with a single sire and up to 100 offsprings. Trait data are available only on the offspring. In contrast with the two-locus linkage example above, no information is given on the mothers (dams) of these offspring and hence the maternal segregations are all ignored.

The phenotype record on offspring j of sire i is a realisation of the random variable Y_{ij}. The effect of the unobserved genotype at the QTL is q_{ij}, where q_{ij} can have three possible values μ_1, μ_2, μ_3 corresponding to each of the three genotypes. A normal linear mixed model for the data is

$$Y_{ij} = Z_i + q_{ij} + E_{ij},$$

where Z_i represents the average additive genetic effect of the ith sire on the phenotypes of his offspring, which cannot be explained by the QTL. Let σ_a^2 be the total additive genetic variance unexplained by the QTL and σ_e^2 be the environmental variance. We have that $Z_i \sim N(0, \sigma_z^2) \forall i$ where $\sigma_z^2 = \frac{1}{4}\sigma_a^2$, the sire variance component, since half the genes of an offspring are shared with its sire (Falconer and Mackay, 1996). The remaining unexplained variation is picked up by the residual term, $E_{ij} \sim N(0, \frac{3}{4}\sigma_a^2 + \sigma_e^2)$. Assuming no genetic interference, only one of the unknown marker-QTL recombination fractions, or equivalently the QTL map location λ_Q, is required to parameterise the problem.

Figure 24.13 shows the graphical model for this trivial QTL-mapping problem for one sire and two offspring. The marker loci are labelled α and β while the QTL locus is now δ. The model is essentially an extension of the single-point linkage problem in Figure 24.12 to a two-point problem. Gene nodes are added for the third locus in an analogous fashion with the segregation indicator for inheritance from the sire linked to the previous value
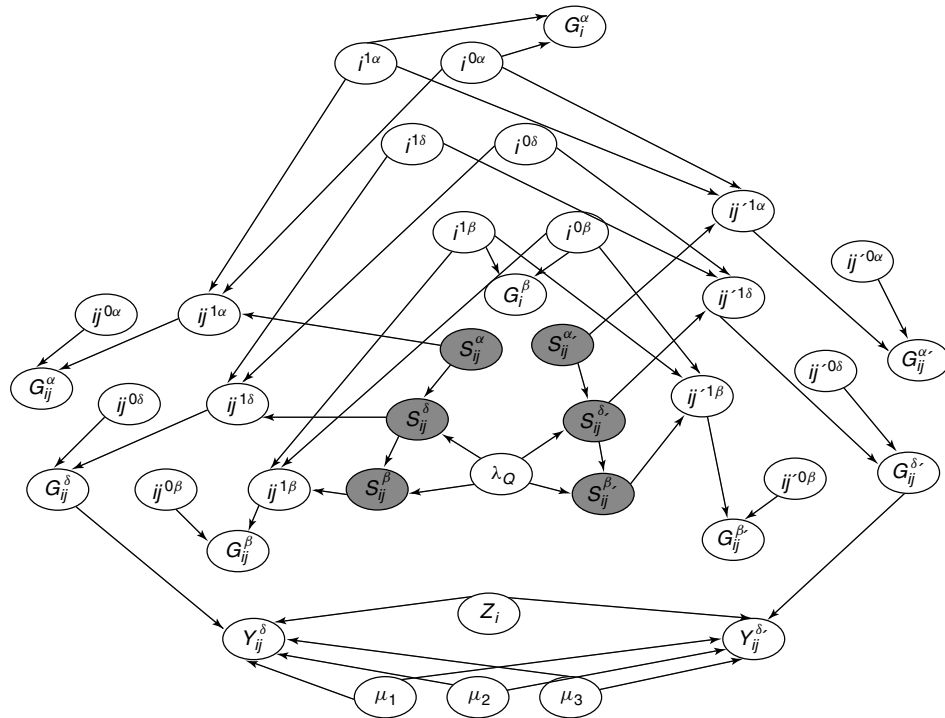
Figure 24.13 The graphical model for the QTL-mapping problem depicting a sire, i with two daughters, ij and ij'. [Reproduced from Sheehan *et al.* (2002). International Statistical Review, 68:83–110 with permission from Blackwell Publishing.]

via the recombination fraction between the second and third loci, as described in (24.17) above. This assumes that there is no genetic interference and that recombinations in adjacent intervals are independent. Maternal genes in the offspring are randomly drawn from the population as there is no information on the dams. Sire and offspring have marker genotype nodes while only offspring have trait genotype and phenotype nodes. Covariance between the offspring is reflected in the genes they share with their sire, and they are duly connected by the sire effect node. Note that this creates a cycle in the graph which becomes increasingly complex computationally when more offspring are added and, for a typical half-sib design with a sire having up to 100 offspring, the relevant Bayesian network for this problem features many long cycles despite the simplicity of the pedigree structure and the mapping problem under consideration (Sheehan *et al.*, 2002). The half-sib design is a zero-loop pedigree (or *tree*) and genetic mapping problems do not get any simpler than this one. The graphical model highlights the computational complexity implicit in the mixed-model approach to this analysis and clarifies why these analyses are challenging on more complex problems.

24.5.3 Pedigree Uncertainty

The flexibility of a graphical modelling approach to applications in genetics is powerfully demonstrated when the pedigree is not fixed and known, or when other circumstances

should be integrated into the analysis. Hansen and Pedersen (1994) elegantly handle incomplete paternity information in a two-locus inheritance model for fur colour in foxes from pedigrees supplied by Scandinavian fur farms where there is uncertainty with some of the litter paternities. It is common breeding practice to mate a female with two males in order to increase the chances of fertilisation and hence, it is not always possible to determine which male actually fathered the resulting pups. Indeed, two males could father pups in the same litter. The pedigree declares the most likely candidate as the father (to all pups) and registers the second sire as an alternative whenever there is doubt.

The phenotypic record on each fox is a subjective classification of fur colour on a scale from $1 - 8$. From analogous models for mice and sheep (Adalsteinsson *et al.*, 1987), a model for genetic inheritance of fur colour was proposed involving two diallelic loci, α (with alleles A and a) and ε (E, e), possibly on the same chromosome with unknown recombination fraction.

Although there are some loops, the fox pedigrees are generally small enough for exact likelihood calculation with simple models (Skjøth *et al.*, 1994). The usual method for handling paternity uncertainty is to compare the likelihoods for all possible pedigrees (Thompson, 1986). There is only one alternative father for each of a small number of litters, but as each pup in the litter could have been fathered by either of the two candidates, this problem would require the consideration of 2^{21} pedigrees. Skjøth *et al.* (1994) circumvent this problem by estimating paternal genotypes from the phenotypic information and choosing the most likely individual but this is not very satisfactory. Although standard statistical genetics programs will not accept a pedigree where an individual can have more than one biological father, this presents no difficulty for a general graphical model. Hansen and Pedersen (1994) exploit this to incorporate all the paternity information by defining a binary node, W_i, for each queried pup:

$$W_i = \begin{cases} 1 & \text{if stated father is the true father} \\ 0 & \text{if alternative father is the true father,} \end{cases}$$

and $W_i \sim Ber(p_w)$, where p_w is to be estimated.

Again, the graphical model for this problem is essentially the two-locus linkage model described above (Section 24.5.1) *except* for those cases where paternity is uncertain. We thus focus on the quadruplet comprising the mother, 1, declared father, 2, offspring, 3, and alternative father, 4. For illustration, we have simplified the model by assuming a known penetrance matrix and known allele frequencies at both loci. As before, paternal and maternal alleles, $i^{1\alpha}, i^{0\alpha}, i^{1\varepsilon}, i^{0\varepsilon}$, are assigned to the founders $i = 1, 2,$ and 4 at both loci. Mother–offspring segregation indicators are assigned for each locus, $S_{1,3}^{\alpha}, S_{1,3}^{\varepsilon}$, and indicators for paternal inheritance are $S_{f_3,3}^{\alpha}, S_{f_3,3}^{\varepsilon}$, where $f_3 = 2$ if $W_3 = 1$ and $f_3 = 4$, otherwise. In contrast with our representation of Section 24.5.1, Hansen and Pedersen (1994) consider segregation at both loci jointly, so we have phase indicators for inheritance, $S_{1,3}$ and $S_{f_3,3}$, where

$$S_{1,3}, S_{f_3,3} = \begin{cases} (0,0) & \text{with probability} & (1\text{-}r)/2 \\ (0,1) & \text{with probability} & r/2 \\ (1,0) & \text{with probability} & r/2 \\ (1,1) & \text{with probability} & (1\text{-}r)/2 \end{cases}.$$
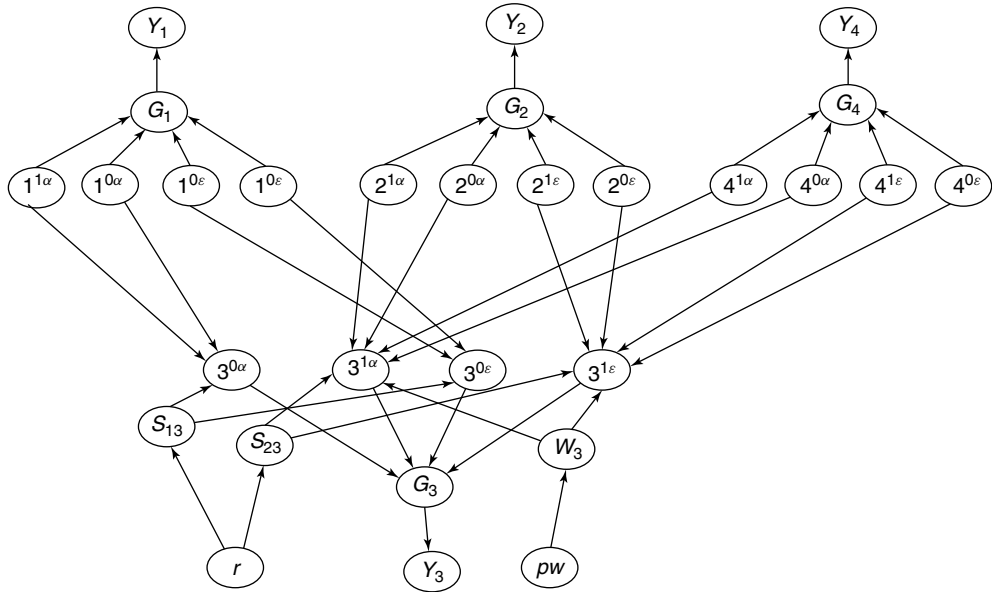
Figure 24.14 The graphical model for the fox data depicting a mother, 1, father 2, offspring 3, and alternative father 4. [Adapted from Hansen and Pedersen (1994).]

Note that this is the same model as described in (24.16) and (24.17), since $S_{1,3} = (S_{1,3}^{\alpha}, S_{1,3}^{\varepsilon})$. The earlier parameterisation, involving more nodes with fewer states, is more flexible when considering more than two loci and is generally better for computational purposes. All four alleles of individual i are graph parents of the node G_i representing the two-locus genotype and this, in turn, is a graph parent of Y_i, the fur colour phenotype. Figure 24.14 shows the corresponding graphical model for this problem.

Despite the simplifications, the graph in Figure 24.14 is more complicated than those shown earlier in this section in that it has many more cycles. The advantage is that questions about paternity, genetic inheritance, and linkage can all be addressed from this one graph, whereas these would typically require separate considerations using standard pedigree software.

An interesting class of questions focuses on the determination of pedigrees from observable genetic information, see, e.g. Egeland *et al.* (2000), Cowell and Mostad (2003), Steel and Hein (2006), and Sheehan and Egeland (2007).

24.5.4 Forensic Applications

Graphical models, or *probabilistic expert systems*, have been shown to be particularly useful in forensic applications (Dawid *et al.*, 2002; Cowell, 2003; Taroni *et al.*, 2006) and have been adapted to handle a wide range of routine problems in forensic inference. The central problem here is to infer the identity of an individual based on the given evidence which possibly includes DNA profile information. This can be done by calculating relative likelihoods for the various competing hypotheses (Dawid and Mortera, 1996), but these become computationally intensive when information is imperfect or missing (Dawid and
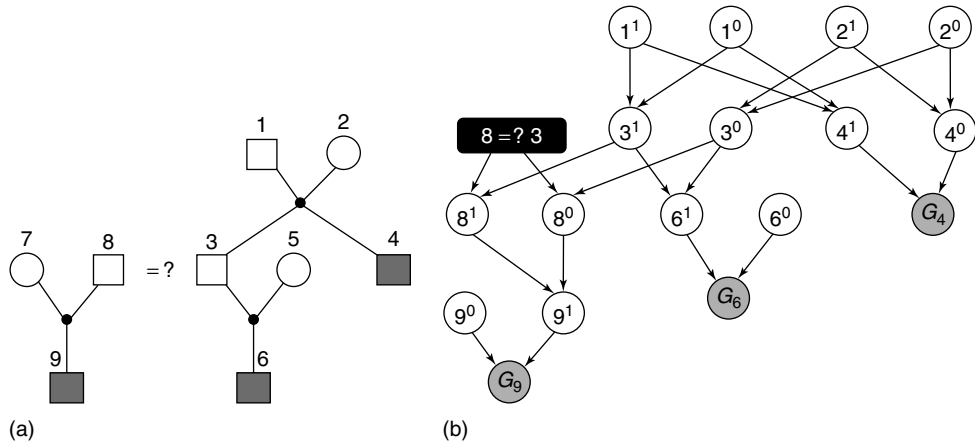
Figure 24.15 The simple paternity problem of Dawid *et al.* (2002) represented here (a) by two marriage node graphs, where individuals shaded in grey denote those for whom DNA evidence is available, and (b) as a graphical model. The three grey nodes in (b) represent the observed genotypes and the black node is the 'query' node.

Mortera, 1998) and especially when the possibility of observing a mutation from one generation to the next is considered (Dawid *et al.*, 2001).

Consider the inheritance claim case in Figure 24.15(a) from Dawid *et al.* (2002), where a man, 9, claims to be the son of the diseased individual 3 and hence entitlement to part of his estate. It is known that 3 had a (undisputed) child, 6, and that 6 and 9 had different mothers. There is no DNA information on 3 since he is dead and buried, nor on either of the two mothers, but we have DNA profile samples from both offspring 6 and 9, and from the brother of the diseased, 4. This problem can be formally expressed as a case of disputed paternity with 3 as the putative father and 9 as the disputed child. As is common for such applications, attention is on just two competing hypotheses: either the true father, 8, of the disputed child, 9, really is one and the same as the putative father, 3, *or* the true father can be considered as randomly drawn from the general population.

Forensic markers are usually selected to be unlinked so we only need a model for a single marker and the overall likelihood is the product over all markers. Figure 24.15(b) shows the single marker representation used by Dawid *et al.* (2002). This uses the allele network rather than the segregation network but, as indicated earlier, the latter is superfluous in the absence of linkage. Our notation is as before with the marker labels omitted, so i^0 and i^1 represent the random variables assigning maternal and paternal genes of the individual i, and G_i assigns the genotype of i at the marker. Untyped individuals who are not directly of interest (i.e. 1, 2, 5, and 7) are only represented by the genes they contribute to the next generation, which, in the absence of any information, are assumed on the contrary to be randomly drawn from the population.

The black node in Figure 24.15(b) which is a (graph) parent of both genes in 8, is the 'query' or 'target' node (Dawid *et al.*, 2002). This is a binary node and is 1 if the true father of the disputed child is the putative father, i.e. if individual 8 is the same as 3 in Figure 24.15(a) and the two genes in 8 are hence copies of the corresponding two in 3. Otherwise, the men are different individuals and the genes of 8 are drawn randomly from

the population. The advantage of using the query node for this application is that the quantity of interest to the court–the likelihood ratio in favour of paternity–can be read off directly. However, we note that despite the different emphasis of the analysis, this node is essentially the same as the paternity indicator, W_i, in Figure 24.14 determining which of the two possible alternatives fathered the individual i. For the forensic example, although a specific alternative is often not available, determination of paternity is crucial: for the fox example, a genetic model for inheritance of fur colour was the focus and the uncertain paternity was a nuisance factor that had to be taken into account. The same graphical modelling approach can be used to address both questions.

The area of forensic genetics yields a variety of problems which can clearly benefit from the flexibility and modularity of a graphical modelling environment. For example, in criminal cases such as assault and rape cases, it is not uncommon to observe crime trace evidence which represents a *mixture* of DNA from an unknown number of individuals. Mortera *et al.* (2003) consider an example where there is a victim 1, a suspect 2, and exactly two contributors, 6 and 7, to the trace evidence (i.e. there were three alleles present for at least one marker in the mix). To complicate matters, the suspect has left the country and is not available for typing. However, his brother 3 has been found and has given a sample. Figure 24.16 shows the allele network used by Mortera *et al.* (2003) for this identification problem, where individuals 4 and 5 are the parents of the two brothers. Four standard competing hypotheses concerning the makeup of the set contributing to the mixture (Weir *et al.*, 1997) are as follows:

1. the suspect and victim have both contributed;

2. the suspect and an unknown contributor are represented;

3. the victim and an unknown contributor are represented;

4. both contributors to the mixture are unknown.

Figure 24.16 shows that we have genotype data on 1 and 3. Unobserved genotypes of the two contributors, 6 and 7, are graph parents of the node representing the observed
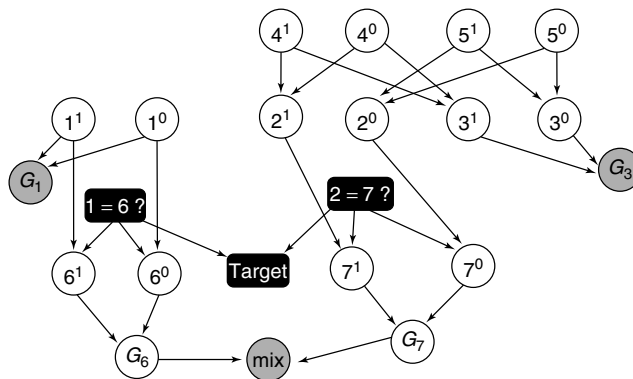
Figure 24.16 The mixture network of Mortera *et al.* (2003). The three grey nodes represent the observed genotypes and the black nodes are the 'query' and 'target' nodes respectively. [Reprinted from Theoretical Population Biology, 63(3), Mortera J. *et al.*, Probabilistic expert systems, 191–205, copyright (2003) with permission from Elsevier.]

mixed trace. We now have two binary query nodes defined exactly as before which address the two questions: is the victim the first contributor ($1 = 6$?) and is the suspect the second ($2 = 7$?). If the answer is 'yes', the genes of the two individuals are identical. Otherwise, the genes of the contributor are drawn randomly from the population gene pool. The 'target' node is the logical conjunction of these two, and has four states corresponding to the four hypotheses above. Thus, competing standard hypotheses can all be considered from the same network and separate pairwise comparisons, as are routinely performed, are not necessary. In principle, this approach can be extended to deal with multiple contributors, possible silent alleles in the mixture, and multiple missing individuals (Mortera *et al.*, 2003), and information on the amount of DNA contributed by each individual can be exploited for separation of DNA profiles (Cowell *et al.*, 2007).

24.5.5 Bayesian Approaches

These models lend themselves readily to Bayesian analysis and interpretation (Spiegelhalter, 1998), where all unknown quantities such as data, latent variables, and model parameters can be regarded as random variables and thus represented as extra nodes in the graph with their associated distributions. Exact local computation algorithms fail as more complex distributional types need to be accommodated, so alternative methods for calculating the quantities of interest, such as MCMC (Hastings, 1970; Metropolis *et al.*, 1953) must be entertained. It is well known that the popular single-site Gibbs sampler Geman and Geman (1984) can mix very slowly in complex models involving both discrete and continuous nodes, even when the sampler is theoretically irreducible (Janss *et al.*, 1995; Heath, 1997; Jensen and Kong, 1999; Lund and Jensen, 1999). Some kind of blocking or joint updating of variables is hence required in order to sample more efficiently. Lund and Jensen (1999), e.g. use graphical models for a Bayesian formulation of a mixed inheritance model, Sheehan *et al.* (2002) extend the model in Figure 24.13 to a full Bayesian analysis for the QTL mapping problem, and Hansen and Pedersen (1994) take a Bayesian approach to the fox problem of Figure 24.14. They all invoke *random propagation* as described in Section 24.4.3 as an essential part of the associated block update for the discrete part of the model.

24.6   CAUSAL INFERENCE

Graphical models, especially DAGs, have natural causal interpretations and hence provide a formal framework to discuss causal concepts. We will illustrate this using *Mendelian randomisation*, an approach to understanding aetiological relationships in observational studies, by way of an example. Inferring causation from observed associations is often a problem with epidemiological data as it is not always clear which of two variables is the cause, which is the effect, or whether both are common effects of a third unobserved variable or confounder. In the case of experimental data, causal inference is facilitated either by using randomisation or experimental control. In many biological settings, it is not possible to randomly assign values of a hypothesised 'cause' to experimental units for ethical, financial, or practical reasons. In epidemiological applications, for example, randomised controlled trials (RCTs) to evaluate the effects of exposures such as smoking,

alcohol consumption, physical activity, and complex nutritional regimes are unlikely to be carried out.

24.6.1 Causal Concepts

As in Pearl (1995), Lauritzen (2000), and Dawid (2002; 2003), we will regard causal inference to be about predicting the effect of *interventions* in a given system. If X is the cause under investigation and Y the response, the question of interest is whether intervening on X has an effect on Y. By intervening on X, we mean that we can set X (or more generally its distribution) to any value we choose without affecting the distributions of the other variables in the system, except through the resulting changes in X. This is clearly idealistic and may not always be justifiable. The *causal effect* of X on Y is a function of the distributions of Y under different interventions in X. It is well known that this is not necessarily equal to the usual conditional distribution $P(Y|X = x)$ which just describes a statistical dependence (Pearl, 2000; Lauritzen, 2000). We will follow Pearl (2000) and use the notation $P(Y|\mathrm{do}(X = x))$ to clarify that conditioning is on intervention in X.

The average causal effect (ACE) is defined as the difference in expectations under different settings of X:

$$ACE(x_1, x_2) = E(Y|\mathrm{do}(X = x_1)) - E(Y|\mathrm{do}(X = x_2)). \qquad (24.18)$$

X is regarded as causal for Y if the ACE is non-zero for some values x_1, x_2. If X is binary, the unique ACE is given by $E(Y|\mathrm{do}(X = 1)) - E(Y|\mathrm{do}(X = 0))$. If Y is continuous, a popular assumption is that the causal dependence of Y on X is linear (possibly after suitable transformations), i.e. $E(Y|\mathrm{do}(X = x)) = \alpha + \beta x$. In this case, the ACE is $\beta(x_1 - x_2)$ and can simply be summarised by β, which is now interpreted as the average effect of increasing X by one unit through some intervention. In the more general cases of more than two categories and/or nonlinear dependency, the ACE is not necessarily summarised by a single parameter, and one may want to choose a different causal parameter altogether (Didelez and Sheehan, 2007).

A causal parameter is *identifiable* if it can be estimated consistently from obtainable information on the joint distribution of the observed variables. Mathematically, this amounts to being able to express the parameter in terms that do not involve the intervention (i.e. the 'do' operation) by only using 'observational' terms that can be estimated from data. In the presence of unknown confounders, e.g. parameters of $P(Y|\mathrm{do}(X = x))$ cannot be estimated directly from observations that represent $P(Y|X = x)$. In the rare case of known confounders, it can be shown that the intervention distribution can be re-expressed in observational terms and can thus be estimated from the observed data by adjusting for these confounders (Pearl, 1995; 2000; Lauritzen, 2000; Dawid, 2002).

24.6.2 Mendelian Randomisation

Mendelian randomisation has been proposed as a method to test for, or estimate, the causal effect of an exposure or phenotype on a disease when confounding is believed to be likely and not fully understood (Davey Smith and Ebrahim, 2003; Katan, 2004). It exploits the idea that a well-understood genotype, known to affect levels of the exposure, affects the disease status only indirectly and is assigned randomly (given the parents' genes) at meiosis, independently of the possible confounding factors. It is well known

in the econometrics and causal literature (Bowden and Turkington, 1984; Angrist *et al.*, 1996; Pearl, 2000; Greenland, 2000) that these properties define an *instrumental* variable (IV), but they are *minimal* conditions in the sense that unique identification of the causal effect of the phenotype on the disease status is possible only in the presence of additional, fairly strong assumptions.

The core conditions that characterise an IV have been given in many different forms, using counterfactual variables (Angrist *et al.*, 1996; Robins, 1997), linear structural equations (Goldberger, 1972; Pearl, 2000, Chapter 7), or conditional independence statements, as we will use here. Our notation and terminology closely follow Greenland *et al.* (1999) and Dawid (2002). We now present these conditions together with a graphical way of depicting and checking the relevant conditional independencies.

Let X and Y be defined as above with the causal effect of X on Y being of primary interest. Furthermore, let G be the variable that we want to use as the instrument (the genotype in our case) and let U be an unobservable variable that will represent the confounding between X and Y. The 'core conditions' that G has to satisfy are the following

1. $G \perp\!\!\!\perp U$, i.e. G must be (marginally) independent of the confounding between X and Y;

2. $G \not\!\perp\!\!\!\perp X$, i.e. G must not be (marginally) independent of X; and

3. $Y \perp\!\!\!\perp G \mid (X, U)$, i.e. conditional on X and the confounder U, the instrument and the response are independent.

Because U is not observable, these assumptions cannot be formally tested and have to be justified on the basis of subject- matter or background knowledge. Moreover, the above assumptions do not imply any testable conditional independencies regarding the instrument G. Figure 24.17 shows the unique DAG involving G, X, Y, and U that satisfies the core conditions 1–3 with corresponding factorisation

$$p(y, x, u, g) = p(y|u, x)\,p(x|u, g)\,p(u)\,p(g). \tag{24.19}$$

Note that DAGs only represent conditional dependencies and independencies: they are not causal in themselves despite the arrow suggesting a 'direction' of dependence. We say that the DAG has a *causal* interpretation with respect to the relationship between X and Y, or, more specifically, the DAG is causal with respect to intervention in X, if we believe that an intervention in X does not change any of the other factors in the joint
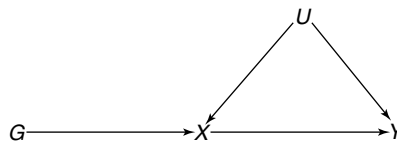
Figure 24.17   The directed acyclic graph (DAG) representing the core conditions for G to be an instrument. [Reproduced from Didelez, V. and Sheehan, N. A. (2007) Mendelian randomisation as an instrumental variable approach to causal inference. Statistical Methods in Medical Research, by permission of Sage Publication Ltd.]

distribution (24.19) (see Pearl, 2000, page 23). This means that

$$p(y, u, g|\mathrm{do}(X = x_0)) = p(y|u, x_0)p(u)p(g), \qquad (24.20)$$

assuming that $p(y|u, x_0) = p(y|u, \mathrm{do}(X = x_0))$.

The limitations of Mendelian randomisation, from the perspective of complicating features leading to poor estimation of the required genotype–phenotype and genotype–disease associations, have been discussed in detail in several places in the literature (Davey Smith and Ebrahim, 2003; 2004; Thomas and Conti, 2004; Davey Smith $et\ al.$, 2005). More crucially, biological complications can sometimes violate one or more of the core conditions 1–3 so that Figure 24.17 no longer applies. In order to understand what any added complexity implies with regard to meeting these conditions, the relevant conditional independencies can be easily checked using DAGs that are ideally dictated by the biology. For instance, when our chosen gene G_1 is in linkage disequilibrium with another gene G_2 which has a direct or indirect influence on the disease Y, condition 3, $Y \perp\!\!\!\perp G_1|(X, U)$, might be violated as shown in Figure 24.18(a), or else condition 1, $G \perp\!\!\!\perp U$, might be violated as shown in Figure 24.18(b).

It is possible that the core conditions may still hold for the chosen instrument G_1 in the presence of genetic heterogeneity, as illustrated in Figure 24.19, if none of the other genes influence Y in any way other than via their effect on X. If instead, the situation is similar to Figure 24.18(b), however, the core conditions may be violated as already explained. Note that in Figure 24.19, genetic heterogeneity could weaken the $G_1 - X$ association and thus G_1 would be a poor instrument. If the chosen instrument G has pleiotropic effects and, in particular, if it is associated with another intermediate phenotype which also affects the disease Y (Figure 24.20(a)), condition 3, $Y \perp\!\!\!\perp G|(X_1, U)$, is again violated if we do not also condition on X_2. Moreover, a genetic polymorphism under study might have pleiotropic effects that influence confounding factors like consumption of tobacco or alcohol, for example, (Davey Smith and Ebrahim, 2003). This is represented in Figure 24.20(b) and violates condition 1. In the presence of population stratification, we see in Figure 24.21(a) that condition 3, $Y \perp\!\!\!\perp G_1|(X, U)$, is again violated: we need to condition on the population subgroup as well. However, if the effect of population stratification is to cause an association between allele frequencies and phenotype levels, as in Figure 24.21(b), all conditions for G to be an instrument are still satisfied, and, in this situation, the $G - X$ association may in fact be strengthened, as a result.
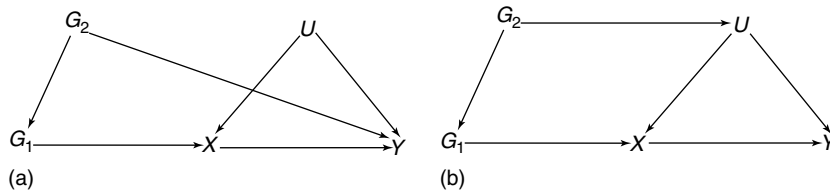
Figure 24.18 Linkage disequilibrium where the chosen instrument G_1 is associated with another genotype G_2 which directly influences the outcome Y, as in (a), or influences Y indirectly via the confounder U, as in (b). [Reproduced from Didelez, V. and Sheehan, N. A. (2007) Mendelian randomisation as an instrumental variable approach to causal inference. Statistical Methods in Medical Research, by permission of Sage Publication Ltd.]
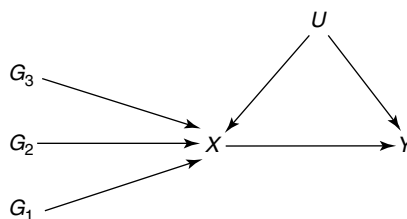
Figure 24.19 Genetic heterogeneity showing three genes which are all associated with the intermediate phenotype X, but none of which has an effect on the disease Y except through X. [Reproduced from Didelez, V. and Sheehan, N. A. (2007) Mendelian randomisation as an instrumental variable approach to causal inference. Statistical Methods in Medical Research, by permission of Sage Publication Ltd.]
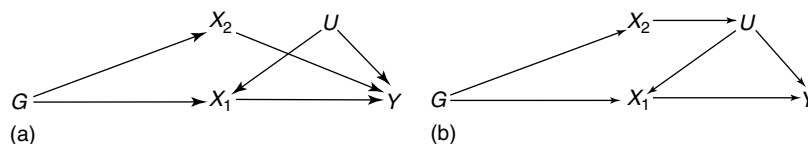
Figure 24.20 An example of pleiotropy where the instrument G is associated with both X_1 and X_2 and (a) both have a direct effect on the outcome Y of interest, or (b) where X_1 has a direct effect but X_2 has an indirect effect via the confounder U. [Reproduced from Didelez, V. and Sheehan, N. A. (2007) Mendelian randomisation as an instrumental variable approach to causal inference. Statistical Methods in Medical Research, by permission of Sage Publication Ltd.]
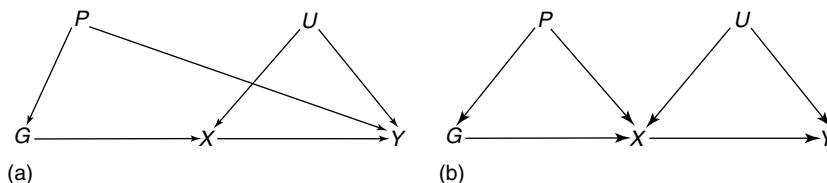
Figure 24.21 Two examples of population stratification where (a) one of the conditions for G to be an instrument is violated and (b) all conditions are satisfied. [Reproduced from Didelez, V. and Sheehan, N. A. (2007) Mendelian randomisation as an instrumental variable approach to causal inference. Statistical Methods in Medical Research, by permission of Sage Publication Ltd.]

24.7   OTHER APPLICATIONS

24.7.1   Graph Learning for Genome-wide Associations

The above examples all assume that the graphical structure of the problem is known. It is also possible to learn from the data and estimate the graphical model from observations of a joint distribution (Cowell *et al.*, 1999). Modelling linkage disequilibrium, or the tendency for alleles observed at one genetic locus to be correlated with those observed at another locus, is an important issue in statistical genetics with the increasing availability

of single nucleotide polymorphism (SNP) data over dense marker maps for large numbers of individuals. In particular, the search for genetic determinants of complex diseases in genome-wide association studies is based on the idea that the ancestral disease-bearing mutations in a population will be flanked by segments of chromosome that will show less variability amongst those individuals with the disease than amongst those unaffected. Linkage disequilibrium is a function of physical distance between loci on a chromosome, but much heterogeneity has been observed in the correlations between adjacent loci, long range correlations are quite common, and completely uncorrelated loci can be interspersed between regions of tightly linked loci. Existing methods for multilocus haplotype analysis exploiting excess sharing amongst affected individuals around a disease locus do not scale up to the datasets that we can now reasonably anticipate (Verzilli *et al.*, 2006).

Thomas and Camp (2004) suggest the use of undirected graphs to model dependencies between genetic loci allowing for higher order interactions. Their method assumes proximal loci (e.g. SNPs within a single gene) and requires haploid data. The space of decomposable, or triangulated, graphs is searched to find the best-fitting model within that class using a simulated annealing algorithm, where the objective function is the maximised log-likelihood penalised according to the number of degrees of freedom. A simple perturbation rule for searching this space is provided by which two vertices of the existing graph are selected, and are connected if they are not already connected, disconnected otherwise. However, the authors acknowledge that mixing properties are greatly enhanced when other perturbation rules are included. The method is extended to deal with unphased diploid data (i.e. genotype data) by Thomas (2005), where the haplotype estimation step is incorporated iteratively with the estimation of the graphical model. Verzilli *et al.* (2006) apply these ideas to identify patterns of multilocus genotypes around a disease locus and thereby avoid the problems inherent in the estimation of haplotypes for large numbers of loci. They consider case-control data where the unphased genotypes, together with a binary disease status indicator, form the vertices of the graph. Edges between genotypes thus reflect the linkage disequilibrium structure, and edges between genotypes and the disease status indicator would suggest the presence of a disease susceptibility locus somewhere near these loci. Computational efficiency is achieved by the restriction of the search space to the set of triangulated graphs and by putting a limit on the size of the cliques. The set of possible graphs is restricted even further by imposing a prior that restricts the physical distance between clique members. Proposed moves around this space act via changes to the set of cliques and separators in the current graph thus avoiding the need to check that the new proposal is decomposable.

Although some restrictions on the set of graphs connecting over 500 000 genotypes and a disease status indicator are obviously required to make the graph learning exercise feasible, the above restrictions, as conceded by all the authors, are mainly driven by the desire to facilitate the computations and provide a solution in reasonable time. Consequently, all kinds of configurations are ruled out by these restrictions which may not be biologically implausible. For example, there does not seem to be any strong reason to believe that the graph in Figure 24.22(b) is an unreasonable model for four linked loci, where the locus at each end is associated with the disease. The complexity of the biology underlying most complex diseases is still not well understood, however, and it is thus difficult to suggest a more sensible restriction on the set of models to be explored.
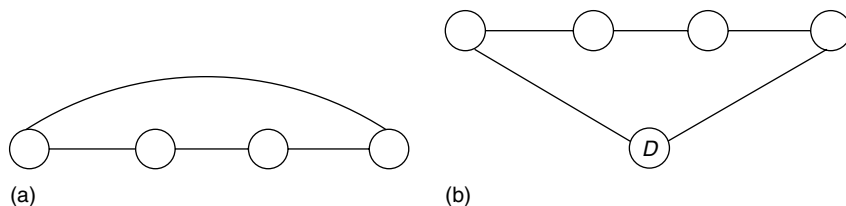
Figure 24.22 Two graphical models reflecting (a) associations between genotypes and (b) associations between genotypes and a disease indicator. Both are disallowed because they are not decomposable.

24.7.2   Gene Networks

Other recent and important applications of graphical models in genetics are concerned with inferring regulatory mechanisms involving genes, typically based on expression data measured using microarray technology. The graphical relationships involved in this type of model describe regulatory mechanisms, in principle, based on a causal interpretation of the relevant graphical models , such as those described in Section 24.6. The applications of graphical models in this context are generally exploratory and serve to exploit the massive data available to conjecture potential relationships which must subsequently be investigated further using other forms of biological subject-matter knowledge and experiments. One direction of research uses relatively sophisticated model specifications with some simplifying structures (Friedman, 2004; Segal *et al.*, 2003a; 2003b), exploiting the notion of probabilistic relational models (Friedman *et al.*, 1999) which add repeated relational structure to the conditional independence structure of graphical models.

Another approach uses less sophisticated models, essentially undirected Gaussian graphical models, exploiting the simplicity and detailed statistical understanding of these to make efficient and well-founded search procedures, either using Bayesian ideas (Jones *et al.*, 2005) or other graphical model selection algorithms (Spirtes *et al.*, 1993) to conjecture interesting relationships between genes under study. For an early example of this type of research, we refer to West *et al.* (2001). The methodology is currently developing rapidly, see, e.g. Schäfer and Strimmer (2005) and the corresponding software implementation as described in Schäfer *et al.* (2006). Markovetz (2006) provides an up-to-date online bibliography of the area.

REFERENCES

Adalsteinsson, S., Hersteinsson, P. and Gunnarsson, E. (1987). Fox colors in relation to colors in mice and sheep. *The Journal of Heredity* **78**, 235–237.

Andersen, S.K., Olesen, K.G., Jensen, F.V. and Jensen, F. (1989). HUGIN – a shell for building Bayesian belief universes for expert systems. *Proceedings of the 11th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers, San Mateo, CA, pp. 1080–1085.

Angrist, J., Imbens, G. and Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**(434), 444–455.

Baum, L.E. (1972). An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* **3**, 1–8.

Berry, A., Bordat, J.-P. and Cogis, O. (2000). Generating all the minimal separators of a graph. *International Journal of Foundations of Computer Science* **11**, 397–403.

Bouchitté, V. and Todinca, I. (2001). Treewidth and minimum fill-in: grouping the minimal separators. *SIAM Journal on Computing* **31**, 212–232.

Bowden, R. and Turkington, D. (1984). *Instrumental Variables*. Cambridge University Press, Cambridge.

Cannings, C., Thompson, E.A. and Skolnick, M.H. (1978). Probability functions on complex pedigrees. *Advances in Applied Probability* **10**, 26–61.

Cowell, R.G. (2003). A probabilistic expert system for forensic identification. *Forensic Science International* **134**, 196–206.

Cowell, R.G., Dawid, A.P., Lauritzen, S.L. and Spiegelhalter, D.J. (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York.

Cowell, R.G., Lauritzen, S.L. and Mortera, J. (2007). Identification and separation of DNA mixtures using peak area information. *Forensic Science International* **166**, 28–34.

Cowell, R.G. and Mostad, P. (2003). A clustering algorithm using DNA marker information for subpedigree reconstruction. *Journal of Forensic Sciences* **48**, 1239–1248.

Davey Smith, G. and Ebrahim, S. (2003). Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**, 1–22.

Davey Smith, G. and Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* **33**, 30–42.

Davey Smith, G., Ebrahim, S., Lewis, S., Hansell, A., Palmer, L. and Burton, P. (2005). Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* **366**, 1484–1498.

Dawid, A.P. (1992). Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing* **2**, 25–36.

Dawid, A.P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review* **70**, 161–189.

Dawid, A.P. (2003). Causal inference using influence diagrams: the problem of partial compliance. In *Highly Structured Stochastic Systems*, P.J. Green, N.L. Hjort and S. Richardson, eds. Oxford University Press, Oxford, pp. 45–81.

Dawid, A.P. and Mortera, J. (1996). Coherent analysis of forensic identification evidence. *Journal of the Royal Statistical Society, Series B* **58**, 425–443.

Dawid, A.P. and Mortera, J. (1998). Forensic identification with imperfect evidence. *Biometrika* **85**, 835–849.

Dawid, A.P., Mortera, J. and Pascali, V.L. (2001). Non-fatherhood or mutation? A probabilistic approach to parental exclusion in paternity testing. *Forensic Science International* **124**, 55–61.

Dawid, A.P., Mortera, J., Pascali, V.L. and van Boxel, D. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics* **29**, 577–595.

Dawid, A.P., Mortera, J. and Vicard, P. (2007). Object-oriented Bayesian networks for complex forensic DNA profiling problems. *Forensic Science International* **169**, 195–205.

Didelez, V. and Sheehan, N.A. (2007). Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical methods in Medical Research* (in press).

Egeland, T., Mostad, P.F., Mervåg, B. and Stenersen, M. (2000). Beyond traditional paternity and identification cases. Selecting the most probable pedigree. *Forensic Science International* **110**, 47–59.

Elston, R.C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523–542.

Falconer, D.S. and Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*, 4th edition. Longman Group Ltd.

Fishelson, M. and Geiger, D. (2002). Exact genetic linkage computations for general pedigrees. *Bioinformatics* **18**, S189–S198.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **303**(5659), 799–805.

Friedman, N., Getoor, L., Koller, D. and Pfeffer, A. (1999). Learning probabilistic relational models. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm, Sweden*. Morgan Kaufman, San Francisco, CA, pp. 1300–1307.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Goldberger, A. (1972). Structural equation methods in the social sciences. *Econometrica* **40**, 979–1001.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* **29**, 722–729.

Greenland, S., Robins, J.M. and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science* **14**, 29–46.

Hansen, B. and Pedersen, C.B. (1994). Analysing complex pedigrees using Gibbs sampling. A theoretical and empirical investigation. Technical Report R-94-2032, Institute for Electronic Systems, Aalborg University.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Heath, S.C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oliogenic models. *American Journal of Human Genetics* **61**, 748–760.

Heath, S.C. (2003). Genetic linkage analysis using Markov chain Monte Carlo techniques. In *Highly Structured Stochastic Systems*, P.J. Green, N.L. Hjort and S. Richardson, eds. Oxford University Press, Oxford.

Janss, L.L.G., Thompson, R. and Van Arendonk, J.A.M. (1995). Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. *Theoretical and Applied Genetics* **91**, 1137–1147.

Jensen, C.S. (1997). Blocking Gibbs sampling for inference in large and complex Bayesian networks with applications in genetics. Ph.D. thesis, Aalborg University, Aalborg.

Jensen, F. (2002). *HUGIN API Reference Manual Version 5.4*. HUGIN Expert Ltd., Aalborg.

Jensen, C.S. and Kong, A. (1999). Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops. *American Journal of Human Genetics* **65**, 885–901.

Jensen, F.V., Lauritzen, S.L. and Olesen, K.G. (1990). Bayesian updating in causal probabilistic networks by local computation. *Computational Statistics Quarterly* **4**, 269–282.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science* **20**, 388–400.

Katan, M.B. (2004). Commentary: Mendelian randomization, 18 years on. *International Journal of Epidemiology* **33**, 10–11.

Koller, D. and Pfeffer, A. (1997). Object-oriented Bayesian networks. In *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence*, D. Geiger and P. Shenoy, eds. Morgan Kaufmann Publishers, San Francisco, CA, pp. 302–313.

Kong, A. (1991). Efficient methods for computing linkage likelihoods of recessive diseases in inbred pedigrees. *Genetic Epidemiology* **8**, 81–103.

Lander, E.S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 2363–2367.

Lange, K. and Elston, R.C. (1975). Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. *Human Heredity* **25**, 95–105.

Lauritzen, S.L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Lauritzen, S.L. (2000). Causal inference from graphical models. In *Complex Stochastic Systems*, Chapter 2, O.E. Barndorff-Nielsen, D.R. Cox and C. Kluppelberg, eds. Chapman Hall, pp. 63–107.

Lauritzen, S.L. and Jensen, F.V. (1997). Local computation with valuations from a commutative semigroup. *Annals of Mathematics and Artificial Intelligence* **21**, 51–69.

Lauritzen, S.L. and Sheehan, N.A. (2003). Graphical models for genetic analysis. *Statistical Science* **18**, 489–514.

Lauritzen, S.L. and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B* **50**, 157–224.

Lund, M.S. and Jensen, C.S. (1999). Blocking Gibbs sampling in the mixed inheritance model using graph theory. *Genetics, Selection, Evolution* **31**, 3–24.

Markovetz, F. (2006). A bibliography on learning causal networks of gene interactions. Manuscript: `http://www.molgen.mpg.de/~markowet/docs/network-bib.pdf`.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Mortera, J., Dawid, A.P. and Lauritzen, S.L. (2003). Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology* **63**, 191–205.

Pearl, J. (1988). *Probabilistic Inference in Intelligent Systems*. Morgan Kaufmann Publishers, San Mateo, CA.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82**, 669–710.

Pearl, J. (2000). *Causality*. Cambridge University Press.

Robins, J. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling with Applications to Causality*, M. Berkane, ed. Springer-Verlag, New York, pp. 69–117.

Schäfer, J., Opgen-Rhein, R. and Strimmer, K. (2006). Reverse engineering genetic networks using the GeneNet package. *R News* **6**, 50–53.

Schäfer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21**(6), 754–764.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003a). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* **34**(2), 166–176.

Segal, E., Wang, H. and Koller, D. (2003b). Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* **19**(Suppl. 1), i264–i272.

Sheehan, N.A. (2000). On the application of Markov chain Monte Carlo methods to genetic analyses on complex pedigrees. *International Statistical Review* **68**, 83–110.

Sheehan, N. and Egeland, T. (2007). Structured incorporation of prior information in relationship estimation problems. *Annals of Human Genetics* (in press).

Sheehan, N.A., Guldbrandtsen, B., Lund, M.S. and Sorensen, D.A. (2002). Bayesian MCMC mapping of quantitative trait loci in a half-sib design: a graphical model perspective. *International Statistical Review* **70**, 241–267.

Shenoy, P.P. and Shafer, G. (1990). Axioms for probability and belief-function propagation. In *Uncertainty in Artificial Intelligence 4*, R.D. Shachter, T.S. Levitt, L.N. Kanal and J.F. Lemmer, eds. North-Holland, Amsterdam, pp. 169–198.

Shoiket, K. and Geiger, D. (1997). A practical algorithm for finding optimal triangulations. *Proceedings of the Fourteenth National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA, pp. 185–190.

Skjøth, F., Lohi, O. and Thomas, A.W. (1994). Genetic models for the inheritance of the silver mutation of foxes. *Genetical Research* **64**, 11–18.

Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics* **58**, 1323–1337.

Spiegelhalter, D.J. (1990). Fast algorithms for probabilistic reasoning in influence diagrams, with applications in genetics and expert systems. In *Influence Diagrams, Belief Nets and Decision Analysis*, R.M. Oliver and J.Q. Smith, eds. John Wiley & Sons, Chichester, pp. 361–384.

Spiegelhalter, D.J. (1998). Bayesian graphical modelling: a case-study in monitoring health outcomes. *Applied Statistics* **47**, 115–133.

Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, Prediction and Search*. Springer-Verlag, New York. Reprinted by MIT Press.

Steel, M. and Hein, J. (2006). Reconstructing pedigrees: a combinatorial perspective. *Journal of Theoretical Biology* **240**, 360–367.

Taroni, F., Aitken, C., Garbolino, P. and Biedermann, A. (2006). *Bayesian Networks and Probabilistic Inference in Forensic Science*. John Wiley & Sons, Chichester.

Thomas, A. (1985). Data structures, methods of approximation and optimal computation for pedigree analysis. Ph.D. thesis, Cambridge University.

Thomas, A. (2005). Characterizing alleleic associations from unphased diploid data by graphical modeling. *Genetic Epidemiology* **29**, 23–35.

Thomas, A. and Camp, N.J. (2004). Graphical modeling of the joint distribution of alleles at associated loci. *American Journal of Human Genetics* **74**, 1088–1101.

Thomas, D. and Conti, D. (2004). Commentary: the concept of "Mendelian randomization". *International Journal of Epidemiology* **33**, 21–25.

Thomas, A., Gutin, A., Abkevich, V. and Bansal, A. (2000). Multilocus linkage analysis by blocked Gibbs sampling. *Statistics and Computing* **10**, 259–269.

Thompson, E.A. (1981). Pedigree analysis of Hodgkin's disease in a newfoundland genealogy. *Annals of Human Genetics* **45**, 279–292.

Thompson, E.A. (1986). *Pedigree Analysis in Human Genetics*. The Johns Hopkins University Press, Baltimore, MD.

Thompson, E.A. (1994). Monte Carlo likelihood in genetic mapping. *Statistical Science* **9**(3), 355–366.

Thompson, E.A. (2000). *Statistical Inference from Genetic Data on Pedigrees*, *Volume 6 of NSF-CBMS regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and the American Statistical Association.

Thompson, E.A. (2001). Monte Carlo methods on genetic structures. In *Complex Stochastic Systems*, O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg, eds. Chapman Hall/CRC Press, London/Boca Raton, FL, pp. 176–218.

Thompson, E.A. and Heath, S.C. (2000). Estimation of conditional multilocus gene identity among relatives. In *Statistics in Molecular Biology and Genetics*, F., Seiller-Moiseiwitsch, ed. *IMS Lecture Notes*. Institute of Mathematical Statistics, American Mathematical Society, pp. 95–113.

Verzilli, C.J., Stallord, N. and Whittaker, J.C. (2006). Bayesian graphical models for genomewide association studies. *American Journal of Human Genetics* **79**, 100–112.

Weir, B.S., Triggs, C.M., Starling, L., Stowell, L.I., Walsh, K.A.J. and Buckleton, J.S. (1997). Interpreting DNA mixtures. *Journal of Forensic Sciences* **42**, 213–222.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.A., Marks, J.R. and Nevins, J.R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 11462–11467.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20**, 557–585.

Wright, S. (1923). The theory of path coefficients: a reply to Niles' criticism. *Genetics* **8**, 239–255.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics* **5**, 161–215.

Yannakakis, M. (1981). Computing the minimum fill-in is NP-complete. *SIAM Journal on Algebraic and Discrete Methods* **2**, 77–79.

25

Coalescent Theory

M. Nordborg

Molecular and Computational Biology, University of Southern California, Los Angeles, CA, USA

The coalescent process is a powerful modeling tool for population genetics. The allelic states of all homologous gene copies in a population are determined by the genealogical and mutational history of these copies. The coalescent approach is based on the realization that the genealogy is usually easier to model backward in time, and that selectively neutral mutations can then be superimposed afterwards. A wide range of biological phenomena can be modeled using this approach.

   Whereas almost all of classical population genetics considers the future of a population given a starting point, the coalescent considers the present, while taking the past into account. This allows the calculation of probabilities of sample configurations under the stationary distribution of various population genetic models, and makes full likelihood analysis of polymorphism data possible. It also leads to extremely efficient computer algorithms for generating simulated data from such distributions, data which can then be compared with observations as a form of exploratory data analysis.

25.1 INTRODUCTION

The stochastic process known as 'the coalescent' has played a central role in population genetics since the early 1980s, and results based on it are now used routinely to analyze DNA sequence polymorphism data. In spite of this, there is no comprehensive textbook treatment of coalescent theory. For biologists, the most widely used source of information is probably Hudson's seminal review (Hudson, 1990), which, along with a few other book chapters (Donnelly and Tavaré, 1995; Hudson, 1993; Li, 1997) and various unpublished lecture notes, is all that is available beyond the primary literature. Furthermore, since the field is very active, many relevant results are not generally available because they have not yet been published. They may be due to appear sometime in the indefinite future in a mathematical journal or obscure conference volume, or they may simply never have been written down. As a result of all this, there is a considerable gap between the theory that is available, and the theory that is being used to analyze data.

The present chapter is intended as an up-to-date introduction suitable for a wider audience. The focus is on the stochastic process itself, and especially on how it can be used to model a wide variety of biological phenomena. I consider a basic understanding of coalescent theory to be extremely valuable – even essential – for anyone analyzing genetic polymorphism data from populations, and will try to defend this view throughout. First of all, such an understanding can in many cases provide an intuitive feeling for how informative polymorphism data are likely to be (the answer is typically 'Not very'). When intuition is not enough, the coalescent provides a simple and powerful tool for exploratory data analysis through the generation of simulated data. The efficacy of the coalescent as a simulation tool is also the basis for promising statistical methods that use rejection algorithms to compute likelihoods (e.g., Weiss and von Haeseler, 1998). Various more sophisticated inference methods are described in **Chapter 26**.

25.2   THE COALESCENT

The word 'coalescent' is used in several ways in the literature, and it will also be used in several ways here. Hopefully, the meaning will be clear from the context. The coalescent – or, perhaps more appropriately, the coalescent approach – is based on two fundamental insights, which are the topic of Section 25.2.1. Section 25.2.2 then describes the stochastic process known as the coalescent, or sometimes Kingman's coalescent in honor of its discoverer (Kingman, 1982a; 1982b; 1982c). This process results from combining the two fundamental insights with a convenient limit approximation.

The coalescent will be introduced in the setting of the Wright–Fisher model of neutral evolution, but it applies more generally. This is one of the main topics for the remainder of the chapter. First of all, many different neutral models can be shown to converge to Kingman's coalescent. Second, more complex neutral models often converge to coalescent processes analogous to Kingman's coalescent.

The coalescent was described by Kingman (1982a; 1982b; 1982c), but it was also discovered independently by Hudson (1983) and by Tajima (1983). Indeed, arguments anticipating it had been used several times in population genetics (reviewed by Tavaré, 1984).

25.2.1   The Fundamental Insights

The first insight is that since selectively neutral variants by definition do not affect reproductive success, it is possible to separate the neutral mutation process from the genealogical process. In classical terms, 'state' can be separated from 'descent'.

To see how this works, consider a population of N clonal organisms that reproduce according to the neutral Wright–Fisher model, that is to say, generations are discrete, and each new generation is formed by randomly sampling N parents with replacement from the current generation. The number of offspring contributed by a particular individual is thus binomially distributed with parameters N (the number of trials) and $1/N$ (the probability of being chosen), and the joint distribution of the numbers of offspring produced by all N individuals is symmetrically multinomial. Now consider the random genealogical relationships (i.e. 'who begat whom') that result from reproduction in this setting. These can be represented graphically, as shown in Figure 25.1. Going forward in
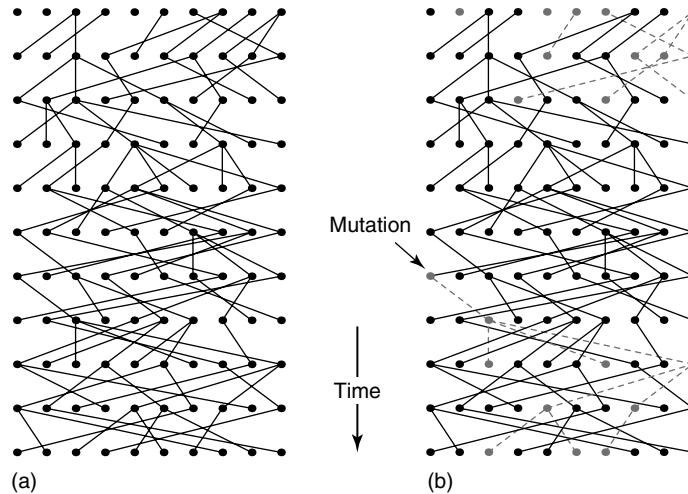
Figure 25.1 The neutral mutation process can be separated from the genealogical process. The genealogical relationships in a particular 10-generation realization of the neutral Wright–Fisher model (with population size $N = 10$) are shown on the left. On the right, allelic states of have been superimposed (so-called 'gene dropping').

time, lineages branch whenever an individual produces two or more offspring, and end when there is no offspring. Going backward in time, lineages coalesce whenever two or more individuals were produced by the same parent. They never end. If we trace the ancestry of a group of individuals back through time, the number of distinct lineages will decrease and eventually reach one, when the most recent common ancestor (MRCA) of the individuals in question is encountered. None of this is affected by neutral genetic differences between the individuals.

As a consequence, the evolutionary dynamics of neutral allelic variants can be modeled through so-called 'gene dropping' ('mutation dropping' would be more accurate): given a realization of the genealogical process, allelic states are assigned to the original generation in a suitable manner, and the lines of descent then simply followed forward in time, using the rule that offspring inherit the allelic state of their parent unless there is a mutation (which occurs with some probability each generation). In particular, the allelic states of any group of individuals (for instance, all the members of a given generation) can be generated by assigning an allelic state to their MRCA and then 'dropping' mutations along the branches of the genealogical tree that leads to them. Most of the genealogical history of the population is then irrelevant (cf. Figures 25.1 and 25.2).

The second insight is that it is possible to model the genealogy of a group of individuals backward in time without worrying about the rest of the population. It is a general consequence of the assumption of selective neutrality that each individual in a generation can be viewed as 'picking' its parent at random from the previous generation. It follows that the genealogy of a group of individuals may be generated by simply tracing the lineages back in time, generation by generation, keeping track of coalescences between lineages, until eventually the MRCA is found. It is particularly easy to see how this is done for the Wright–Fisher model, where individuals pick their parents independently of each other.
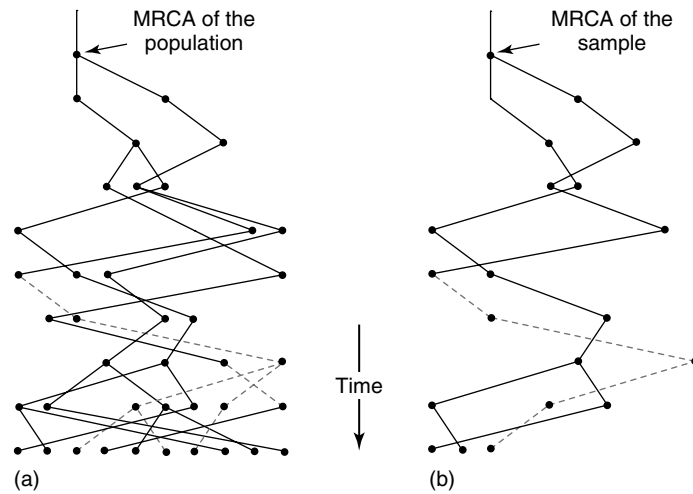
Figure 25.2 The genetic composition of a group of individuals is completely determined by the group's genealogy and the mutations that occur on it. The genealogy of the final generation in Figure 25.1 is shown on the left, and the genealogy of a sample from this generation is shown on the right. These trees could have been generated backward in time without generating the rest of Figure 25.1.

In summary, the joint effects of random reproduction (which causes 'genetic drift') and random neutral mutations in determining the genetic composition of a group of clonal individuals (such as a generation or a sample thereof) may be modeled by first generating the random genealogy of the individuals backward in time, and then superimposing mutations forward in time. This approach leads directly to extremely efficient computer algorithms (cf. the 'classical' approach which is to simulate the *entire*, usually very large population forward in time for a long period of time, and then to look at the final generation). It is also mathematically elegant, as Section 25.2.2 will show. However, its greatest value may be heuristic: the realization that the pattern of neutral variation observed in a population can be viewed as the result of random mutations on a random tree is a powerful one, which profoundly affects the way we think about data.

In particular, we are almost always interested in biological phenomena that affect the genealogical process, but do not affect the mutation process (e.g. population subdivision). From the point of view of inference about such phenomena, the observed polymorphisms are only of interest because they contain information about the unobserved underlying genealogy. Furthermore, the underlying genealogy is only of interest because it contains information about the evolutionary process that gave rise to it. In statistical terms, almost all inference problems that arise from polymorphism data can be seen as 'missing data' problems.

It is crucial to understand this, because no matter how many individuals we sample, there is still only a *single* underlying genealogy to estimate. It could of course be that this single genealogy contains a lot of information about the interesting aspect of the evolutionary process, but if it does not, then our inferences will be as good as one would normally expect from a sample of size one!

Another consequence of the above is that it is usually possible to understand how model parameters affect polymorphism data by understanding how they affect genealogies. For this reason, I will focus on the genealogical process and only discuss the neutral mutation process briefly toward the end of the chapter.

25.2.2   The Coalescent Approximation

The previous subsection described the conceptual insights behind the coalescent approach. The sample genealogies central to this approach can be conveniently modeled using a continuous-time Markov process known as the coalescent (or Kingman's coalescent, or sometimes 'the n-coalescent' to emphasize the dependence on the sample size). We will now describe the coalescent and show how it arises naturally as a large-population approximation to the Wright–Fisher model. Its relationship to other models will be discussed later.

Figure 25.2 is needlessly complicated because the identity (i.e. the horizontal position) of all ancestors is maintained. In order to superimpose mutations, all we need to know is which lineage coalesces with which, and when. In other words, we need to know the topology, and the branch lengths. The topology is easy to model: because of neutrality, individuals are equally likely to reproduce; therefore all lineages must be equally likely to coalesce. It is convenient to represent the topology as a sequence of coalescing equivalence classes: two members of the original sample are equivalent at a certain point in time if and only if they have a common ancestor at that time (see Figure 25.3). But what about the branch lengths, that is, the coalescence times?

Follow two lineages back in time. We have seen that offspring pick their parents randomly from the previous generation, and that, under the Wright–Fisher model, they do so independently of each other. Thus, the probability that the two lineages pick the same parent and coalesce is $1/N$, and the probability that they pick different parents and remain distinct is $1 - 1/N$. Since generations are independent, the probability that they remain distinct more than t generations into the past is $(1 - 1/N)^t$. The expected coalescence time is N generations. This suggests a standard continuous-time diffusion
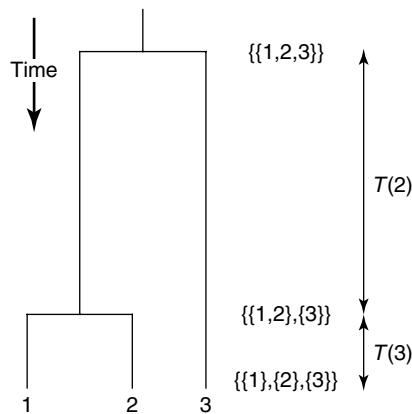
Figure 25.3   The genealogy of a sample can be described in terms of its topology and branch lengths. The topology can be represented using equivalence classes for ancestors. The branch lengths are given by the waiting times between successive coalescence events.

approximation, which is good as long as N is reasonably large (see **Chapter 22**). Rescale time so that one unit of scaled time corresponds to N generations. Then the probability that the two lineages remain distinct for more than τ units of scaled time is

$$\left(1 - \frac{1}{N}\right)^{[N\tau]} \longrightarrow e^{-\tau}, \tag{25.1}$$

as N goes to infinity ($[N\tau]$ is the largest integer less than or equal to $N\tau$). Thus, in the limit, the coalescence time for a pair of lineages is exponentially distributed with mean 1.

Now consider k lineages. The probability that none of them coalesce in the previous generation is

$$\prod_{i=0}^{k-1} \frac{N-i}{N} = \prod_{i=1}^{k-1}\left(1 - \frac{i}{N}\right) = 1 - \frac{\binom{k}{2}}{N} + O\left(\frac{1}{N^2}\right), \tag{25.2}$$

and the probability that more than two do so is $O(1/N^2)$. Let $T(k)$ be the (scaled) time till the first coalescence event, given that there are currently k lineages. By the same argument as above, $T(k)$ is in the limit exponentially distributed with mean $2/[(k(k-1))]$. Furthermore, the probability that more than two lineages coalesce in the same generation can be neglected. Thus, under the coalescent approximation, the number of distinct lineages in the ancestry of a sample of (finite) size n decreases in steps of one back in time, so $T(k)$ is the time from k to $k-1$ lineages (see Figure 25.3).

In summary, the coalescent models the genealogy of a sample of n haploid individuals as a random bifurcating tree, where the $n-1$ coalescence times $T(n), T(n-1), \ldots, T(2)$ are mutually independent, exponentially distributed random variables. Each pair of lineages coalesces independently at rate 1, so the total rate of coalescence when there are k lineages is 'k choose 2'. A concise (and rather abstract) way of describing the coalescent is as a continuous-time Markov process with state space $\mathcal{E}_n$ given by the set of all equivalence relations on $\{1, \ldots, n\}$, and infinitesimal generator $Q = (q_{\xi\eta})_{\xi,\eta\in\mathcal{E}_n}$ given by

$$q_{\xi\eta} := \begin{cases} -k(k-1)/2, & \text{if } \xi = \eta, \\ 1, & \text{if } \xi \prec \eta, \\ 0, & \text{otherwise,} \end{cases} \tag{25.3}$$

where $k := |\xi|$ is the number of equivalence classes in ξ, and $\xi \prec \eta$ if and only if η is obtained from ξ by coalescing two equivalence classes of ξ.

It is worth emphasizing just how efficient the coalescent is as a simulation tool. In order to generate a sample genealogy under the Wright–Fisher model as described in Section 25.2.1, we would have to go back in time some N generations, checking for coalescences in each of them. Under the coalescent approximation, we simply generate $n-1$ independent exponential random numbers and, independently of these, a random bifurcating topology.

What do typical coalescence trees look like? Figure 25.4 shows four examples. It is clear that the trees are extremely variable, both with respect to topology and branch lengths. This should come as no surprise considering the description of the coalescent just given: the topology is independent of the branch lengths; the branch lengths are

independent, exponential random variables; and the topology is generated by randomly picking lineages to coalesce (in this sense all topologies are equally likely).

Note that the trees tend to be dominated by the deep branches, when there are few ancestors left. Because lineages coalesce at rate 'k choose 2', coalescence events occur much more rapidly when there are many lineages (intuitively speaking, it is easier for lineages to find each other then). Indeed, the expected time to the MRCA (the height of the tree) is

$$E\left[\sum_{k=2}^{n} T(k)\right] = \sum_{k=2}^{n} E[T(k)] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\left(1 - \frac{1}{n}\right), \tag{25.4}$$

while $E[T(2)] = 1$, so the expected time during which there are only two branches is greater than half the expected total tree height. Furthermore, the variability in $T(2)$ accounts for most of the variability in tree height. The dependence on the deep branches becomes increasingly apparent as n increases, as can be seen by comparing Figures 25.4 and 25.5.

The importance of realizing that there is only a single underlying genealogy was emphasized earlier. As a consequence of the single genealogy, sampled gene copies from a population must almost always be treated as dependent, and increasing the sample size is therefore often surprisingly ineffective (the point is well made by Donnelly, 1996). Important examples of this follow directly from the basic properties of the coalescent. Consider first the MRCA of the population. One might think that a large sample is needed to ensure that the deepest split is included, but it can be shown (this and related results can be found in Saunders *et al.*, 1984) that the probability that a sample of size n contains the MRCA of the whole population is $(n - 1)/(n + 1)$. Thus even a small sample is likely to
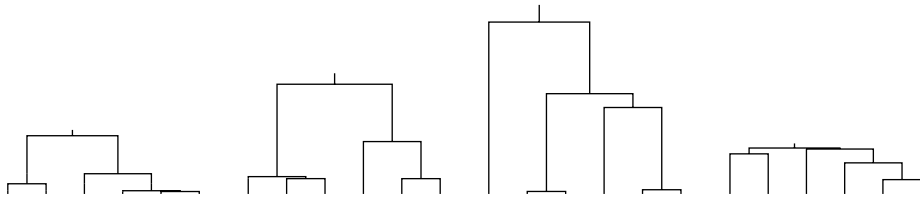
Figure 25.4 Four realizations of the coalescent for $n = 6$, drawn on the same scale (the labels 1–6 should be assigned randomly to the tips).
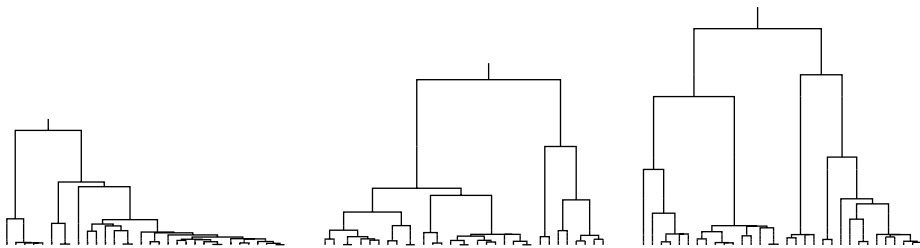
Figure 25.5 Three realizations of the coalescent for $n = 32$, drawn on the same scale (the labels 1–32 should be assigned randomly to the tips).

contain it and the total tree height will quickly stop growing as n increases. Second, the number of distinct lineages decreases rapidly as we go back in time. This severely limits inferences about ancient demography (e.g. Nordborg, 1998). Third, since increasing the sample size only adds short twigs to the tree (cf. Figure 25.5), the expected total branch length of the tree, $T_{tot}(n)$ grows very slowly with n. We have

$$\mathrm{E}[T_{tot}(n)] = \mathrm{E}\left[\sum_{k=2}^{n} kT(k)\right] = \sum_{k=1}^{n-1}\frac{2}{k} \sim 2(\gamma + \log n), \qquad (25.5)$$

as $n \to \infty$ ($\gamma \approx 0.577\,216$ is Euler's constant). Since the number of mutations that are expected to occur in a tree is proportional to $\mathrm{E}[T_{tot}(n)]$, this has important consequences for estimating the mutation rate, as well as for inferences that depend on estimates of the mutation rate. Loosely speaking, it turns out that a sample of n copies of a gene often has the statistical properties one would expect of a random sample of size $\log n$, or even of size 1 (which is not much worse than $\log n$ in practice).

25.3   GENERALIZING THE COALESCENT

This section will present ideas and concepts that are important for generalizing the coalescent. The following sections will then illustrate how these can be used to incorporate greater biological realism.

25.3.1   Robustness and Scaling

We have seen that the coalescent arises naturally as an approximation to the Wright–Fisher model, and that it has convenient mathematical properties. However, the real importance of the coalescent stems from the fact that it arises as a limiting process for a wide range of neutral models, *provided time is scaled appropriately* (Kingman, 1982b; 1982c; Möhle, 1998b; 1999). It is thus robust in this sense.

This is best explained through an example. Recall that the number of offspring contributed by each individual in the Wright–Fisher model is binomially distributed with parameters N and $1/N$. The mean is thus 1, and the variance is $1 - 1/N \to 1$, as $N \to \infty$. Now consider a generalized version of this model in which the mean number of offspring is still 1 (as it must be for the population size to remain constant), but the limiting variance is σ^2, $0 < \sigma^2 < \infty$ (perhaps giants step on 90 % of the individuals before they reach reproductive age). It can be shown that this process also converges to the coalescent, provided time is measured in units of N/σ^2 generations. We could also measure time in units of N generations as before, but then $\mathrm{E}[T(2)] = 1/\sigma^2$ instead of $\mathrm{E}[T(2)] = 1$, and so on. Either way, the expected coalescence time for a pair of lineages is N/σ^2 generations. The intuition behind this is clear: increased variance in reproductive success causes coalescence to occur faster (at a higher rate). In classical terms, 'genetic drift' operates faster. By changing the way we measure time, this can be taken into account, and the standard coalescent process obtained.

The remarkable fact is that a very wide range of biological phenomena (overlapping generations, separate sexes, mating systems – several examples will be given below) can

likewise be treated as a simple linear change in the time scale of the coalescent. This has important implications for data analysis. The good news is that we may often be able to justify using the coalescent process even though 'our' species almost certainly does not reproduce according to a Wright–Fisher model (few species do). The bad news is that biological phenomena that can be modeled this way will never be amenable to inference based on polymorphism data alone. For example, σ^2 in the model above could never be estimated from polymorphism data unless we had independent information about N (and vice versa).

Of course, we could not even estimate N/σ^2 without external data. It is important to realize that all parameters in coalescent models are scaled, and that only scaled parameters can be directly estimated from the data. In order to make any kind of statement about unscaled quantities, such as population numbers, or ages in years or generations, external information is needed. This adds considerable uncertainty to the analysis. For example, an often used source of external information is an estimate of the neutral mutation probability per generation. Roughly speaking, this estimate is obtained by measuring sequence divergence between species, and dividing by the estimated species divergence time (Li, 1997). The latter is in turn obtained from the fossil record and a rough guess of the generation length. It should be clear that it is not appropriate to treat such an estimate as a known parameter when analyzing polymorphism data. However, it should also be noted that interesting conclusions can often be drawn directly from scaled parameters (e.g. by looking at relative values). Such analyses are likely to be more robust, given the robustness of the coalescent.

Because the generalized model above converges with the same scaling as a Wright–Fisher model with a population size of N/σ^2, it is sometimes said that it has an 'effective population size', $N_e = N/\sigma^2$. Models that scale differently would then have other effective population sizes. Although convenient, this terminology is unfortunate for at least two reasons. First, the classical population genetics literature is full of variously defined 'effective population sizes', only some of which are effective population sizes in the sense used here. For example, populations that are subdivided or vary in size cannot in general be modeled as a linear change in the time scale of the coalescent. Second, the term is inevitably associated with real population sizes, even though it is simply a scaling factor. To be sure, N_e is always a function of the real demographic parameters, but there is no direct relationship with the total population size (which may be smaller as well as much, much larger). Indeed, as we shall see in Section 25.7, it is now clear that N_e must vary between chromosomal regions in the same organism!

25.3.2 Variable Population Size

Real populations vary in size over time. Although the coalescent is not robust to variation in the population size in the sense described above (i.e. there is no 'effective population size'), it is nonetheless easy to incorporate changes in the population size, at least if we are willing to assume that we know what they were – that is, if we assume that the variation can be treated deterministically. Since a rigorous treatment of these results can be found in the review by Donnelly and Tavaré (1995), also in **Chapter 22** this volume, I will try to give an intuitive explanation.

Imagine a population that evolves according to the Wright–Fisher model, but with a different population size in each generation. If we know how the size has changed over time, we can trace the genealogy of a sample precisely as before. Let $N(t)$ be the population size t generations ago. Going back in time, lineages are more likely to coalesce

in generations when the population is small than in generations when the population is large. In order to describe the genealogy by a continuous-time process analogous to the coalescent, we must therefore allow the rate of coalescence to change over time. However, since the time scale used in the coalescent directly reflects the rate of coalescence, we may instead let this scaling change over time. In the standard coalescent, t generations ago corresponds to t/N units of coalescence time, and τ units of coalescence time ago corresponds to $[N\tau]$ generations. When the population size is changing, we find instead that t generations ago corresponds to

$$g(t) := \sum_{i=1}^{t} \frac{1}{N(i)} \tag{25.6}$$

units of coalescence time, and τ units of coalescence time ago corresponds to $[g^{-1}(\tau)]$ generations (g^{-1} denotes the inverse function of g). It is clear from equation (25.6) that many generations go by without much coalescence time passing when the population size is large, and, conversely, that much coalescence time passes each generation when the population is small. Let $N(0)$ go to infinity, and assume that $N(t)/N(0)$ converges to a finite number for each t, to ensure that the population size becomes large in every generation. It can be shown that the variable population size model converges to a coalescent process with a *nonlinear* time scale in this limit (Griffiths and Tavaré, 1994). The scaling is given by (25.6). Thus, a sample genealogy from the coalescent with variable population size can be generated by simply applying g^{-1} to the coalescence times of a genealogy generated under the standard coalescent.

An example will make this clearer. Consider a population that has grown exponentially, so that, backwards in time, it shrinks according to $N(t) = N(0)e^{-\beta t}$ (note that this violates the assumption that the population size be large in every generation – this turns out not to matter greatly). Then

$$g(t) \approx \int_0^t \frac{1}{N(s)} \, ds = \frac{e^{\beta t} - 1}{N(0)\beta}, \tag{25.7}$$

and

$$g^{-1}(\tau) \approx \frac{\log(1 + N(0)\beta\tau)}{\beta}. \tag{25.8}$$

The difference between this model and one with a constant population size is shown in Figure 25.6. When the population size is constant, there is a linear relationship between real and scaled time. The genealogical trees will tend to look like those in Figures 25.4 and 25.5. When the population size is changing, the relationship between real and scaled time is nonlinear, because coalescences occur very slowly when the population was large, and more rapidly when the population was small. Genealogies in an exponentially growing population will tend to have most coalescences early in the history. Since all branches will then be of roughly equal length, the genealogy is said to be 'starlike'.

Models of exponential population growth have often been used in the context of human evolution (e.g. Rogers and Harpending, 1992; Slatkin and Hudson 1991). Marjoram and Donnelly (1997) have pointed out that some of the predictions from such models (e.g. the starlike genealogies) depend crucially on exponential growth from a *very* small
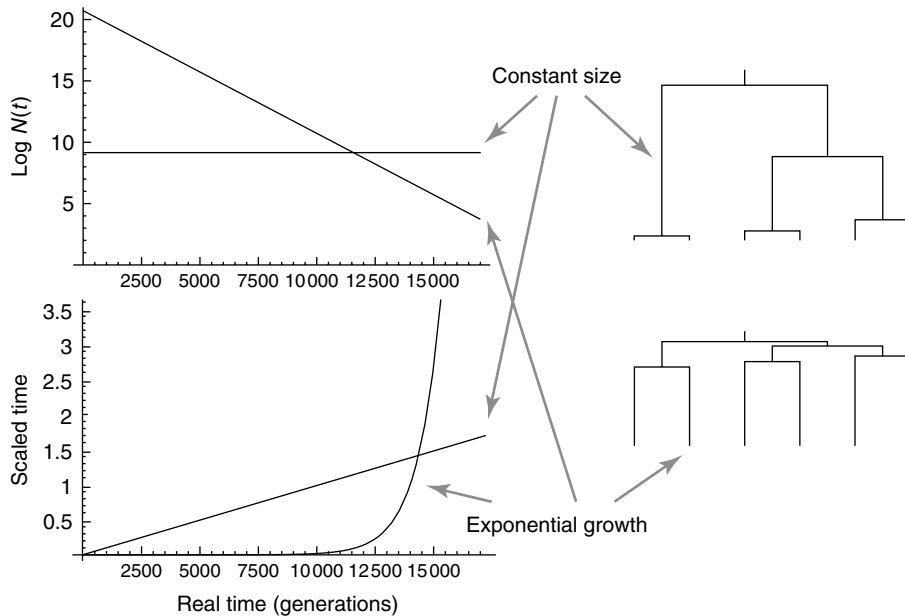
Figure 25.6 Variable population size can be modeled as a standard coalescent with a nonlinear time scale. Here, a constant population is compared to one that has grown exponentially. As the latter population shrinks backward in time, the scaled time begins to run faster, reflecting the fact that coalescences are more likely to have taken place when the population was small. Note that the trees are topologically equivalent and differ only in the branch lengths.

size – unrealistically small for humans. However, other predictions are more robust. For example, the argument in the previous paragraph explains why it may be reasonable to ignore growth altogether when modeling human evolution, even though growth has clearly taken place: if the growth was rapid and recent enough, no scaled time would pass, and no coalescence occur. In classical terms, exponential growth stops genetic drift.

Finally, it should be pointed that it is not entirely clear how general the nonlinear scaling approach to variable population sizes is. It relies, of course, on knowing the historical population sizes, but it also requires assumptions about the type of density regulation (Marjoram and Donnelly, 1997).

25.3.3 Population Structure on Different Time Scales

Real populations are also often spatially structured, and it is obviously important to be able to incorporate this in our models. However, structured models turn out to be even more important than one might have expected from this, because many biological phenomena can be thought of as analogous to population structure (Nordborg, 1997; Rousset, 1999a). Examples range from the obvious, like age structure, to the more abstract, like diploidy and allelic classes.

The following model, which may be called the 'structured Wright–Fisher model', turns out to be very useful in this context. Consider a clonal population of size N, as before, but let it be subdivided into patches of fixed sizes N_i, $i \in \{1, \ldots, M\}$, so that $\sum_i N_i = N$. In

every generation, each individual produces an effectively infinite number of propagules. These propagules then migrate among the patches independently of each other, so that with probability m_{ij}, $i, j \in \{1, \ldots, M\}$, a propagule produced in patch i ends up in patch j. We also define the 'backward migration' probability, b_{ij}, $i, j \in \{1, \ldots, M\}$, that a randomly chosen propagule in patch i after dispersal was produced in patch j; it is easy to show that

$$b_{ij} = \frac{N_j m_{ji}}{\sum_k N_k m_{ki}}. \tag{25.9}$$

The next generation of adults in each patch is then formed by random sampling from the available propagules.

Thus the number of offspring a particular individual in patch i contributes to the next generation in patch j is binomially distributed with parameters N_j and $b_{ji}N_i^{-1}$. The joint distribution of the numbers of offspring contributed to the next generation in patch j by *all* individuals in the current generation is multinomial (but no longer symmetric).

Just like the unstructured Wright–Fisher model, the genealogy of a finite sample in this model can be described by a discrete-time Markov process. Lineages coalesce in the previous generation if and only if they pick the same parental patch, and the same parental individual within that patch. A lineage currently in i and a lineage currently in j 'migrate' (backward in time) to k and coalesce there with probability $b_{ik}b_{jk}N_k^{-1}$.

It is also possible to approximate the model by a continuous-time Markov process. The general idea is to let the total population size, N, go to infinity with time scaled appropriately, precisely as before. However, we now also need to decide how M, N_i, and b_{ij} scale with N. Different biological scenarios lead to very different choices in this respect, and it is often possible to utilize convergence results based on separation of time scales (Möhle, 1998a; Kaj, *et al.*, 1991; Nordborg, 1997; 1999; Nordborg and Donnelly, 1997; Nordborg and Krone, 2002; Wakeley, 1999). This technique will be exemplified in what follows.

25.4   GEOGRAPHICAL STRUCTURE

Genealogical models of population structure have a long history. The classical work on identity coefficients (see **Chapter 28**) concerns genealogies when $n = 2$, and the coalescent was also quickly used for this purpose (for early work see Slatkin, 1987; Strobeck, 1987; Tajima, 1989a; Takahata, 1988).

Since geographical structure is reviewed in **Chapter 28**, we will mainly use it to introduce some of the scaling ideas that are central to the coalescent. The discussion will be limited to the structured Wright–Fisher model (which is a matrix migration model when viewed as a model of geographic subdivision). Most coalescent modeling has been done in this setting (reviewed in Wilkinson-Herbots, 1998 and Hudson, 1998). For time-scale approximations different from the ones discussed below, see Takahata (1991) and Wakeley (1999). An important variant of the model considers isolation: gene flow which

stopped completely at some point in the past, for example due to speciation (e.g. Wakeley, 1996). For models of continuous environments and isolation-by-distance, see Barton and Wilson (1995) and Wilkins and Wakeley (2002).

25.4.1 The Structured Coalescent

Assume that M, $c_i := N_i/N$, and $B_{ij} := 2Nb_{ij}$, $i \neq j$, all remain constant as N goes to infinity. Then, with time measured in units of N generations, the process converges to the so-called 'structured coalescent', in which each pair of lineages in patch i coalesces independently at rate $1/c_i$, and each lineage in i 'migrates' (backward in time) independently to j at rate $B_{ij}/2$ (Herbots, 1994; Notohara, 1990; Wilkinson-Herbots, 1998). The intuition behind this is as follows (an excellent discussion of how the scaled parameters should be interpreted can be found in **Chapter 22**). By assuming that B_{ij} remains constant, we assure that the backward per-generation probabilities of *leaving* a patch (b_{ij}, $i \neq j$), are $O(1/N)$. Similarly, by assuming that c_i remains constant, we assure that all per-generation coalescence probabilities are $O(1/N)$. Thus, in any given generation, the probability that all lineages remain in their patch, without coalescing, is $1 - O(1/N)$. Furthermore, the probabilities that more than two lineages coalesce, that more than one lineage migrates, and that lineages both migrate and coalesce, are all $O(1/N^2)$ or smaller. In the limit $N \to \infty$, the only possible events are pairwise coalescences within patches, and single migrations between patches.

These events occur according to independent Poisson processes, which means the following. Let k_i denote the number of lineages currently in patch i. Then the waiting time till the first event is exponentially distributed with rate given by the sum of the rates of all possible events, that is,

$$ h(k_1, \ldots, k_M) = \sum_i \left(\frac{\binom{k_i}{2}}{c_i} + \sum_{j \neq i} k_i \frac{B_{ij}}{2} \right). \tag{25.10} $$

When an event occurs, it is a coalescence in patch i with probability

$$ \frac{\binom{k_i}{2}/c_i}{h(k_1, \ldots, k_M)}, \tag{25.11} $$

and a migration from i to j with probability

$$ \frac{k_i B_{ij}/2}{h(k_1, \ldots, k_M)}. \tag{25.12} $$

In the former case, a random pair of lineages in i coalesces, and k_i decreases by one. In the latter case, a random lineage moves from i to j, k_i decreases by one, and k_j increases by one. A simulation algorithm would stop when the MRCA is found, but note that this single remaining lineage would continue migrating between patches if followed further back in time.

Structured coalescent trees generally look different from standard coalescent trees. Whereas variable population size only altered the branch lengths of the trees, population structure also affects the topology. If migration rates are low, lineages sampled from the

same patch will tend to coalesce with each other, and a substantial amount of time can then pass before migration allows the ancestral lineages to coalesce (see Figure 25.7). Structure will often increase the mean and, equally importantly, the variance in time to the MRCA considerably (discussed in the context of human evolution by Marjoram and Donnelly, 1997).

25.4.2 The Strong-migration Limit

It is intuitive that weak migration, which corresponds to strong population subdivision, can have a large effect on genealogies. Conversely, we would expect genealogies in models with strong migration to look much like standard coalescent trees. This intuition turns out to be correct, except for one important difference: the scaling changes. Strong migration is thus one of the phenomena that can be modeled as a simple linear change in the time scale of the coalescent. It is important to understand why this happens.

Formally, the strong-migration limit means that $\lim_{N \to \infty} Nb_{ij} = \infty$ because the per-generation migration probabilities, b_{ij}, are not $O(1/N)$. Since the coalescence probabilities *are* $O(1/N)$, this means that, for large N, migration will be much more likely than coalescence. As $N \to \infty$, there will in effect be infinitely many migration events between coalescence events. This is known as separation of time scales: migration occurs on a faster time scale than does coalescence. However, coalescences can of course still only occur when two lineages pick a parent in the same patch. How often does this happen? Because lineages jump between patches infinitely fast on the coalescence time scale, this is determined by the stationary distribution of the migration process (strictly speaking, this assumes that the migration matrix is ergodic). Let π_i be the stationary probability that a lineage is in patch i. A given pair of lineages then co-occur in i a fraction π_i^2 of the time. Coalescence in this patch occurs at rate $1/c_i$. Thus the total rate at which pairs of lineages coalesce is $\alpha := \sum_i \pi_i^2/c_i$. Pairs coalesce independently of each other just as in the standard model, so the total rate when there are k lineages is $\binom{k}{2}\alpha$. If time is measured in units of $N_e = N/\alpha$ generations, the standard coalescent is retrieved (Nagylaki, 1980; Notohara, 1993).

It can be shown that $\alpha \geq 1$, with equality if and only if $\sum_{j \neq i} N_i b_{ij} = \sum_{j \neq i} N_j b_{ji}$ for all i. This condition means that, going forward in time, the number of emigrants equals the number of immigrants in all populations, a condition known as 'conservative
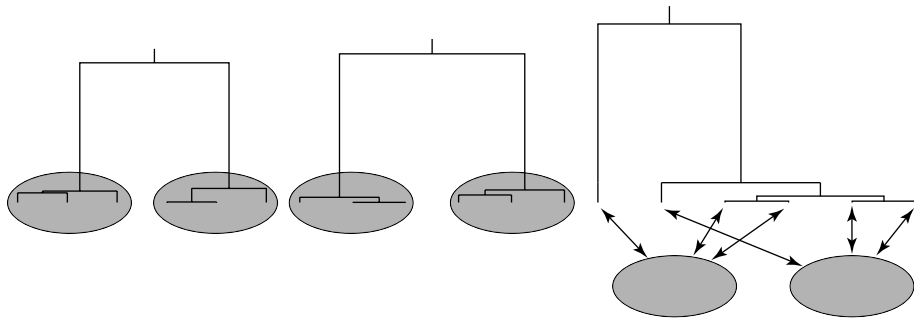
Figure 25.7 Three realizations of the structured coalescent in a symmetric model with two patches, and $n = 3$ in each patch (labels should be assigned randomly within patches). Lineages tend to coalesce within patches – but not always, as shown by the rightmost tree.

migration' (Nagylaki, 1980). Thus we see that, unless migration is conservative, the effective population size with strong migration is smaller than the total population size. The intuitive reason for this is that when migration is nonconservative, some individuals occupy 'better' patches than others, and this increases the variance in reproductive success among individuals. The environment has 'sources' and 'sinks' (Pulliam, 1988; Rousset, 1999b). Conservative migration models (like Wright's island model) have many simple properties that do not hold generally (Nagylaki, 1982; 1998; Nordborg, 1997).

25.5   SEGREGATION

Because everything so far has been done in an asexual setting, it has not been necessary to distinguish between the genealogy of an organism and that of its genome. This becomes necessary in sexual organisms. Most obviously, a diploid organism that was produced sexually has two parents, and each chromosome came from one of them. The genealogy of the genes is thus different from the genealogy (the pedigree) of the individuals: the latter describes the *possible* routes the genes could have taken (and is largely irrelevant – cf. Figure 25.9). This is simply Mendelian segregation viewed backwards in time, and it is the topic of this section. It is usually said that diploidy can be taken into account by simply changing the scaling from N to $2N$; it will become clear from what follows why, and in what sense, this is true.

The other facet of sexual reproduction, genetic recombination, turns out to have much more important effects. Genetic recombination causes ancestral lineages to branch, so that the genealogy of a sample can no longer be represented by a single tree: instead it becomes a collection of trees, or a single, more general type of graph. We will continue to ignore recombination until Section 25.6 (it makes sense to discuss diploidy first).

Sex takes many forms: I will first consider organisms that are hermaphroditic and therefore potentially capable of fertilizing themselves (this includes most higher plants and many mollusks), and thereafter discuss organisms with separate sexes (which includes most animals and many plants). Further examples can be found in Nordborg and Krone (2002).

25.5.1   Hermaphrodites

The key to modeling diploid populations is the realization that a diploid population of size N can be thought of as a haploid population of size $2N$, divided into N patches of size 2. In the notation of the structured Wright–Fisher model above, $M = N$, $N_i = 2$, and $c_i = 2/N$. Thus, in contrast to the assumptions for the structured coalescent, both M and c_i depend on N. This leads to a convenient convergence result based on separation of time scales (Nordborg and Donnelly, 1997; for a formal proof, see Möhle, 1998a), that can be described as follows (cf. Figure 25.8).

If time is scaled in units of $2N$ generations, then each pair of lineages 'coalesces' into the same individual at rate 2. Whenever this happens, there are two possibilities: either the two lineages pick the same of the two available (haploid) parents, or they pick different ones. The former event, which occurs with probability $\frac{1}{2}$, results in a real coalescence, whereas the latter event, which also occurs with probability $\frac{1}{2}$, simply results in the two distinct lineages temporarily occupying the same individual. Let S be the probability
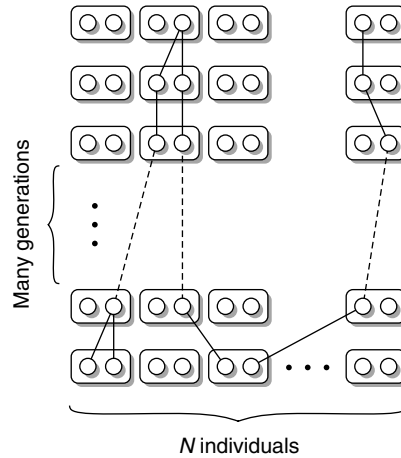
Figure 25.8 The coalescent with selfing. On the coalescent time scale, lineages within individuals instantaneously coalesce (probability F), or end up in different individuals (probability $1 - F$).

that a fertilization occurs through selfing, and $1 - S$ the probability that it occurs through outcrossing. If the individual harboring two distinct lineages was produced through selfing (probability S), then the two lineages must have come from the same individual in the previous generation, and again pick different parents with probability $\frac{1}{2}$ or coalesce with probability $\frac{1}{2}$. If the individual was produced through outcrossing, the two lineages revert to occupying distinct individuals. Thus the two lineages will rapidly either coalesce or end up in different individuals. The probability of the former outcome is

$$\frac{S/2}{S/2 + 1 - S} = \frac{S}{2 - S} =: F, \tag{25.13}$$

and that of the latter $1 - F$. Thus each time a pair of lineages coalesces into the same individual, the total probability that this results in a coalescence event is $\frac{1}{2} \times 1 + \frac{1}{2} \times F = (1 + F)/2$, and since pairs of lineages coalesce into the same individual at rate 2, the rate of coalescence is $1 + F$. On the chosen time scale, all states that involve two or more pairs occupying the same individual are instantaneous.

Thus, the genealogy of a random sample of gene copies from a population of hermaphrodites can be described by the standard coalescent if time is scaled in units of

$$2N_e = \frac{2N}{1 + F} \tag{25.14}$$

generations (cf. Pollak, 1987). If individuals are obligate outcrossers, $F = 0$, and the correct scaling is $2N$.

It should be noted that a sample from a diploid population is not a random sample of gene copies, because both copies in each individual are sampled. This is easily taken into account. It follows from the above that the two copies sampled from the same individual will instantaneously coalesce with probability F, and end up in different individuals with probability $1 - F$. The number of distinct lineages in a sample of $2n$ gene copies from n

individuals is thus $2n - X$, where X is as a binomially distributed random variable with parameters n and F. This corresponds to the well-known increase in the frequency of homozygous individuals predicted by classical population genetics. Note that this initial 'instantaneous' process has much nicer statistical properties than the coalescent, and that most of the information about the degree of selfing comes from the distribution of variability within and between individuals (Nordborg and Donnelly, 1997).

25.5.2 Males and Females

Next consider a diploid population that consists of N_m breeding males and N_f breeding females so that $N = N_m + N_f$. The discussion will be limited to *autosomal* genes, that is, genes that are not sex-linked. With respect to the genealogy of such genes, the total population can be thought of as a haploid population of size $2N$, divided into two patches of size $2N_m$ and $2N_f$, respectively, each of which is further divided into patches of size 2, as in the previous section. Clearly, a lineage currently in a male has probability $\frac{1}{2}$ of coming from a male in the previous generation, and probability $\frac{1}{2}$ of coming from a female. Within a sex, all individuals are equally likely to be chosen. The model looks a lot like a structured Wright–Fisher model with $M = 2$, $c_m = N_m/N$, $c_f = N_f/N$, and $b_{mf} = b_{fm} = \frac{1}{2}$, the only difference being that two distinct lineages in the same individual must have come from individuals of different sexes in the previous generation, and thus do not migrate independently of each other. However, because states involving two distinct lineages in the same individual are instantaneous, this difference can be shown to be irrelevant. Pairs of lineages in different individuals (regardless of sex) coalesce in the previous generation if and only if both members of the pair came from: (a) the same sex; (b) the same diploid individual within that sex; and (c) the same haploid parent within that individual. This occurs with probability

$$\frac{1}{4} \times \frac{1}{N_m} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{N_f} \times \frac{1}{2} = \frac{N_m + N_f}{8N_m N_f}, \qquad (25.15)$$

or, in the limit $N \to \infty$, with time measured in units of $2N$ generations, and c_m and c_f held constant, at rate $\alpha = (4c_m c_f)^{-1}$ (in accordance with the strong-migration limit result above). Alternatively, if time is measured in units of

$$2N_e = \frac{2N}{\alpha} = \frac{8N_m N_f}{N_m + N_f}, \qquad (25.16)$$

generations, the standard coalescent is obtained (cf. Wright, 1931). Note that if $N_m = N_f = N/2$, the correct scaling is again the standard one of $2N$.

25.6 RECOMBINATION

In the era of genomic polymorphism data, the importance of modeling recombination can hardly be overemphasized. When viewed backward in time, recombination (in the broad sense that includes phenomena such as gene conversion and bacterial conjugation in
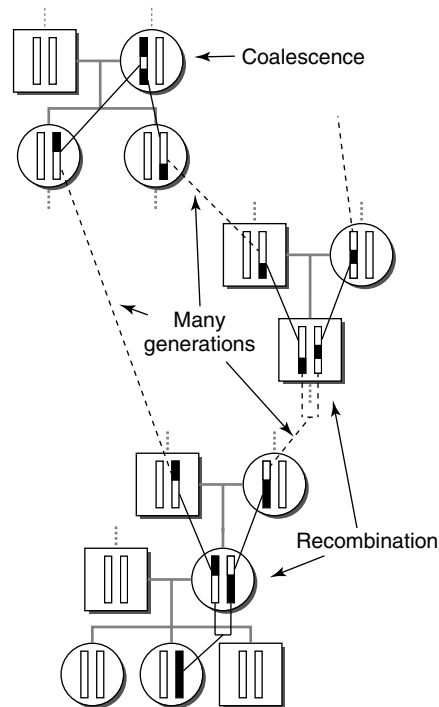
Figure 25.9 The genealogy of a DNA segment (colored black) subject to recombination both branches and coalesces. Note also that the genealogy of the sexually produced *individuals* (the pedigree) is very different from the genealogy of their *genes*.

addition to crossing over) causes the ancestry of a chromosome to spread out over many chromosomes in many individuals. The lineages branch, as illustrated in Figure 25.9. The genealogy of a sample of recombining DNA sequences can thus no longer be represented by a single tree: it becomes a graph instead. Alternatively, since the genealogy of each point in the genome (each base pair, say) *can* be represented by a tree, the genealogy of a sample of sequences may be envisioned as a 'walk through tree space'.

25.6.1 The Ancestral Recombination Graph

As was first shown by Hudson (1983), incorporating recombination into the coalescent framework is in principle straightforward. The following description is based on the elegant 'ancestral recombination graph' of Griffiths and Marjoram (1996; 1997), which is closely related to Hudson's original formulation (for other approaches, see Simonsen and Churchill, 1997; Wiuf and Hein, 1999b).

Consider first the ancestry of a single ($n = 1$) chromosomal segment from a diploid species with two sexes and an even sex ratio. As shown in Figure 25.9, each recombination event (depicted here as crossing over at a point – we will return to whether this is reasonable below) in its ancestry means that a lineage splits into two, when going backward in time. Recombination spreads the ancestry of the segment over many chromosomes, or rather over many 'chromosomal lineages'. However, as also shown in

Figure 25.9, these lineages will coalesce in the normal fashion, and this will tend to bring the ancestral material back together on the same chromosome (Wiuf and Hein, 1997).

To model this, let the per-generation probability of recombination in the segment be r, define $\rho := \lim_{N \to \infty} 4Nr$, and measure time in units of $2N$ generations. Then the (scaled) time till the first recombination event is exponentially distributed with rate $\rho/2$ in the limit as N goes to infinity. Furthermore, once recombination has created two or more lineages, we find that these lineages undergo recombination independently of one another, and that simultaneous events can be neglected. This follows from standard coalescent arguments analogous to those presented for migration above. The only thing that may be slightly nonintuitive about recombination is that *the lineages we follow never recombine with each other* (the probability of such an event is vanishingly small): they always recombine with the (infinitely many) nonancestral chromosomes.

Each recombination event increases the number of lineages by one, and because lineages recombine independently, the total rate of recombination when there are k lineages is $k\rho/2$. Each coalescence event decreases the number of lineages by one, and the total rate of coalescence when there are k lineages is $k(k-1)/2$, as we have seen previously. Since lineages are 'born' at a linear rate, and 'die' at a quadratic rate, the number of lineages is guaranteed to stay finite and will even hit one occasionally – there will then temporarily be a single ancestral chromosome again (Wiuf and Hein, 1997).

A sample of n lineages behaves in the same way. Each lineage recombines independently at rate $\rho/2$, and each pair of lineages coalesces independently at rate 1. The number of lineages *will* hit one occasionally. The segment in which this first occurs is known as the 'Ultimate' MRCA, because, as we shall see, each point in the sample may well have a younger MRCA.

The genealogy of a sample of n lineages back to the Ultimate MRCA can thus be described by a branching and coalescing graph (an 'ancestral recombination graph') that is analogous to the standard coalescent. A realization for $n = 6$ is shown in Figure 25.10.

What does a lineage in the graph look like? For each point in the segment under study, it must contain information about *which* (if any) sample members it is ancestral to. It is convenient to represent the segment as a (0,1) interval (this is just a coordinate system that can be translated into base pairs or whatever is appropriate). An ancestral lineage can then be represented as a set of elements of the form {interval, labels}, where the intervals are those resulting from all recombinational breakpoints in the history of the sample (Fisher's 'junctions' (Fisher, 1965) for aficionados of classical population genetics) and the labels denote the descendants of that segment (using the 'equivalence class' notation introduced previously). An example of this notation is given in Figure 25.10. Note that pieces of a given chromosomal lineage will often be ancestral to no one in the sample. Indeed, recombination in a nonancestral piece may result in an entirely nonancestral lineage!

So far nothing has been said about where or how recombination breakpoints occur. This has been intentional, to emphasize that the ancestral recombination graph does not depend on (most) details of recombination. It is possible to model almost any kind of recombination (including, for example, various forms of gene conversion) in this framework. But of course the graph has no meaning unless we interpret the recombination events somehow. To proceed, we will assume that each recombination event results in crossing over at a point, x, somewhere in (0,1). How x is chosen is again up to the modeler: it could be a fixed point; it could be a uniform random variable; or it could be drawn from some other distribution (perhaps centered around a 'hotspot'). In any case, a
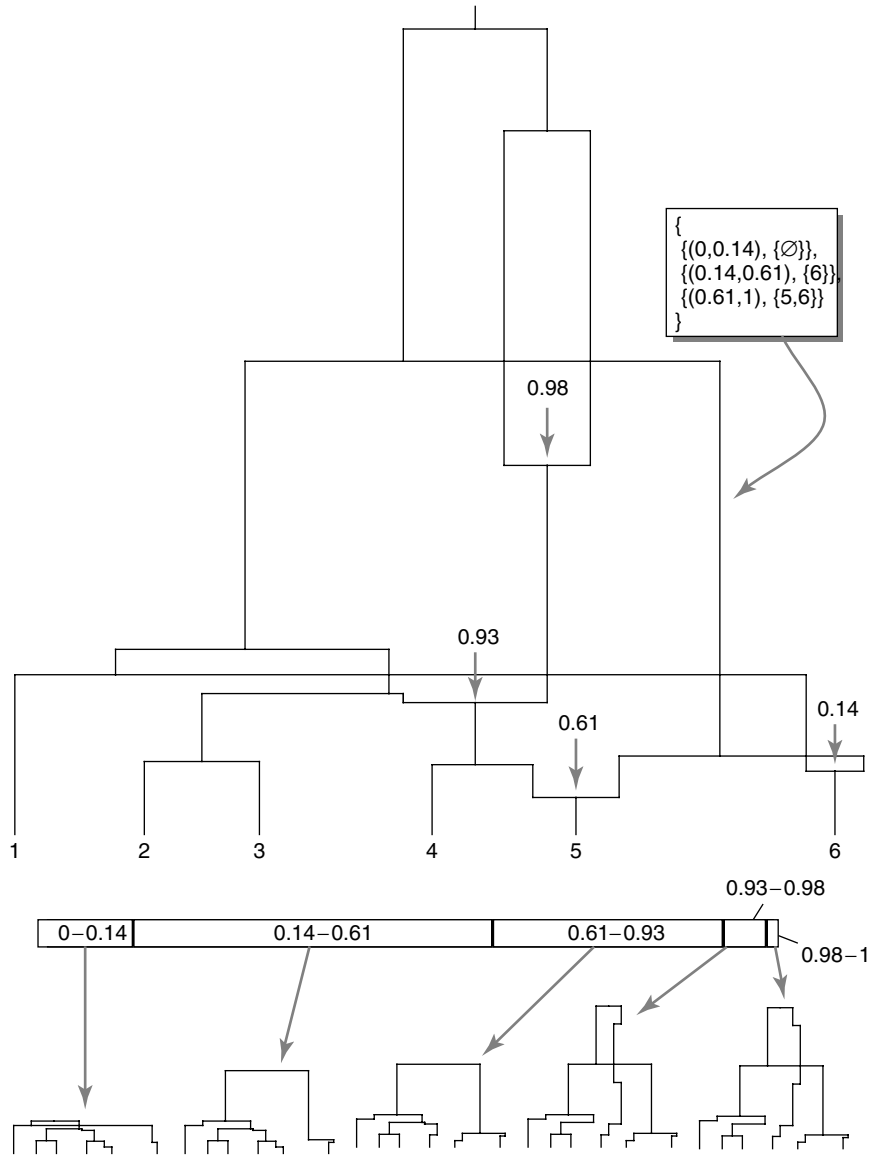
Figure 25.10 A realization of an ancestral recombination graph for $n = 6$. There were four recombination events, which implies $6 + 4 - 1 = 9$ coalescence events. Each recombination was assumed to lead to crossing over at a point, which was chosen randomly in $(0, 1)$. Four breakpoints (or 'junctions') implies five embedded trees, which are shown underneath. The tree for a particular chromosomal point is extracted from the graph by choosing the appropriate path at each recombination event. I have followed the convention that one should 'go left' if the point is located 'to the left' of (is less than) the breakpoint. Note that the two rightmost trees are identical. The box illustrates notation that may be used to represent ancestral lineages in the graph. The lineage pointed to is ancestral to: no (sampled) segment for the interval $(0, 0.14)$; segment 6 for the interval $(0.14, 0.61)$; and segments 5 and 6 for the interval $(0.61, 1)$.

breakpoint needs to be generated for each recombination event in the graph. We also need to know which branch in the graph carries which recombination 'product' (remember that we are going backward in time). With breaks affecting a point, a suitable rule is that the left branch carries the material to the 'left' of the breakpoint (i.e. in $(0, x)$), and the right branch carries the material to the 'right' (i.e. in $(x, 1)$).

Once recombination breakpoints have been added to the graph, it becomes possible to extract the genealogy for any given point by simply following the appropriate branches. Figure 25.10 illustrates how this is done. An ancestral recombination graph contains a number of embedded genealogical trees, each of which can be described by the standard coalescent, but which are obviously not independent of each other. An alternative way of viewing this process is thus as a 'walk through tree space' along the chromosome (Wiuf and Hein, 1999). The strength of the correlation between the genealogies for linked points depends on the scaled genetic distance between them, and goes to zero as this distance goes to infinity. The number of embedded trees equals the number of breakpoints plus one, but many of these trees may (usually will) be identical (cf. the two rightmost trees in Figure 25.10). Note also that the embedded trees vary greatly in height. This means that some pieces will have found their MRCA long before others. Indeed, it is quite possible for every piece to have found its MRCA long before the Ultimate MRCA. A number of interesting results concerning the number of recombination events and the properties of the embedded trees are available (see Griffiths and Marjoram, 1996; 1997; Hudson, 1983; 1987; Hudson and Kaplan, 1985; Kaplan and Hudson, 1985; Pluzhnikov and Donnelly, 1996; Simonsen and Churchill, 1997; Wiuf and Hein, 1999a,b).

25.6.2 Properties and Effects of Recombination

It probably does not need to be pointed out that the stochastic process just described is extremely complicated. At least I have found that whereas it is possible to develop a fairly good intuitive understanding of the random trees generated by the standard coalescent, the behavior of the random recombination graphs continues to surprise. It may therefore be worth questioning first of all whether it is necessary to incorporate recombination. It would seem reasonable that recombination could be ignored if it is sufficiently rare in the segment studied (e.g. if the segment is very short). But what is 'sufficiently rare'? Consider a pair of segments. The probability that they coalesce before either recombines is

$$\frac{1}{1 + 2(\rho/2)} = \frac{1}{1 + \rho} \tag{25.17}$$

(cf. equation (25.11)). In order for recombination not to matter, we would need to have $\rho \approx 0$. It is thus the *scaled* recombination rate that matters, not the per-generation recombination probability. Estimates based on comparing genetic and physical maps indicate that the average per-generation per-nucleotide probability of recombination is of roughly the same order of magnitude as the average per-generation per-nucleotide probability of mutation (which can be estimated in various ways). This means that the scaled mutation and recombination rates will also be of the same order of magnitude, and, thus, that recombination can be ignored when mutation can be ignored. In other words, as long we restrict our attention to segments short enough not to be polymorphic, we do not need to worry about recombination!

Of course, both recombination and mutation rates vary widely over the genome, so regions where recombination can be ignored almost certainly exist. Unfortunately,

whereas direct estimates of recombination probabilities (genetic distances) are restricted to large scales, estimates of the recombination rate from polymorphism data are extremely unreliable (Griffiths and Marjoram, 1996; Hudson, 1987; Hudson and Kaplan, 1985; Wakeley, 1997; Wall, 2000; Fearnhead and Donnelly, 2001; McVean *et al.*, 2002). The latter problem is unavoidable. The main reason is the usual one that there is only a single realization of the underlying genealogy. Thus, for example, numerous recombination events in a particular region of a gene do not necessarily mean that it is a recombinational hotspot: it could just be that that region has a deep enough genealogy for multiple recombination events to have had time to occur. This is the same problem that affects estimates of the mutation rate.

However, there are also problems peculiar to recombination. It is important to realize that most recombination events are undetectable (Hudson and Kaplan, 1985). Recombination in sequence data has often been inferred by identifying 'tracts' that have obviously moved from one sequence to another. The presence of such tracts is actually indicative of *low* rather than of high recombination rates (Maynard Smith, 1999). Even a moderate amount of recombination will wipe out the tracts. Recombination can then only be 'detected' through the 'four-gamete test' (Hudson and Kaplan, 1985): the four linkage configurations AB, Ab, aB, and ab for two linked loci can only arise through recombination or repeated mutation (which is more likely is debatable (Eyre-Walker *et al.*, 1999; Templeton *et al.*, 2000)). Furthermore, recombination events can clearly only be detected if there is sufficient polymorphism. However, many recombination events can *never* be detected even with infinite amounts of polymorphism (Griffiths and Marjoram, 1997; Hudson and Kaplan, 1985; Nordborg, 2000). Consider, for example, the two rightmost trees in Figure 25.10. These trees are identical. This means that the recombination event that gave rise to them cannot possibly leave any trace.

The phenomenon of undetectable breakpoints turns out to have special relevance for models with inbreeding. The 'forward' intuition that corresponds to undetectable recombination events is that these events took place in homozygous individuals. Inbreeding increases the frequency of homozygous individuals, and can therefore have a considerable effect on the recombination graph. It turns out that this effect can also be modeled as a scaling change, but this time of the recombination rate. Thus, for example, the recombination graph in a partially selfing hermaphrodite reduces to the standard recombination graph if we introduce an 'effective recombination rate', $\rho_e :=$ $\rho(1 - F)$ (Nordborg, 2000). Recombination breaks up haplotypes much less efficiently in inbreeders.

So far, we have only discussed the problems associated with recombination. It must be remembered that recombination is the only thing that allows us to get around the 'single underlying genealogy'. Unlinked loci will, with respect to most questions, provide independent samples. Of course this also applies within a segment: if ρ were infinite, then each base pair would be an independent locus (Pluzhnikov and Donnelly, 1996). High rates of recombination are thus an enormous advantage for many purposes.

Finally, it should be noted that since crossing over is mechanistically tied to gene conversion, there is reason to question the applicability of the simple model used above at the intragenic scale (Andolfatto and Nordborg, 1998; Nordborg, 2000). However, the ancestral recombination graph is quite general, and more realistic recombination

models have been developed (Wiuf, 2000; Wiuf and Hein, 2000). Models of other kinds of recombination, such as bacterial transformation (Hudson, 1994) and intergenic gene conversion (Bahlo, 1998), also exist.

25.7 SELECTION

The coalescent depends crucially on the assumption of selective neutrality, because if the allelic state of a lineage influences its reproductive success, it is not possible to separate 'descent' from 'state'. Nonetheless, it turns out that it is possible to circumvent this problem, and incorporate selection into the coalescent framework. Two distinct approaches have been used. The first is an elegant extension of the coalescent process, known as the 'ancestral selection graph' (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997). The genealogy is generated backward in time, as in the standard coalescent, but it contains branching as well as coalescence events. The result is a genealogical graph that is superficially similar to the one generated by recombination. Mutations are then superimposed forward in time, and, with knowledge of the state of each branch, the graph is 'pruned' to a tree by preferentially removing bad branches (i.e. those carrying selectively inferior alleles). In a sense, the ancestral selection graph allows the separation of descent from state by including 'potential' descent: lineages that might have lived, had their state allowed it.

The second approach is based on two insights. First, a polymorphic population may be thought of as subdivided into *allelic classes* within which there is no selection. Second, if we know the historical sizes of these classes, then they may be modeled as analogous to patches, using the machinery described above. Lineages then 'mutate between classes' rather than 'migrate between patches'. This approach was pioneered in the context of the coalescent by Kaplan *et al.* (1988). Knowing the past class sizes is the same as knowing the past allele frequencies, so it is obviously not possible to study the dynamics of the selectively different alleles themselves using this approach. However, it is possible to study the effects of selection on the underlying genealogical structure, which is relevant if we wish to understand how linked neutral variants are affected.

It is not entirely clear how the two approaches relate to each other. Since the second approach requires knowledge of the past allele frequencies, it may be viewed as some kind of limiting (strong selection) or, alternatively, conditional version of the selection graph (Nordborg, 1999). However, whereas the second approach would be most appropriate for very strong, deterministic selection, the selection graph requires all selection coefficients to be $O(1/N)$. This is an area of active research.

The ancestral selection graph is described in **Chapter 22**, and will not be discussed here. The second approach, which might be called the 'conditional structured coalescent', will be illustrated through three simple but very different examples.

25.7.1 Balancing Selection

By 'balancing selection' is meant any kind of selection that tends to maintain two or more alleles in the population. The effect of such selection on genealogies has been studied by a number of authors (Barton and Navarro, 2002; Hey, 1991; Hudson and Kaplan, 1988; Kaplan *et al.*, 1988; Kaplan *et al.*, 1991; Kelly and Wade, 2000; Navarro and

Barton, 2002; Nordborg, 1997; 1999; Schierup *et al.*, 2001; Takahata, 1990; Vekemans and Slatkin, 1994). We will limit ourselves to the case of two alleles, A_1 and A_2, maintained at constant frequencies p_1 and $p_2 = 1 - p_1$ by strong selection. Alleles mutate to the other type with some small probability v per generation, and we define the scaled rate $v := 4Nv$. Reproduction occurs according to a diploid Wright–Fisher model, as for the recombination graph above.

Consider a segment of length ρ that contains the selected locus. Depending on the allelic state at the locus, the segment belongs to either the A_1 or the A_2 allelic class. Say that it belongs to the A_1 allelic class. Trace the ancestry of the segment a single generation back in time. It is easy to see that its creation involved an $A_2 \rightarrow A_1$ mutation with probability

$$\frac{vp_2}{vp_2 + (1-v)p_1} = \frac{v}{4N} \times \frac{p_2}{p_1} + O\left(\frac{1}{N^2}\right), \tag{25.18}$$

(cf. equation (25.9)), and involved recombination with probability $r = \rho/(4N)$. Thus the probability that neither happens is $1 - O(1/N)$, and the probability of two events, for example both mutation and recombination, is $O(1/N^2)$ and can be neglected. If nothing happens, then the lineage remains in the A_1 class. If there was a mutation, the lineage 'mutates' to the A_2 allelic class. If there was a recombination event, we have to know the genotype of the individual in which the event took place.

Because the lineage we are following is A_1, we know that the individual must have been either an A_1A_1 homozygote or an A_1A_2 heterozygote. What fraction of the A_1 alleles was produced by each genotype? In general, this will depend on their relative fitness as well as their frequencies. Let x_{ij} be the frequency of A_iA_j individuals, and w_{ij} their relative fitness. Then the probability that an A_1 lineage was produced in a heterozygote is

$$\frac{w_{12}x_{12}/2}{w_{12}x_{12}/2 + w_{11}x_{11}}. \tag{25.19}$$

If we can ignore the differences in fitness, and assume Hardy–Weinberg equilibrium (see Nordborg, 1999, for more on this), (25.19) simplifies to

$$\frac{p_1p_2}{p_1p_2 + p_1^2} = p_2. \tag{25.20}$$

Thus the probability that an A_1 lineage 'meets' and recombines with an A_2 segment is equal to the frequency of A_2 segments, which is intuitive. The analogous reasoning applies to A_2 lineages, which recombine with A_1 segments with probability p_1, and with members of their own class with probability p_2. It should be noted that the above can be made rigorous using a model that treats genotypes as well as individuals as population structure (Nordborg, 1999).

What happens when the lineage undergoes recombination? If it recombines in a homozygote, then both branches remain in the A_1 allelic class. However, if it recombines in a heterozygote, then one of the branches (the one *not* carrying the ancestry of the selected locus) will 'jump' to the A_2 allelic class. The other branch remains in the A_1 allelic class.

When more than two lineages exist, coalescences may occur, but only within allelic classes (remember that since mutation is $O(1/N)$ it is impossible for lineages to mutate and coalesce in the same generation).

If time is measured in units of $2N$ generations, and we let N go to infinity, the model converges to a coalescent process with the following types of events:
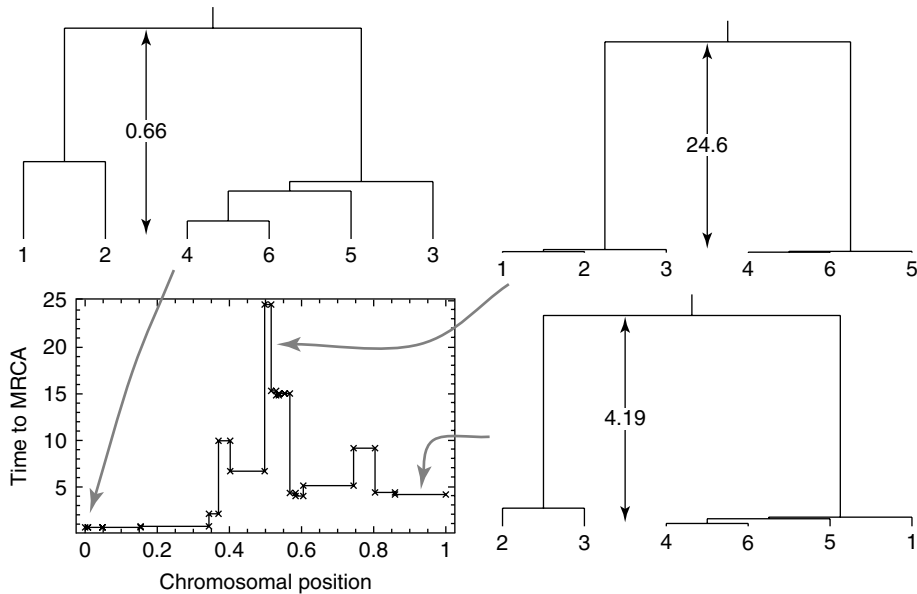
Figure 25.11 Selection will have a local effect on genealogies. A realization of the coalescent with recombination and strong balancing selection is shown. Lineages 1–3 belong to one allelic class, and lineages 4–6 to the other. The selected locus is located in the middle of the region. The plot shows how the time to the MRCA varies along the chromosome (the crosses denote cross-over points). The three extracted trees exemplify how the topology and branch lengths are affected by linkage to the selected locus. Note that the trees are not drawn to scale (the numbers on the arrows give the heights).

- each pair of lineages in the A_i allelic class coalesces independently at rate $1/p_i$;

- each lineage in A_i recombines with a segment in class j at rate ρp_j;

- each lineage in A_i mutates to A_j, $j \neq i$, at rate $v p_j/p_i$.

The process may be stopped either when the Ultimate MRCA is reached, or when all points have found their MRCA.

This model has some very interesting properties. Consider a sample that contains both types of alleles. Since coalescence is only possible within allelic classes, the selected locus (in the strict sense of the word, i.e. the 'point' in the segment where the selectively important difference lies) cannot coalesce without at least one mutation event. If mutations are rare, then this may have occurred a very long time ago. In other words, the polymorphism may be ancient. All coalescences will occur within allelic classes before mutation allows the final two lineages to coalesce. The situation is similar to strong population subdivision (see Figure 25.7). However, this is only true for the locus itself: linked pieces may 'recombine away' and coalesce much earlier. This will usually result in a local increase in the time to MRCA centered around the selected locus, as illustrated in Figure 25.11. Because the expected number of mutations is proportional to the height of the tree, this may lead to a 'peak of polymorphism' (Hudson and Kaplan, 1988).

25.7.2   Selective Sweeps

Next consider a population in which favorable alleles arise infrequently at a locus, and are rapidly driven to fixation by strong selection. Each such fixation is known as a 'selective sweep' for reasons that will become apparent. This process can be modeled using the framework developed above, if we know how the allele frequencies have changed over time. Of course we do not know this, but if the selection is strong enough, it may be reasonable to model the increase in frequency of a favorable allele deterministically (Kaplan *et al.*, 1989).

Consider a population that is currently not polymorphic, but in which a selective sweep recently took place. During the sweep, there were two allelic classes just as in the balancing selection model above. The difference is that these classes changed in size over time. In particular, the class corresponding to the allele that is currently fixed in the population will *shrink* rapidly back in time. The genealogy of the selected locus itself (in the 'point' sense used above) will therefore behave as if it were part of a population that has expanded from a very small size (cf. Figure 25.6). Indeed, unlike 'real' populations, the allelic class *will* have grown from a size of 1. A linked point must have grown in the same way, unless recombination in a heterozygote took place between the point and the selected locus. Whether this happens or not will depend on how quickly the new allele increased. Typically, it depends on the ratio r/s, where s is the selective advantage of the new allele, and r is the relevant recombination probability.

The result of such a fixation event is thus to cause a local 'genealogical distortion', just like balancing selection. However, whereas the distortion in the case of balancing selection looks like population subdivision, the distortion caused by a fixation event looks like population growth. Close to the selected site, coalescence times will have a tendency to be short, and the genealogy will have a tendency to be starlike (cf. Figure 25.6). Note that a single recombination event in the history of the sample can change this, and that the variance will consequently be enormous (note the variance in time to MRCA in Figure 25.11). Shorter coalescence times mean less time for mutations to occur, so a local reduction in variability is expected. This is obvious: when the new allele sweeps through the population and fixes, it causes linked neutral alleles to 'hitchhike' along and also fix (Maynard Smith and Haigh, 1974). Repeated selective sweeps can thus decrease the variability in a genomic region (Kaplan *et al.*, 1989; Simonsen *et al.*, 1995; Kim and Stephan, 2002; Przeworski, 2002). Because each sweep is expected to affect a bigger region the lower the rate of recombination is, this has been proposed as an explanation for the correlation between polymorphism and local rate of recombination that is observed in many organisms (Begun and Aquadro, 1992; Nachman, 1997; Nachman *et al.*, 1998).

25.7.3   Background Selection

We have seen that selection can affect genealogies in ways reminiscent of strong population subdivision and of population growth. It is often difficult to distinguish statistically between selection and demography for precisely this reason (Fu and Li, 1993; Tajima, 1989b). It is also possible for selection to affect genealogies in a way that is completely undetectable, that is, as a linear change in time scale. This appears to be the case for selection against deleterious mutations, at least under some circumstances (Charlesworth *et al.*, 1995; Hudson and Kaplan, 1994; 1995; Nordborg, 1997; Nordborg *et al.*, 1996).

The basic reason for this is the following. Strongly deleterious mutations are rapidly removed by selection. Looking backward in time, this means that each lineage that carries a deleterious mutation must have a nonmutant ancestor in the near past. On the coalescent time scale, lineages in the deleterious allelic class will 'mutate' (backward in time) to the 'wild-type' allelic class instantaneously. The process looks like a strong-migration model, with the wild-type class as the source environment, and the deleterious class as the sink environment: the presence of deleterious mutations increases the variance in reproductive success. The resulting reduction in the effective population size is known as 'background selection' (Charlesworth *et al.*, 1993).

More realistic models with multiple loci subject to deleterious mutations, recombination, and several mutational classes turn out to behave similarly. The strength of the background selection effect at a given genomic position will depend strongly on the local rate of recombination, which determines how many mutable loci influence a given point. Thus, deleterious mutations have also been proposed as an explanation for the correlation between polymorphism and local rate of recombination referred to above (Charlesworth *et al.*, 1993). The 'effective population size' would thus depend on the mutation, selection, and recombination parameters in each genomic region.

It should be pointed out that, unlike the many limit approximations presented in this chapter, the idea that background selection can be modeled as a simple scaling is not mathematically rigorous. However, we would rather hope that selection against deleterious mutations can be taken care of this way, because given that amino acid sequences are conserved over evolutionary time, practically all of population genetics theory would be in trouble otherwise!

25.8 NEUTRAL MUTATIONS

Not much has been said about the neutral mutation process because it is trivial from a mathematical point of view. Once we know how to generate the genealogy, mutations can be added afterwards according to a Poisson process with rate $\theta/2$, where θ is the scaled per-generation mutation probability. Thus, if a particular branch has length τ units of scaled time, the number of mutations that occur on it will be Poisson-distributed with mean $\tau\theta/2$ (and they occur with uniform probability along the branch). It is also possible to add mutations while the genealogy is being created, instead of afterwards. This can in some circumstances lead to much more efficient algorithms (see, for example, the 'urn scheme' described by Donnelly and Tavaré, 1995), although from the point of view of simulating samples, all coalescent algorithms are so efficient that such fine-tuning does not matter. However, it can matter greatly for the kinds of inference methods described in **Chapter 26**.

It should be noted that the mutation process is just as general as the recombination process. Almost any neutral mutation model can be used. A useful trick is so-called 'Poissonization': let mutation events occur according to a simple Poisson process with rate $\theta/2$, but once an event occurs, determine the *type* of event through some kind of transition matrix which includes mutation back to self (i.e. there was no mutation). This allows models where the mutation probability depends on the current allelic state.

The only restriction is that in order to interpret samples generated by the coalescent as samples from the relevant stationary distribution (which incorporates demography,

migration, selection at linked sites, etc.), we need to be able to choose the type of the MRCA from the stationary distribution of the mutation process (alone, since demography, for example, does not affect samples of size $n = 1$). In many cases, such as the infinite-alleles model (each mutation gives rise to a new allele) or the infinite-sites model (each mutation affects a new site), the state of the MRCA does not matter, since all we are interested in is the number of mutational changes.

25.9 CONCLUSION

25.9.1 The Coalescent and 'Classical' Population Genetics

The differences between coalescent theory and 'classical' population genetics have frequently been exaggerated or misunderstood. First, the basic models do not differ. The coalescent is essentially a diffusion model of lines of descent. This can be done forward in time, for the whole population (e.g. Griffiths, 1980), but it was realized in the early 1980s that it is easier to do it backward in time. Second, the coalescent is not limited to finite samples. Everything above has been limited to finite samples because it is mathematically much easier, but it is likely that all of it could be extended to the whole population. Of course, it is essential for the independence of events that the number of lineages be finite, but in the whole-population coalescent the number of lineages becomes finite infinitely fast (it is an 'entrance boundary', e.g. Griffiths, 1984). Third, classical population genetics is not limited to the whole population. A sample of size $n = 6$ from a K-allele model, say, could be obtained either through the coalescent, or by first drawing a population from the stationary distribution found by Wright (1949), and then drawing six alleles conditional on this population. Note, however, that it would be rather more difficult (read 'impossible') to use the second approach for most models. Fourth, the coalescent is in no sense tied to sequence data: any mutation model can be used. The impression that it is came about doubtless because models for sequence evolution such as the infinite-sites model are indeed impossibly hard to analyze using classical methods (Ethier and Griffiths, 1987).

I would argue that the real difference is more philosophical. As has been pointed out by Ewens (1979; 1990), essentially all of classical population genetics is 'prospective', looking forward in time. Another way of saying this is that it is conditional: given the state in a particular generation, what will happen? This approach is fine when modeling is done to determine 'how evolution might work' (which is what most classical population genetics was about). It is usually not suitable for statistical analysis of data, however. Wright considered how 'heterozygosity' would decay from the same starting point in infinitely many identical populations, that is to say, he took the expectation over evolutionary realizations. Data, alas, come from a single time-slice of one such realization. The coalescent forces us to acknowledge this, and allows the utilization of modern statistical methods, such as the calculation of likelihoods for samples.

25.9.2 The Coalescent and Phylogenetics

If the differences between coalescent theory and classical population genetics have sometimes been exaggerated, the differences between coalescent theory and phylogenetics

have not always been fully appreciated. The central role played by trees in both turns out to be very misleading.

To be able to compare them, we need to model speciation. This has usually been done using an 'isolation' model in which randomly mating populations split into two completely isolated ones at fixed times in the past. The result is a 'species tree', within which we find 'gene trees' (see Figure 25.12). The model is quite simple: lineages will tend to coalesce within their species, and can only coalesce with lineages from other species back in the ancestral species.

Molecular phylogenetics attempts to estimate the species tree by estimating the genealogy of homologous sequences from the different species, that is, by estimating the gene tree. The species tree is assumed to exist and is treated as a model parameter.

In addition, the standard methods rely on all branches in the species tree being very long compared to within-species coalescence times. This means that the coalescent can be ignored: regardless of how we sample, all (neutral) gene genealogies will rapidly coalesce within their species, and thereafter have the same topology as the species tree. Furthermore, the variation in the branch lengths caused by different coalescence times in the ancestral species will be negligible compared to the lengths of the interspecific branches. There is no need to sample more than one individual per species, and recombination is completely irrelevant. Gene trees perfectly reflect the species tree.

It is of course widely acknowledged that gene trees and species trees may differ (see Avise, 1994; Avise and Ball, 1990; Hey, 1994; Hudson, 1992; Li, 1997; Nei, 1987; Takahata, 1989; Wu, 1991; phylogenetic methods are discussed in **Chapters 15** and **16**). Nonetheless, phylogenetic methods would not work unless the interspecific branches usually were long enough for the gene trees to reflect the species tree closely. Indeed, in many situations, the problem is the opposite: the branches are so long that repeated mutations have erased much of the phylogenetic information.
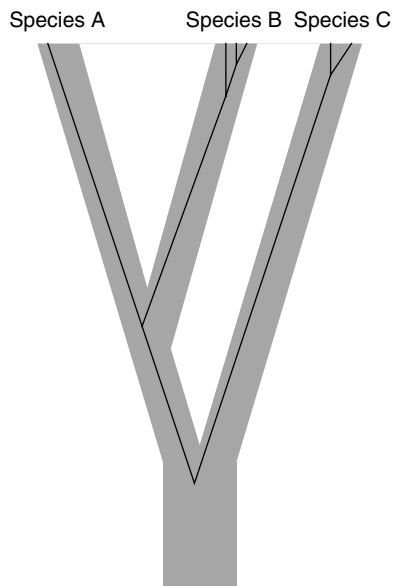
Figure 25.12 A gene tree within a species tree.

Phylogenetic inference can thus be viewed as a 'missing data' problem just like population genetic inference: polymorphisms contain information about an unobserved genealogy, which in turn provides information about an evolutionary model. However, note that in phylogenetics, there is relatively little doubt about what the right model is (it is typically an isolation model that gives rise to a species tree, as in Figure 25.12). Furthermore, because of the long branch lengths, the gene genealogies, although random variables with a coalescent distribution under the model, can be treated as parameters (which, among other things, means that we do not need to know the sizes of ancestral populations to estimate divergence times). None of this is true when analyzing population genetic data (which, strictly speaking, means any data for which the 'long branch' assumption above is not fulfilled). Unfortunately, the considerable success and popularity of phylogenetics (coupled with the ready availability of user-friendly software) has sometimes led to the inappropriate application of phylogenetic methods. It is important to remember that a genealogical tree from a population (or several populations that have not been isolated for a very long time) does not have an obvious interpretation: it certainly contains information about the process that gave rise to it, but usually less than we would hope (see Rosenberg and Nordborg, 2002).

25.9.3   Prospects

A theme of this review has been the versatility and generality of the coalescent model. Considerable theoretical progress is being made, especially in the areas of statistical inference and modeling selection. At the same time, we are entering the era of genomic polymorphism data. This wealth of information will make it increasingly possible to evaluate whether the models constructed by population genetics over the years actually fit the data. It seems likely that we will find that the data is in many ways much less informative than imagined (as happened before in population genetics; see Lewontin, 1991); it also seems likely that we will discover new phenomena that require new models. Either way, the importance of a rigorous statistical approach to analyzing genetic polymorphism data can hardly be overstated.

Acknowledgments

REFERENCES

Andolfatto, P. and Nordborg, M. (1998). The effect of gene conversion on intralocus associations. *Genetics* **148**, 1397–1399.

Avise, J.C. (1994). *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York.

Avise, J.C. and Ball, R.M. (1990). In *Oxford Surveys in Evolutionary Biology*, Vol. 7, D. Futuyama, and J. Antonovics, eds. Oxford University Press, Oxford, pp. 45–67.

Bahlo, M. (1998). Segregating sites in a gene conversion model with mutation. *Theoretical Population Biology* **54**, 243–256.

Barton, N.H. and Navarro, A. (2002). Extending the coalescent to multilocus systems: the case of balancing selection. *Genetical Research* **79**, 129–139.

Barton, N.H. and Wilson, I. (1995). Genealogies and geography. *Proceedings of the Royal Society London Series B* **349**, 49–59.

Begun, D.J. and Aquadro, C.F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. *Nature* **356**, 519–520.

Charlesworth, B., Morgan, M.T. and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.

Charlesworth, D., Charlesworth, B. and Morgan, M.T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics* **141**, 1619–1632.

Donnelly, P. (1996). In *Variation in the Human Genome*, Ciba Foundation Symposium No. 197. John Wiley & Sons, Chichester, pp. 25–50.

Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29**, 401–421.

Ethier, S.N. and Griffiths, R.C. (1987). The infinitely many sites model as a measure valued diffusion. *Annals of Probability* **5**, 515–545.

Ewens, W.J. (1979). *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Ewens, W.J. (1990). *Mathematical and Statistical Development of Evolutionary Theory*, S. Lessard, ed. Kluwer Academic, Dordrecht, pp. 177–227.

Eyre-Walker, A., Smith, N.H. and Maynard Smith, J. (1999). How clonal are human mitochondria? *Proceedings of the Royal Society London Series B* **266**, 477–483.

Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318.

Fisher, R.A. (1965). *Theory of Inbreeding*, 2nd edition. Oliver and Boyd, Edinburgh.

Fu, Y.-X. and Li, W.-H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Griffiths, R.C. (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoretical Population Biology* **17**, 37–50.

Griffiths, R.C. (1984). Asymptotic line-of-descent distributions. *Journal of Mathematical Biology* **21**, 67–75.

Griffiths, R.C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3**, 479–502.

Griffiths, R.C. and Marjoram, P. (1997). In *Progress in Population Genetics and Human Evolution*, P. Donnelly and S. Tavaré, eds. Springer-Verlag, New York, pp. 257–270.

Griffiths, R.C. and Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London Series B* **344**, 403–410.

Griffiths, R.C. and Tavaré, S. (1998). The age of a mutant in a general coalescent tree. *Stochastic Models* **14**, 273–295.

Herbots, H.M. (1994). Stochastic models in population genetics: genealogy and genetic differentiation in structured populations. PhD thesis, University of London.

Hey, J. (1991). A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theoretical Population Biology* **39**, 30–48.

Hey, J. (1994). *Molecular Ecology and Evolution: Approaches and Applications*, B. Schierwater, B. Streit, G.P. Wagner, R.Desalleeds. Birkhäuser, Basel, pp. 435–449.

Hudson, R.R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.

Hudson, R.R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetical Research, (Cambridge)* **50**, 245–250.

Hudson, R.R. (1990). In *Oxford Surveys in Evolutionary Biology*, Vol. 7, D. Futuyma and J. Antonovics, eds. Oxford University Press, Oxford, pp. 1–43.

Hudson, R.R. (1992). Gene trees, species trees and the segregation of ancestral alleles. *Genetics* **131**, 509–512.

Hudson, R.R. (1993). *Mechanisms of Molecular Evolution*, A.G. Takahata and N. Clark, eds. Japan Scientific Societies Press, Tokyo, pp. 23–36.

Hudson, R.R. (1994). Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation. *Journal of Evolutionary Biology* **7**, 535–548.

Hudson, R.R. (1998). Island models and the coalescent process. *Molecular Ecology* **7**, 413–418.

Hudson, R.R. and Kaplan, N.L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.

Hudson, R.R. and Kaplan, N.L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840.

Hudson, R.R. and Kaplan, N.L. (1994). *Non-neutral Evolution: Theories and Molecular Data*, G.B. Golding, ed. Chapman & Hall, New York, pp. 140–153.

Hudson, R.R. and Kaplan, N.L. (1995). Deleterious background selection with recombination. *Genetics* **141**, 1605–1617.

Kaj, I., Krone, S.M. and Lascoux, M. (1991). Coalescent theory for seed bank models. *Journal of Applied Probability* **38**, 285–301.

Kaplan, N.L. and Hudson, R.R. (1985). The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination. *Theoretical Population Biology* **28**, 382–396.

Kaplan, N.L., Darden, T. and Hudson, R.R. (1988). The coalescent process in models with selection. *Genetics* **120**, 819–829.

Kaplan, N.L., Hudson, R.R. and Iizuka, M. (1991). The coalescent process in models with selection, recombination and geographic subdivision. *Genetical Research, Cambridge* **57**, 83–91.

Kaplan, N.L., Hudson, R.R. and Langley, C.H. (1989). The 'hitch-hiking' effect revisited. *Genetics* **123**, 887–899.

Kelly, J.K. and Wade, M.J. (2000). Molecular evolution near a two-locus balanced polymorphism. *Journal of Theoretical Biology* **204**, 83–101.

Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777.

Kingman, J.F.C. (1982a). The coalescent. *Stochastic Processes and their Applications* **13**, 235–248.

Kingman, J.F.C. (1982b). In *Exchangeability in Probability and Statistics*, G. Koch and F. Spizzichino, eds. North-Holland, Amsterdam, pp. 97–112.

Kingman, J.F.C. (1982c). In *Essays in Statistical Science: Papers in Honour of P.A.P. Moran*, J. Gani and E.J. Hannan, eds. (Journal of Applied Probability, special, Vol. 19A) Applied Probability Trust, Sheffield, pp. 27–43.

Krone, S.M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology* **51**, 210–237.

Lewontin, R.C. (1991). Twenty-five years ago in genetics: electrophoresis in the development of evolutionary genetics: milestone or millstone? *Genetics* **128**, 657–662.

Li, W.-H. (1997). *Molecular Evolution*. Sinauer, Sunderland, MA.

Marjoram, P. and Donnelly, P. (1997). *Progress in Population Genetics and Human Evolution*, S. Donnelly, P. Tavaré, eds. Springer-Verlag, New York, pp. 107–131.

Maynard Smith, J. (1999). The detection and measurement of recombination from sequence data. *Genetics* **153**, 1021–1027.

Maynard Smith, J. and Haigh, J. (1974). The hitchhiking effect of a favourable gene. *Genetical Research, Cambridge* **23**, 23–35.

McVean, G., Awadalla, P. and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241.

Möhle, M. (1998a). A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Advances in Applied Probability* **30**, 493–512.

Möhle, M. (1998b). Robustness results for the coalescent. *Journal of Applied Probability* **35**, 438–447.

Möhle, M. (1999). Weak convergence to the coalescent in neutral population models. *Journal of Applied Probability* **36**, 446–460.

Nachman, M.W. (1997). Patterns of DNA variability at X-linked loci in Mus domesticus. *Genetics* **147**, 1303–1316.

Nachman, M.W., Bauer, V.L., Crowell, S.L. and Aquadro, C.F. (1998). DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**, 1133–1141.

Nagylaki, T. (1980). The strong-migration limit in geographically structured populations. *Journal of Mathematical Biology* **9**, 101–114.

Nagylaki, T. (1982). Geographical invariance in population genetics. *Journal of Theoretical Biology* **99**, 159–172.

Nagylaki, T. (1998). The expected number of heterozygous sites in a subdivided population. *Genetics* **149**, 1599–1604.

Navarro, A. and Barton, N.H. (2002). The effects of multilocus balancing selection on neural variability. *Genetics* **161**, 849–863.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Neuhauser, C. and Krone, S.M. (1997). The genealogy of samples in models with selection. *Genetics* **145**, 519–534.

Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics* **146**, 1501–1514.

Nordborg, M. (1998). On the probability of Neanderthal ancestry. *American Journal of Human Genetics* **63**, 1237–1240.

Nordborg, M. (1999). In *Statistics in Molecular Biology and Genetics, IMS Lecture Notes Monograph Series*, Vol. 33, F. Seillier-Moiseiwitsch, ed. Institute of Mathematical Statistics, Hayward, CA, pp. 56–76.

Nordborg, M. (2000). Linkage disequilibrium, gene trees, and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929.

Nordborg, M. and Donnelly, P. (1997). The coalescent process with selfing. *Genetics* **146**, 1185–1195.

Nordborg, M. and Krone, S.M. (2002). *Modern Developments in Theoretical Population Genetics*, M. Slatkin, M. Veuille, eds. Oxford University Press, Oxford, pp. 194–232.

Nordborg, M., Charlesworth, B. and Charlesworth, D. (1996). The effect of recombination on background selection. *Genetical Research, Cambridge* **67**, 159–174.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured populations. *Journal of Mathematical Biology* **29**, 59–75.

Notohara, M. (1993). The strong-migration limit for the genealogical process in geographically structured populations. *Journal of Mathematical Biology* **31**, 115–122.

Pluzhnikov, A. and Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* **144**, 1247–1262.

Pollak, E. (1987). On the theory of partially inbreeding finite populations. I. Partial selfing. *Genetics* **117**, 353–360.

Przeworski, M. (2002). The signature of positive selection and randomly chosen loci. *Genetics* **160**, 1179–1189.

Pulliam, H.R. (1988). Sources, sinks, and population regulation. *American Naturalist* **132**, 652–661.

Rogers, A.R. and Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology Evolution* **9**, 552–569.

Rosenberg, N.A. and Nordborg, M. (2002). Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. *Nature Reviews Genetics* **3**, 380–390.

Rousset, F. (1999a). Genetic differentiation in populations with different classes of individuals. *Theoretical Population Biology* **55**, 297–308.

Rousset, F. (1999b). Genetic differentiation within and between two habitats. *Genetics* **151**, 397–407.

Saunders, I.W., Tavaré, S. and Watterson, G.A. (1984). On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability* **16**, 471–491.

Schierup, M.H., Mikkelsen, A.M. and Hein, J. (2001). Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. *Genetics* **159**, 1833–1844.

Simonsen, K.L. and Churchill, G.A. (1997). A Markov chain model of coalescence with recombination. *Theoretical Population Biology* **52**, 43–59.

Simonsen, K.L., Churchill, G.A. and Aquadro, C.F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.

Slatkin, M. (1987). The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology* **32**, 42–49.

Slatkin, M. (1996). Gene genealogies within mutant allelic classes. *Genetics* **143**, 579–587.

Slatkin, M. and Hudson, R.R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562.

Slatkin, M. and Rannala, B. (1997a). Estimating the age of alleles by use of intraallelic variability. *American Journal of Human Genetics* **60**, 447–458.

Slatkin, M. and Rannala, B. (1997b). The sampling distribution of disease-associated alleles. *Genetics* **147**, 1855–1861.

Strobeck, C. (1987). Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* **117**, 149–153.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.

Tajima, F. (1989a). DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. *Genetics* **123**, 229–240.

Tajima, F. (1989b). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetical Research, Cambridge* **52**, 213–222.

Takahata, N. (1989). Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* **122**, 957–966.

Takahata, N. (1990). A simple genealogical structure of strongly balanced allelic lines and trans-species polymorphism. *Proceedings of the National Academy of Sciences (USA)* **87**, 2419–2423.

Takahata, N. (1991). Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. *Genetics* **129**, 585–595.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetic models. *Theoretical Population Biology* **26**, 119–164.

Templeton, A.R., Clark, A.G., Weiss, K.M., Nickerson, D.A., Boerwinkle, E. and Sing, C.F. (2000). Recombinational and mutational hotspots within the human lipoprotein lipase gene. *American Journal of Human Genetics* **66**, 69–83.

Vekemans, X. and Slatkin, M. (1994). Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics* **137**, 1157–1165.

Wakeley, J. (1996). Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical Population Biology* **49**, 369–386.

Wakeley, J. (1997). Using the variance of pairwise differences to estimate the recombination rate. *Genetical Research, (Cambridge)* **69**, 45–48.

Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics* **153**, 1863–1871.

Wall, J.D. (2000). A comparison of estimators of the population recombination rate. *Molecular Biology Evolution* **17**, 156–163.

Weiss, G. and von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics* **149**, 1539–1546.

Wilkins, J.F. and Wakeley, J. (2002). The coalescent in a continuous, finite, linear population. *Genetics* **161**, 873–888.

Wilkinson-Herbots, H.M. (1998). Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology* **37**, 535–585.

Wiuf, C. (2000). A coalescence approach to gene conversion. *Theoretical Population Biology* **57**, 357–367.

Wiuf, C. and Donnelly, P. (1999). Conditional genealogies and the age of a neutral mutant. *Theoretical Population Biology* **56**, 183–201.

Wiuf, C. and Hein, J. (1997). On the number of ancestors to a DNA sequence. *Genetics* **147**, 1459–1468.

Wiuf, C. and Hein, J. (1999a). The ancestry of a sample of sequences subject to recombination. *Genetics* **151**, 1217–1228.

Wiuf, C. and Hein, J. (1999b). Recombination as a point process along sequences. *Theoretical Population Biology* **55**, 248–259.

Wiuf, C. and Hein, J. (2000). The coalescent with gene conversion. *Genetics* **155**, 451–462.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1949). *Genetics, Palaeontology, and Evolution*, G.L. Jepson, G.G. Simpson and E. Mayr, eds. Princeton University Press, Princeton, NJ, pp. 365–389.

Wu, C.-I. (1991). Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* **127**, 429–435.

Inference Under the Coalescent

M. Stephens

Departments of Statistics and Human Genetics, University of Chicago, Chicago, IL, USA

This chapter introduces some modern statistical methods for inference from molecular population genetic data. The methods are based on the use of the coalescent to model the genealogy relating a random sample of chromosomes from a population, and help us to say something about the demographic and genetic factors that shaped the evolution of populations from which the samples were taken. The chapter focuses primarily on 'full-data' methods, which aim to make full use of the high resolution of modern genetic data, rather than relying on simple summaries such as pair-wise differences, or sample heterozygosity. These methods are often somewhat complex, and the aim of this chapter is to provide practical guidelines for those who wish to apply these methods using existing software, and sufficient theoretical background to understand how the methods work, at least in outline. We also include a brief review of other popular approaches that provide computationally attractive alternatives to full-data methods. In describing these methods, the chapter provides an introduction to importance sampling and Markov chain Monte Carlo – statistical methods that are likely to play a major role in the analyses of future genetic data.

26.1 INTRODUCTION

Genetic data collected from populations are potentially useful for answering a variety of interesting questions. These questions could relate to, for example, the following:

- The genetic forces, such as mutation and recombination, which have affected the evolution of a particular chromosomal segment or locus.

- The historical relationships amongst different subpopulations. In humans this includes the relationships amongst different continental groups, and the times and routes of major migrations.

- The ancestry of the sampled chromosomes, including, for example, the ages of particular mutations, which are carried by some of the chromosomes.

This chapter describes some modern statistical methods for answering these kinds of questions.

Historically, inference in these settings has been based on summaries of the data, such as pair-wise differences, or sample heterozygosity. In contrast, here we focus primarily on methods that aim to make use of the full genetic data in order to provide more accurate inference than is possible using simple summary statistics. These modern methods make use of a theoretical innovation known as the *coalescent* (see **Chapter 25** for an introduction). Both methodological and computational advances have made these 'full-data' methods tractable for some problems, although for other applications, particularly those involving more than a small amount of recombination, full-data approaches remain computationally intractable. In these cases, inference is generally performed using summary statistics, or through other recently developed approximate approaches, some of which we briefly review towards the end of the chapter. Some of these approximate approaches are based on applying full-data approaches to smaller subsets of the data, and then combining information across subsets. Thus, even where the full-data methods themselves are intractable, they can nevertheless help provide a path to an effective practical solution.

The chapter is intended for those who wish to use and understand full-data coalescent-based inference methods. We give practical guidelines, with theoretical background and examples. For ease of exposition, we focus on the simplest models, but most of the principles we discuss will continue to hold in more complex settings. While the focus of this chapter is on inference under the coalescent, many of the concepts and ideas discussed are of more general interest. For example, the next section gives a brief introduction to the ideas underlying likelihood inference and a discussion of the relative merits of maximum-likelihood and Bayesian approaches. Later sections include descriptions of computationally intensive statistical methods, such as importance sampling (IS) and Markov chain Monte Carlo (MCMC), which have proved useful in a wide range of statistical applications, including some related to genetics. Most of the comments we make on these methods apply quite generally, and not only to problems involving the coalescent.

26.1.1   Likelihood-based Inference

Broadly speaking, likelihood-based methods of drawing inference from data proceed by treating the observed data as having arisen from some random process, or *model*, certain aspects of which are unknown. We refer to these unknown quantities as *parameters*. Typically, the aim of a statistical analysis is to use the data to estimate the parameters of the model, and (importantly) to assess the degree of uncertainty associated with these estimates.

More explicitly, likelihood-based inference requires calculation of the probability, $P(\mathcal{D} \mid \psi)$, of observing data $\mathcal{D}$ if the parameters of the model take the value ψ. The likelihood $L(\psi)$ is defined to be this quantity considered as a function of ψ:

$$L(\psi) = P(\mathcal{D} \mid \psi). \tag{26.1}$$

In the population genetics context, the data $\mathcal{D}$ are typically the genetic types of a random sample of chromosomes.[1] from a population Their distribution depends in a complex way on many unknown parameters, relating to both the demographic history of the population (e.g. population size, migration rates) and the underlying genetic mechanisms

---

[1] In this chapter, when we refer to chromosomes we typically mean 'chromosomal segments'.

(e.g. mutation and recombination rates). It is not usually possible to write down an explicit expression for the likelihood of these parameters.

For example, consider an idealised population, evolving according to the *Wright–Fisher model* (see also **Chapter 22**). In other words, the population evolves in non-overlapping generations, with constant size N chromosomes, with the number of offspring of the chromosomes in each generation being symmetric multinomial, independently of their genetic types (i.e. evolution is neutral). Assume a simple mutation model with K possible genetic types, or *alleles*, with the distribution of the type of the offspring of a parent of type i being given by

$$P(\text{offspring of type } j \mid \text{parent of type } i) = (1 - \mu)\delta_{ij} + \mu p_{ij}, \qquad (26.2)$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise, and $P = (p_{ij})$ is a known transition matrix (so p_{ij} is the probability that an allele of type i mutates to type j, given that a mutation occurs). Thus with probability $1 - \mu$ the offspring is an identical copy of the parent, but with probability μ a mutation occurs according to the transition matrix P. Taken together, these demographic and mutation processes specify a model for the evolution of the population, with two unknown parameters, N and μ. The models are amongst the simplest imaginable, avoiding the complications of fluctuating population size, selection or recombination, for example. Nevertheless, if we examine the genetic types $\mathcal{D}$ of a random sample of N chromosomes taken from the population after it has been evolving for a long period of time, and ask, 'how should we calculate the likelihood $L(N, \mu) = P(\mathcal{D} \mid N, \mu)$?' the answer is far from obvious.

In this case it turns out that the probability $P(\mathcal{D} \mid N, \mu)$ actually depends on N and μ through only their product[2] $N\mu$, and not the individual values of N and μ. As a result, it is common to define the *scaled mutation parameter*[3] $\theta = 2N\mu$, and write the probability $P(\mathcal{D} \mid N, \mu)$ as $P(\mathcal{D} \mid \theta)$. We can then concentrate on calculating the likelihood $L(\theta)$ for the single parameter θ. For concreteness, the remainder of the chapter concentrates on this problem, though most of the methods described can be extended to more complex problems with more parameters (see Section 26.4.7 of this chapter).

Although reduction of the model to a single parameter simplifies things slightly, it is still not possible to write down an explicit expression for the likelihood. However, using methods described in later sections we can accurately approximate it. Figure 26.1 shows an approximation of the (log) likelihood surface for data on the β-globin gene from Harding *et al.* (1997). We now consider how such a likelihood surface can be used to estimate the parameter θ and to assess the uncertainty associated with this estimate. There are two common approaches: the maximum-likelihood approach, and the Bayesian approach.

26.1.1.1   *Maximum-likelihood Inference*

Perhaps the most common likelihood-based approach to parameter estimation is to estimate θ by the maximum-likelihood estimate, $\hat{\theta}$, which is that value of θ that maximises

---

[2] Actually this is only an approximation that is derived by assuming N is large, and μ is small, but it is an extremely accurate approximation in practice.

[3] Our definition is based on our earlier definition of N as the number of *chromosomes* in the population. An alternative is to define N to be the number of *individuals* in the population, in which case there are $2N$ chromosomes if the individuals are diploid, and $\theta = 4N\mu$.
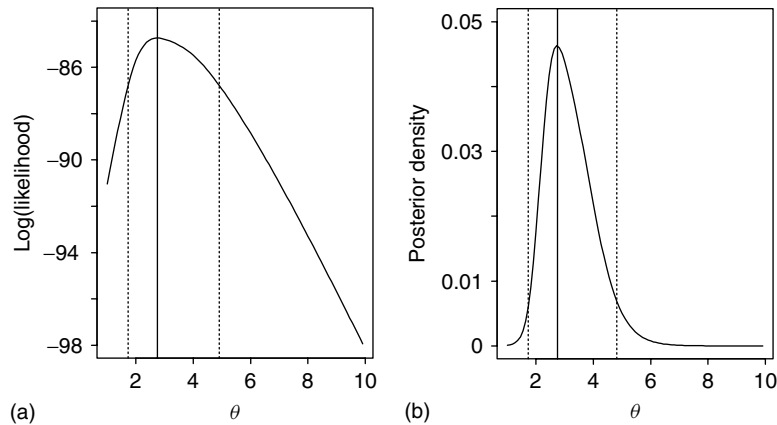
Figure 26.1 (a) Estimated log likelihood for the β-globin data from Harding *et al.* (1997). The vertical solid line indicates the maximum-likelihood estimate, while the vertical dotted lines indicate the approximate 95 % confidence interval found by taking values of θ for which the likelihood is within 2 log likelihood units of the maximum. (b) Posterior density of θ for the same data, with uniform prior on θ. The vertical solid line indicates the posterior mode, while the vertical dotted lines indicate a 95 % credible region. In this case, the 95 % confidence interval and the 95 % credible region are almost identical.

the likelihood $L(\theta)$. The uncertainty associated with this estimate is typically expressed by giving a 95 % *confidence region* for θ, which is a set of values of θ that has the rather tricky and non-intuitive interpretation that if you sampled data from the model repeatedly, and created confidence regions in the same way, for 95 % of datasets, the region would contain the true value of θ. The trickiness of this interpretation must be the source of more confusion than almost any other basic statistical concept. Without dwelling on the issue, we note that it is *not* correct to say that the probability that θ lies in the confidence region is 0.95, which is the intuitive interpretation most people would like to place on the interval: in order to make such statements, one must take a Bayesian approach, as outlined below.

There is a rather general and elegant statistical theory underlying the maximum-likelihood approach to inference. The theory states that *under suitable conditions* $\widehat{\theta}$ will be asymptotically[4] normally distributed with mean being the 'true' value of θ, and that the log likelihood ratio statistic

$$\Lambda = -2 \log \frac{L(\theta_0)}{L(\widehat{\theta})}, \tag{26.3}$$

has asymptotically a χ-squared distribution, under repeated sampling from the model, if θ_0 is the 'true' value of θ. A useful consequence of this is that an approximate 95 % confidence interval for a one-dimensional parameter θ may be obtained by considering those θ for which the log likelihood is within two units of the maximum log likelihood.

---

[4] The term *asymptotically* here means 'as the amount of data tends to infinity'. The implicit hope is that asymptotic results will be good approximations for reasonably large amounts of data. This is, of course, not always the case.

The maximum likelihood estimate of θ for the β-globin data (Figure 26.1) is $\theta = 2.75$, and an approximate 95 % confidence interval found in this way is [1.73, 4.91].

Note that we have emphasised the phrase 'under suitable conditions' in the previous paragraph. This is because in the types of applications considered here, these suitable conditions often fail to hold, destroying the validity of the asymptotic theory, and making the construction of valid confidence intervals much more difficult. The reason the conditions often fail to hold in genetics applications is that, unlike most statistical problems where observed data are independent, the types of randomly sampled chromosomes are *not* independent, owing to the fact that they share ancestry. In the same way that you and your parents are related, and your genetic types are not independent, randomly sampled chromosomes are also (more distantly) related, and their types are not independent. These relationships between sampled chromosomes are of course exactly those that lead to the coalescent. While this lack of independence does not in itself necessarily preclude the standard theory from applying, in many settings it is known that the standard theory does not hold, and in other settings it is unclear. The standard method for obtaining confidence intervals from the likelihood, and indeed any procedure that relies on the asymptotic χ-squared distribution of the likelihood ratio statistic, must therefore be used with caution in population genetics applications.

Proponents of the maximum-likelihood approach sometimes claim that it has the advantage (over the Bayesian approach described below) of 'objectivity', in that the opinion of the researcher should not affect the results obtained. However, this is only partially true since specification of the likelihood function itself typically requires subjective judgements to be made about what to include in the model. For example, is it plausible to assume that all sites mutate at an equal rate? Should we assume that the population has grown linearly or exponentially? Can we ignore selection and/or recombination? The answers to such questions are often unclear, and in practice it is necessary to make some subjective modelling assumptions in order to proceed. Furthermore, the maximum-likelihood approach provides no coherent way for making use of information we have about parameters in the model from sources other than the data: from archaeological or anthropological sources, or previous genetics studies for example. The Bayesian approach to inference allows such information to be included in the analysis.

26.1.1.2 Bayesian Inference

Bayesian methods also make use of the likelihood in performing inference for parameters in the model, but they allow (indeed require) the incorporation of 'prior information' about the model parameters. Formally, information about the parameters θ must be expressed by specifying a *prior* or *pre-data* distribution $P(\theta)$ for θ. The distribution $P(\theta)$ is weighted towards those θ that are considered most likely according to our prior information. This prior information is then combined with the likelihood by multiplying them together to give the *posterior* or *post-data* distribution $P(\theta \mid \mathcal{D})$:

$$P(\theta \mid \mathcal{D}) = L(\theta)P(\theta)/P(\mathcal{D}). \tag{26.4}$$

From this equation we see that $P(\theta \mid \mathcal{D})$ will be large for values of θ that are both well-supported by the data (i.e. have high likelihood) *and* are consistent with our prior information (high $P(\theta)$).

The post-data distribution represents our beliefs about the parameters, taking into account both our prior information and the observed data. It is often convenient to summarise these beliefs in some way. For example, it is common practice to report a point estimate for θ (usually the mode of the post-data distribution) and to specify a 95 % credible region for θ, which is a set of values Θ satisfying $P(\theta \in \Theta \mid \mathcal{D}) = 0.95$. Unlike a confidence region, a 95 % credible region has the natural interpretation that the probability that θ is in the region is 0.95. Another nice feature is that the Bayesian approach does not rely on asymptotic arguments, and so is valid in settings where the standard likelihood theory fails.

For the β-globin data, taking our prior distribution on θ to be uniform in the range $0-10$, thus favouring no particular value of θ in that range above any others,[5] the posterior distribution of θ is as shown in Figure 26.1. The mode of the distribution is at $\theta = 2.75$, and the 95 % credible region with the highest posterior probability is [1.73, 4.82]. Thus, in this case, the Bayesian and maximum-likelihood approaches provide similar results. In later examples, results from the two approaches differ rather more than they do here.

Despite the apparent advantages of the Bayesian approach over the Maximum-likelihood approach, there is still considerable resistance to it from some quarters. Concerns are often centred on the 'subjective' nature of the Bayesian method. Since the post-data distribution depends on the specified prior distribution, different researchers with differing prior beliefs may come to differing conclusions. In some cases, given enough data, the post-data distribution is dominated by the likelihood, and the conclusions drawn become relatively insensitive to the prior distribution used. Nevertheless, in practice it is often the case that conclusions *do* depend quite sensitively on the prior distribution, and it is important to recognise this fact. Such a result may seem unsatisfactory, but is simply a reflection of the fact that there is sometimes insufficient information in the data to make robust inferential statements.

It may be clear from the above discussion that the author's preferences lean towards the Bayesian approach to inference. However, the methods we examine may be used to compute likelihood surfaces, and thus allow either a Bayesian or maximum-likelihood approach to be taken.

26.2   THE LIKELIHOOD AND THE COALESCENT

In the following sections, we consider methods for accurately approximating the likelihood $L(\theta)$. These methods make use of ideas from coalescent theory, and those unfamiliar with these ideas will find it helpful to study the companion **Chapter 25**, before proceeding.

The coalescent, and related processes, describe the distribution of the unknown tree $\mathcal{T}$ relating a random sample of chromosomes. Here we wish to be deliberately vague about what we mean by the 'tree': it might be simply the genealogy relating the sampled chromosomes (see, e.g. Figure 26.3 and accompanying text in **Chapter 25**), or it might also include the mutations on the branches of the genealogy. What matters

---

[5] This prior is convenient for the purposes of illustration. As we discuss later, rather than placing a prior distribution on $\theta = 2N\mu$, it makes more sense to consider priors for N and μ separately.

is that if we knew the tree then we could calculate[6] the probability of the data $P(\mathcal{D} \mid \mathcal{T}, \theta)$.

We emphasise that although the tree relating the sampled individuals plays a major role in the computational methods we are considering here, it is a role that is different to that of the tree in phylogenetic analyses (see for example Huelsenbeck, 2001 this volume). In population genetics applications, interest typically focuses on the genetic and evolutionary forces that have affected the evolution of the genetic region under study, and the tree relating randomly sampled individuals from those populations is of interest only in so far as it informs us about these parameters. Nevertheless, many population genetics studies have focused on the problem of estimating quantities relating to the tree (i.e. ancestral inference). In particular, estimating the time since the most recent common ancestor, T_{MRCA}, of the sampled chromosomes, and the ages of mutations in the tree, have become common objectives.[7] Formally, these questions involve the post-data distribution[8] of the tree, $P(\mathcal{T} \mid \mathcal{D})$, which is often easy to find using methods we describe later. We anticipate that in the future focus will shift away from estimating the tree, and towards estimating parameters in a model for the demographic history of populations under study (e.g. times of major migration events in human history).

In attempting to approximate the likelihood $L(\theta)$, it turns out to be helpful to write it as an average over all possible trees:

$$L(\theta) = P(\mathcal{D} \mid \theta) = \sum_{\mathcal{T}} P(\mathcal{D} \mid \mathcal{T}, \theta) P(\mathcal{T} \mid \theta). \tag{26.5}$$

This is helpful because every term in this sum is easy to calculate: we noted earlier that we can calculate $P(\mathcal{D} \mid \mathcal{T}, \theta)$, and the pre-data distribution of the tree relating the sampled individuals, $P(\mathcal{T} \mid \theta)$, is given by basic coalescent theory (**Chapter 25**). However, in most cases (26.5) cannot be directly used to evaluate the likelihood, because the number of terms in the sum is so huge that it simply cannot be calculated in a reasonable amount of time. For example, the number of tree topologies relating 10 chromosomes is 2571912000.

In fact, since the branch lengths of the tree $\mathcal{T}$ are generally continuous quantities, the sum above ought really to be written as an integral:

$$L(\theta) = P(\mathcal{D} \mid \theta) = \int P(\mathcal{D} \mid \mathcal{T}, \theta) P(\mathcal{T} \mid \theta) \, d\mathcal{T}. \tag{26.6}$$

This makes exact computation of the likelihood even harder, and most of the remainder of this chapter is spent examining efficient methods of approximating this integral, and others like it. Note that the integration takes place over the space of all possible $\mathcal{T}$, which is a big space! Numerical integration methods, such as Gaussian quadrature, which typically work well for integrals in one or two dimensions, are difficult to apply to problems such as

---

[6] In the case where $\mathcal{T}$ is the genealogy of the sampled chromosomes this calculation can be done using the *peeling* algorithm (Felsenstein, 1981).

[7] As noted by Wilson and Balding (1998), while the T_{MRCA} of a sample of chromosomes may not be the most important time in human history, it has become central to the interpretation of genetic samples through its widespread use.

[8] In the coalescent framework, the tree $\mathcal{T}$ is the result of a random process (the evolution of the population), and not a parameter in the model. This is true even when taking the maximum-likelihood approach to inference. Thus, it is technically correct to talk of the distribution of $\mathcal{T}$, rather than the likelihood of $\mathcal{T}$.

these, and the approximation of such high-dimensional integrals is now most commonly performed using Monte Carlo (i.e. simulation-based) methods.

A naive Monte Carlo method of approximating (26.6) is based on the idea that, if a random variable X has density $f_X(x)$, then the mean of any function of X, $g(X)$ say, may be approximated by simulating many values $X^{(1)}, X^{(2)}, \ldots, X^{(M)}$ from the distribution with density f_X, and forming the average:

$$E(g(X)) = \int g(x) f_X(x) \, dx \approx \frac{1}{M} \sum_{i=1}^{M} g(X^{(i)}).$$ (26.7)

With suitable conditions on the function g, this approximation will be good provided M is sufficiently large (formally the error tends to 0 as M tends to infinity). Applying this idea to (26.6) gives

$$L(\theta) = \int P(\mathcal{D} \mid \mathcal{T}, \theta) P(\mathcal{T} \mid \theta) \, d\mathcal{T} \approx \frac{1}{M} \sum_{i=1}^{M} P(\mathcal{D} \mid \mathcal{T}^{(i)}, \theta),$$ (26.8)

where[9] $\mathcal{T}^{(1)}, \mathcal{T}^{(2)}, \ldots, \mathcal{T}^{(M)} \sim P(\mathcal{T} \mid \theta)$. The approximation (26.8) is easy to implement, since $P(\mathcal{T} \mid \theta)$ can be simulated from using coalescent methods, and $P(\mathcal{D} \mid \mathcal{T}^{(i)}, \theta)$ is easily calculated. Unfortunately, the approximation is also almost useless for problems involving samples of more than a few chromosomes. The reason for this is that $P(\mathcal{D} \mid \mathcal{T}^{(i)}, \theta)$ will be very small for all but a few of the $\mathcal{T}^{(i)}$ we simulate. Only the few larger values will contribute significantly to the sum (26.8), but almost all of our effort will be spent calculating the very small terms that contribute negligible amounts to the sum. Typically, the proportion of trees that actually contribute significantly to the sum is less than one in a million, and in such cases the method becomes hopelessly inefficient.

26.3 IMPORTANCE SAMPLING

Importance sampling (IS) (see Ripley, 1987, for background) is a standard statistical method for improving the efficiency of Monte Carlo integration. The idea is simple, and based on rectifying the inefficiency of the approximation (26.8) by concentrating simulation and computational effort on the more 'important' trees, which are those trees for which $P(\mathcal{D} \mid \mathcal{T}, \theta)$ is relatively large, or in other words those trees that are more consistent with the observed data. IS was first used in this context by Griffiths and Tavaré (1994a; 1994b; 1994c), though they derived their method in a rather different way, as a method of solving recursive equations. Similar methods have since been applied, by themselves and others, to a variety of genetic systems and demographic models (see, for example, Griffiths and Marjoram, 1996; Griffiths and Tavaré, 1997; 1999; Nielsen, 1997; Bahlo and Griffiths, 2000). The connection between the Griffiths–Tavaré method and IS was pointed out by Felsenstein *et al.* (1999), and Stephens and Donnelly (2000) show how this observation can be exploited to develop substantially more efficient algorithms.

---

[9] The notation $\sim$ in what follows means 'are distributed as'.

The IS method is based on rewriting the integral (26.6) as

$$L(\theta) = \int P(\mathcal{D} \mid \mathcal{T}, \theta) P(\mathcal{T} \mid \theta) \, d\mathcal{T}$$

$$= \int P(\mathcal{D} \mid \mathcal{T}, \theta) \frac{P(\mathcal{T} \mid \theta)}{Q(\mathcal{T})} Q(\mathcal{T}) \, d\mathcal{T}, \tag{26.9}$$

where $Q(\cdot)$ is any distribution satisfying the condition

$$Q(\mathcal{T}) > 0 \text{ whenever } P(\mathcal{D} \mid \mathcal{T}, \theta) P(\mathcal{T} \mid \theta) > 0. \tag{26.10}$$

(This condition is required when multiplying by $Q(\mathcal{T})/Q(\mathcal{T})$ to obtain (26.9) above, to avoid multiplying by 0/0, which is undefined.) The distribution $Q(\cdot)$ is referred to as the IS distribution, or proposal distribution. Straightforward Monte Carlo approximation of (26.9) gives

$$L(\theta) \approx \frac{1}{M} \sum_{i=1}^{M} P(\mathcal{D} \mid \mathcal{T}^{(i)}, \theta) \frac{P(\mathcal{T}^{(i)} \mid \theta)}{Q(\mathcal{T}^{(i)})}, \tag{26.11}$$

where $\mathcal{T}^{(1)}, \mathcal{T}^{(2)}, \ldots, \mathcal{T}^{(M)} \sim Q(\mathcal{T})$.

By choosing the distribution Q carefully, this method of approximation is much more efficient than (26.8). For example, when estimating $L(\theta)$ the optimal choice Q_θ^* for Q depends on θ and is the post-data distribution of the tree $\mathcal{T}$ given the genetic data $\mathcal{D}$ and the parameter θ:

$$Q_\theta^*(\mathcal{T}) = P(\mathcal{T} \mid \mathcal{D}, \theta) = \frac{P(\mathcal{T} \mid \theta) P(\mathcal{D} \mid \mathcal{T}, \theta)}{P(\mathcal{D} \mid \theta)} = \frac{P(\mathcal{T}, \mathcal{D} \mid \theta)}{P(\mathcal{D} \mid \theta)}. \tag{26.12}$$

Intuitively this choice of Q directs more sampling and computational effort towards trees that have larger values of $P(\mathcal{D} \mid \mathcal{T}, \theta)$, and thus avoids the inefficiency associated with the naive approximation (26.8) discussed earlier. In fact, for $Q = Q_\theta^*$, every term in the sum (26.11) is the same:

$$\frac{P(\mathcal{D} \mid \mathcal{T}^{(i)}, \theta) P(\mathcal{T}^{(i)} \mid \theta)}{Q_\theta^*(\mathcal{T}^{(i)})} = \frac{P(\mathcal{D}, \mathcal{T}^{(i)} \mid \theta)}{P(\mathcal{D}, \mathcal{T}^{(i)} \mid \theta)/P(\mathcal{D} \mid \theta)} = L(\theta), \tag{26.13}$$

and the variance of the estimator (26.11) is 0, or in other words the approximation becomes exact. This seems almost too good to be true, and in most situations this is indeed the case, the problem being that *we do not know the distribution required in* (26.12). In fact, to implement the estimate (26.11) we must be able to do two things:

1. Simulate (directly) from $Q(\cdot)$.

2. Calculate $Q(\mathcal{T}^{(i)})$ for each $\mathcal{T}^{(i)}$.

For $Q = Q_\theta^*$ we are not able to do either of these things.

Nevertheless, the observation that $Q_\theta^*(\mathcal{T}) = P(\mathcal{T} \mid \mathcal{D}, \theta)$ is helpful, as it gives us an insight into which choices of Q may be sensible. In particular, it seems that when attempting to estimate $L(\theta)$ at a particular value of θ, a good strategy would be to choose Q to closely approximate Q_θ^*. This is the strategy pursued in Stephens and Donnelly

(2000), who found it to be a very successful way of developing good IS methods in many cases.

26.3.1 Likelihood Surfaces

The approximation (26.11) allows us to approximate the likelihood surface $L(\theta)$ at all values of θ, using samples from a single $Q(\mathcal{T})$. However, since the optimal IS function (26.12) depends on θ, it is clear that no one choice of Q can be optimal for all θ simultaneously, and that the efficiency of the resulting estimator for $L(\theta)$ may vary with θ. Suppose that we have constructed an IS function $Q_{\theta_0}(\mathcal{T})$ to approximate $Q_{\theta_0}^*(\mathcal{T})$, for some fixed value θ_0, and consider using this IS function to estimate the likelihood curve $L(\theta)$ at all values of θ, using (26.11). A strategy along these lines was adopted in Griffiths and Tavaré (1994c), which refers to θ_0 as the 'driving value' for θ, and has since been frequently employed in implementations of IS schemes. Intuitively we might expect this IS function to be most efficient in estimating $L(\theta)$ for values of θ near θ_0, and less efficient for θ far away from θ_0. For reasons discussed in Stephens and Donnelly (2000), this can tend to cause such methods to underestimate the likelihood for values of θ away from the driving value θ_0. This can have two rather undesirable consequences. First, it can cause the estimate of the likelihood surface to have its maximum near θ_0, even when the true likelihood surface has its maximum elsewhere. Second, if the driving value θ_0 is near the maximum of the true likelihood surface then the estimate of the likelihood may tend to be more highly peaked about θ_0 than the true likelihood.[10] The severity of these effects is rather difficult to predict: in some cases they cause few problems, while in others the effects can be extreme. It is important to bear this in mind when using these methods.

In order to avoid such problems we might try using a different IS function to estimate the likelihood at each value of θ. In other words, we might use an IS function Q_{θ_i} to estimate the likelihood $L(\theta_i)$ on a grid of values, $\theta = \theta_1, \theta_2, \ldots, \theta_R$, say. In order to obtain efficient estimates at all values of θ we need Q_{θ_i} to be a good approximation to the optimal IS function $Q_{\theta_i}^*$. This 'point-wise' approach to estimating the likelihood surface is somewhat wasteful in that it requires a set of samples from Q_{θ_i} for each i, but at each point in the grid it uses only one of these samples to estimate the likelihood. A more efficient approach, which is in some sense a combination of the 'point-wise' and 'driving value' approaches, is to use a sample from a single IS function

$$Q(\mathcal{T}) = \frac{1}{R} \sum_{i=1}^{R} Q_{\theta_i}(\mathcal{T}), \qquad (26.14)$$

to estimate the likelihood surface at a grid of values of θ. A stratified sample from this IS function could be obtained by pooling samples of equal size from each of the $Q_{\theta_i}(\mathcal{T})$. It can be shown that using the IS function (26.14) is guaranteed to be more efficient (by some reasonable criterion) than using the point-wise approach. However, while this more efficient approach is straightforward in principle, it needs to be coded into the software being used. Where this is not the case, it is advisable to at least compare estimates of the likelihood surfaces using different driving values in order to check that this is not causing

---

[10] This is important because it would cause estimated confidence intervals, or credible regions to be too small.

difficulties. If the surfaces found using different driving values are very different, then a 'point-wise' approach may be safer, if rather time consuming.

26.3.2   Ancestral Inference

IS schemes also provide a straightforward way of approximating the post-data distribution of the tree $P(\mathcal{T} \mid \mathcal{D}, \theta)$. Define the *weight* $w^{(i)}$ associated with tree $\mathcal{T}^{(i)}$ by

$$w^{(i)} = \frac{W^{(i)}}{\sum_{j=1}^{M} W^{(j)}}, \tag{26.15}$$

where

$$W^{(i)} = P(\mathcal{D} \mid \mathcal{T}^{(i)}, \theta) \frac{P(\mathcal{T}^{(i)} \mid \theta)}{Q(\mathcal{T}^{(i)})}. \tag{26.16}$$

It is easy to show that the distribution with weight $w^{(i)}$ on tree $\mathcal{T}^{(i)}$ is an approximation to the post-data distribution $P(\mathcal{T} \mid \mathcal{D}, \theta)$. As a result it is straightforward to approximate quantities of interest relating to the tree by forming a weighted average of these quantities over the sampled trees. For example, the expected value of T_{MRCA} can be approximated by

$$E(T_{\text{MRCA}}) \approx \sum_{i} w^{(i)} T_{\text{MRCA}}(\mathcal{T}^{(i)}). \tag{26.17}$$

26.3.3   Application and Assessing Reliability

One of the most important considerations when applying computationally intensive methods such as these is whether the results obtained are reliable. When applying IS methods, this comes down to whether the number of iterations M is sufficiently large for the approximations being made to be reasonably accurate. Unfortunately the value of M required varies drastically from problem to problem, and so no simple answer exists to the question, 'How many iterations must I use?'. However, a simple and generally effective procedure is to run the algorithm several times with different seeds for the pseudo-random number generator, and to check that the same results are obtained each time. If the results differ greatly with each run then the runs are too short. Here we confine ourselves to a single illustrative example: a more detailed discussion and more examples can be found in Stephens and Donnelly (2000), and in Fearnhead and Donnelly (2001), both of which give illuminating examples of how inaccurate results can be, even with large amounts of computing time, and how these inaccuracies are often evident in multiple runs of an algorithm.

Our illustration comes from applying the Griffiths–Tavaré method, implemented in the program `genetree`, to the β-globin data from Harding *et al.* (1997), where a more comprehensive analysis and discussion of the biological significance of the results can be found. Five different approximations to the likelihood surface, based on runs of length $M = 10\,000$, using a driving value[11] of $\theta = 5.0$, are shown in Figure 26.2(a). The

---

[11] Earlier we warned that care must be taken when using a 'driving value' to approximate the likelihood surface. In this case we checked that using point-wise estimates, or different driving values, gives very similar answers. In general using a driving value approach with `genetree` tends to give reasonable results in models without migration or growth. However, once growth and migration are included in a model more care may be required.
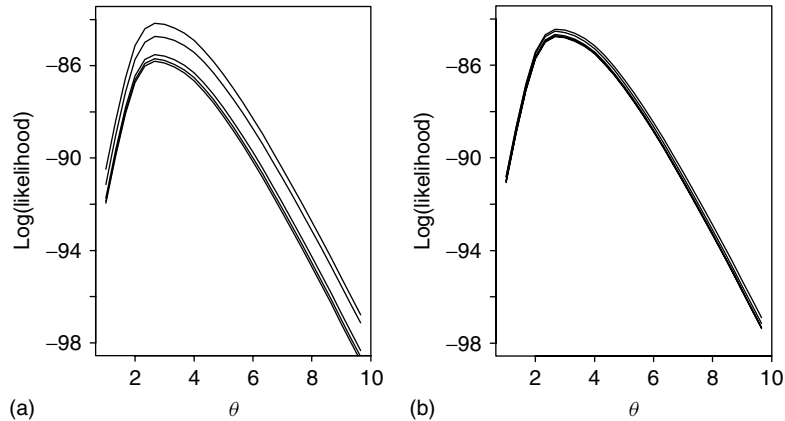
Figure 26.2 Estimated log likelihoods for the β-globin data from Harding *et al.* (1997). (a) Five different curves, each obtained using a different seed for the random number generator and $M = 10\,000$ iterations. (b) Five different curves, each obtained using a different seed for the random number generator and $M = 1\,000\,000$ iterations.

reasonably large differences between the runs indicate that the runs are perhaps rather too short to give reliable results. Five approximations based on runs of length $1\,000\,000$ shown in part (b) of the figure show much less variation. We conclude that for these data `genetree` requires perhaps millions of iterations to deliver reliable results. This sort of requirement is typical. Indeed for problems involving migration between subpopulations and/or growth, many more iterations are typically required, which can result in the method becoming prohibitively time consuming (perhaps months or even years of computer time). In order to tackle such problems, improved IS methods along the lines of Stephens and Donnelly (2000) may be required.

A notable feature of the curves in Figure 26.2 is that although many of the curves are at different *heights*, they are all a similar 'shape', and peak at the same place. It is not really well understood why exactly this occurs, but it often appears to be the case, at least in settings not involving population growth, migration or recombination. It is important to stress that when sufficient iterations have been performed, all curves should in theory be both the same height *and* shape. If they are not all the same height, it is a sign that the algorithm has not been run for long enough. In some cases, as here, the results obtained when the algorithm *has* been run long enough do not differ in most practical terms from the results obtained from a short run. However, in other cases they may differ, and you can only find out by doing the longer runs!

26.4 MARKOV CHAIN MONTE CARLO

26.4.1 Introduction

MCMC has recently become a very popular technique in computational statistics. Essentially it is a method for producing samples from a distribution that is not easy to simulate directly. For example, we noted that to answer questions about the tree

relating the sampled chromosomes, it would be helpful to simulate from the conditional distribution of the tree given the data, $P(\mathcal{T} \mid \mathcal{D})$. Although this is difficult to achieve directly, it is relatively straightforward, at least in principle, using MCMC methods. Further, as we will also see later, these methods can also be used to approximate the likelihood surface for θ. However, we begin by explaining the basics of MCMC techniques.

Suppose we wish to generate trees from some distribution $\pi(\mathcal{T})$ that is difficult (or impossible) to simulate from directly (e.g. $\pi(\mathcal{T}) = P(\mathcal{T} \mid \mathcal{D})$). Informally, MCMC methods achieve this by starting at some initial tree $\mathcal{T}^{(0)}$, and moving randomly from tree to tree in such a way that, in the long run, the frequency with which trees are visited is proportional to $\pi(\mathcal{T})$. The question that MCMC methods answer is how can we move randomly from genealogy to genealogy so that this is achieved? Here we consider perhaps the most important and widely used MCMC method: the Metropolis–Hastings algorithm. The idea is that rather arbitrary methods of moving randomly from genealogy to genealogy can be modified to create an algorithm with the required properties. Let $\mathcal{Q}$ denote some method of moving randomly from tree to tree, with $\mathcal{Q}(\mathcal{T} \to \mathcal{T}')$ being the probability that, if we start at $\mathcal{T}$ then we move to $\mathcal{T}'$. The following algorithm ensures that, in the long run, the frequency with which trees are visited is proportional to $\pi(\mathcal{T})$.

Starting at some initial point, iterate the following steps:

1. Given the ith genealogy $\mathcal{T}^{(i)}$, draw a genealogy $\mathcal{T}'$ from $\mathcal{Q}(\mathcal{T}^{(i)} \to \mathcal{T}')$.

2. With probability A, given by

$$A = \min\left(1, \frac{\pi(\mathcal{T}')\mathcal{Q}(\mathcal{T}' \to \mathcal{T}^{(i)})}{\pi(\mathcal{T}^{(i)})\mathcal{Q}(\mathcal{T}^{(i)} \to \mathcal{T}')}\right), \tag{26.18}$$

accept the proposed genealogy, and set $\mathcal{T}^{(i+1)} = \mathcal{T}'$. Otherwise *reject* the proposed genealogy and set $\mathcal{T}^{(i+1)} = \mathcal{T}^{(i)}$.

Note that when a proposed genealogy is 'rejected' in the above algorithm, the new tree is the same as the current tree. It is important that these duplicate trees are counted, and not simply discarded. The distribution $\mathcal{Q}$ is often referred to as the *proposal distribution*, and the probability A is referred to as the *acceptance probability*. Calculating A requires being able to calculate the transition probabilites $\mathcal{Q}$ (which is usually easy) and the ratio $\pi(\mathcal{T}')/\pi(\mathcal{T})$ which is also often easy, even if $\pi(\mathcal{T})$ itself cannot be calculated. For example, in the case $\pi(\mathcal{T}) = P(\mathcal{T} \mid \mathcal{D}, \theta)$ we have

$$\frac{\pi(\mathcal{T}')}{\pi(\mathcal{T})} = \frac{P(\mathcal{T}' \mid \mathcal{D}, \theta)}{P(\mathcal{T} \mid \mathcal{D}, \theta)} = \frac{P(\mathcal{T}', \mathcal{D} \mid \theta)}{P(\mathcal{T}, \mathcal{D} \mid \theta)} = \frac{P(\mathcal{T}' \mid \theta)P(\mathcal{D} \mid \mathcal{T}', \theta)}{P(\mathcal{T} \mid \theta)P(\mathcal{D} \mid \mathcal{T}, \theta)}, \tag{26.19}$$

which we can calculate just as we could calculate the terms of the sum (26.5).

Under relatively weak conditions on $\mathcal{Q}$ (see Gilks *et al.*, 1996, for example), which are usually easily satisfied in practice, it is straightforward to show that $\mathcal{T}^{(1)}, \mathcal{T}^{(2)}, \mathcal{T}^{(3)}, \ldots$ form a Markov Chain with stationary distribution $\pi(\cdot)$, and that $\mathcal{T}^{(b)}$, for sufficiently large b, is approximately a sample from $\pi(\cdot)$, the distribution we wished to sample from. Furthermore, for sufficiently large k, $\mathcal{T}^{(b)}, \mathcal{T}^{(b+k)}, \mathcal{T}^{(b+2k)}, \cdots$ may be treated as

approximately independent samples from $\pi(\cdot)$. The value b is known as the *burn-in*, while k is referred to as the *thinning interval*.

Why does this algorithm work? Some helpful intuition can be obtained by considering the special case where the proposal distribution Q is symmetric, in that $Q(T \to T') = Q(T' \to T)$ for all T and T'. In this case, the acceptance probability for moving from T to T' becomes $A = \min(1, \pi(T')/\pi(T))$. Thus, moves that are accepted tend to be *to* trees for which $\pi(T')$ is relatively large, and moves that are rejected tend to be *from* trees for which $\pi(T)$ is relatively large. As a Result, the algorithm tends to spend more time exploring those trees with large values of π than those with small values, which is at least a necessary condition for it to work.

26.4.2 Choosing a Good Proposal Distribution

Although in theory the above algorithm works for *any* choice of proposal distribution Q that satisfies the required conditions, in practice choice of Q can have a very strong effect on the efficiency of the method, or in other words on what values of b and k are 'sufficiently large'. Even for good choices of Q, the values of b and k required can be large for problems of moderate size: tens of thousands, or millions, for example. For bad choices of Q, the values of b and k required are so large as to make the approach infeasible. How to choose a good proposal Q is therefore an important issue.

Unfortunately, it is often difficult to predict in advance whether one particular Q will perform better than another, although there are some general guidelines. For example, it is desirable that Q at least occasionally proposes moves to 'good' trees, which are those with a high value for π. Otherwise almost every proposed tree will be rejected, and the algorithm will remain stuck where it is. Similarly, algorithms that always propose only small changes to the current tree will tend to explore the set of all possible trees rather slowly. Algorithms that exhibit this kind of behaviour are sometimes called *sticky*. Conversely, algorithms that move freely between very different trees are said to 'mix well'. Ideally then we want our proposal distribution to propose moves that (1) are to trees that are very different from the current tree, and (2) have a high probability of acceptance. It is typically difficult to find a proposal that achieves both these aims: proposals that propose big moves tend to have low acceptance rates. A common strategy is to simply experiment with different schemes, trying various *ad hoc* modifications until a scheme that performs well (in that it moves quickly between different plausible trees) is found. This strategy has practical merit, and is acceptable since in principle these *ad hoc* modifications do not affect the validity of the algorithm. Nevertheless it often remains tricky to assess how well a scheme is performing, even after it has been implemented and the results can be examined.

26.4.3 Likelihood Surfaces

There are many ways of using MCMC methods to calculate likelihood surfaces for θ, and we now examine some of these in detail. Although it is possible to use MCMC methods to estimate the likelihood surface itself, in practice it is much easier to estimate a *relative* likelihood surface, that is, a function $\widetilde{L}(\theta)$ that satisfies $\widetilde{L}(\theta) = \alpha L(\theta)$ for some unknown constant α. Knowledge of a relative likelihood surface is sufficient for most applications: in particular it allows the parameters to be estimated using either the maximum-likelihood or Bayesian approach. We distinguish between two types of methods here: those that keep θ fixed and those that allow θ to vary.

26.4.3.1 θ Fixed

Suppose $\mathcal{T}^{(1)}, \mathcal{T}^{(2)}, \ldots, \mathcal{T}^{(M)}$ is an MCMC sample from $P(\mathcal{T} \mid \mathcal{D}, \theta_0)$ for some fixed θ_0. Then we might attempt to estimate the likelihood surface using $P(\mathcal{T} \mid \mathcal{D}, \theta_0)$ as an IS distribution:

$$
\begin{aligned}
L(\theta) &= \int P(\mathcal{D} \mid \theta, \mathcal{T}) P(\mathcal{T} \mid \theta) \, \mathrm{d}\mathcal{T} \\
&= \int \frac{P(\mathcal{D} \mid \theta, \mathcal{T}) P(\mathcal{T} \mid \theta)}{P(\mathcal{T} \mid \mathcal{D}, \theta_0)} P(\mathcal{T} \mid \mathcal{D}, \theta_0) \, \mathrm{d}\mathcal{T} \\
&\approx \frac{1}{M} \sum_{i=1}^{M} \frac{P(\mathcal{D} \mid \theta, \mathcal{T}^{(i)}) P(\mathcal{T}^{(i)} \mid \theta)}{P(\mathcal{T}^{(i)} \mid \mathcal{D}, \theta_0)}.
\end{aligned}
\tag{26.20}
$$

However, this estimator is of little use as we cannot evaluate $P(\mathcal{T}^{(i)} \mid \mathcal{D}, \theta_0)$. In fact we noted earlier that we could not use $P(\mathcal{T} \mid \mathcal{D}, \theta_0)$ as an IS function since we could neither simulate from it directly, nor calculate it explicitly. MCMC methods allow us to simulate from it indirectly, but do not solve the problem of calculating it explicitly. However, by writing

$$
P(\mathcal{T}^{(i)} \mid \mathcal{D}, \theta_0) = \frac{P(\mathcal{T}^{(i)}, \mathcal{D} \mid \theta_0)}{P(\mathcal{D} \mid \theta_0)} = \frac{P(\mathcal{T}^{(i)}, \mathcal{D} \mid \theta_0)}{L(\theta_0)},
\tag{26.21}
$$

and substituting into (26.20) we obtain

$$
\frac{L(\theta)}{L(\theta_0)} \approx \frac{1}{M} \sum_{i=1}^{M} \frac{P(\mathcal{D}, \mathcal{T}^{(i)} \mid \theta)}{P(\mathcal{D}, \mathcal{T}^{(i)} \mid \theta_0)},
\tag{26.22}
$$

which we *can* evaluate, and which gives us an estimate of a relative likelihood surface, $\widetilde{L}(\theta) = \alpha L(\theta)$, where the unknown constant $\alpha = [L(\theta_0)]^{-1}$.

What can be said of the efficiency of the estimate (26.22)? Well, recall that $P(\mathcal{T} \mid \mathcal{D}, \theta_0)$ was identified as the optimal IS function for estimating $L(\theta)$ at $\theta = \theta_0$, but that this efficiency will be reduced for other values of θ. Similarly it is the optimal IS function for estimating $\widetilde{L}(\theta)$ at $\theta = \theta_0$, and efficiency will be reduced for other values of θ. However, the value of $\widetilde{L}(\theta)$ at $\theta = \theta_0$ is known by definition to be $L(\theta_0)/L(\theta_0) = 1.0$, so 'optimality' at this point is very easy to achieve. What is of more concern is the potential lack of efficiency for other values of θ, which can cause severe problems. We repeat the warning that lack of efficiency in these methods needs to be taken seriously, as it can easily lead to estimates of the (relative) likelihood that are wrong by orders of magnitude. In particular, likelihood surfaces estimated in this way may have a peak near θ_0 even when the true likelihood surface has its peak away from θ_0, or, if the true likelihood surface has a peak near θ_0, may be more sharply peaked about θ_0 than the true surface. A short theoretical exploration of this problem in a simple context is given in Stephens (1999).

One approach to addressing this problem is based on the fact that the estimator (26.22) for $\widetilde{L}(\theta)$ will be most efficient, and hence most reliable, for values of θ that are 'close' to θ_0. Suppose we wish to estimate the relative likelihood on a grid of values of $\theta = (\theta_1, \theta_2, \ldots, \theta_R)$, where $\theta_1 < \theta_2 < \cdots < \theta_R$. If θ_2 is close enough to θ_1

then we can accurately estimate $L(\theta_2)/L(\theta_1)$ using (26.22) and a sample of trees from $P(\mathcal{T} \mid \mathcal{D}, \theta_1)$. Similarly, if θ_3 is close enough to θ_2 then we can accurately estimate $L(\theta_3)/L(\theta_2)$ using (26.22) and a sample of trees from $P(\mathcal{T} \mid \mathcal{D}, \theta_2)$. We can then estimate

$$\frac{L(\theta_3)}{L(\theta_1)} = \frac{(L(\theta_3)}{L(\theta_2))} \times \frac{(L(\theta_2)}{L(\theta_1))},$$

and continuing in this way can estimate $L(\theta_i)/L(\theta_1)$ for $i = 1, 2, \ldots, R$, and thus construct an estimate of the relative likelihood surface.

Note that the above procedure requires MCMC samples from $P(\mathcal{T} \mid \mathcal{D}, \theta_i)$ for $i = 1, 2, \ldots, (R-1)$, and thus requires us to run $R-1$ independent Markov chain simulations. In fact, given that we have to generate these samples anyway, there is a more efficient way of combining them than the naive method we have just described. It is an iterative method due to Geyer (1991), who named it 'reverse logistic regression'. Its implementation in this context is described by Kuhner *et al.* (1995).

26.4.3.2 θ Varying

MCMC schemes can also be used to produce an MCMC sample

$$(\theta^{(1)}, \mathcal{T}^{(1)}), (\theta^{(2)}, \mathcal{T}^{(2)}), \ldots, (\theta^{(M)}, \mathcal{T}^{(M)}),$$

from $P(\theta, \mathcal{T} \mid \mathcal{D})$ for any given prior distribution on θ. Instead of simply wandering around different $\mathcal{T}$, now the algorithm must also explore different possible values for θ. One way to achieve this is to define a proposal distribution $\mathcal{Q}$ that proposes a move from the current pair of values $(\theta, \mathcal{T})$ to a new pair of values $(\theta', \mathcal{T}')$, and then accepts or rejects this proposal in the same way as before (replacing $\pi(\mathcal{T})$ with $\pi(\mathcal{T}, \theta)$). More commonly, methods that alternately propose changing $\mathcal{T}$ and changing θ can be developed, as in Wilson and Balding (1998).

This approach arises most naturally from the Bayesian approach to inference, when we would be interested in the post-data distribution for θ. However, since the post-data distribution for θ is proportional to the prior $\pi(\theta)$ times the likelihood $L(\theta)$, it is straightforward to obtain an estimate of a relative likelihood surface from an estimate of the post-data density, simply by dividing by the prior density (which is usually easy to calculate). The simplest (though not the best) method of estimating the post-data density of θ is to plot a histogram of the θ values in an MCMC sample $(\theta^{(1)}, \mathcal{T}^{(1)}), (\theta^{(2)}, \mathcal{T}^{(2)}), \ldots, (\theta^{(M)}, \mathcal{T}^{(M)})$. Thus these methods can provide a useful computational tool for likelihood-based inference in general, and need not be confined to Bayesian analyses. (However, it is worth noting that this approach to obtaining a likelihood surface is probably practical only if the likelihood involves at most two parameters–for larger numbers of parameters it will typically be difficult to obtain reliable estimates of the post-data density from an MCMC sample.)

26.4.4 Ancestral Inference

The use of MCMC methods to perform ancestral inference is relatively straightforward. For example, if $\mathcal{T}^{(1)}, \mathcal{T}^{(2)}, \ldots, \mathcal{T}^{(M)}$ is an MCMC sample from $P(\mathcal{T} \mid \mathcal{D}, \theta)$ for some fixed

value of θ, then the expectation of T_{MRCA} given this value of θ can be approximated by the straightforward Monte Carlo method:

$$E(T_{\mathrm{MRCA}} \mid \mathcal{D}, \theta) = \int T_{\mathrm{MRCA}}(\mathcal{T}) P(\mathcal{T} \mid \mathcal{D}, \theta)\, d\mathcal{T} \qquad (26.23)$$

$$\approx \frac{1}{M} \sum_{i=1}^{M} T_{\mathrm{MRCA}}(\mathcal{T}^{(i)}). \qquad (26.24)$$

Provided the MCMC scheme mixes reasonably well, this approximation will be accurate for moderate values of M. Other quantities of interest relating to the ancestry of the sample, such as the expected age of a particular mutation, can be approximated in a similar way. There is though a slight complication: what value of θ should be used to make these estimates? In a maximum-likelihood framework, it is natural to use the maximum-likelihood estimate for θ. However, this approach ignores the fact that in practice we do not know the true value of θ, and there may be considerable uncertainty associated with any estimate of θ. There does not seem to be an obvious way around this problem without moving to the Bayesian framework, which provides a coherent way of taking into account the uncertainty in our estimate of θ. For example, if $(\mathcal{T}^{(1)}, \theta^{(1)}), (\mathcal{T}^{(2)}, \theta^{(2)}), \ldots, (\mathcal{T}^{(M)}, \theta^{(M)})$ is an MCMC sample from $P(\mathcal{T}, \theta \mid \mathcal{D})$, then

$$E(T_{\mathrm{MRCA}} \mid \mathcal{D}) = \int T_{\mathrm{MRCA}}(\mathcal{T}) P(\mathcal{T} \mid \mathcal{D})\, d\mathcal{T} \qquad (26.25)$$

$$\approx \frac{1}{M} \sum_{i=1}^{M} T_{\mathrm{MRCA}}(\mathcal{T}^{(i)}), \qquad (26.26)$$

is an estimate of the expected value of T_{MRCA} that takes into account the uncertainty associated with θ.

26.4.5   Example Proposal Distributions

As we noted earlier, in principle there is a huge amount of flexibility in the choice of proposal distribution $\mathcal{Q}$. Not surprisingly then, several different suggestions for the choice of suitable $\mathcal{Q}$ have been made. We examine two of the earliest suggestions here. Examples of other MCMC schemes include Beaumont (1999), and Nielsen (2000).

The examples we discuss below are illustrated in Figures 26.3 and 26.4. These figures show trees relating a sample of five individuals, each typed at a single micro-satellite locus, for which each individual's genetic type may be represented by an integer (shown at the bottom of the tree). In what follows, it is worth remembering that alleles that are closer numerically, are likely to be more closely related (so a '7' is likely to be more closely related to a '6' than to a '3').

26.4.5.1   *The Beerli–Kuhner–Yamato–Felsenstein (BKYF) 'Conditional Coalescent' Proposal*

This is the proposal described in Beerli and Felsenstein (1999), and Felsenstein *et al.* (1999). They refer to it as a 'conditional coalescent' proposal, because a new tree is formed from the current tree, by removing one of the branches, and resimulating it from
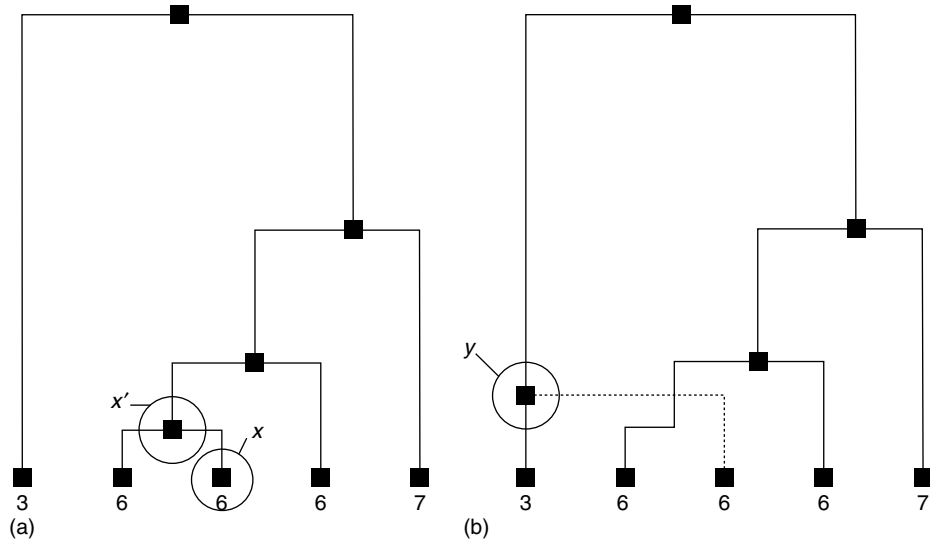
Figure 26.3 Illustration of the BKYF 'conditional coalescent' proposal, which is a proposal distribution for moving from one tree (a) to another (b), as explained in the text.

the appropriate coalescent process, *conditional* on all the other branches staying fixed. For the case of samples from a single panmictic population, it proceeds (in outline) as follows:

1. Choose a node, x, uniformly at random from all nodes other than the root on the current tree.

2. Remove the parent node of x (marked x' in Figure 26.3) from the tree, together with the lineage joining x to x'.

3. Starting from x, simulate a lineage towards the root of the tree. At any given time this lineage coalesces at constant rate with each lineage existing in the tree at that time. (If necessary, the lineage above the current root is extended backwards in time.)

4. Eventually the branch from x will coalesce with another branch, to create a new node (marked y in Figure 26.3).

A more detailed description of this proposal, including a natural extension for samples from several different populations, is given in Beerli and Felsenstein (1999).

26.4.5.2 The Wilson and Balding (WB) 'Branch-swapping' Proposal

In order to design their proposal distribution, Wilson and Balding (1998) augmented their tree to include details of the genetic types at every node. In outline, the proposal distribution proceeds as follows:

1. Choose a node, x, uniformly at random from all nodes other than the root on the current tree.
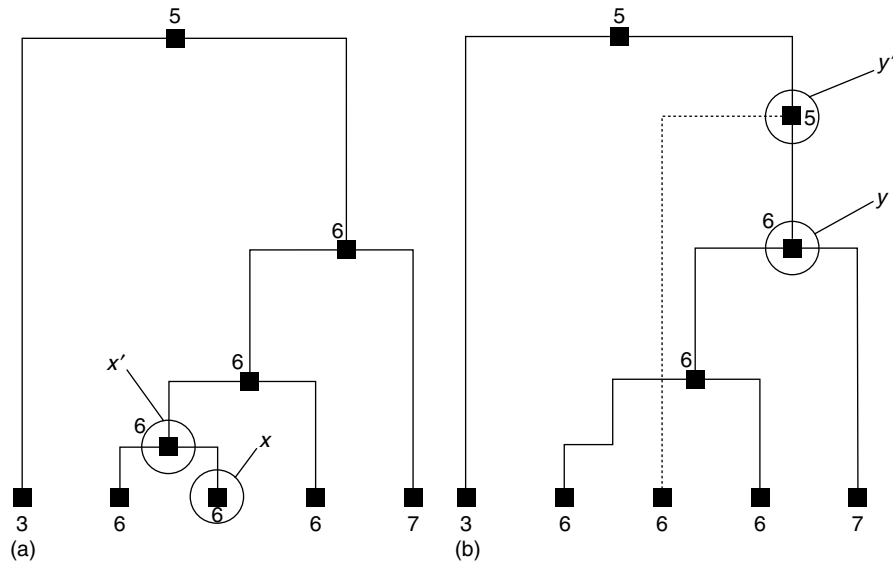
Figure 26.4 Illustration of the Wilson and Balding 'branch-swapping' algorithm, which is a proposal distribution for moving from one tree (a) to another (b), as explained in the text.

2. Remove the parent node of x (marked x' in Figure 26.3) from the tree, together with the lineage joining x to x'.

3. Choose a node, y, above which to attach the branch that formerly attached x to the tree, weighting the choice of y towards nodes that have a type that is similar to the type at x.

4. Attach the branch leading from x somewhere above y, creating a new node (labelled y' in Figure 26.4) where the join occurs.

5. Choose a type label for the newly created node, weighting this choice towards those types that are close to the types at both x and y (i.e. favouring types between the types at x and y).

26.4.5.3 Comparing the BKYF and WB Proposals

The Beerli–Kuhner–Yamato–Felsenstein (BKYF) and Wilson and Balding (WB) proposal distributions have an important qualitative difference. By augmenting the genealogy to include the genetic types at the internal nodes, Wilson and Balding were able to design an 'intelligent' proposal distribution that tends to make sensible choices for destinations to which branches may be moved. In particular, step 3, which tends to reattach the branch leading to x at a sensible point in the tree, would be impossible without information on the types at the internal nodes. In contrast, the behaviour of the BKYF proposal does not even depend on the observed genetic types at the tips of the tree, and as a consequence moves to very 'unlikely' trees may often be proposed (and will usually be rejected). For example, the move illustrated in Figure 26.3 is to a tree that is rather inconsistent with the observed data, and so would most likely be rejected. Such

a move has reasonable probability under the BKYF proposal, but small probability under the WB proposal. One might wonder whether there is ever a good reason for choosing a less 'intelligent' scheme. The answer is that the intelligence of the WB scheme comes at a cost: the expanded sample space, which results from including the types at the internal nodes. In order for the MCMC scheme to explore the space properly it is now necessary for it to explore not only all plausible trees, but also all plausible configurations for the internal node types. Another potential drawback of more 'intelligent' schemes is that more computation may be required to propose each move, and so fewer moves are proposed in a fixed time.

The trick then is to get a good balance between the sizes of the space to be explored, the computational effort required for each proposal, and the 'intelligence' of the proposal distribution. In the author's experience the WB proposal seems to strike this balance nicely for micro-satellite data. In order to apply a similar approach to sequence data, it might be worth including the types at the internal nodes at the segregating sites only, as this should allow intelligent proposals to be designed while keeping the sample space down to a manageable size.

26.4.6 Application and Assessing Reliability

The application of an MCMC scheme is illustrated using the `micsat` program of Wilson and Balding (1998), and the example files (`cooper.inp` and `cooper.dat`), which are supplied with `micsat`. The data is from Cooper *et al.* (1996), and pertains to 60 males from Nigeria, Sardinia and East Anglia, typed at five micro-satellites on the Y chromosome. Wilson and Balding (1998) refers to these data as the NSE dataset.

Wilson and Balding (1998) use `micsat` to provide a Bayesian analysis of these data. In particular, they estimate the post-data distributions of θ and T_{MRCA} given certain priors on N and μ. Histograms of these distributions, based on the output obtained from the example files supplied with `micsat`, are shown in Figure 26.5. (We show the distribution of T_{MRCA} measured in units of generations. This can be converted to years by multiplying by a generation time, as we see later.) As might be expected, the results are very similar
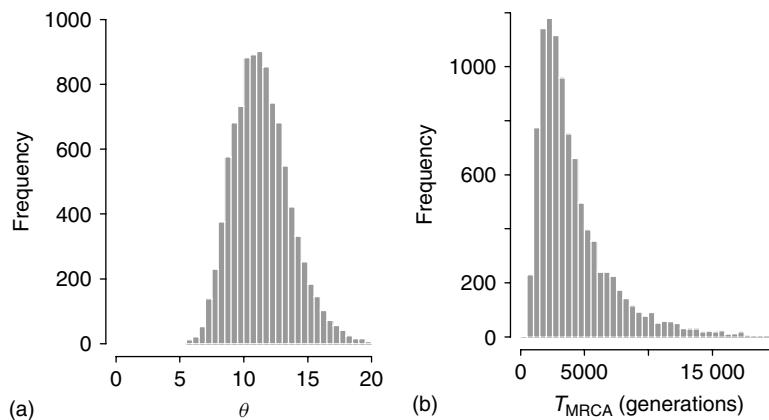
Figure 26.5 Results of a Bayesian analysis of the NSE data, based on output of `micsat` using example files `cooper.inp` and `cooper.dat`. (a) Histogram of post-data distribution of θ. (b) Histogram of post-data distribution of T_{MRCA}.
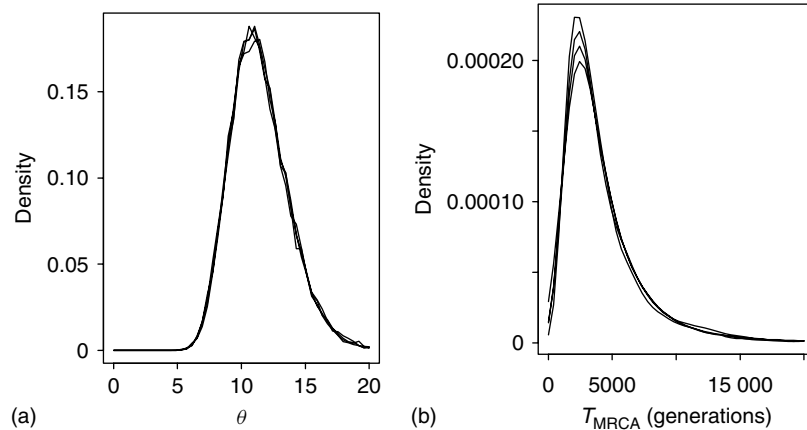
Figure 26.6   Estimates of post-data distribution of (a) θ and (b) T_{MRCA} based on output of `micsat` with example files `cooper.inp` and `cooper.dat`. Each of the four estimates shown is based on a run using a different seed for the pseudo-random number generator (which is set in the file `cooper.inp`). The similarity of the estimates to one another gives grounds for optimism that the MCMC runs are sufficiently long.

to those in Wilson and Balding (1998), which the reader should consult for a discussion of their biological significance. Here we concentrate on some of the statistical issues, the first and most important being *how can we be sure that the MCMC scheme has been run for long enough to give reliable results?* This question should always be asked when results from an MCMC method are being considered. Perhaps the most straightforward and practical method of investigating this is to re-run the algorithm several times, using different seeds for the pseudo-random number generator, and different initial trees, and compare the results obtained. Although more sophisticated diagnostic methods exist (see Brooks and Roberts, 1998, for example), none are foolproof, and this is the approach the author would recommend for the novice user. It has the practical advantage that it can be applied even when the program implementing the MCMC scheme gives the user only a summary of the results, which is sometimes the case (though `micsat` allows you to inspect the raw MCMC sample). Figure 26.6 shows estimated post-data distributions for θ and T_{MRCA} from four further runs of `micsat`, each using a different seed. Very similar results are obtained in each case, giving grounds for optimism that the MCMC runs are sufficiently long.

Another issue is to what extent the results depend on the use of the Bayesian method, and choice of prior. Wilson and Balding (1998) show how their results vary with different choices of prior for the parameters N and μ. It is natural to specify independent priors for N and μ, rather than a prior for θ directly, in view of the following:

1. There is often information about likely values for μ from sources other than the data being considered. For example, Wilson and Balding (1998) based their prior for μ on pedigree studies, in which 3 mutations were observed in 1491 meioses.

2. There is often information about likely values for N from studies at other loci: under the assumption of neutrality all loci should share the same value for N (except of

course that the X chromosome and Y chromosome should have values for N that are respectively 3/4 and 1/4 of the value of N for autosomal loci).

Note also that the priors for N and μ affect the distribution of T_{MRCA}, and that T_{MRCA} cannot be estimated unless there is some information from another source on either N or μ.

Here we compare the estimated distribution of T_{MRCA} obtained with the 'high-variance' priors from Wilson and Balding (1998), with that obtained by a maximum-likelihood approach. An estimate of the (relative) likelihood for θ can be obtained by taking an estimate of the post-data density for θ from the histogram in Figure 26.5, and dividing it by the prior density for θ. Maximising this gives an estimate of the maximum-likelihood estimate for θ, which for these data is around $\theta = 11.0$. In this case, the maximum likelihood estimate is very close to the post-data mode (the peak of the post-data density) for θ, but this is not always the case. On the basis of the observation of 3 mutations in 1491 meioses in a pedigree, the maximum-likelihood estimate of μ is $3/1491 \approx 0.002$, and substituting these point estimates into $\theta = 2N\mu$ gives a point estimate of 2734 for N. The natural maximum likelihood approach to estimating the distribution of T_{MRCA} is to take these estimates as given, which can be achieved using micsat by placing priors on N and μ, which are very peaked about these point estimates.

Finally, in order to convert T_{MRCA} from generations to years, it is necessary to multiply by what we believe to be the generation time. Opinions differ on what the appropriate value for generation time should be. Although 20 years has been commonly used, there is evidence suggesting that the appropriate value may be nearer 27 Weiss (1973) or even higher Tremblay and Vézina (2000). The uncertainty in this value ought really to be taken into account in a proper analysis, and in a Bayesian framework this can be achieved straightforwardly by placing on the generation time. As an illustration, Figure 26.7 shows the distribution of T_{MRCA} if we assume a uniform prior distribution on generation time between 25 and 35 years. This distribution is rather less peaked than is obtained using a point estimate of 27.5 years for the generation time in the maximum-likelihood framework (same figure). This is a consequence of the maximum likelihood method ignoring the inherent uncertainty in estimates of N and μ and in the estimated generation time. For these reasons, the author endorses a Bayesian approach for such questions.

26.4.7 Extensions to More Complex Demographic and Genetic Models

The examples given above have all been restricted to the assumption of a constant-sized panmictic population. Fortunately, almost everything that we have said about the simple case also applies to these more complex cases. Typically, the number of parameters in these more complex models is larger, which makes plotting likelihood surfaces (for example) rather harder, but the basic principles underlying IS and MCMC are unchanged. When dealing with these more complex models, some important questions to bear in mind are as follows:

- Have I run the method long enough to obtain reliable results? As we have stressed, multiple runs using different seeds and starting points can be a big help here.

- Is my model appropriate, and will it allow me to answer the questions I am interested in? For example, if you are interested in historical rates of migration between subpopulations, it makes sense to try to fit a model with migration, rather than a model in which populations split some time in the past, with no subsequent migration.
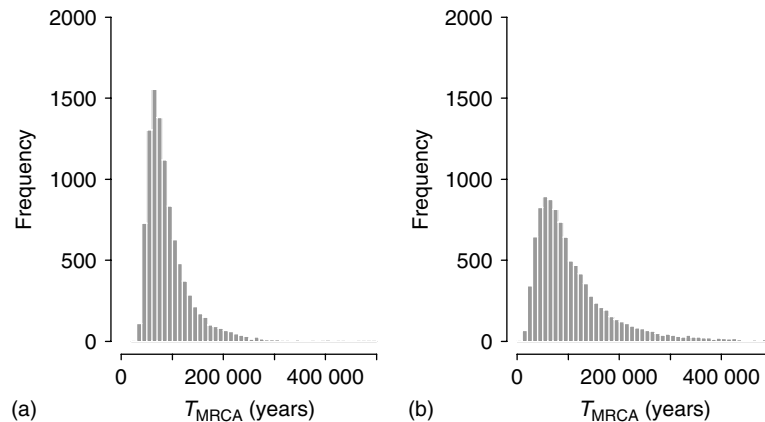
Figure 26.7 Comparison of the post-data distribution of T_{MRCA} for the NSE dataset using maximum-likelihood and Bayesian approaches. (a) Post-data distribution of T_{MRCA} using the maximum-likelihood point estimates, $\theta = 11.0$ and $\mu = 0.02$, and a generation time of 27.5 years. (b) Post-data distribution of T_{MRCA} using a fully Bayesian approach, with the priors on N and μ used by Wilson and Balding (1998), and a uniform distribution on generation time between 20 and 35 years. The Bayesian approach allows for uncertainty in the estimates of μ, N and the generation time, and as a result gives a more diffuse distribution for T_{MRCA}.

26.5   OTHER APPROACHES

As noted in the introduction, the full-data methods that are the main focus of this chapter remain computationally demanding. In many settings they remain computationally intractable, and in other settings they may be tractable, but no software is available implementing them. As a result, there remains considerable interest in alternative approaches that are computationally simpler and (often) easier to implement, without sacrificing too much in the way of accuracy. Here we review some of these methods (see also Marjoram and Tavaré, 2003).

26.5.1   Rejection Sampling and Approximate Bayesian Computation

Rejection sampling provides a convenient way to perform inference based on summaries of the observed data – see Tavaré *et al.* (1997), Li and Fu (1999), and Pritchard *et al.* (1999) for examples of this approach within the context of coalescent-based inference. Rejection sampling has the advantage of being relatively easy to implement, provided one has the ability to simulate data from models of interest. Several software packages are now available for simulating population data under a wide range of different modelling assumptions, including models involving factors such as selection, which can be difficult to incorporate into full-data inference methods. See the section on software for a selection of simulation packages.

In outline, the rejection sampling approach is as follows:

1.   Simulate a value for the parameter θ from its prior distribution.

2.   Simulate data given the value of θ obtained in step 1, from $P(\mathcal{D} \mid \theta)$.

3. Compute a summary, S, of the simulated data, and compare this summary with the same summary computed for the observed data (S_{obs} say). If $S = S_{obs}$ 'accept' (i.e. save) the simulated value of θ; otherwise reject it.

The values of θ that are saved (rather than rejected) in step 3 provide a sample of draws from the posterior distribution of θ given S_{obs}. (The intuition here is that the values are simulated from the prior distribution, and accepted with probability given by the likelihood $P(S_{obs}|\theta)$, resulting in samples from the posterior, which is proportional to the prior times the likelihood.) The above steps are repeated until sufficient samples of θ have been obtained to perform desired inference.

Note that the *statistical* efficiency of this approach depends on whether the summary used contains most of the information in the data regarding θ. In most cases, there is little theory to guide appropriate choice of summary, and we are left with intuition, and experimentation, in order to find 'good' summaries. Typically the choice of good summary will depend heavily on which parameters are of primary interest. For example, the number of polymorphic sites in a region seems likely to contain substantial information about the mutation rate in the region, but less information about the recombination rate. In principle, one can use as complex a summary of the data as one would like – it could involve multiple summary statistics, up to the whole data itself. However, the *computational* efficiency of the approach becomes very poor as the summary S becomes more complex. This is because if the summary is of very high dimension then one will never obtain $S = S_{obs}$ in step 3, and so the algorithm will never produce any accepted values of θ. Computational efficiency also depends on the choice of prior distribution, and will be greatest when the prior distribution is concentrated on areas of the parameter space that are consistent with the observed data. Of course, one should not change one's prior distribution in order to make it consistent with the observed data: the point of the prior is that it encapsulates information available *before* one has observed the data. However, this consideration does mean that one might want to avoid using simple but unrealistic prior distributions that include many values that the user would consider implausible.

Depending on the choice of summary statistic, the rejection sampling approach can be a very simple and effective approach to inference. However, in many cases one will run into the problem, as noted above, of very rarely accepting any simulated values of θ, which makes the approach computationally impractical. One way to alleviate this problem is to relax the condition in step 3 that $S = S_{obs}$, and to accept values of θ where S is 'close to' S_{obs}. In other words, to accept θ when $d(S, S_{obs}) < \delta$ where $d(\cdot, \cdot)$ is some measure of distance between two values of the summary statistic, and δ is some threshold that controls how similar the values need to be in order to accept θ. Typically, one would choose δ as small as possible while keeping the acceptance rate high enough to make the algorithm computationally tractable. This modified rejection sampling scheme is probably more widely used in practice than the standard rejection scheme. Formally, it provides samples from the distribution $P(\theta|d(S, S_{obs}) < \delta)$, rather than the distribution $P(\theta|S = S_{obs})$ sampled from by the standard scheme.

Two other approaches to inference, related to the above rejection sampling scheme, are worth noting. The first, due to Beaumont *et al.* (2002) (see also **Chapter 30**), is based on the ideas of adjusting, and weighting sampled values of θ. The adjustment and weights are designed both to correct the sample so that it more closely follows the distribution $P(\theta|S = S_{obs})$ rather than $P(\theta|d(S, S_{obs}) < \delta)$, and to improve computational efficiency (by weighting the samples, rather than simply accepting or rejecting them). The second,

from Marjoram *et al.* (2003), incorporates the accept/reject steps into an MCMC scheme. One motivation for doing this is to try to avoid problems suffered by rejection sampling in situations where the prior distribution is very different from the posterior distribution for θ. In such cases, as noted above, the standard rejection scheme becomes very inefficient, spending most of its time simulating values of θ from the prior that are completely inconsistent with the observed data, and then rejecting them. The MCMC version can avoid this problem because once it has found values of θ that are consistent with the data it concentrates on proposing (and then accepting or rejecting) near-by values of θ, which will also tend to be somewhat consistent with the data. However, this MCMC approach also introduces potentially serious problems of its own: the acceptance rate for proposed moves can become very small when exploring the tails of the distribution of θ (see Sisson *et al.* (2007), who also suggest a way to circumvent this problem).

26.5.2   Composite Likelihood Methods

Composite likelihood methods are being widely used in recent years, particularly for estimating recombination rates from population data under coalescent-based models. The use of composite likelihoods in this context was pioneered by Hudson (2001), who suggested estimating recombination rates in a region by computing likelihoods for all pairs of available markers, and then multiplying these likelihoods together to form what is known as a '*composite*' *likelihood*. In other words, the composite likelihood for the population recombination rate, ρ, given observed genotype data G, is

$$L_{\text{comp}}(\rho; G) = \prod_{i,j} P(G_i, G_j \mid \rho), \qquad (26.27)$$

where G_i denotes the genotype data at marker i. The great advantage of this approach is that it requires only the ability to compute the probability of data at two markers, which is considerably easier than computing the probability of data at several markers. Indeed, with current computing power, one can accurately estimate the probability of any given sample configuration at two markers in n individuals (at least if these markers are bi-allelic) by simply simulating two-marker genotypes in n individuals huge numbers of times, and measuring how often each configuration is observed (see Hudson, 2001). This kind of approach is impractical for several markers, because the number of configurations grows to be too large; but for two markers the number of possible configurations is manageable.

The composite likelihood (26.27) is sometimes referred to as the 'pair-wise' composite likelihood, as it involves a product over all pairs of sites. Alternative composite likelihood approaches are also possible: for example, Fearnhead and Donnelly (2002) form a composite likelihood for a region by multiplying together likelihoods obtained from non-overlapping subregions. Each of these non-overlapping subregions itself contains multiple markers, and so they use the kind of IS methods described above to approximate the likelihood for each subregion. The advantage of this approach over simply using IS to approximate the likelihood for the whole region is that the likelihood for each subregion is easier to accurately approximate (because it contains fewer markers and less recombination). Compared to the pair-wise composite likelihood method, this approach typically requires considerably more computation and is much more complicated to implement; however, it has the potential advantage that it takes into account information at multiple markers simultaneously, which could lead to gains in efficiency.

Both the pair-wise and region-wise composite likelihoods share the property that they are obtained by multiplying together likelihoods obtained from subsets of the data that are not actually independent. One consequence of ignoring dependence in this way is that composite likelihoods are typically more peaked about their maximum than they 'should' be (i.e. than they would be if the dependence was taken into account). As a result, it can be difficult to obtain confidence intervals or other measures of uncertainty for parameter estimates; however, the parameter estimates themselves appear to perform well in practice (see Fearnhead, 2003 for a theoretical treatment), and are widely used. See, for example, McVean *et al.* (2004), Myers *et al.* (2005), and **Chapter 27**.

26.5.3 Product of Approximate Conditionals (PAC) Models

Another approach to inference from population data was introduced by Li and Stephens (2003), who suggested approximating the likelihood of the observed data $\mathcal{D}$ by first writing this likelihood as a product of conditional distributions, and then developing approximations to these conditional distributions that are easy to compute. Specifically, in the context of estimating the population-scaled recombination rate ρ, from observed haplotypes $H_1, \ldots, H_n$, they write

$$L(\rho) = P(H_1, \ldots, H_n \mid \rho) = P(H_1|\rho)P(H_2|H_1, \rho) \ldots P(H_n|H_1, H_2, \ldots, H_{n-1}, \rho),$$
(26.28)

and then propose approximations to the conditional distributions on the right-hand side of this equation, which can be used to create an approximate likelihood, which they call aproduct of approximate conditionals (PAC) likelihood. This can be viewed as an attempt to approximate inference under a coalescent-like model, in that the approximate conditional distributions used are motivated by attempting to approximate the true conditional distributions under a coalescent model. One advantage of this kind of approach over the composite likelihood methods is that (26.28) is actually based on a probability distribution, and so, unlike composite likelihoods, the PAC likelihood has about the right amount of peakedness about its maximum (see Li and Stephens, 2003). This makes it possible to obtain (approximate) interval estimates for parameters (e.g. confidence or credible intervals). One disadvantage of the PAC compared with the pair-wise composite likelihood is that it requires haplotypes to be known, or estimated, and cannot be applied directly to unphased genotype data. However, actually the PAC model by itself provides a way to estimate haplotypes from genotype data (Stephens and Scheet, 2005). Further, and perhaps unexpectedly, the pair-wise composite likelihood appears to be more accurate when applied to haplotypes estimated using the PAC model than when applied directly to unphased genotypes (Smith and Fearnhead, 2005).

In addition to estimating recombination rates, the PAC approach has also been used to estimate the population-scaled mutation rate for micro-satellites (Cornuet and Beaumont, 2007; Chaudhuri and Stephens, 2006), and it seems likely that it could also be helpful in other contexts involving, for example, migration or selection.

26.6 SOFTWARE AND WEB RESOURCES

Among the methods discussed above, the rejection sampling methods based on summary statistics are perhaps the only ones that are algorithmically simple enough to be

implemented, and adapted to specific contexts, by researchers with limited computing experience. Most of the other methods are more complex, and so choice of method in a particular context will depend largely on the availability of software that can handle the types of data and models of interest. Where more than one program is available, a conservative recommendation would be to use all available methods to check that they give the same answers. This can give a useful check that the software is behaving properly, and that you are using it correctly.

Here we give a brief selection of the software currently available on the World Wide Web, implementing some of the methods described in this chapter, and their extensions to more complex genetic and demographic settings. Some of these packages are also included in the review by Excoffier and Heckel (2006).

26.6.1 Population Genetic Simulations

Simulating data from coalescent models is a key step in implementing the rejection sampling approaches outlined above, and can also be useful for testing other more complex approaches to inference. A number of programs are available.

- `ms` (Hudson, 2002), available from `http://home.uchicago.edu/~rhudson1/source.html`, can perform coalescent-based simulation of sequence data under a range of neutral models, including models with migration, population expansion and bottlenecks, with recombination and/or gene conversion.

- `FREGENE`, available from, `http://www.ebi.ac.uk/projects/BARGEN`, uses forwards simulation to simulate data from coalescent-like models.

- `cosi`, (Schaffner *et al.*, 2005), available from `http://www.broad.mit.edu/~sfs/cosi/`, performs coalescent-based simulations. A novel feature is the availability of parameter values that aim to generate data that match, in certain ways, data from modern human populations.

- `Selsim`, (Spencer and Coop, 2004), available from `http://www.stats.ox.ac.uk/~spencer/SelSim/Controlfile.html`, simulates data under a coalescent model with recombination, allowing for the presence of a single bi-allelic site that has experienced natural selection.

26.6.2 Inference Methods

26.6.2.1 General

- The program `micsat` (Wilson and Balding, 1998) is available from `http://www.mas.ncl.ac.uk/~nijw/`. It is an MCMC scheme for a Bayesian analysis of microsatellite data from a constant-sized panmictic population, based on the proposal distribution outlined earlier. A more powerful and flexible MCMC program, `batwing`, is also now available from the same address, and can deal with single nucleotide polymorphisms as well as micro-satellites, and more complicated demographic models involving populations splitting from each other some time in the past, with no subsequent migration.

- The programs `Lamarc`, available from `http://evolution.genetics.washington.edu/lamarc.html`, implements MCMC approaches for estimating

population growth rates and recombination and migration rates from both sequence and micro-satellite data. Earlier versions of the software were based on keeping parameters of the model fixed, and use methods described in Section 26.4.3.1 to calculate likelihood surfaces. Current versions are also able to perform Bayesian analyses; parameter estimates obtained using the Bayesian approach may be more reliable in some contexts (Stephens, 1999; Beerli, 2006).

- The program `genetree` (see for example Griffiths and Tavaré, 1994a) is available from `http://www.stats.ox.ac.uk/mathgen/software.html`. The software implements an IS algorithm, and includes options allowing estimation of migration and growth rates in structured populations.

- The program `IM`, available from `http://lifesci.rutgers.edu/~heylab/`, implements an MCMC approach to the estimation of divergence times and migration rates between populations.

26.6.2.2   Estimating Recombination Rates

Although some of the above packages (e.g. Lamarc) incorporate recombination, most of the above packages are most appropriate either for genomic regions that are assumed to contain little or no recombination (eg mitochondrial DNA, Y chromosome, and possibly small autosomal regions), or for analysing multiple unlinked loci. In recent years, there has been considerable interest in estimating fine-scale recombination rates across larger genomic regions using population data (see **Chapter 27**). The following list is a selection of 'coalescent-based' software available.

- The programs `maxdip` and `maxhap`, available from `http://home.uchicago.edu/~rhudson1/source.html`, implement the pair-wise composite likelihood approach to estimating recombination rates (both crossover and gene conversion) from unphased and phased single nucleotide polymorphism (SNP) data. These programs assume that the recombination rates across the region to be analysed are constant.

- The program `LDhat`, implementing the composite likelihood approach for estimating (varying) recombination rates from McVean *et al.* (2004), is available from `http://www.stats.ox.ac.uk/~mcvean/LDhat/`.

- The program `PHASE`, available from `http://www.stat.washington.edu/stephens/software.html`, implements a PAC likelihood approach to estimating (varying) recombination rates and identifying recombination hotspots.

- There are several programs implementing IS and composite likelihood methods for estimating recombination rates available from Paul Fearnhead at `http://www.maths.lancs.ac.uk/~fearnhea/software/`.

REFERENCES

Bahlo, M. and Griffiths, R.C. (2000). Inference from gene trees in a subdivided population. *Theoretical Population Biology* **57**, 79–95.

Beaumont, M. (1999). Detecting population expansion and decline using microsatellites. *Genetics* **153**, 2013–2029.

Beaumont, M.A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**(4), 2025–2035.

Beerli, P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**(3), 341–345.

Beerli, P. and Felsenstein, J. (1999). Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**(2), 763–773.

Brooks, S.P. and Roberts, G.O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing* **8**, 319–335.

Chaudhuri, A.R. and Stephens, M. (2006). Fast and accurate estimation of the population scaled mutation parameters, θ, from microsatellite data. *Genetics*, in press.

Cooper, G., Amos, W., Hoffman, D. and Rubinsztein, D.C. (1996). Network analysis of human Y microsatellite haplotypes. *Human Molecular Genetics* **5**, 1759–1766.

Cornuet, J.M. and Beaumont, M.A. (2007). A note on the accuracy of pac-likelihood inference with microsatellite data. *Theoretical Population Biology* **71**(1), 12–19.

Excoffier, L. and Heckel, G. (2006). Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics* **7**(10), 745–758.

Fearnhead, P. (2003). Consistency of estimators of the population-scaled recombination rate. *Theoretical Population Biology* **64**(1), 67–79.

Fearnhead, P.N. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318.

Fearnhead, P.N. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society Series B* **64**, 657–680.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Felsenstein, J., Kuhner, M.K., Yamato, J. and Beerli, P. (1999). Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In *Statistics in Molecular Biology and Genetics*, Volume 33 of *IMS Lecture Notes–Monograph Series*, F. Seillier-Moiseiwitsch, ed. Institute of Mathematical Statistics and American Mathematical Society, Hayward, CA, pp. 163–185.

Geyer, C. (1991). *Reweighting Monte Carlo mixtures*. Technical Report No. 568, School of Statistics, University of Minnesota. Available from `http://stat.umn.edu/PAPERS/tr568r.html`.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson and D.J. Spiegelhalter, eds. Chapman & Hall, London.

Griffiths, R.C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3**, 479–502.

Griffiths, R.C. and Tavaré, S. (1994a). Ancestral inference in population genetics. *Statistical Science* **9**, 307–319.

Griffiths, R.C. and Tavaré, S. (1994b). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society London Series B* **344**, 403–410.

Griffiths, R.C. and Tavaré, S. (1994c). Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131–159.

Griffiths, R.C. and Tavaré, S. (1997). Computational methods for the coalescent. In *Progress in Population Genetics and Human Evolution*, Chapter 10, Volume 87 of *IMA Volumes in*

Mathematics and its Applications, P. Donnelly and S. Tavaré, eds. Springer Verlag, Berlin, pp. 165–182.

Griffiths, R.C. and Tavaré, S. (1999). The ages of mutations in gene trees. *Annals of Applied Probability* **9**, 567–590.

Harding, R.M., Fullerton, S.M., Griffiths, R.C., Bond, J., Cox, M.J., Schneider, J.A., Moulin, D.S. and Clegg, J.B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *American Journal of Human Genetics* **60**, 772–789.

Hudson, R.R. (2001). Two-locus sampling distribution and their application. *Genetics* **159**, 1805–1817.

Hudson, R.R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338.

Huelsenbeck, J. (2001). Likelihood analysis of phylogenetic trees. In *Application of the Likelihood Function in Phylogenetic Analysis*. John Wiley & Sons.

Kuhner, M.K., Yamato, J. and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* **140**, 1421–1430.

Li, N. and Stephens, M. (2003). Modelling linkage disequilibrium, and identifying recombination hotspots using snp data. *Genetics* **165**, 2213–2233.

Li, W.-H. and Fu, Y.-X. (1999). Coalescent theory and its application in population genetics. In *Statistics in Genetics*, M.E. Halloran and S. Geisser, eds. Springer, pp. 45–80.

Marjoram, P., Molitor, J., Plagnol, V. and Tavare, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **100**(26), 15324–15328.

Marjoram, P. and Tavaré, S. (2006). Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics* **7**(10), 759–770.

McVean, G.A.T., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R. and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584.

Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**(5746), 321–324.

Nielsen, R. (1997). A likelihood approach to population samples of microsatellite alleles. *Genetics* **146**, 711–716.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942.

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A. and Feldman, M.W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**(12), 1791–1798.

Ripley, B.D. (1987). *Stochastic Simulation*. John Wiley & Sons, New York.

Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J. and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* **15**(11), 1576–1583.

Sisson, S.A., Fan, Y. and Tanaka, M.M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **104**, 1760–1765.

Smith, N. and Fearnhead, P. (2005). A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness. *Genetics* **171**(4), 2051–2062.

Spencer, C.C. and Coop, G. Selsim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* **20**(18), 3673–3675.

Stephens, M. (1999). Problems with computational methods in population genetics, 273–276. Available from http://www.stats.ox.ac.uk/~stephens.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society Series B* **62**, 605–655.

Stephens, M. and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing data imputation. *American Journal of Human Genetics* **76**, 449–462.

Tavaré, S., Balding, D.J., Griffiths, R.C. and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518.

Tremblay, M. and Vézina, H. (2000). New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *American Journal of Human Genetics* **66**, 651–658.

Weiss, K. (1973). Demographic models for anthropology. *American Antiquity* **38**, 1–186.

Wilson, I.J. and Balding, D.J. (1998). Genealogical inference from microsatellite data. *Genetics* **150**, 499–510.

Linkage Disequilibrium, Recombination and Selection

G. McVean

Department of Statistics, Oxford University, Oxford, UK

Every chromosome carries a unique sequence of DNA; however, certain combinations of variants are shared between individuals. The extent of this sharing is referred to as *linkage disequilibrium* and its distribution in natural populations can be informative about diverse processes, from the distribution of recombination to the influence of adaptive evolution. This chapter aims to provide a foundation for the empirical analysis of linkage disequilibrium discussing how it can be measured, how it relates to underlying genealogical processes and how to perform inference about underlying molecular, historical and evolutionary processes. A central idea is that linkage disequilibrium is most naturally understood in terms of the genealogical structure underlying a sample of chromosomes.

27.1   WHAT IS LINKAGE DISEQUILIBRIUM?

Every human genome has a unique DNA sequence. This is, in part, because each individual inherits a few hundred novel mutations that occurred in the germ lines of their parents (Nachman and Crowell, 2000). But it is also because the meiotic processes of chromosomal segregation and recombination shuffle existing variation: the result of mutations in our ancestors' germ lines. Consequently, while every genome may be unique, certain combinations of variants are shared: sometimes by just a few individuals, sometimes by a large fraction of the population. The term *linkage disequilibrium* (LD) is broadly used to refer to the non-random sharing (or lack thereof) of combinations of variants. It would, perhaps, be better to talk about 'haplotype structure' or 'allelic association' (both terms also used in this chapter). Nevertheless, although LD neither requires linkage (physical association on a chromosome) nor is particularly a disequilibrium (e.g. one can discuss the equilibrium level of LD), the term *LD* has stuck.

   To be clear about what LD means here, consider the three example data sets in Figure 27.1. In each case the marginal allele frequencies at each polymorphic site are approximately equal; what is different between the data sets is the degree of structuring
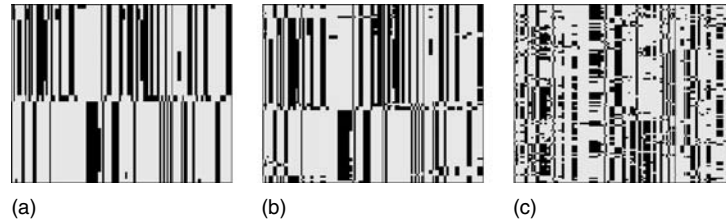
Figure 27.1 Haplotype patterns with different levels of linkage disequilibrium. Each panel shows a sample of 100 chromosomes drawn from a population with (a) no, (b) low and (c) high crossover. Each row is a chromosome and the rarer allele at each site has been shaded black. Data sets all have approximately the same set of marginal allele frequencies, so the only difference is in terms of the degree of structuring, or linkage disequilibrium.

of the variation or LD. Specifically, Figure 27.1(a) shows a very high level of structuring, Figure 27.1(c) shows a very low level of structuring and Figure 27.1(b) shows something in between. Such differences in how genetic variation is structured point to differences in the underlying biological processes experienced by the populations from which the samples are drawn. Here, the primary difference is in the crossing-over rate; the data have been simulated with zero, some and lots of crossing over respectively. However, other molecular and historical processes, including mutation, natural selection, geographical isolation and changes in population size, will also influence the structuring of genetic variation in populations. It is the goal of population genetics to make inference about such processes from observations of genetic variation in contemporary populations. Naturally, we wish to use the relevant information contained in patterns of LD. The aim of this chapter is to explore how we can understand patterns of LD observed in empirical data.

It is also worthwhile to give a more formal definition of LD. Consider a sample of chromosomes where polymorphism has been observed at a series of three loci, x, y and z. For simplicity each locus is assumed to have only two alleles (A/a, B/b and C/c respectively). The obvious description of the sample is in terms of the number of times we observe each haplotype, n_{ABC}, n_{abc}, etc., or alternatively their sample proportions, f_{ABC}, f_{abc}, etc. However, we can equivalently describe the data in terms of the marginal allele frequencies at each locus, f_A, f_a, etc. and a series of terms that reflect the extent to which combinations of alleles are found more or less frequently than expected assuming independence. For example:

$$f_{ABC} = f_A f_B f_C + f_A D_{BC} + f_B D_{AC} + f_C D_{AB} + D_{ABC}, \tag{27.1}$$

where

$$D_{AB} = f_{AB} - f_A f_B$$
$$D_{AC} = f_{AC} - f_A f_C$$
$$D_{BC} = f_{BC} - f_B f_C$$
$$D_{ABC} = f_{ABC} - f_A D_{BC} - f_B D_{AC} - f_C D_{AB} - f_A f_B f_C. \tag{27.2}$$

The D terms, which are referred to as LD coefficients, therefore measure the difference between the observed frequency of pairs or triples of alleles and that expected from the marginal allele frequencies and other D terms of lower order (e.g. the D terms for triples contain the D terms for pairs, etc.). Similar expressions apply to any data set of any

complexity (in terms of number of loci and number of alleles at each locus). However, the number of terms clearly explodes as the number of loci increases. Nevertheless, the point should be clear: patterns of genetic variation can be described in terms of the marginal allele frequencies and a series of terms relating to the degree of association between pairs, triples, etc. of alleles. These terms are broadly referred to as *LD*.

Before progressing, there is a rather subtle (but ultimately profound) point to make. The previous paragraph deals with how to describe genetic variation within a sample of chromosomes. Historically, population genetics has focused more on describing genetic variation within *populations*: idealised entities consisting of (effectively) infinite numbers of individuals whose genetic composition can be described in terms of allele frequencies and coefficients of LD, just as in the sample. While the notion of a population is very helpful (and is used extensively in this chapter) a focus on the sample has three benefits. First, the sample is all that we have; although we can of course use the sample to make inferences about populations. Second, in reality there is no such thing as a population, just a set of individuals related through a complex and unknown pedigree. Third, thinking about the history of the sample, specifically the genealogical history, provides a coherent way of linking what we observe in patterns of genetic variation to underlying biological and historical processes (Hudson, 1990). For these reasons, this chapter focusses heavily on the interpretation of LD within a sample.

The rest of this chapter is divided into three parts. In the first, the method of summarising LD in empirical data is explored. In the second, how simple probabilistic models of genealogical history can be used to explore the effects of various molecular and historical processes on patterns of LD is dealt with. Finally, in the third part, the problem of inference is considered: how we can learn about such processes from empirical data.

27.2   MEASURING LINKAGE DISEQUILIBRIUM

As stated above, our ultimate aim is to make inference about underlying biological and historical processes from patterns of genetic variation, including the structuring of variants, or LD. To motivate the problem, consider the three data sets shown in Figure 27.2. Each panel shows the inferred haplotypes for single nucleotide polymorphism (SNP) data from the same 100 kb surrounding the *Lactase* gene on human chromosome 2, but from three different samples of individuals: individuals of European ancestry living in Utah (referred to as *CEU*), individuals of Yoruba origin from Nigeria (referred to as *YRI*) and a combination of Han Chinese from Beijing and Japanese people living in Tokyo (referred to collectively as *CHB/JPT*); the data is from the International HapMap Project (The International HapMap Consortium, 2005). This gene is important because a particular mutation in the promoter, found at high frequencies in Europe, is associated (in Europe) with the ability to digest lactose in milk persisting until adulthood, whereas in most non-pastoralist populations from Africa, Asia and the Americas this ability ceases between 5 and 10 years. The high frequency of this mutation is thought to have been the result of strong selection for lactose persistence associated with the innovation and subsequent importance of dairy farming. This hypothesis is strongly supported by the genetic variation data, which shows European populations dominated by a single haplotype in marked comparison to the other HapMap populations (Bersaglieri *et al.*, 2004; Hollox *et al.*, 2001). Such patterns are suggestive of a recent selective sweep (Maynard Smith
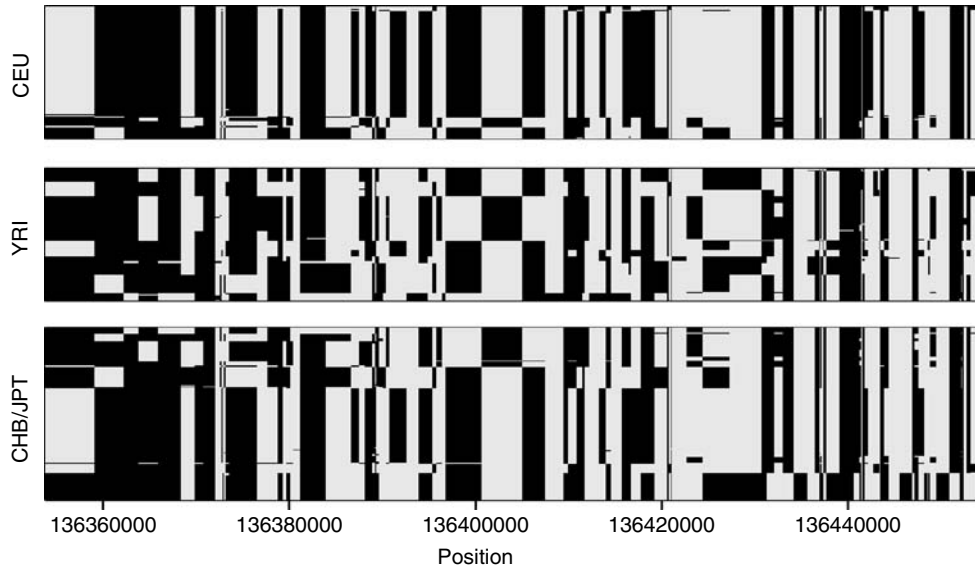
Figure 27.2 Haplotype structure in a 100-kb region surrounding the *Lactase* gene on human chromosome 2 for the four HapMap populations (The International HapMap Consortium, 2005). CEU = Individuals of European origin from Utah; YRI = Yoruba from Nigeria; CHB/JPT = Han Chinese from Beijing and Japanese from the Tokyo region (60 unrelated individuals in each of CEU and YRI, 90 in CHB/JPT). The CEU panel is dominated by a single haplotype that extends over the entire region. Much higher haplotype diversity is found in the other populations (see also Table 27.1).

and Haigh, 1974) in which the haplotype on which the beneficial mutation arose was also dragged to high frequency. However, it is worth noting that other pastoralist populations, such as the Fulani and bedouin, where lactose persistence also occurs at high frequency, do not show the same patterns and most likely carry other mutations within the *Lactase* region (Ingram *et al.*, 2007).

However, suppose we knew nothing about the gene and its function. Faced with such haplotype data, how might we begin to assemble a coherent picture of the underlying processes? By way of an aside it should be noted that, throughout, the haplotype 'phase' is typically assumed to be known (i.e. for diploid species the sequence of each chromosome in an individual is known separately), through a combination of experimental techniques, genotyping in pedigrees and/or statistical analysis (e.g., Marchini *et al.*, 2006). Our first approach might be to present the data graphically, as in Figure 27.2. This shows the marked differences between populations and is clearly the most complete representation of the genetic variation data. Nevertheless, it would also be useful to have some low-dimensional summaries of the data that could be used to compare populations, or perhaps this region to some other region in the genome.

The aim of this section is to introduce a range of low-dimensional summaries of LD that could be applied to such data. In reality, we would also use low-dimensional summaries of the data that are functions of the SNP allele frequencies rather than their structuring. For example, statistics like Tajima's D (Tajima, 1989), Fay and Wu's H (Fay and Wu, 2000), Fu and Li's D (Fu and Li, 1993), F_{ST}, etc. Indeed, it doesn't make much sense

to separate out the analysis of LD from the analysis of allele frequencies. However, the diversity of measures of LD is sufficient to justify separate consideration. One thing that must be stressed, however, is that no single summary of LD is 'best' in the sense that it captures all information about underlying processes (e.g. no single-number summary is *sufficient* in the statistical sense that it captures all information about some parameter of interest). Different summaries are more or less useful for identifying the effects of different underlying processes. In the analysis of empirical data it is therefore important to stress the use of multiple summaries, each of which may give some more insight.

27.2.1 Single-number Summaries of LD

A natural starting point in the analysis of LD would be to ask whether there are any single-number statistics of the data that are informative about LD in the same way that, for example, the average pairwise diversity or Tajima's D are used in the analysis of allele frequency data. If the region is relatively short, one useful summary is simply the number of unique haplotypes observed (the more LD the fewer haplotypes). Similarly, haplotype homozygosity (the probability that two haplotypes picked without replacement from the sample are identical) is an indication of how skewed the haplotype frequencies are. There is, however, a problem with such summaries. As the length of the region surveyed increases (or more SNPs are typed), we would eventually expect to reach a point where every haplotype is unique (and haplotype homozygosity is 0). We might therefore want a summary whose value is not so arbitrarily determined by the length of the region analysed.

There are many possibilities for such statistics. One approach is to attempt to break the data up into a series of blocks, each of which represents a region with low numbers of haplotypes and high haplotype homozygosity (Anderson and Novembre, 2003; Gabriel *et al.*, 2002; Wang *et al.*, 2002; Zhang *et al.*, 2002). The number of such blocks is therefore a measure of the degree of structure in the data. However, any choice about how to break the data up is arbitrary and the concept of such 'haplotype blocks' has been little used in recent years. A related idea, motivated by the design of association studies to map the genetic basis of phenotypic variation, is to identify 'tag' SNPs that capture (in ways that are discussed below) variation within a region (Carlson *et al.*, 2004; de Bakker *et al.*, 2005; Johnson *et al.*, 2001). The number of tag SNPs required for a region is therefore a measure of how structured the variation is (e.g. if only two distinct haplotypes were observed, only one tag SNP would be required). However, there is no single most useful measure of how well variation is 'captured' and differences in experimental design mean that estimated numbers of tag SNPs are hard to compare between studies.

Another approach to summarising LD within a region is to estimate the influence of recombination. For example, non-parametric techniques (Hudson and Kaplan, 1985; Myers and Griffiths, 2003; Song *et al.*, 2005) can be used to estimate the minimum number of historical recombination events consistent with the data under the infinite-sites assumption. Similarly, parametric, coalescent-based techniques (discussed later) can be used to estimate the 'population recombination rate', or more accurately the 'population crossover rate', $4N_ec$, under the standard neutral model, where c is the genetic distance across the region and N_e is the effective population size (Stumpf and McVean, 2003). Broadly speaking, a high crossover rate will tend to result in little genetic structuring (like Figure 27.1c), while low rates will tend to result in data sets like that in Figure 27.1(a). However, the association is far from perfect, particularly if the region of interest has experienced adaptive evolution, the demographic history is not well approximated by a

randomly mating population of constant size (e.g. there have been dramatic changes in population size or geographical subdivision) or there is considerable gene conversion.

By way of example, single-number summaries of LD for the *Lactase* gene region of Figure 27.2 are presented in Table 27.1. The strong structuring of the CEU sample is shown clearly in the reduced number of haplotypes, the increased homozygosity and the smaller number of detectable recombination events relative to the other populations. The similarity in haplotype numbers and detectable recombination events between YRI and CHB/JPT is complicated by the larger sample size of the latter; sub-samples of 120 chromosomes from CHB/JPT typically show values intermediate between YRI and CEU. Interestingly, the parametric estimates of the population crossover rates show the rate in CEU to be slightly higher than CHB/JPT, a pattern generally seen across the genome (Myers *et al.*, 2006; The International HapMap Consortium, 2005) and perhaps not one expected for a gene where natural selection has acted so strongly. It is important to note that the model-based estimates of the population crossover rate typically assume neutrality (and a very simple demographic history). The action of natural selection is just one of many historical forces that can result in different summaries of LD giving apparently conflicting indications as to the relative amount of LD.

27.2.2   The Spatial Distribution of LD

Any single-number summary of LD will fail to capture heterogeneity in the observed structuring of variation. Some genomic regions may have greater structuring than others or there may be systematic patterns in the relationship between genomic location and LD. For example, crossing over during meiosis will tend to lead to systematically lower levels of LD for variants at distantly separated loci compared to that for those at closely situated ones. Alternatively, gene conversion and/or mutational hotspots might create variants that are much more randomly distributed than their neighbours.

There are two approaches to summarizing the spatial distribution of LD (i.e. how it changes along the chromosome). One possibility is to make inferences about the spatial nature of the underlying biological or evolutionary processes (crossing over, gene conversion, mutation, natural selection, etc.). For example, LD in humans is strongly influenced by the concentration of meiotic crossing-over events into short regions called *recombination hotspots*. Consequently, inferences about the underlying recombination landscape reflect, at least in part, the distribution of LD along a chromosome (Jeffreys *et al.*, 2001; 2005; McVean *et al.*, 2005; The International HapMap Consortium, 2005).

Table 27.1   Statistics of LD for the *Lactase* gene.

	YRI	CEU	CHB/JPT
Number of chromosomes	120	120	180
Number of unique haplotypes	34	18	35
Haplotype homozygosity	0.05	0.53	0.15
Recombination events*	23	10	23
Estimated $4N_e c$/kb[†]	0.12	0.10	0.07

* Lower bound on the minimum number of recombination events estimated by the method of Myers and Griffiths (2003).
[†] Estimated using the method of McVean *et al.* (2002) assuming a constant crossover rate and $\theta = 0.001$ per site.

The alternative is to make summaries of LD for subsets of the data (e.g. pairs of sites) and show, usually graphically, spatial patterns in these summaries. The following discussion focuses on two-locus summaries of LD as these are the most widely used summaries of LD for genetic variation data.

Consider a pair of loci, at which exactly two different alleles are observed in the population; these being A/a at the first locus and B/b at the second. These are most naturally thought of as SNPs, but they might also be insertion–deletion polymorphisms or restriction fragment length polymorphism (RFLPs). For the moment assume that the haplotype phase of the alleles is known. As described above, the standard coefficient of LD between the alleles at the two loci is defined as

$$
\begin{aligned}
D_{AB} &= f_{AB} - f_A f_B \\
&= f_{AB} f_{ab} - f_{Ab} f_{aB}
\end{aligned}, \tag{27.3}
$$

where f_{AB} is the frequency of haplotypes carrying the A and B alleles and f_A is the marginal allele frequency of allele A. D_{AB} therefore measures the difference between the frequency of the AB haplotype and that expected if the haplotype frequencies were simply given by the product of the marginal allele frequencies. Any deviation from this expectation results in a non-zero value for D_{AB}, with a positive value indicating that the AB haplotype is found more often than expected assuming independence and a negative value indicating that it is found less frequently than expected. Although (27.3) focuses on the AB haplotype, the coefficient of LD for any other haplotype is given by the simple relationship $D_{AB} = -D_{aB} = -D_{Ab} = D_{ab}$.

As described above, the coefficient is computed from the sample haplotype frequencies. However, we might also be interested in asking how the sample coefficient relates to that of the population (if we believe that one exists). If we let D_{AB} be the population coefficient, the sample coefficient $\hat{D}_{AB}$ has the properties (Hill, 1974)

$$
\hat{D}_{AB} = \hat{f}_{AB} - \hat{f}_A \hat{f}_B
$$

$$
E\left[\hat{D}_{AB}\right] = \left(\frac{n-1}{n}\right) D_{AB}
$$

$$
\text{Var}(\hat{D}_{AB}) = \frac{1}{n}[f_A f_a f_B f_b + (f_A - f_a)(f_B - f_b)D_{AB} - D_{AB}^2]. \tag{27.4}
$$

Here $\hat{f}_{AB}$ means the obvious estimate of f_{AB} (the population frequency) from the sample; i.e. n_{AB}/n, where n_{AB} is the number of AB haplotypes in the sample. The most important point about (27.4) is that the variance in the estimate is strongly influenced by the allele frequencies at the two loci. Furthermore, the range of values $\hat{D}_{AB}$ can take is strongly influenced by the allele frequencies. If we arbitrarily define the A and B alleles to be the rarer alleles at each locus and enforce (without loss of generality) $\hat{f}_B \leq \hat{f}_A$, it follows that

$$
-\hat{f}_A \hat{f}_B \leq \hat{D}_{AB} \leq \hat{f}_a \hat{f}_B. \tag{27.5}
$$

The strong dependency on allele frequency of the standard coefficient of LD is an undesirable property because it makes comparison between pairs of alleles with different allele frequencies difficult. Consequently, several other measures of LD have been proposed that (at least in some ways) are less sensitive to marginal allele frequencies (Hedrick, 1987).

The most useful of these is the r^2 measure (Hill and Robertson, 1968). Consider assigning an allelic value, X_A, which is 1 if the allele at the first locus is A and 0 if

the allele is a. Also assign an allelic value, X_B, with equivalent properties at the second locus. The quantity measured by (27.3) can then be interpreted as the covariance in allelic value between the loci. A standard way to transform the covariance is to measure the Pearson correlation coefficient,

$$r_{AB} = \frac{\text{Cov}(X_A, X_B)}{\sqrt{\text{Var}(X_A)\text{Var}(X_B)}} = \frac{D_{AB}}{\sqrt{f_A f_a f_B f_b}}. \tag{27.6}$$

In fact, for several reasons (not least because (27.6) has an arbitrary sign depending on how the allelic values are assigned), it is actually more useful to consider the square of the correlation coefficient,

$$r^2 = \frac{D^2}{f_A f_a f_B f_b}. \tag{27.7}$$

The r^2 measure has many useful properties. First, as indicated by the lack of subscripts for D in (27.7), it has the same value however the alleles at the two loci are labelled. Second, as described later, there are simple relationships between r^2 and two features of interest, the power of association studies (Chapman *et al.*, 2003; Pritchard and Przeworski, 2001) and properties of the underlying genealogical history (McVean, 2002). Third, there is a direct relationship between the sample estimate of r^2, obtained by replacing population values by the sample values in (27.7), and the power to detect significant association, i.e. to reject the null hypothesis $H_0\colon D = 0$. An obvious test to consider is the contingency table test where, under the null hypothesis, the test statistic

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{27.8}$$

is asymptotically χ^2 distributed with 1 df as the sample size tends to infinity. Here O_{ij} and E_{ij} are, respectively, the observed and expected counts of the ij haplotype where the expectation is calculated assuming independence between the loci. The relationship between (27.8) and r^2 is

$$X^2 = n\hat{r}^2. \tag{27.9}$$

Consequently the null hypothesis of no association can be rejected at a specified level, α, if $n\hat{r}^2$ is greater than the appropriate critical value of the test statistic. Another test that might be considered is the likelihood ratio test, where the test statistic

$$\Lambda = 2\log\left(\frac{L(D = \hat{D})}{L(D = 0)}\right), \tag{27.10}$$

is also asymptotically χ^2 distributed with 1 df under the null hypothesis. Here $L(D)$ indicates the likelihood of the LD coefficient calculated using the multinomial distribution and the specified value of D. Using a Taylor expansion to approximate the logarithm, it can be shown that

$$\Lambda = 2n \sum_{ij} (\hat{f}_i \hat{f}_j + \hat{D}_{ij}) \log\left(1 + \frac{\hat{D}_{ij}}{\hat{f}_i \hat{f}_j}\right)$$

$$= n\hat{r}^2 + o(D^3). \tag{27.11}$$

Consequently the test statistics (and hence the power) of the contingency table and likelihood ratio tests are approximately equal and a function only of the sample size and observed r^2. Note that for small sample sizes the χ^2 approximation is unlikely to hold. In these circumstances it is possible to use standard permutation procedures or exact tests to estimate the significance of the observed correlation.

Although r^2 has many useful properties, it is far from the only statistic in use. For example, the $|D'|$ measure (Lewontin, 1964) is defined as the absolute value of the ratio of the observed D to the most extreme value it could take given the observed allele frequencies

$$|D'| = \begin{cases} \dfrac{-\hat{D}_{AB}}{\min(\hat{f}_A\hat{f}_B, \hat{f}_a\hat{f}_b)} & \hat{D}_{AB} < 0 \\[3mm] \dfrac{\hat{D}_{AB}}{\min(\hat{f}_A\hat{f}_b, \hat{f}_a\hat{f}_B)} & \hat{D}_{AB} > 0 \end{cases}. \tag{27.12}$$

The main use of $|D'|$ is that it measures the evidence for recombination between the loci. A feature of (27.12) is that $|D'|$ can only be less than 1 if all four possible haplotypes are observed in the sample. If the mutation rate is low, such that repeat or back mutation is unlikely, then if all four possible haplotypes are observed, it can be inferred that at least one recombination event must have occurred in the history of the sample (Hudson and Kaplan, 1985). Conversely, if anything less than the four combinations are observed the data are compatible with a history in which no recombination has occurred. Furthermore, the greater the rate of recombination between the loci, the more likely the alleles are to be in linkage equilibrium. So a value of $|D'| = 1$ can be interpreted as evidence for no recombination, while a value near 0 can be interpreted as evidence for considerable recombination. There is, however, a problem with such an interpretation for rare alleles (Figure 27.3). Even if all four combinations are present in the population, it may be unlikely to see all four in a finite sample if at least one haplotype is at low frequency (Devlin and Risch, 1995; Hedrick, 1987; Lewontin, 1988). For this reason, the interpretation of a $|D'|$ of 1 is highly dependent on the sample allele frequencies and the construction of confidence intervals for $|D'|$ is highly recommended (Gabriel *et al.*, 2002). Furthermore, if the primary interest of a study is to learn about recombination, it makes considerably more sense to use non-parametric or parametric approaches to learn about recombination directly.

As discussed above, an informative approach to summarising the spatial structure of LD is to compute two-locus statistics for all pairs of polymorphic loci and to represent these values graphically. Figure 27.4 shows three example data sets and their corresponding LD plots that demonstrate how the spatial distribution of LD can vary. In Figure 27.4(a)

Figure 27.3 The relationship between sample configuration and the r^2 and $|D'|$ measures of two-locus LD. Each panel shows a configuration that corresponds to either high or low values of the two measures. Each bar is a chromosome and each circle represents an allele. For the diagonal plots the measures agree. However, for (b), r^2 is near 0 while $|D'|$ is 1, demonstrating how the two measures focus on different aspects of the sample configuration.
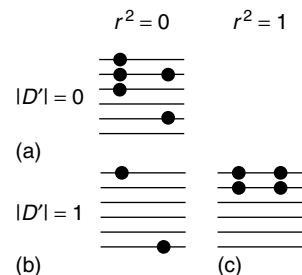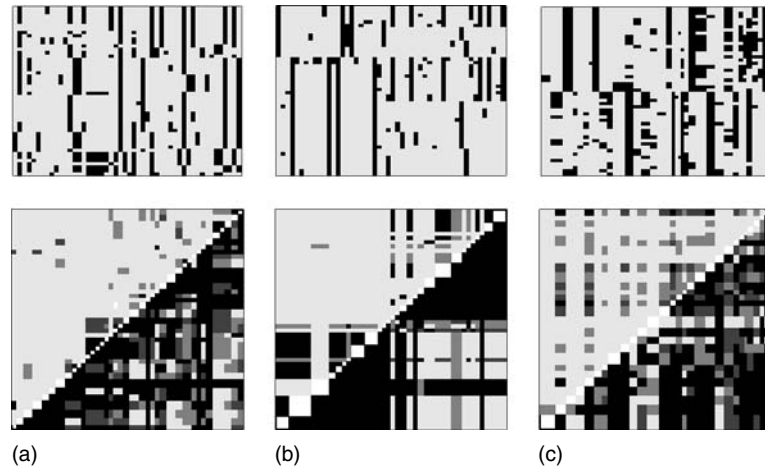
Figure 27.4 The spatial structure of LD. In each panel the upper plot shows the haplotypes and the lower plot shows a matrix of pairwise r^2 (upper left) and $|D'|$ values (bottom right) shaded by magnitude from black (values near one) to light grey (values near zero). (a) In a region of constant crossover rate sites close to each other show strong LD, which gradually decreases the further apart they are ($n = 50, \theta = 10, C = 30$). (b) In a region with a central crossover hotspot, LD appears blocklike, with regions of very high LD separated by points at which the association breaks down ($n = 50, \theta = 10, C = 50$ concentrated on a single central hotspot). (c) Where LD is generated by the mixture of individuals from two differentiated populations there is no spatial structure to LD; sites near or far can be in strong LD ($n = 25$ in each population, 80 loci, data simulated under a β-binomial model of differentiation (Balding and Nichols, 1995) with $F_{ST} = 0.9$).

there is a tendency for closely situated alleles to show strong to moderate LD, while more distant ones show much weaker LD. In Figure 27.4(b) there are two strong blocks of LD separated by a point at which LD breaks down almost completely. In Figure 27.4(c) there is no apparent spatial structure to LD: alleles can be in strong association whether they are near each other or far away. These differences in LD patterns reflect different underlying processes: a region with a moderate and constant crossover rate, a region with a strong crossover hotspot and low background crossover rate and a series of unlinked loci sampled from two highly differentiated populations, respectively. Although the pictures are somewhat noisy, it is clear that an understanding of the spatial distribution of LD can greatly help in the interpretation of underlying processes.

27.2.3 Various Extensions of Two-locus LD Measures

The previous situation considered how to measure LD in the setting where loci are bi-allelic and the allelic phase is known. But in many situations neither may be true. If the sampled individuals are diploid and the haplotype phase of alleles is unknown, it is possible to estimate the haplotype frequencies, e.g. by maximum likelihood (Weir, 1996). For two bi-allelic loci, lack of phase information adds remarkably little uncertainty to estimates of LD (Hill, 1974), because the only two-locus genotype where phase cannot be accurately inferred is when the individual is heterozygous at both loci. For multi-allelic loci (such as microsatellites), haplotype estimation can also be achieved by maximum likelihood, e.g. by the expectation maximisation (EM) algorithm.

The key problem for multi-allelic systems is how to summarise LD. One approach, motivated by the relationship between r^2 and the χ-square test in the bi-allelic case, is to use the statistic (Hill, 1975)

$$Q = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\hat{D}_{ij}^2}{\hat{f}_i \hat{f}_j}. \tag{27.13}$$

Here, as before, O_{ij} and E_{ij} are the observed and expected haplotype counts when there are k and l alleles respectively at the two loci. Again, under the null hypothesis $H_0 : D_{ij} = 0$, for all i and j, Qn will be approximately χ^2 distributed, though with $(k-1)(l-1)$ degrees of freedom. Of course, as the number of alleles increases, so that the expected haplotype counts tend to decrease, it becomes more important to use permutation methods or exact tests, rather than the χ^2 approximation, to assess significance. There are also multi-allele versions of $|D'|$ (Hedrick, 1987). Again, it is informative to present the result of such analyses graphically.

27.2.4 The Relationship between r^2 and Power in Association Studies

Before trying to understand what influences the distribution of LD (and how we can learn about these processes), it is worth discussing an important, though perhaps rather misunderstood, relationship between measures of LD and the power of association studies (Chapman *et al.*, 2003; Pritchard and Przeworski, 2001). An association study is an experiment that aims to map the genetic basis of phenotypic variation by comparing the genetic variation between individuals with a disease (cases) and without the disease (controls) with the aim of identifying variants that are enriched among sufferers of the disease; similar approaches can be taken for quantitative traits (see **Chapter 37**). Typically, only a subset of all polymorphic loci will have been analysed. Nevertheless, if the causative locus does not happen to be among those analysed, there is still some hope of identifying the locus through the LD that might exist between the causative variant and nearby variants that have been analysed in the study (Hirschhorn and Daly, 2005).

To see what reduction in power this leads to we need to specify a way in which the statistical analysis of the data will be performed. The simplest (though almost certainly not the most powerful) approach is to measure the difference in allele frequencies between cases and controls and assess the significance by means of a contingency table analysis. For simplicity, the diploid nature of most eukaryotes is ignored, though the results shown here generalise to the diploid case for diseases with additive (or multiplicative) risk (Chapman *et al.*, 2003). Table 27.2 shows the frequency of each case/allele combination for a simple disease model where the risk allele, A, increases the chance of getting the disease by a factor e^γ. The population frequencies of the two alleles are f_A and f_a and the proportion of the subjects who are cases and controls is ϕ_D and ϕ_C, respectively. Under this model, the sample counts are drawn from these frequencies (usually with fixed row totals) and the association between the disease and genetic variation at the locus could be summarised by the sample covariance between allelic status and disease status, a statistic called T here

$$T = \text{Cov}(A, D) = \tfrac{1}{n^2}(n_{AD}n_{aC} - n_{AC}n_{aD}). \tag{27.14}$$

Where n_{AD} is the number of individuals with the disease and the A allele, etc., and n is the total number of individuals. Note that this statistic is also equivalent to the standard

Table 27.2 Frequencies of each case/allele status under
a simple disease model.

Status	Allele		Totals
	A	a	
Disease	$\dfrac{\phi_{\mathrm{D}} f_A \mathrm{e}^{\gamma}}{f_a + f_A \mathrm{e}^{\gamma}}$	$\dfrac{\phi_{\mathrm{D}} f_a}{f_a + f_A \mathrm{e}^{\gamma}}$	ϕ_{D}
Control*	$\approx \phi_{\mathrm{C}} f_A$	$\approx \phi_{\mathrm{C}} f_a$	ϕ_{C}

* It is assumed that the allele risk and disease prevalence are
sufficiently small so as not to create a skew in allele frequencies
in the control set.

coefficient of LD between the allelic and disease status (27.3). However, we also wish
to know whether the observed value of T lies outside the range we would expect by
chance if there were no causative association (i.e. whether we can reject the null model
at some specified significance level). Under the null hypothesis, $\gamma = 0$ (i.e. when there is
no causal association), we know from (27.4) that

$$E[T] = 0$$

$$\mathrm{Var}(T) = \frac{f_A f_a \phi_D \phi_C}{n}. \tag{27.15}$$

So under the null hypothesis the statistic

$$Z = \frac{T - E[T]}{\sqrt{\mathrm{Var}(T)}} = \frac{T}{\sqrt{\hat{f}_A \hat{f}_a \phi_D \phi_C}} n^{\frac{1}{2}} \tag{27.16}$$

is asymptotically normally distributed with mean 0 and variance 1 as the sample size
tends to infinity. Note the hats on the frequencies indicate that these are estimates from the
sample. Under the alternative, $\gamma \neq 0$, but assuming that γ is small such that $e^{\gamma} \approx 1 + \gamma$
and the variance of the test statistic does not change appreciably, it follows that

$$E[T] \approx \gamma \times f_A f_a \phi_D \phi_C$$

$$Z = \frac{T}{\sqrt{\hat{f}_A \hat{f}_a \phi_D \phi_C}} n^{\frac{1}{2}} \sim \mathrm{Normal}\left(n^{\frac{1}{2}} [f_A f_a \phi_D \phi_C]^{\frac{1}{2}} \gamma, 1 \right). \tag{27.17}$$

Equivalently, the standard test statistic for the 2×2 contingency table, Z^2, is
approximated by the χ^2 distribution with 1 df and a non-centrality parameter of
$n^{\frac{1}{2}} [f_A f_a \phi_D \phi_C]^{\frac{1}{2}} \gamma$. Now consider a marker locus with alleles B and b. The entries in
Table 27.2 can be updated in a simple manner by noting that

$$\Pr(B|D) = \frac{\Pr(D \ \& \ B)}{\Pr(D)}$$

$$= \frac{\Pr(B)[\Pr(A|B)\Pr(D|A) + \Pr(a|B)\Pr(D|a)]}{\Pr(D)}$$

$$= \frac{f_{AB}\mathrm{e}^{\gamma} + f_{aB}}{f_A\mathrm{e}^{\gamma} + f_a}$$

$$\approx f_B + D_{AB}\gamma. \tag{27.18}$$

Substituting this term into Table 27.2 and an equivalent one for the b allele, $\Pr(b|D) \approx f_b - D_{AB}\gamma$, shows that the test statistic at the marker locus

$$Z = \frac{\hat{f}_{BD}\hat{f}_{bC} - \hat{f}_{BC}\hat{f}_{bD}}{\sqrt{\hat{f}_B\hat{f}_b\phi_D\phi_C}}n^{\frac{1}{2}} \tag{27.19}$$

is approximately normally distributed

$$Z \sim \mathrm{Normal}\left(n^{\frac{1}{2}}[\phi_D\phi_C]^{\frac{1}{2}}\gamma\frac{D_{AB}}{\sqrt{f_Bf_b}}, 1\right)$$

$$\sim \mathrm{Normal}\left(n^{\frac{1}{2}}[f_Af_a\phi_D\phi_C]^{\frac{1}{2}}\gamma\frac{D_{AB}}{\sqrt{f_Af_af_Bf_b}}, 1\right) \tag{27.20}$$

The last term in the mean of the distribution should look familiar as (27.6), the correlation coefficient between the A and B alleles at the two loci. Finally, imagine two experiments. In the first, we type the causative locus in a total of n_1 individuals. In the second, we type the marker locus in n_2 individuals. From (27.17) and (27.20) it should be clear that the distributions of the test statistics across replicates of the two experiments will be the same if and only if

$$n_2 = \frac{n_1}{r^2}. \tag{27.21}$$

The subscript on the square of the correlation coefficient has again been dropped to indicate that it is the same for all pairs of alleles. Equation (27.21) implies that to achieve the same power in the experiment at the marker locus the sample size has to be increased by a factor of $1/r^2$.

This is a very elegant result and clearly has implications for the design of association studies. Nevertheless, two critical issues need to be appreciated. The first is that while the r^2 result can be used to define experiments with equal power, it does not follow that in a given experiment typing a marker locus relative to a causative locus results in a loss of power of $100 \times (1 - r)\,\%$. The relationship between the non-centrality parameter and the power is non-linear. Consequently, analysing a marker with $r^2 = 0.5$ to an unanalysed causative allele may result in a drop in power of much less than 50 % if the power to detect the disease allele (if typed) is very high or a drop of much more than 50 % if the power is intermediate or low. The second point is that the r^2 result, at least as stated, only applies to one, probably suboptimal, way of analysing the data for association. Indeed, it only considers an experiment where a single marker has been studied. In reality, more complex models will be fitted to substantially more complex data sets in which the multiple-testing issue becomes important (because many markers will be analysed). It is therefore clear that the magnitude of pairwise r^2 values between alleles is only one factor in determining the power of an association study.

27.3 MODELLING LD AND GENEALOGICAL HISTORY

It should be clear by now that many different forces can influence the distribution of LD. These include molecular processes such as mutation and recombination, historical processes such as natural selection and population history and various aspects of experimental design (marker and subject ascertainment). If we are to have any hope of making useful inferences about the underlying processes from empirical data, we need to have a coherent framework within which to assess the way in which they can affect the patterns of variation we observe. One approach is to use simple probabilistic models to explore the distribution of patterns of LD we might observe under different scenarios. The aim of this section is to introduce such models and illustrate how they can be used to provide insights from empirical data. A key feature is the idea that genealogical models, specifically the coalescent with recombination (see **Chapter 25**), provide a flexible and intuitive approach to modelling genetic variation. The relation between features of the underlying genealogical history and properties of LD is also shown.

27.3.1 A Historical Perspective

Before introducing the coalescent perspective it is useful to provide a brief historical sketch of mathematical treatments of LD (see also **Chapter 22**). These approaches have given considerable insight into the nature of LD and, in contrast to most of the Monte Carlo based coalescent work, do provide simple analytical results about quantities of interest. The description given below is not chronological. All of these models assume a population of constant size and random mating.

27.3.1.1 The Relationship between LD and Two-locus IBD

Intuitively, the level of LD observed between two loci must be related to the extent to which they share a common ancestry (i.e. the extent to which the two loci have been co-transmitted in genealogical history). Indeed this is really the point of coalescent modelling. Completely linked loci share exactly the same ancestry and typically show high levels of LD, while unlinked loci have independent ancestries and typically show low LD. One way of quantifying the degree of shared ancestry between two loci is two-locus identity by descent (IBD). Single-locus IBD measures the probability that two chromosomes sampled at random from a population share a common ancestor before some defined point in the past (note that all chromosomes ultimately share a common ancestor so the equilibrium value of IBD is 1). The two-locus version simply extends the notion to measure the probability that two chromosomes sampled at random share a single common ancestor at both loci before some defined point in the past (and that there has been no crossover on either pathway from the sample chromosomes to the common ancestor). Note that IBD does not refer to identity in state (i.e. whether the two chromosomes carry the same alleles) rather it refers to relatedness. Over time, IBD within a population increases through genetic drift and decreases through recombination. An expression for the change in two-locus IBD, $Q = \Pr(\text{same ancestor})$, over time can be obtained (Sved, 1971) as a function of the diploid population size, N, and the per generation probability of a crossing-over event occurring between the two loci, c. Note that in the literature c can refer to either the genetic map distance between two loci (for details of how genetic

maps are constructed see **Chapter 1**) or the probability of crossing over (which cannot take a value greater than 0.5). In all cases considered here, c is sufficiently small that the genetic map distance and probability of crossover are approximately the same.

$$Q_{t+1} = \frac{1}{2N}(1-c)^2 + \left(1 - \frac{1}{2N}\right) Q_t (1-c)^2. \tag{27.22}$$

Solving for the equilibrium state gives

$$\tilde{Q} \approx \frac{1}{1+C}, \tag{27.23}$$

where $C = 4Nc$. Note that the same result can be obtained more naturally by thinking of the process backwards in time. Two-locus IBD is just the probability that the first 'event' that happens to the chromosomal segment flanked by the two loci on two randomly sampled chromosomes is a common ancestor event. That is, they coalesce before any crossover event occurs.

How does two-locus IBD relate to LD? Sved (1971) proposed the following argument. Consider sampling two chromosomes known to be identical in state at one locus proportionally to the allele frequencies at that locus. Both loci are bi-allelic with alleles A/a at the first and B/b at the second. It can be shown that the probability that these chromosomes are also identical at the second locus, P_H, is a function only of the homozygosity at the second locus, $F_B = f_B^2 + f_b^2$ where f_B is the frequency of the B allele etc. and the r^2 measure of LD between the alleles at the two loci

$$P_H = r^2 + (1-r^2)F_B. \tag{27.24}$$

If the crossover rate is large relative to the mutation rate then two chromosomes that show IBD at both loci are likely to also be homozygous at both loci (i.e. under these conditions IBD also implies identity in state). Consequently,

$$P_H \approx \tilde{Q} + (1 - \tilde{Q})F_B. \tag{27.25}$$

Combining expressions gives

$$r^2 \approx \tilde{Q} = \frac{1}{1+C}. \tag{27.26}$$

In short, the argument suggests that the expected value of r^2 is near 1 for very small crossover rates and approaches $1/C$ for $C \gg 1$.

Although this approximation is a useful heuristic (Chakravarti *et al.*, 1984) and is widely quoted (Jobling *et al.*, 2004), it is limited in application (Weir and Hill, 1986) because two chromosomes may share a common ancestor yet be different in allelic state due to a more recent mutation. Implicit within (27.25) is the assumption that allele frequencies do not change over time. For these reasons, (27.26) will only be a good approximation when the time scale over which two chromosomes at two loci may share a common ancestor before recombining is very short, which is only true for large values of C.

27.3.1.2 Matrix Methods and Diffusion Approximations

There are two alternative, and rather more rigorous, approaches to obtaining results about the expected value of LD statistics under simple population models. One approach is to use matrix recursions to describe the change in moments of LD statistics over time (Hill, 1975; 1977; Hill and Robertson, 1966; 1968). The other is to use a diffusion approximation (see also **Chapter 22**), replacing the discrete nature of genes in populations by a continuous space of allele frequencies (Ohta and Kimura, 1969a; 1969b; 1971). Although these methods appear somewhat different at first, they are actually closely related and can be used to examine both the dynamics of change in LD over time and to obtain expressions for quantities of interest at equilibrium. For example, although it is not possible to calculate the expected value of r^2 at equilibrium, it is possible to calculate a related quantity

$$\sigma_d^2 = \frac{E[D_{AB}^2]}{E[f_A(1 - f_A)f_B(1 - f_B)]},\tag{27.27}$$

for a pair (or a set of equivalent pairs) of bi-allelic loci, where f_A and f_B are the allele frequencies at two loci. Under the infinite-sites model (Karlin and McGregor, 1967; Kimura, 1969), the diffusion approximation leads to the solution (Ohta and Kimura, 1971)

$$\sigma_d^2 = \frac{10 + C}{22 + 13C + C^2}.\tag{27.28}$$

The same result can be obtained as the limit under low mutation rate from models of bi-allelic loci with a symmetric rate of mutation between alleles (Hill, 1975; Ohta and Kimura, 1969a). Like the expression of Sved this result predicts that for large C, the expected value of r^2 is approximately $1/C$. The main difference between the predictions of (27.26) and (27.28) is for small C where (27.28) predicts a value considerably less than 1 (a value of 5/11 for $C = 0$). Figure 27.5 compares estimates from Monte Carlo coalescent simulation. Neither approximation provides a particularly accurate prediction for the expected value of r^2, unless rare variants (loci where the rare variant is less than 10 % in frequency) are excluded. Nevertheless, (27.28) does predict the general shape of the decrease in average r^2 with increasing C.

27.3.2 Coalescent Modelling

The single most striking feature about Figure 27.5 is just how noisy LD is; the mean value of r^2 between loci at a given genetic distance captures very little of the complexity of the full distribution. This has two implications. First, it is hard to obtain an intuitive understanding of LD by thinking about 'expected' values of LD statistics. Second, the analysis of empirical data by comparing observed LD statistics to their 'expected' values is likely to be only weakly informative.

In order to capture the full complexity of LD patterns it is necessary to use stochastic modelling techniques to simulate the types of patterns one might observe under difference scenarios. The advent of coalescent modelling (Hudson, 1983b; Kingman, 1982), and specifically the coalescent with recombination (Hudson, 1983a), has led to a revolution in the way genetic variation data is approached. Coalescent models focus on properties of the sample by considering the genealogical history that relates a set of chromosomes to each other (see **Chapter 25**). For recombining data, the ancestral recombination
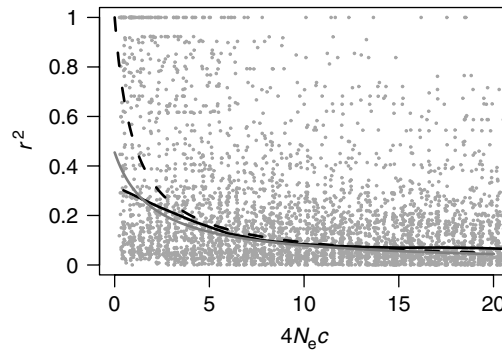
Figure 27.5 Analytical approximations for the expected r^2 between pairs of alleles as a function of the population genetic distance, $4N_ec$, between sites. Sved's approximation, (27.26): dashed black line. Ohta and Kimura's approximation, (27.28): solid grey line. Also shown (grey dots) are the values of r^2 between pairs of alleles at the corresponding genetic distance obtained from a single infinite-sites coalescent simulation with 50 chromosomes, $\theta = 100$, $C = 100$ and the sliding average (solid black line) for all pairs of sites where the minor allele is at frequency of at least 10 %. Although the Ohta and Kimura approximation performs quite well at predicting the mean r^2, its predictive power for any pair of sites is extremely poor due to high variance.

graph (ARG) generalises the idea of a 'tree' that relates individuals to each other to a complex network in which the trees at different positions along a chromosome are embedded (Griffiths, 1991). Figure 27.6 shows an ARG for two loci. Looking back in time there are two types of events: coalescent events, in which two lineages find a common ancestor, and recombination events, in which an ancestral lineage splits. The marginal trees at the two loci are embedded in the larger graph and the trees at the two loci, though different, do share some clades. Mutations that lead to variation within the sample occur along the branches of the ARG. Consequently, the structure of LD reflects the correlation structure of the underlying genealogical history. In other words, from a coalescent perspective the best way of understanding genetic variation is to think about the structure of the underlying genealogy of the sample. Different evolutionary forces have different effects on the shape and correlation structure of these underlying genealogies. Although this view is strongly driven by a neutral perspective (the mutations we observe have not themselves influenced genealogical history), the effects of certain types of natural selection, such as selective sweeps or balancing selection, on patterns of linked neutral variation can also be considered from a coalescent or genealogical viewpoint.

In addition to the various theoretical insights that coalescent theory has made possible, a genealogical approach has greatly enabled the analysis of genetic variation through the ability to simulate data under various evolutionary models. In the next three sections the use of a genealogical framework to understand the distribution of LD is described.

27.3.2.1 *LD Patterns in the Absence of Recombination*

It may sound strange, but we can actually learn a lot about LD by studying its behaviour in regions where there is no recombination (Slatkin, 1994). Consider the three data
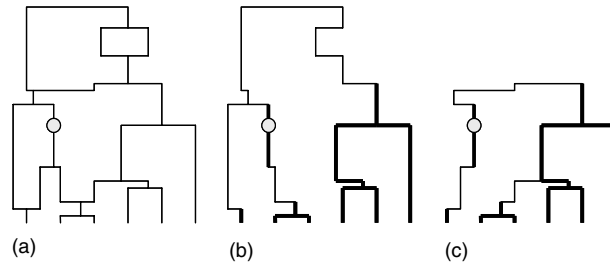
Figure 27.6 An example ancestral recombination graph for two loci and six chromosomes. (a) The history of the chromosomes can be traced back in time to coalescent events (where pairs meet their common ancestors) and crossover events, which split the lineage into two. The resulting trees at the two loci, (b) and (c), share some branches (indicated by the thick lines). Because of recombination, a mutation on a particular shared internal branch, indicated by the circle, will be present in different members of the sample depending on which locus it occurs at.
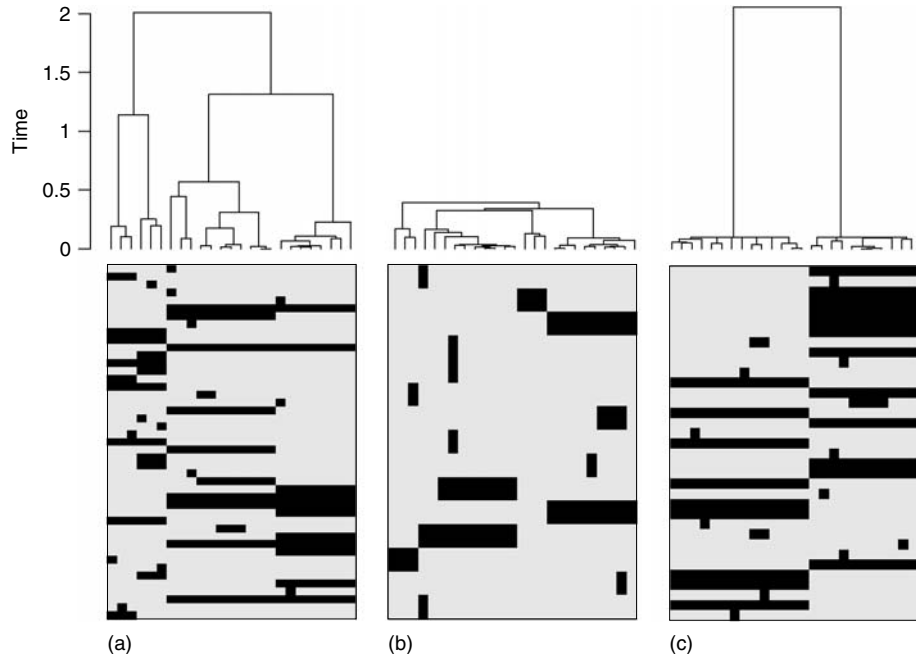
Figure 27.7 Linkage disequilibrium in non-recombining regions. Despite each example having no recombination, there can be low, moderate or high LD depending on the structure of the underlying genealogy. Data simulated with either (a) a constant population size, (b) a strongly growing population or (c) a population that has experienced a severe recent bottleneck.

sets and their corresponding genealogies in Figure 27.7 corresponding respectively to simulations with a constant population size, a growing population and a population that has experienced a strong recent bottleneck (though not one strong enough to wipe out all pre-existing variation). Apart from the differences in the numbers of polymorphic sites, there are also strong differences in their structuring. First note that any two mutations
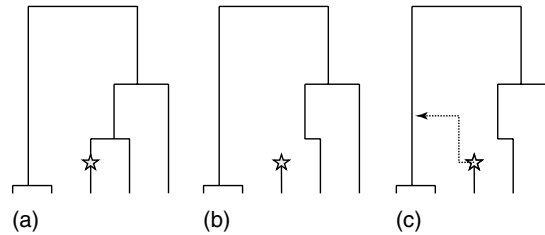
Figure 27.8 Crossover events as a point process along the sequence (Wiuf and Hein, 1999). (a) At any point along the sequence there is an underlying genealogical tree. The distance to the next crossover event is exponentially distributed with rate dependent on the total branch length of the current tree and the crossover rate. Having chosen the physical position of the crossover event, the position on the tree, indicated by the star, is chosen uniformly along the branch lengths. (b) In an approximation to the coalescent process, called the *SMC* (McVean and Cardin, 2005), the branch immediately above the recombination point is erased leaving a floating lineage. (c) The floating lineage coalesces back into the remaining tree and the process is repeated. In the coalescent, no branch can be erased.

at two distinct sites that occur on the same branch in the tree will occur in exactly the same members of the sample, and hence be in 'perfect' LD, i.e. $r^2 = 1$. Next note that mutations on different branches will be only very weakly correlated, $r^2 << 1$, if the branches are in different parts of the tree and particularly if one or both are near the tips. It follows that the extent of LD reflects the extent to which the tree shape is dominated by a few long branches, as in the case of the bottlenecked population (on which lots of mutations occur and which are therefore in perfect association), or is not, as in the case of the growing population (mutations occur on branches in different parts of the tree, particularly towards the tips, and typically show weak association). It follows that (at least for r^2) LD is strongest for the bottlenecked population and weakest for the growing one. The constant-size population shows a mixture of highly correlated and weakly correlated alleles, reflecting the distribution of mutations all across the tree. Of course, because of the inherent stochasticity in the coalescent process it would be unwise to make inferences about population history from a single non-recombining locus that showed the patterns in Figure 27.8(b or c).

Digressing slightly, it is interesting to note how classical population genetics theory concerning the distribution of genetic variation under the infinite-alleles model can be related to study the structure of LD in infinite-sites models without recombination. For example, the number of distinct haplotypes in the sample, H, is equivalent to the number of distinct alleles, k. The classic result then gives an expectation for these quantities for the case of constant-sized neutral populations (Ewens, 1972),

$$E[k] = \sum_{i=0}^{n-1} \frac{\theta}{i + \theta}. \qquad (27.29)$$

Here $\theta = 4N_e u L$ is the population mutation rate over the region of interest, where u is the per site, per generation mutation rate and L is the number of sites. Similarly, the Ewens sampling formula (Ewens, 1972) describes the distribution of the numbers of each

distinct haplotype conditional on the total number of observed haplotypes

$$\Pr(n_1, n_2, \ldots, n_k | k, n) \propto \frac{1}{n_1 n_2 \ldots n_k}. \tag{27.30}$$

This result inspired the first statistical test for the hypothesis that the region of interest is evolving neutrally: the Ewens–Watterson homozygosity test (Watterson, 1977; 1978), which compares the observed (haplotype) homozygosity to the distribution expected from the above formula. Indeed, thinking about the effect of recombination on this test first led to the development of the coalescent model with recombination (Hudson, 1983a; Strobeck and Morgan, 1978). Recombination tends to increase the number of observed alleles (haplotypes) and reduces the skew in allele (haplotype) frequency resulting in a systematic decrease in (haplotype) homozygosity.

27.3.2.2   LD in Recombining Regions

When recombination occurs within a region, different positions may have different, though correlated, trees (Figure 27.6). This raises a series of questions. How does recombination change tree structure? What is the relationship between correlation in tree structure and LD? How should we measure the correlation in trees? To answer these questions it is first helpful to think about how trees 'evolve' along a sequence through recombination (Wiuf and Hein, 1999). The following argument gives a sense of how genealogical history changes because of recombination, specifically crossover events, by considering how to simulate a sequence of trees along a unit region (i.e. a region where the start is labelled 0, and the end labelled 1) over which there is a constant crossover rate of C. Readers should be aware that what follows is an approximation, referred to as the *spatially Markov coalescent* (SMC) (McVean and Cardin, 2005). Nevertheless, the distribution of data generated by the SMC process is almost indistinguishable from that generated by the true coalescent process (Marjoram and Wall, 2006; McVean and Cardin, 2005).

1. Simulate a coalescent tree (i.e. no recombination) at the far left-hand edge of the region. The total tree length (in units of $2N_e$ generations) is T_L.

2. The distance along the region to the next crossover event is exponentially distributed with rate $T_L C/2$. If the position of the next crossover event lies within the unit region, continue, otherwise stop.

3. Choose a point within the tree to recombine uniformly along its branches. Erase the remainder of the branch immediately above the chosen point.

4. Allow the recombined lineage to coalesce back to the remaining lineages at a rate proportional to the number of non-recombined lineages present (note this has to be updated if there are coalescent events among these). Note also that the recombined lineage could coalesce beyond the most recent common ancestor (MRCA) of the current tree.

5. Now assign the total length of the new tree to T_L. Return to step 2.

These steps are illustrated in Figure 27.8. Several features should be clear from this approximation. First, trees change over the region by small steps. Of course, lots of steps

(resulting from large crossover rates) result in lots of changes, so the tree at the right hand of the sequence looks very different from that at the left hand. Second, because of the structure of coalescent trees, many of the crossover events occur deep in the tree when there are relatively few lineages. These often have remarkably little effect on the distribution of genetic variation and crossover events during the phase when only two lineages are present are essentially undetectable. Third, if a pair or group of sequences shares a very recent common ancestor, this part of the tree will persist over considerable genetic distances. Consequently, it is often possible to identify pairs or small groups of sequences that are identical over extremely long regions. To illustrate the effect consider the statistic T_S, the sum of the lengths of the branches shared by the trees at the start and end of a sequence. The expectation of T_S under the SMC is

$$E[T_S] = 2 \sum_{i=1}^{n-1} \frac{1}{i + C}. \tag{27.31}$$

For $n > C$ the proportion of the total expected tree length that is shared is approximately $1 - \log(C)/\log(n)$ (note that in the coalescent the shared time is likely to be slightly higher). Consequently, for large sample sizes a considerable proportion of the tree can be shared even when the total crossover rate is very high. For example, with 1000 sequences, over 15 % of the total expected time is shared by points separated by $C = 400$, which in humans of European origin corresponds to about 1 cM. Of course, if the tree shape is strongly dominated by large recent clades, as e.g. happens if there has been a recent and strong but partial selective sweep, such parts of the tree can persist over much greater genetic distances. These considerations emphasise the fact that LD is very heterogeneous, not just between genomic regions but also between different individuals within a sample.

27.3.2.3 LD in Populations with Geographical Subdivision and/or Admixture

Although the notion of a tree changing along the sequence provides useful insights into the nature of LD, there is one aspect that, at least at first glance, it fails to describe: LD between alleles at unlinked loci (e.g. on different chromosomes) in structured populations. As demonstrated in Figure 27.4, when the sample contains individuals from one or more populations LD can exist even between unlinked loci because of differences in allele frequencies between populations. More generally, any deviation from random mating will lead to systematic spatial structuring of genetic variation and LD even between unlinked loci. How can we understand this phenomenon within a genealogical context?

The answer is simple. LD is determined by the correlation in genealogical history along a chromosome. Recombination acts to break down such correlations, but if there are biases induced by geographic or cultural factors in terms of which lineages can coalesce, some correlation will persist indefinitely. Figure 27.9 illustrates this idea by considering independent coalescent trees sampled from a pair of populations that diverged some time ago. Despite their independence, both trees show a strong clustering of individuals from the same populations. In short, while directly sharing genealogical trees results in
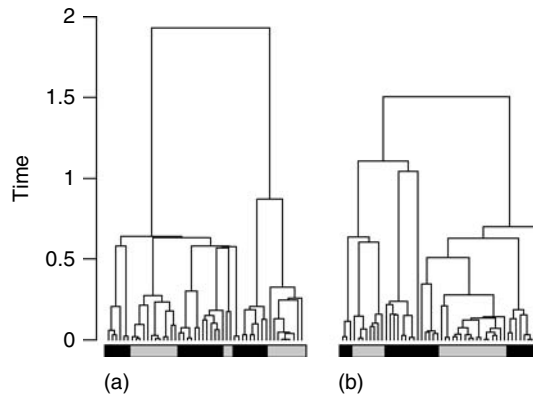
Figure 27.9 Correlations in genealogical history at unlinked loci. Individuals from two populations, indicated by the grey and black bars below, have been sampled at two unlinked loci. Although the trees are independent, there is nevertheless genealogical correlation because a pair of individuals sampled from within a population are likely to have a more recent common ancestor at both loci than a pair of chromosomes where one from each population has been sampled. Data simulated with 25 chromosomes sampled from each of two populations of equal size that diverged exactly N_e generations ago from an ancestral population of the same size.

genealogical correlation and hence LD, genealogical correlations can also arise indirectly by forces shaping the nature of coalescence.

27.3.3 Relating Genealogical History to LD

It should be clear by now that LD is a reflection of correlation in the genealogical history of samples. Informally, if the coalescence time for a pair of sequences sampled at a given locus is informative about the coalescence time for the same pair of sequences at a second locus (relative to the sample as a whole), we expect variation at the two loci to exhibit significant LD. But exactly what is the relationship? Is it possible to be more quantitative about which aspects of genealogical correlation relate to which measures of LD?

A partial answer to this question comes from studying the quantity σ_d^2; see (27.27). It is well known that D^2 between alleles at a pair of loci, x and y, can be written in terms of two-locus identity coefficients (Hudson, 1985; Strobeck and Morgan, 1978).

$$D_{xy}^2 = F(C_{ij}^x, C_{ij}^y) - 2F(C_{ij}^x, C_{ik}^y) + F(C_{ij}^x, C_{kl}^y). \tag{27.32}$$

The three terms relate to the probabilities of identity at the two loci for pairs of chromosomes sampled in different ways. Consider sampling four chromosomes from a population and labelling them i, j, k and l. The first identity coefficient compares chromosomes i and j at both the x and the y loci. The second compares chromosomes i and j at the x locus, but i and k at the y locus. The third compares chromosomes i and j at the x locus and k and l at the y locus. The expectation of these identity coefficients over evolutionary replicates therefore determines the average strength of the LD.

What is the relationship between identity coefficients and genealogical history? The key point is that each identity coefficient relates to the probability that no mutation has
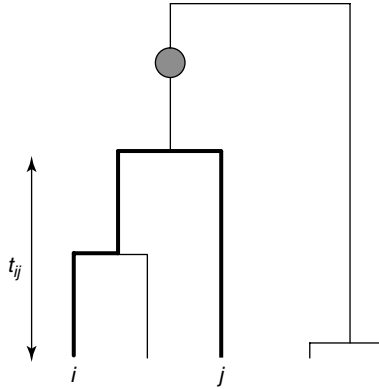
Figure 27.10 Identity in state and genealogical history. Two chromosomes, i and j, will be identical in state if no mutation occurs on those branches of the tree that lead to their common ancestor(s), at time t_{ij}. It follows that two-locus identity in state reflects whether mutations at the two trees fall on the branches to the pair's common ancestor. In the limit of a low mutation rate, but conditioning on segregation at the two loci in a sample of size n, identity in state can be written as a function of the correlations in coalescence time at the two loci, see (27.34).

occurred on the branches of the genealogy that link two chromosomes to their common ancestor (see Figure 27.10). If we condition on a single mutation occurring at both loci within a sample of size n and let the mutation rate tend towards 0, we get the result (McVean, 2002)

$$
\begin{aligned}
E[F(C_{ij}^x, C_{ij}^y)] &= \lim_{\theta \to 0} \frac{E[(T_x - 2t_{ij}^x)(T_y - 2t_{ij}^y)e^{-\theta(T_x + T_y)/2}]}{E[T_x T_y e^{-\theta(T_x + T_y)/2}]} \\
&= \frac{E[(T_x - 2t_{ij}^x)(T_y - 2t_{ij}^y)]}{E[T_x T_y]},
\end{aligned}
\tag{27.33}
$$

where T_x is the sum of the branch lengths in the tree at locus x and t_{ij}^x is the coalescent time for chromosomes i and j at locus x. Similar results can be found for the other identity coefficients and for the denominator in (27.27). Combining these equations, we get the following result:

$$
\sigma_d^2 = \frac{\rho(t_{ij}^x, t_{ij}^y) - 2\rho(t_{ij}^x, t_{ik}^y) - \rho(t_{ij}^x, t_{kl}^y)}{E[t]^2/\text{Var}(t) + \rho(t_{ij}^x, t_{kl}^y)},
\tag{27.34}
$$

where $\rho(t_{ij}^x, t_{kl}^y)$ indicates the Pearson correlation coefficient between the coalescent time for chromosomes i and j at locus x and the coalescent time for chromosomes k and l at locus y and the first term in the denominator is the ratio of the square of the expected coalescence time for a pair of sequences to its variance.

Equation (27.34) has two important features. First, the required correlations in time to the common ancestor can be obtained from coalescent theory (Griffiths, 1981; Pluzhnikov and Donnelly, 1996), replicating the result of (27.28). More importantly, (27.34) gives a way of understanding the behaviour of LD, or perhaps more appropriately r^2, under more complex population genetic models. For example, the increase in LD

that accompanies population bottlenecks can largely be understood through its effect on the ratio of the mean coalescence time to its standard deviation. Bottlenecks increase the variance in coalescent time considerably, leading to a reduction in the denominator of (27.34) and an increase in LD (McVean, 2002). It is also possible to describe the behaviour of LD under models of population structure (McVean, 2002; Wakeley and Lessard, 2003) and even selective sweeps (McVean, 2007) using the same approach.

27.4   INFERENCE

Coalescent modelling provides a framework within which to understand patterns of LD (see also **Chapter 26**). But what we are usually interested in is making inferences about underlying processes from the patterns of genetic variation observed in an experiment. The aim of this section is to introduce key concepts in statistical inference and to explore how these relate to understanding patterns of genetic variation.

27.4.1   Formulating the Hypotheses

Let us return to the example of the *Lactase* gene. Various summaries of the data have been presented: graphically in Figure 27.2 and numerically in Table 27.1. While the observed patterns are suggestive of various underlying processes, we need to approach the analysis of the data in a more rigorous fashion. Indeed, we need to start by asking ourselves why we are analysing this region in the first place. Do we want to categorically prove or disprove the hypothesis that it has experienced a selective sweep? Do we want to identify the simplest model that provides an adequate explanation for the data? Do we want to use these data as an aid to design an experiment to understand the genetic basis of lactose intolerance? The motivation determines, and is inseparable from, the approach to statistical inference.

Suppose that our goal is to determine whether a selective sweep has happened in the region. How might we go about doing this? Suppose that we knew exactly the process that shapes genetic variation in the absence of selection, which is referred to as Ω (a complete description of mutation rates, crossover and gene conversion rates, changes in population size, geographic structuring and experimental design). In effect this means being able to accurately characterise the distribution of sampled genetic variation in the absence of selection, from which we can calculate the probability (density) of the collected data (the likelihood of Ω). Suppose, in addition, that we can calculate the likelihood under a model that includes Ω, but also a selective sweep, which is referred to as Π (in what follows it is important that both Ω and Π have specified, though possibly multidimensional, values). Then the Neyman–Pearson lemma (e.g., Casella and Berger, 1990) tells us that the most powerful test compares the likelihood of the data with the sweep to the likelihood of the data without the sweep. Specifically, the test compares the likelihood ratio test statistic against the quantiles of the null distribution such that the null is rejected if

$$\Lambda(x) = \frac{L(\Pi, \Omega; x)}{L(\Omega; x)} > k, \text{ where } \Pr(\Lambda(X) > k | H_0) = \alpha. \qquad (27.35)$$

Here x refers to the data, X refers to the random variable of which x is a realisation and α is the size of specified rejection region (e.g. $\alpha = 0.01$).

Although (27.35) looks useful, unfortunately careful consideration reveals many problems. First, it is very unlikely that we can exactly specify the component values of Ω. In fact, the only source of information about many of its details comes from the data we have just collected. Second, a related problem is that we would not want to restrict ourselves to considering one specific value of Π, rather we would like to estimate the time of origin of the mutation, its location in the sequence and the strength and nature of the selection pressure from the data. This raises the problem that estimation of Π may influence estimates of Ω. Third, there is no simple way of calculating the likelihood given the data. Fourth, we can, of course, never really know Ω, all we can hope to do is to capture its key features. Finally, though perhaps least importantly, even if we could estimate all these parameters from the data, (27.35) does not actually tell us which hypothesis test or other method of model choice is the most powerful in practice (e.g. maximum likelihood, etc.).

So how might we proceed? Although it seems daunting, we have to try to learn about the components of Ω from the data and through any other sources of information. For example, we can refer back to the experimenter's notebook to see how the experiment was designed. We can also learn about aspects of Ω from genetic variation in other regions of the genome. For example, the population history, which should affect all loci more or less equally, and the mutation rate, which is largely predictable from the DNA sequence (see **Chapter 31**). To learn about some processes, such as the fine-scale structure of recombination rate variation, we may have to use the data, but in the case of *Lactase* we have three populations, only one of which is hypothesised to have experienced selection. In general, when estimating the components of Ω we are looking for estimators with four key properties:

- Statistical sufficiency: the estimator uses as much of the information in the data about the parameter of interest as possible.
- Relative efficiency: the estimator with the lowest mean-square error is preferred. Accurate estimates of uncertainty are also important.
- Robustness: the methods are not strongly influenced by deviations from the underlying assumptions.
- Computational tractability: estimation can be achieved within a reasonable time frame (perhaps no more than a month!).

It should be stressed that in most population genetics problems fully efficient inference is currently beyond the limits of computational tractability (see **Chapter 26**). For this reason, most current approaches to estimating parameters from genetic variation use methods that are less efficient but computationally tractable.

27.4.2 Parameter Estimation

To illustrate various approaches to parameter estimation in the context of understanding LD the problem of how to estimate the crossover rate, or rather the population crossover rate, over a region is considered. All of the methods described consider only a neutrally evolving population of constant size. The point of covering the various estimators is to give an indication of the range of possible approaches.

27.4.2.1   Moment Methods

The first population genetic method for estimating the population crossover rate (Hudson, 1987; Wakeley, 1997) used a method of moments approach. Using the results mentioned above that relate LD coefficients to two-locus sampling identities, Hudson derived an expression for the expected sample variance of the number of nucleotide differences between pairs of sequences under the infinite-sites model. If the number of pairwise differences between sequences i and j is π_{ij} then under the infinite-sites model

$$\bar{\pi} = \frac{1}{n^2} \sum_{i,j}^{n} \pi_{ij}$$

$$S_\pi^2 = \frac{1}{n^2} \sum_{i,j}^{n} \pi_{ij}^2 - \bar{\pi}^2$$

$$E[\bar{\pi}] = (1 - 1/n)\theta$$

$$E[S_\pi^2] = f(\theta, C, n), \tag{27.36}$$

where $f(\theta, C, n)$ is a known function of the two unknown parameters and the sample size (Hudson, 1987). To obtain a point estimate of C, θ is replaced with a point estimate from the sample (also obtained by the method of moments) and the equation is solved (if possible) for C. Similar approaches could be constructed for other single-number summaries of the data (e.g. those in Table 27.1), although Monte Carlo methods would have to be used to obtain estimates.

The great strength of this estimator is its simplicity. Unfortunately, the estimator also has very poor properties (bias, high variance, lack of statistical consistency, undefined values). This is partly because of the inherent stochasticity of the coalescent process; so for any given value of C and θ, there is a huge amount of variation in the observed patterns of variation (hence any estimator of C is expected to have considerable variance). But it is partly also because it only uses a small fraction of the total information about crossover present in the data. It is not known how well moment estimators from other sample properties might perform.

27.4.2.2   Likelihood Methods

As suggested above, a useful quantity to compute is the likelihood of the parameter (proportional to the probability, or probability density, of observing the data given the specified parameter value). Ideally, we would like to calculate the probability of observing the data given the coalescent model and specified values of the population parameters, perhaps choosing the values that maximise the likelihood as point estimates. While this problem has received considerable attention (Fearnhead and Donnelly, 2001; Kuhner et al., 2000; Nielsen, 2000), currently such approaches are only computationally feasible for small to moderate sized data sets (Fearnhead et al., 2004). A common feature of all these methods is the use of Monte Carlo techniques (particularly Markov chain Monte Carlo and importance sampling). For example, in importance sampling, a proposal function, Q, is used to generate coalescent histories, H (a series of coalescent, mutation

and recombination events), compatible with the data for given values of θ and C. The likelihood of the data can be estimated from

$$L(\theta, C; x) = E_Q \left[\frac{\Pr(H|\theta, C)}{Q(H|\theta, C, x)} \Pr(x|H) \right], \tag{27.37}$$

where $\Pr(H|\theta, C)$ is the coalescent probability of the history (note that this is not a function of the data), $\Pr(x|H)$ is the probability of the data given the history (which is always exactly one here) and $Q(H|\theta, C, x)$ is the proposal probability for the history (note that this is a function of the data). The main problem is that unless Q is close to the optimal proposal scheme,

$$Q_{\text{OPT}} = \Pr(H|\theta, C, x), \tag{27.38}$$

the large variance in likelihood estimates across simulations makes obtaining accurate estimates of the likelihood nearly impossible.

For larger data sets a more practical alternative is to calculate the likelihood from some informative summary (or summaries) of the data using Monte Carlo techniques (Beaumont *et al.*, 2002; Marjoram *et al.*, 2003; Padhukasahasram *et al.*, 2006; Wall, 2000). Wall's estimator, based on the number of haplotypes and a non-parametric estimate of the minimum number of detectable recombination events, while conditioning on the number of segregating sites, is good in terms of having low bias and variance comparable to the best alternatives. A great strength of these methods is that they can provide useful summaries of uncertainty in estimates, e.g. through estimating the Bayesian posterior distribution of the parameter. However, because some information is thrown away, the resulting uncertainty will be greater than the true uncertainty. Furthermore, although likelihood surfaces computed from a summary of the data are expected to mirror likelihood surfaces computed from full data, such agreement is not guaranteed and, for some pathological data sets, they may disagree considerably.

27.4.2.3 *Approximating the Likelihood*

Although it may not be practical to calculate the coalescent likelihood of the entire data, it is possible to approximate the likelihood function through multiplying the likelihoods for subsets of the data (Fearnhead *et al.*, 2004; Hey and Wakeley, 1997; Hudson, 2001; McVean *et al.*, 2002; 2004; Wall, 2004), the resulting quantity being referred to as a *composite likelihood*. In effect, the idea is to treat subsets of the data as if they were independent of each other whereas in reality these subsets (pairs, triples or non-overlapping sets) are clearly not. Nevertheless, finding the value of the crossover parameter that maximises this function appears to provide estimates that are at least as accurate as any other approximate approach.

Perhaps the greatest strength of the composite likelihood approaches lies in their flexibility, particularly in the use of estimating variable crossover rates and identifying crossover hotspots (Fearnhead and Donnelly, 2002; Fearnhead and Smith, 2005; McVean *et al.*, 2004; Myers *et al.*, 2005). For example, consider Hudson's composite likelihood approach, which considers pairs of sites. For each pair of sites it is possible to pre-compute the coalescent likelihood for a specified value of θ per site and a range of crossover rates between the sites, e.g. using the importance sampling approach (McVean *et al.*, 2002).

The likelihood of a genetic map $g = \{g_1, g_2, \ldots, g_m\}$, where each g_i is the map position of the ith site in a set of m ordered sites, is approximated by

$$L_C(\theta, g; x) = \prod_{i,j>i} L(\theta, g_j - g_i; x_{ij}). \tag{27.39}$$

Here L_C indicates the composite likelihood, L indicates the coalescent likelihood for the genetic distance between loci i and j and x_{ij} is the data at those two sites. In practice, the value of θ is estimated previously using a moment method. Because of the pre-computation, searching over the space of g is computationally feasible. McVean *et al.* (2004) used a Monte Carlo technique called *reversible jump Markov chain Monte Carlo* (Green, 1995) to explore the space of possible genetic maps. The greatest weakness of this approach is the difficulty in assessing uncertainty. Specifically, because of non-independence between the pairs of sites, the composite likelihood surface is typically more sharply peaked compared to the true likelihood surface (even though it uses less information), resulting in considerable underestimation of uncertainty.

27.4.2.4 Approximating the Coalescent

An alternative approach to approximating the coalescent likelihood function is to devise an alternative model for the data, motivated by an understanding of the coalescent, but under which it is straightforward to calculate the likelihood (Crawford *et al.*, 2004; Li and Stephens, 2003). The product of approximate conditionals (PAC) scheme uses the following decomposition to motivate an approximate model (see **Chapter 26**). Consider the data $x = \{x_1, x_2, \ldots, x_n\}$, where x_i is the ith haplotype in a sample of size n. The likelihood function can be written as the product of a series of conditional distributions (the dependence on θ has been dropped for simplicity)

$$L(C; x) = \Pr(x_1|C) \Pr(x_2|x_1, C) \ldots \Pr(x_n|x_1, x_2, \ldots, x_{n-1}, C). \tag{27.40}$$

Of course, knowing the conditional probabilities is equivalent to knowing the coalescent likelihood. However, the PAC scheme approximates the conditional probabilities by considering the kth haplotype as an imperfect mosaic of the previous $k-1$. Specifically, the model employed has the structure of a hidden Markov model in which the underlying state at a given nucleotide position refers to which of the $k-1$ other chromosomes the kth is derived from. The transition probabilities are a function of the crossover rate and the emission probabilities are a function of the mutation rate and the sample size. This model captures many of the key features of genetic variation, such as the relatedness between chromosomes, how this changes through crossover and how, as the sample size increases, additional chromosomes tend to look more and more like existing ones, but it is entirely non-genealogical.

This approach was actually derived from the work mentioned above that uses impor-tance sampling to calculate the full coalescent likelihood of the data (Fearnhead and Donnelly, 2001; Stephens and Donnelly, 2000). Because of the Markovian structure of the coalescent, each history that is compatible with the data can be broken down into a series of events that can be sampled sequentially. The optimal proposal density chooses an

event (coalescence, mutation and crossover), e, according to its probability given the data

$$Q_{\text{OPT}}(e|\theta, C, x) = \Pr(e|\theta, C)\frac{\Pr(x + e|\theta, C)}{\Pr(x|\theta, C)}, \tag{27.41}$$

where $\Pr(e|\theta, C)$ is the coalescent probability of the event and $x + e$ is the original data, x, modified by event e. The key insight is to note that $x + e$ and x are very similar. For example, if e is a coalescent event between the ith and jth sequences (which are also identical) it follows that

$$\frac{\Pr(x + e|\theta, C)}{\Pr(x|\theta, C)} = \frac{\Pr(x_1, \ldots, x_i, \ldots, x_k|\theta, C)}{\Pr(x_1, \ldots, x_i, x_j, \ldots, x_k|\theta, C)}$$

$$= \frac{1}{\Pr(x_j|x_1, \ldots, x_i, \ldots, x_k, \theta, C)}. \tag{27.42}$$

This is exactly the same conditional probability approximated in the PAC scheme. For other types of events, similar expressions involving the conditional distributions can be found (Fearnhead and Donnelly, 2001).

The strengths of the PAC scheme are its computational tractability and the sense that it uses much of the information in the data about recombination. It is also very flexible and can potentially be extended to include features such as geographical structuring and gene conversion. Nevertheless, because the model is not the coalescent, parameter estimates are typically biased in ways that are unpredictable (and the estimated uncertainty in estimates may not necessarily reflect the 'true' uncertainty). The approximation also introduces an order dependency into the likelihood function (the true conditionals would, of course, give rise to the same likelihood no matter how the sequences were considered). This can be partly overcome by averaging inference over multiple orderings.

27.4.3 Hypothesis Testing

Aside from estimating parameters, it is also often of interest to make categorical statements about whether particular hypotheses can be rejected. For example, in the case of *Lactase*, we want to attempt to categorically reject the hypothesis that the European population has not experienced adaptive evolution. More generally, we would like to make some choice about which model provides a better explanation of the data: one with selection or one without it.

Historically, most approaches to detecting the influence of natural selection in population genetic data have used a slightly different approach. Specifically, a battery of neutrality tests including Tajima's D (Tajima, 1989), Fu and Li's D (Fu and Li, 1993), Fay and Wu's H (Fay and Wu, 2000), the Hudson, Kreitman, Aguadé (HKA) test (Hudson *et al.*, 1987), the haplotype partition test (Hudson *et al.*, 1994), the extended haplotype homozygosity test (Sabeti *et al.*, 2002; Voight *et al.*, 2006) and more are used to reject the null hypothesis of a neutral model, without rigorously demonstrating that a model that includes natural selection is a better fit to the data. All of these methods are designed to look for patterns in the data that are inconsistent with neutral evolution (Sabeti *et al.*, 2006). Furthermore, Monte Carlo simulation under different scenarios can be used to assess power and whether the methods are robust to deviations from the assumed model. A second class of methods look for large between-population differences in genetic variation; e.g. striking changes in allele frequency (Akey *et al.*, 2002; 2004; Lewontin and

Krakauer, 1973). Such methods aim to identify geographically restricted selective sweeps, as thought to be the case for *Lactase*. However, their major drawback is that it is very hard to accurately specify the null model, because this requires accurate specification of the demographic histories of the populations under study. For this reason, population differentiation–based tests typically rely on comparing a locus to an empirical distribution derived from the analysis of multiple genomic regions, with the implicit assumption that loci with unusual patterns have been targeted by selection. It is currently unclear how well empirical methods perform at identifying true selected loci (Teshima *et al.*, 2006), though such methods have identified strong candidates for local selective sweeps (McVean *et al.*, 2005; Sabeti *et al.*, 2006).

There is, however, a class of methods that explicitly attempt to fit a model of a selective sweep to patterns of genetic variation (Coop and Griffiths, 2004; Kim and Stephan, 2002; Nielsen *et al.*, 2005; Przeworski, 2003). For example, one method uses a theoretical approximation to predict how allele frequencies are distorted around a selected allele, which is fitted to data by a composite likelihood approach (Nielsen *et al.*, 2005). There are also related methods that attempt to date the age of a potentially selected mutation, e.g. by examining the amount of variation found 'under' a mutation (Slatkin, 2000; Slatkin and Rannala, 2000; Toomajian *et al.*, 2003). In effect, the age of a mutation is all that can ever be estimated from genetic variation. The evidence for adaptive evolution comes from observing recent mutations that exist at much higher frequency in a population than expected from neutral mutations of the same age.

To understand the signal these methods are looking for, consider the genealogy in Figure 27.11, in which a recent mutation has risen to high frequency over a short timescale. The effect is to create drastically different genealogical histories for different subsets of the chromosomes. Those under the selected mutation have a very short time to their MRCA compared to the much deeper tree for the other sequences. The impact of this genealogical distortion on genetic variation will be to dramatically reduce genetic variation among those chromosomes that carry the beneficial mutation. Furthermore, because of the short amount of time in the tree under the mutation, it will be some considerable genetic distance from the selected site before crossover has allowed the genealogy to return to the neutral distribution. These features are particularly striking in the *Lactase* example, where the haplotype structure around the selected mutation persists for several hundred kilobases from the gene. However, it is also a region of very low crossover, which is probably why the signal is so remarkable. In regions of high crossover the power to detect recent adaptive evolution may be considerably lower.

27.5   PROSPECTS

The aim of this chapter has been to introduce a way of thinking about patterns of genetic variation – the coalescent with recombination – and to explore how to analyse such data. The reader should be aware that the field of population genetic inference is extremely dynamic, with new methods appearing at a high rate (Marjoram and Tavare, 2006). The reader should also be aware that all the methods discussed above for estimating crossover rates and detecting natural selection have considerable limitations and none are optimal. It seems likely that the next few years will see intensive research in the development of more sophisticated and perhaps fully genealogical approaches to the analysis of genetic
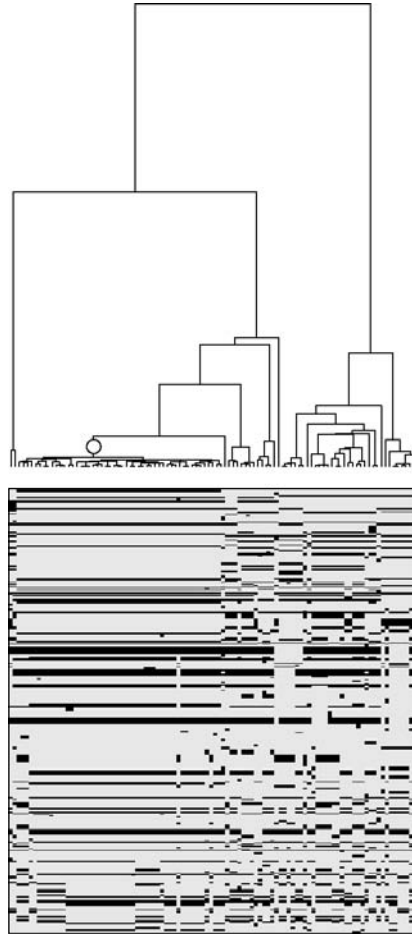
Figure 27.11 The effect of a partial selective sweep on linked genetic variation. A recent selected mutation distorts genealogical history creating a clade within the tree at the selected site with a very recent common ancestor. This clade 'persists' for long genetic distances leading to a marked haplotype structuring amongst those individuals that carry the beneficial mutation. The signature is, however, slowly eroded by crossover. Here the haplotypes are shown vertically, with the genealogy at the selected locus at the top and the selected mutation indicated by the circle. Data were simulated using the program SelSim (Spencer and Coop, 2004) for 100 chromosomes, 50 of which carry the beneficial mutation which has a scaled selective advantage of $4N_\mathrm{e}s = 400, \theta = 200, C = 50$.

variation. Indeed, some genealogical methods are already appearing, particularly in the field of association mapping (Larribe *et al.*, 2002; Minichiello and Durbin, 2006; Zollner and Pritchard, 2005).

Acknowledgments

RELATED CHAPTERS

Chapter 1; **Chapter 22**; **Chapter 25**; **Chapter 26**; **Chapter 31**; and **Chapter 37**.

REFERENCES

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A. and Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology* **2**, e286.

Akey, J.M., Zhang, G., Zhang, K., Jin, L. and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Research* **12**, 1805–1814.

Anderson, E.C. and Novembre, J. (2003). Finding haplotype block boundaries by using the minimum-description-length principle. *American Journal of Human Genetics* **73**, 336–354.

de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J. and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics* **37**, 1217–1223.

Balding, D.J. and Nichols, R.A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3–12.

Beaumont, M.A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E. and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics* **74**, 1111–1120.

Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* **74**, 106–120.

Casella, G. and Berger, R.L. (1990). *Statistical Inference*. Duxbury, Belmont, California.

Chakravarti, A., Buetow, K.H., Antonarakis, S.E., Waber, P.G., Boehm, C.D. and Kazazian, H.H. (1984). Nonuniform recombination within the human beta-globin gene cluster. *American Journal of Human Genetics* **36**, 1239–1258.

Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human Heredity* **56**, 18–31.

Coop, G. and Griffiths, R.C. (2004). Ancestral inference on gene trees under selection. *Theoretical Population Biology* **66**, 219–232.

Crawford, D.C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M.J., Nickerson, D.A. and Stephens, M. (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics* **36**, 700–706.

Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322.

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Fay, J.C. and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413.

Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318.

Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society, Series B* **64**, 657–680.

Fearnhead, P., Harding, R.M., Schneider, J.A., Myers, S. and Donnelly, P. (2004). Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics* **167**, 2067–2081.

Fearnhead, P. and Smith, N.G. (2005). A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. *American Journal of Human Genetics* **77**, 781–794.

Fu, Y.X. and Li, W.H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N, Rotimi, M., Adeyemo, A., Cooper, R, Ward, R., Lander, E.S., Daly, M.J., Atshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* **296**(5576), 2225–2229.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Griffiths, R.C. (1981). Neutral two-locus multiple allele models with recombination. *Theoretical Population Biology* **19**, 169–186.

Griffiths, R.C. (1991). The two-locus ancestral graph. In *Selected Proceedings on the Symposium on Applied Probability*, I.V. Basawa and R.L. Taylor, eds. Institute of Mathematical Statistics, Hayward, CA, pp. 100–117.

Hedrick, P.W. (1987). Gametic disequilibrium measures: proceed with caution. *Genetics* **117**, 331–341.

Hey, J. and Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics* **145**, 833–846.

Hill, W.G. (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**, 229–239.

Hill, W.G. (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8**, 117–126.

Hill, W.G. (1977). Correlation of gene frequencies between neutral linked genes in finite populations. *Theoretical Population Biology* **11**, 239–248.

Hill, W.G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294.

Hill, W.G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226–231.

Hirschhorn, J.N. and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95–108.

Hollox, E.J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A.I. and Swallow, D.M. (2001). Lactase haplotype diversity in the Old World. *American Journal of Human Genetics* **68**, 160–172.

Hudson, R.R. (1983a). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.

Hudson, R.R. (1983b). Testing the constant-rate neutral model with protein data. *Evolution* **37**, 203–217.

Hudson, R.R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631.

Hudson, R.R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetical Research* **50**, 245–250.

Hudson, R.R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*. Oxford University Press, Oxford, pp. 1–44.

Hudson, R.R. (2001). Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817.

Hudson, R.R., Bailey, K., Skarecky, D., Kwiatowski, J. and Ayala, F.J. (1994). Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. *Genetics* **136**, 1329–1340.

Hudson, R.R. and Kaplan, N.L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.

Hudson, R.R., Kreitman, M. and Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

Ingram, C.J., Elamin, M.F., Mulcare, C.A., Weale, M.E., Tarekegn, A., Raga, T.O., Bekele, E., Elamin, F.M., Thomas, M.G., Bradman, N., Swallow, D.M. (2007). A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? *Human Genetics* **120**(6), 779–788.

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

Jeffreys, A.J., Kauppi, L. and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics* **29**, 217–222.

Jeffreys, A.J., Neumann, R., Panayi, M., Myers, S. and Donnelly, P. (2005). Human recombination hot spots hidden in regions of strong marker association. *Nature Genetics* **37**, 601–606.

Jobling, M.A., Hurles, M.E. and Tyler-Smith, C. (2004). *Human Evolutionary Genetics: Origins, Peoples and Disease*. Garland Science, New York.

Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F., Twells, R.C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S.C., Clayton, D.G., Todd, J.A. (2001). Haplotype tagging for the identification of common disease genes. *Nature Genetics* **29**, 233–237.

Karlin, S. and McGregor, J.L. (1967). The number of mutant forms maintained in a population. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **4**, 415–438.

Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903.

Kingman, J.F. (1982). The coalescent. *Stochastic Processes and their Applications* **13**, 235–248.

Kuhner, M.K., Yamato, J. and Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393–1401.

Larribe, F., Lessard, S. and Schork, N.J. (2002). Gene mapping via the ancestral recombination graph. *Theoretical Population Biology* **62**, 215–229.

Lewontin, R.C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**, 49–67.

Lewontin, R.C. (1988). On measures of gametic disequilibrium. *Genetics* **120**, 849–852.

Lewontin, R.C. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175–195.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Abecassis, H., Munro, H.M. and Donnelly, P. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics* **78**, 437–450.

Marjoram, P., Molitor, J., Plagnol, V. and Tavare, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15324–15328.

Marjoram, P. and Tavare, S. (2006). Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics* **7**, 759–770.

Marjoram, P. and Wall, J.D. (2006). Fast "coalescent" simulation. *BMC Genetics* **7**, 16.

Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research* **23**, 23–35.

McVean, G. (2007). The structure of linkage disequilibrium around a selective sweep. *Genetics* **175**, 1395–1406.

McVean, G.A. (2002). A genealogical interpretation of linkage disequilibrium. *Genetics* **162**, 987–991.

McVean, G., Awadalla, P. and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241.

McVean, G.A. and Cardin, N.J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **360**, 1387–1393.

McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R. and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584.

McVean, G., Spencer, C.C. and Chaix, R. (2005). Perspectives on human genetic variation from the HapMap project. *PLoS Genetics* **1**, e54.

Minichiello, M.J. and Durbin, R. (2006). Mapping trait Loci by use of inferred ancestral recombination graphs. *American Journal of Human Genetics* **79**, 910–922.

Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324.

Myers, S.R. and Griffiths, R.C. (2003). Bounds on the minimum number of recombination events in a sample history. *Genetics* **163**, 375–394.

Myers, S., Spencer, C.C., Auton, A., Bottolo, L., Freeman, C., Donnelly, P. and McVean, G. (2006). The distribution and causes of meiotic recombination in the human genome. *Biochemical Society Transactions* **34**, 526–530.

Nachman, M.W. and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G. and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research* **15**, 1566–1575.

Ohta, T. and Kimura, M. (1969a). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* **63**, 229–238.

Ohta, T. and Kimura, M. (1969b). Linkage disequilibrium due to random genetic drift. *Genetical Research (Cambridge)* **13**, 47–55.

Ohta, T. and Kimura, M. (1971). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* **68**, 571–580.

Padhukasahasram, B., Wall, J., Marjoram, P. and Nordborg, M. (2006). Estimating recombination rates from SNPs using summary statistics. *Genetics* **174**, 1517–1528.

Pluzhnikov, A. and Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* **144**, 1247–1262.

Pritchard, J.K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* **69**, 1–14.

Przeworski, M. (2003). Estimating the time since the fixation of a beneficial allele. *Genetics* **164**, 1667–1676.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonlad, G.J., Ackerman, H.C., Campbell, S.J., Atshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., Lander, E.S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**(6909), 832–837.

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., Lander, E.S. (2006). Positive natural selection in the human lineage. *Science* **312**(5780), 1614–1620.

Slatkin, M. (1994). Linkage disequilibrium in growing and stable populations. *Genetics* **137**, 331–336.

Slatkin, M. (2000). Allele age and a test for selection on rare alleles. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **355**, 1663–1668.

Slatkin, M. and Rannala, B. (2000). Estimating allele age. *Annual Review of Genomics and Human Genetics* **1**, 225–249.

Song, Y.S., Wu, Y. and Gusfield, D. (2005). Efficient computation of close lower and upper bounds on the minimum number of recombinations in biological sequence evolution. *Bioinformatics* **21**(Suppl. 1), i413–i422.

Spencer, C.C. and Coop, G. (2004). SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* **20**, 3673–3675.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society, Series B* **62**, 605–655.

Strobeck, C. and Morgan, K. (1978). The effect of intragenic recombination on the number of alleles in a finite population. *Genetics* **88**, 829–844.

Stumpf, M.P. and McVean, G.A. (2003). Estimating recombination rates from population-genetic data. *Nature Reviews Genetics* **4**, 959–968.

Sved, J.A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2**, 125–141.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Teshima, K.M., Coop, G. and Przeworski, M. (2006). How reliable are empirical genomic scans for selective sweeps? *Genome Research* **16**, 702–712.

Toomajian, C., Ajioka, R.S., Jorde, L.B., Kushner, J.P. and Kreitman, M. (2003). A method for detecting recent selection in the human genome from allele age estimates. *Genetics* **165**, 287–297.

Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. *PLoS Biology* **4**, e72.

Wakeley, J. (1997). Using the variance of pairwise differences to estimate the recombination rate. *Genetical Research* **69**, 45–48.

Wakeley, J. and Lessard, S. (2003). Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics* **164**, 1043–1053.

Wall, J.D. (2000). A comparison of estimators of the population recombination rate. *Molecular Biology and Evolution* **17**, 156–163.

Wall, J.D. (2004). Estimating recombination rates using three-site likelihoods. *Genetics* **167**, 1461–1473.

Wang, N., Akey, J.M., Zhang, K., Chakraborty, R. and Jin, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *American Journal of Human Genetics* **71**, 1227–1234.

Watterson, G.A. (1977). Heterosis or neutrality? *Genetics* **85**, 789–814.

Watterson, G.A. (1978). The homozygosity test of neutrality. *Genetics* **88**, 405–417.

Weir, B.S. (1996). *Genetic Data Analysis II. Methods for Discrete Population Genetic Data*, 2nd edition. Sinauer Associates, Sunderland, MA.

Weir, B.S. and Hill, W.G. (1986). Nonuniform recombination within the human beta-globin gene cluster. *American Journal of Human Genetics* **38**, 776–781.

Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology* **55**, 248–259.

Zhang, K., Deng, M., Chen, T., Waterman, M.S. and Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 7335–7339.

Zollner, S. and Pritchard, J.K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169**, 1071–1092.

Inferences from Spatial Population Genetics

F. Rousset

Laboratoire Génétique et Environnement, Institut des Sciences de l'Évolution, Montpellier, France

This chapter reviews theoretical models and statistical methods for inference from genetic data in subdivided populations. With few exceptions, these methods are based on neutral models of genetic differentiation and have been mainly concerned with estimation of dispersal rates. However, simulation-based methods allow to draw inferences under models involving additional demographic processes such as changes in dispersal rates over time. The formulation and main results of migration matrix, island, and isolation-by-distance models, are briefly described. The definition and basic properties of *F*-statistics are reviewed, and moment methods for their estimation are contrasted with likelihood methods. Then, the application of the different methodologies to simple biological scenarios is reviewed. Their practical performance is discussed in light of comparisons with demographic estimates, as well as of their robustness to different assumptions and of concepts of separation of timescale.

28.1 INTRODUCTION

Since the advent of molecular markers in population genetics, there have been many efforts to define methods of inference from the spatial genetic structure of populations. This chapter can only review a small selection of them including, in particular, some recent developments of simulation-based likelihood methods, and also of less sophisticated methods in so far as they provide analytical insight and proven performance in realistic conditions. With few exceptions, I will focus on allele frequency data; some methods for other types of data are described in **Chapter 29**.

The perspective taken in this review is that studies of spatial population structure are conducted in order to make inferences about parameters considered important for the evolution of natural populations, for example, for the dynamics of adaptation. Thus,

all such analyses should ultimately be based on models of evolution in subdivided populations. This would lead to the identification of important parameters in such processes and to the formulation of appropriate statistical models to estimate them (assuming it is useful to estimate them in order to test the models). In this perspective, the material reviewed below may seem imperfect not only because the statistical models are approximate but also because the important evolutionary parameters are not always clearly identified.

In all inferences, we will consider a total sample from a population structured by restricted dispersal in a number of demes (a technical term used in the analysis of the models) or subpopulations (a somewhat looser term). The population concept must be carefully distinguished from another concept of 'population' often considered in statistics, which actually refers to the probability distribution of samples under some model. In general the value of a variable in the biological population is not the expected value of this variable in this statistical 'population', in other words, this is not an expected value in a theoretical model. In practice the word parameter is used for both, but here it will be used only for the value in the theoretical model.

A statistical corollary is that by sampling only one locus, one may compute estimates which will approach the value in the biological population, rather than the parameter value, as more individuals are sampled. In other words, it will approach a value that will depend on the realized genealogy in the biological population, and this will be a random variable. The usual solution to this problem is to analyze several loci with different genealogies, assumed independent. For a nonrecombining DNA, it may not be very useful to sequence longer fragments: Since the whole DNA has the same genealogy, any estimate will depend on the single realized random genealogy in the biological population sampled (see **Chapter 25**).

28.2   NEUTRAL MODELS OF GEOGRAPHICAL VARIATION

The major models considered for statistical analysis describe the evolution of neutral genetic polymorphisms; among models of selected markers, statistical analysis will be considered only for clines.

28.2.1   Assumptions and Parameters

We consider a set of subpopulations each with N_i adults, and with dispersal rates m_{ij} from subpopulation j to subpopulation i. These dispersal rates are defined as the probability that an offspring had its parent in some subpopulation: Thus they are defined by looking backward in time (backward dispersal rates), rather than by looking where offspring go (forward dispersal rates). Forward and backward rates will differ, for example, when individuals that disperse at longer distances have a higher probability of dying before reproduction.

These models are known as *migration-matrix models*, with migration matrix (m_{ij}). Limit cases of these models when all deme sizes $\to \infty$, all backward rates $\to 0$, with their products $N_i m_{ij}$ remaining finite, have been described as 'structured coalescents' (see e.g. **Chapter 25**). With many subpopulations, the number of parameters may be large. However, some symmetric structure is usually assumed, as in the island and

isolation-by-distance models developed below. Further, the migration matrix, as well as the subpopulation sizes N_i, are supposed to be invariant in time. These assumptions allow for more detailed mathematical analysis. Simulation-based methods have allowed to investigate more complex historical scenarios involving range expansions, interruptions of gene flow, and so on. A relatively well-worked case is the isolation-with-migration model (Nielsen and Wakeley, 2001), according to which an initially panmictic population differentiates at some time T in the past into two subpopulations that will keep on exchanging migrants at rate m until the time of sampling.

The island model (Wright, 1931a) with n_d subpopulations is the simplest form of migration-matrix model: For different subpopulations i, j, the dispersal rate is supposed to be independent of i, j and may be written as $m_{ij} = m/(n_d - 1)$ where m is the total dispersal rate; $m_{ii} = 1 - m$. The subpopulation sizes $N_i = N$ are also supposed to be independent of i. The infinite island model is the limit process as $n_d \to \infty$. This is the most often considered model, because of its ease of analysis. However, it should be noticed that most of the results of the infinite island model with $N_i = N$ can easily be extended to infinite island model with N_i different for different subpopulations and with total dispersal rate into each subpopulation i being a function of i (see the discussion of equation (28B.1)). Thus the main defining assumption of such island models is that immigrants may come with equal probability from any of the other subpopulations.

Dispersal is often localized in space, so that immigrants preferentially come from close populations. Two kinds of models that take this into account have been considered, one for demes on a discrete lattice, and one for 'continuous' populations (e.g. Malécot, 1951; 1967). In a continuous population, the local density may fluctuate in space and time, but there is no rigorous mathematical analysis of models incorporating such fluctuations. In the lattice models, different demes are arranged on a regularly spaced lattice and the dispersal rates are a function of the distance between demes. There is a fixed number of adults, N, in every generation on each node of the lattice. Thus the position of individuals is rigidly fixed and density does not fluctuate. The island model may be recovered as a specific case.

In models of isolation by distance, the parameter σ^2 often appears (e.g. Malécot, 1967; Nagylaki, 1976; Sawyer, 1977). This is an average squared distance between parent and offspring. In two dimensions this is the average square of the projection of the two-dimensional (vectorial) distance on an axis, also known as the *axial distance* (Figure 28.1). This parameter is a measure of the speed at which two lineages descending from a common ancestor depart from each other in space. The models as formulated above may be generalized to include age- or stage-structure, and it is possible to generalize some of the results for island and isolation by distance models given below in terms of concept of effective dispersal rate and effective deme size or effective population density, albeit through some approximations (Rousset, 1999a; 2004, Chapter 9). Then, effective dispersal is the asymptotic rate of increase of the second moment of distance between two independently dispersing gene lineages per unit time. The definition of population density also needs to be generalized. First, it is actually not simply a density but a rate of coalescence per surface and per unit time (Rousset, 1999a). In the basic models, it can be computed as the expected number of coalescence events per generation among all pairs of genes in the total population, divided by the total surface occupied by the population (or habitat length in linear habitats). With age structure, it can be computed as a weighted
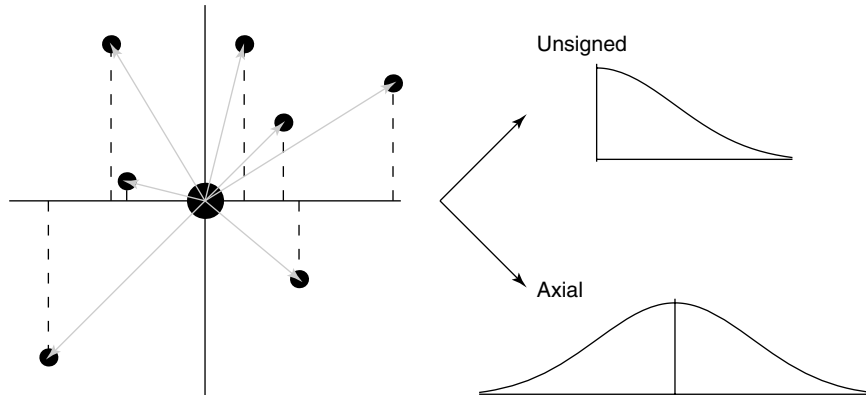
Figure 28.1 σ^2 in two dimensions. One considers the two-dimensional dispersal distances (gray arrows) between one parent (large central dot) and different offspring (or different parent–offspring pairs). The projection of these vectors on two axes yield signed axial distances on each axis. In terms of variance, σ^2 is the variance of the distribution of one such axial dispersal distance (bottom right). This is *not* the variance of the unsigned dispersal distance (top right).

average of such among all pairs, these being weighted by reproductive value weights as in the computation of the effective dispersal parameter.

28.3   METHODS OF INFERENCE

With few exceptions, explicit formulas for the likelihood of samples under the models formulated above are not available. This section therefore focuses on moment methods for which explicit analytical results are available, and on simulation methods for likelihood inference.

28.3.1   *F*-statistics

Moment methods are based on the analysis of moments of order k of allele frequencies. By far the most common of them (analysis of variance) consider only squares of allele frequencies or equivalently frequencies of identical pairs of genes. This is the basis for the theory of F-statistics in population genetics. Autocorrelation methods (e.g. Sokal and Wartenberg, 1983; Epperson and Li, 1997) are constructed from pair-wise comparisons of genes or genotypes, hence there should be essentially the same information in such statistics as in the more standard moment methods. The relationship between autocorrelation methods and some of the methods described below is discussed by Hardy and Vekemans (1999).

28.3.1.1   *Probabilities of Identity and F-statistics*

To define genetic identity, we consider a pair of homologous genes and ask whether they descend without mutation from their most recent common ancestor. If no mutation has occurred since the coalescence of ancestral lineages, there is *identity by descent* (IBD).

By *identity in state* (IIS) of a pair of genes we simply consider whether they have the same sequence (if the alleles are distinguished by their sequence), the same length (if the alleles are distinguished by the number of repeats of a microsatellite motif), the same electrophoretic mobility, etc. In short, we only look at the allelic state of a gene. IBD is a specific case of IIS for the infinite allele model, in which each allele produced by mutation is considered different from preexisting alleles. The generic notation Q will be used to denote expected values of IIS under any model.

If we consider a population structured in any way (age, geography, etc.), one may always define Q_w, the IIS probability within a class of genes (for example among individuals of some age class, in the same subpopulation, etc.), and Q_b, the IIS probability between two different classes of individuals. In a generic way one may then define:

$$F \equiv \frac{Q_w - Q_b}{1 - Q_b}. \tag{28.1}$$

Such quantities are known as *F-statistics*, but Q_w, Q_b, and F as defined above are parameters. That is, Q_w and Q_b are expectations under independent replicates of the stochastic process considered, and F is a function of these parameters. In other words, they are functions of the parameters that define the model under study, such as subpopulation sizes, mutation rates, migration rates, etc. If (say) deme size is by itself random, then F and the Qs, being expectations in the process considered, are function of the parameters of the distribution of deme size. In models of spatial genetic structure, Q_w and Q_b are generally not 'the value in the (biological) population'. Alternative definitions of F-statistics, as values in biological populations, have been used in the literature (e.g. Nei, 1986; see Nagylaki, 1998 for further discussion), but analytical results below hold only with the present parametric definitions.

Let Q_2 be the IIS probability within subpopulations, and Q_3 be the IIS probability between subpopulations. The well-known F_{ST} parameter, originally considered by Wright, is best defined as

$$F_{ST} \equiv \frac{Q_2 - Q_3}{1 - Q_3}. \tag{28.2}$$

F-statistics may be described as correlations of genes within classes with respect to genes between classes, that is as intraclass correlations (Cockerham and Weir, 1987).

28.3.1.2 Generic Methods for Estimation and Testing

The estimation of F-statistics is described at length in the literature (see e.g. **Chapter 29**, Weir, 1996) so I will confine myself to emphasizing a few easily missed points.

A simple way to estimate parameters such as F_{ST} is to estimate each of the probabilities of identity by the corresponding frequencies $\hat{Q}$ of identical pairs of genes in the sample, computed by simple counting. Thus F_{ST} may be estimated by

$$\hat{F} \equiv \frac{\hat{Q}_2 - \hat{Q}_3}{1 - \hat{Q}_3}, \tag{28.3}$$

where $\hat{Q}_2$ and $\hat{Q}_3$ are by definition the frequencies of identical pair of genes in the sample, within and between deme, respectively. This simple approach to defining estimators may

be easily adapted to a number of different settings, given that the parameters to be estimated may be expressed as functions of probabilities of IIS. With balanced samples, this approach directly yields Cockerham's estimator of F_{ST} (Cockerham, 1973; Weir and Cockerham, 1984). This estimator has been developed by analogy with the methods of analysis of variance, and this analogy has proved difficult to understand. Appendix A details the nature of the analogy and its relationship with 28.3.

It is easy to test for differentiation (nonzero F_{ST}) by the usual exact tests for contingency tables either applied to gametic or genotypic data. These are standard statistical techniques and their application to genetic data has been discussed elsewhere (e.g. Weir, 1996; Goudet et al., 1996; Rousset and Raymond, 1997). A general set of techniques to draw confidence intervals from the moment estimators are the bootstrap (e.g. Efron and Tibshirani, 1993) and related techniques based on the resampling of loci. However, the simplest applications of resampling techniques may be misleading. This may be apparent when they lead to symmetric confidence intervals while the variance of the estimator is expected to be sensitive to the parameter value, in which case more involved uses of the bootstrap (DiCiccio and Efron, 1996) may be required.

28.3.1.3   Why F-statistics?

Wright was the first to note that such measures of genetic structure appear in some theoretical models of adaptation, and his ideas remain among the most influential in population genetics. He used them to quantify his 'shifting balance' model Wright (1931a; 1931b), which remains controversial today (Coyne et al., 1997). Nevertheless, F-statistics are useful descriptors of selection in one-locus models (Rousset, 2004). Wright also used them to estimate demographic parameters (Dobzhansky and Wright, 1941) and they have become a standard tool to 'estimate gene flow' or for merely descriptive studies of genetic population structure. Such studies are not always very convincing and may be questioned on statistical grounds. Two major objections are (1) the connection between such measures and the likelihood-based framework of statistics (e.g. Cox and Hinkley, 1974; Lehmann and Casella, 1998) is not obvious; and (2) although F_{ST} bears a simple relationship with the 'number of migrants' Nm in the infinite island model, it is not always clear how this would extend to more general models of population structure. Also, with the definition given above in terms of IIS, F_{ST} might be expected to depend on mutation processes at the loci considered, and how this affects estimation of dispersal parameters is not clear.

One of the main attractions of F-statistics may be their robustness to several factors. In an infinite island model, the ancestral lineages of two genes sampled within the same deme coalesce within this deme in a recent past with probability $\approx F_{ST}$; with probability $\approx 1 - F_{ST}$ the lineages separate (looking backward) in different demes as a result of immigration and will take a long time to coalesce. This implies that F_{ST} will depend mostly on recent events. Before considering the implications in detail, we will see how this argument can be generalized under isolation by distance.

We consider the probability $c_{j,t}$ that two genes coalesce at time t in the past. The j index corresponds to the type of pair of genes considered (e.g. $j = 2$ or 'w' for genes within demes and $j = 3$ or 'b' between demes as above). Identity by descent, here denoted $\dot{Q}$, has been defined as the probability that there has not been any mutation since the common
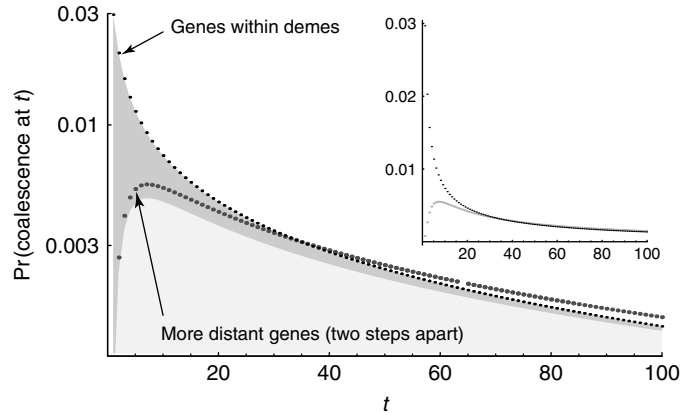
Figure 28.2 This figure compares the distributions of coalescence times in demes two steps apart (thick gray points) and within the same deme (thin black points). The inset shows the distributions on a linear y scale. The distribution for genes within demes is decomposed in two areas, the light gray one whose height is a constant times the height of the other distribution (hence it is shifted on the log scale), and the remainder (dark gray) which is the excess probability of coalescence in recent generations. The distribution were computed for 100 demes of $N = 10$ haploid individuals, with dispersal rate $m = 1/4$.

ancestor. Thus

$$\dot{Q}_j = \sum_{t=1}^{\infty} c_{j,t}(1-u)^{2t}. \tag{28.4}$$

(Malécot, 1975, (28.6); Slatkin, 1991). To understand the properties of F-statistics, we compare the distributions of coalescence times $c_{\mathrm{w},t}$ and $c_{\mathrm{b},t}$ of the pairs of genes that define these parameters. We can view the area covered by the probability distribution of coalescence time of the more related pair of genes as the sum of two 'probability areas', one part which is a smaller copy of the area covered by the probability distribution function of coalescence of less related genes, the other part being the remainder of the area for more related genes (Figure 28.2). This second part decreases faster than probabilities of coalescence (it is approximately $O(c_{.,t}/t)$, Rousset, 2006), and is therefore concentrated on the recent past. As a first approximation, the value of the corresponding F-statistics is this excess probability of recent coalescence. Let us call ω the value of $1 - c_{\mathrm{w},t}/c_{\mathrm{b},t}$ for large t. This will also be the excess probability of recent coalescence (as can be deduced from the fact that both distributions must sum up to 1). Then it can be shown that

$$Q_{\mathrm{w}:k} \approx (1 - \omega)Q_{\mathrm{b}:k} + \omega\pi_k \Rightarrow \frac{Q_{\mathrm{w}:k} - Q_{\mathrm{b}:k}}{\pi_k - Q_{\mathrm{b}:k}} \approx \omega \approx F, \tag{28.5}$$

where π_k is the expected frequency of allele k in the model considered. One may obtain this result by considering that with probability $1 - \omega$, the probability of identity of pairs of genes 'within' is the same as the probability of identity of genes 'between' (this corresponds to the proportional parts of the distributions of coalescence times), and with probability ω (the excess recent probability mass) the coalescence event has occurred recently in a common ancestor, of allelic type k with probability π_k.

The above result should not be overinterpreted. Expressions of the form "probability of identity equals $(1 - F)p^2 + Fp$" for F independent of p are not generally valid unless p is the expectation π_k and the probability of identity is the process expectation (Rousset, 2002), although they also correctly describe the conditional probability of identity given p in the infinite island model, even in cases where p is a random variable.

The above logic will be valid as long as mutations can be neglected within the time span covered by the probability mass. This time span is shorter for higher migration rates m in the island model, or for high σ^2 relative to spatial distance in isolation by distance models, so in practice F-statistics are weakly dependent on mutation rates at small spatial scales. The same argument can also be used to show that F-statistics more quickly recover their stationary values than probabilities of identity under the same conditions after a single demographic perturbation, a fact noticed by several authors (e.g. Crow and Aoki, 1984; Slatkin, 1993; Pannell and Charlesworth, 1999). This kind of approximate independence is important for statistical applications since it makes F-statistics analyses at a small spatial scale interpretable despite the fact that past demographic history and mutation processes are generally not known. At a local scale, F_{ST} is also only weakly dependent on the total population size.

28.3.2 Likelihood Computations

With few exceptions, likelihood computation in population genetics are based on 'coalescent' arguments, i.e. they derive the probability of the sample from consideration of the sequence of states that relate the individuals in the sample to their common ancestor (e.g. Kingman, 1982; Hudson, 1990, see **Chapters 25**, and **26**). This sequence of events may be the genealogy, G, of the sample that includes information about the coalescence time of ancestral lineages and about which lineages coalesce. In other cases it may be a 'gene tree' H, which takes into account the relative timing of coalescence and mutation events, as well as the nature of mutation events, i.e. the states before and after mutation, but which does not take into account the time between events, nor which lineage, among several with identical state, was involved in each event (see **Chapter 26**, Griffiths and Tavaré, 1995).

Coalescent arguments are used to estimate the likelihood by simulation using importance sampling algorithms. In one class of algorithms (see **Chapter 26**, Beerli and Felsenstein, 1999), the likelihood of the parameters $\mathcal{P}$ as a function of the data D may be written as

$$L(\mathcal{P}; D) = \sum_G \Pr(D|G; \mathcal{P}) \Pr(G; \mathcal{P}), \tag{28.6}$$

where the sum is over all possible genealogies G,

$$= \sum_G \Pr(D|G; \mathcal{P}) \frac{\Pr(G; \mathcal{P})}{f(G)} f(G), \tag{28.7}$$

for any distribution $f(G)$ such that $f(G) > 0$ when $\Pr(G; \mathcal{P}) > 0$

$$= \mathcal{E}\left[\Pr(D|G; \mathcal{P}) \frac{\Pr(G; \mathcal{P})}{f(G)} \right], \tag{28.8}$$

where $\mathcal{E}$ is an expectation over sample paths of a Markov chain with stationary distribution $f(G)$

$$\approx \frac{1}{s} \sum_{i=1}^{s} \Pr(D|G(i); \mathcal{P}) \frac{\Pr(G(i); \mathcal{P})}{f(G(i))}, \qquad (28.9)$$

where the sum is over the sample path of a Markov chain with stationary distribution $f(G)$.

In neutral models, given the genealogy G, the data only depend on the mutation process with parameters $\mathcal{N}$, while the genealogy itself does not depend on the mutation process but only on demographic parameters $\mathcal{D}$ (with $\mathcal{P} = (\mathcal{N}, \mathcal{D})$). Thus we may choose the importance sampling function

$$g \equiv \Pr(G|D; \mathcal{P}_0) = \frac{\Pr(G; \mathcal{D}_0) \Pr(D|G; \mathcal{N})}{L(\mathcal{P}_0; D)}, \qquad (28.10)$$

for some value $\mathcal{D}_0$ of $\mathcal{D}$ and for $\mathcal{P}_0 = (\mathcal{N}, \mathcal{D}_0)$. Then from 28.8

$$L(\mathcal{P}; D) = \mathcal{E}\left[L(\mathcal{P}_0; D) \frac{\Pr(G; \mathcal{D}) \Pr(D|G; \mathcal{N})}{\Pr(G; \mathcal{D}_0) \Pr(D|G; \mathcal{N})} \right] = L(\mathcal{P}_0; D) \mathcal{E}\left[\frac{\Pr(G; \mathcal{D})}{\Pr(G; \mathcal{D}_0)} \right], \quad (28.11)$$

for any $\mathcal{N}$. We try to find the maximum likelihood estimate (MLE) of $\mathcal{P}$ or equivalently the maximum value of

$$\frac{L(\mathcal{P}; D)}{L(\mathcal{P}_0; D)} = \mathcal{E}\left[\frac{\Pr(G; \mathcal{D})}{\Pr(G; \mathcal{D}_0)} \right], \qquad (28.12)$$

which may be estimated by

$$\frac{L(\mathcal{P}; D)}{L(\mathcal{P}_0; D)} \approx \frac{1}{s} \sum_{i=1}^{s} \frac{\Pr(G(i); \mathcal{D})}{\Pr(G(i); \mathcal{D}_0)}, \qquad (28.13)$$

where the $G(i)$s are generated by a Markov chain with stationary distribution g. Thus an algorithm to find the maximum must define such a Markov chain (for parameters $\mathcal{P}_0$), and compute $\Pr(G(i); \mathcal{D})/\Pr(G(i); \mathcal{D}_0)$ for different $\mathcal{P}$ values and for this single Markov chain.

Beerli and Felsenstein (1999) have used the importance sampling function (28.10) to estimate the ratio (28.12). They define a Markov chain on genealogies G, and use the Metropolis-Hastings algorithm (Hastings, 1970) to ensure that the importance sampling function g is the stationary distribution of this chain. Their (28.14) shows that the transition probabilities of this chain are determined by the probabilities $\Pr(D|G; \mathcal{N})$, which for sequence data may be computed as previously described (e.g. Swofford and Olsen, 1990).

Griffiths and Tavaré (1994) have proposed a different class of algorithms. They derive recursions for the stationary probability $\Pr(D'|S')$ of a sample D' given sample size S' (here a vector and subsample sizes in different populations) over a time interval (typically the interval between two genealogical or mutation events). For any state D of ancestors in the previous time, $\Pr(D'|S')$ is the probability that a sample of size S' derives from a sample of size S, times the stationary probability of ancestral states $\Pr(D|S)$, times the

forward probability that a sample D' derives from a sample D given the sample sizes:

$$\Pr(D'|S') = \sum_D \Pr(D'|D, S, S') \Pr(S|S') \Pr(D|S). \tag{28.14}$$

It is relatively straightforward to express the backward transition probabilities $\Pr(S|S')$ and the forward transition probabilities $\Pr(D'|D, S, S')$ in terms of model parameters $\mathcal{P}$ (further details are lengthy; see de Iorio and Griffiths, 2004b). An importance sampling algorithm is then derived by writing $\Pr(D'|D, S, S') \Pr(S|S')$ in the form $w(D', D)p(D|D')$ where the $p(D|D')$ define an absorbing Markov chain going backward over possible ancestral histories,

$$\Pr(D') = \sum_D w(D', D)p(D|D') \Pr(D). \tag{28.15}$$

Iterating this recursion until the ancestor of the whole sample shows that

$$\Pr(D) = \mathcal{E}_p \left[\prod_{D_i} w(D_i, D_{i-1}) \right], \tag{28.16}$$

where the product is over successive states D_i of the ancestry of the sample. Then w is an importance sampling weight and p is a proposal distribution (compare 28.8).

Different choices of p, and of the implied w, are possible. Griffiths and Tavaré (1994) describe one such choice and show that the likelihood for different values of $\mathcal{P}$ may be obtained by running the Markov chain for only one value of $\mathcal{P}$. However, more recent works have aimed to optimize the choice of p for each parameter value analyzed so that the variance of $\prod_{D_i} w(D_i, D_{i-1})$ among runs of the Markov chain would be minimal. An optimal choice of the proposal distribution would be such that any realization of the Markov chain would give exactly the likelihood of the sample. This occurs when the proposal distribution has transition probabilities given by the reverse probabilities in the biological process considered, $\Pr(D|D') = \Pr(D'|D) \Pr(D) / \Pr(D')$ (see **Chapter 26**). In cases of interest these reverse probabilities cannot be computed from this formula since the aim is precisely to evaluate the probabilities $\Pr(D)$. Nevertheless, the variance of $\prod_{D_i} w(D_i, D_{i-1})$ should be low if good approximations for $\Pr(D) / \Pr(D')$ are used. de Iorio and Griffiths (2004a; 2004b) write such ratios as simple functions of the probabilities π that an additional gene sampled from a population is of a given allelic type, conditional on the result of previous sampling, and propose approximations $\hat{\pi}$ for them, from which approximations for $\Pr(D) / \Pr(D')$ and for the proposal distribution follow. Their approximation scheme method applies in principle for any stationary migration-matrix model and any Markov mutation model (for allele frequency data). The $\hat{\pi}$ are not given in closed form but as solutions of a system of $n_d \times K$ linear equations for a model of n_d populations and K allelic types. By assuming independence between the mutation and genealogical processes, this can be reduced to a system of n_d equations holding for each of the different eigenvalues of the mutation matrix, a technique used by de Iorio *et al.* (2005) to analyze a two-demes model with stepwise mutation.

Despite substantial improvement over previous proposals, the computation times of the latter algorithm would remain prohibitively long in many practical applications. A method known as product of approximate conditional likelihoods (PAC-likelihood,

Li and Stephens, 2003) has been proposed to derive heuristic approximations of the likelihood estimation. Here a sample of genotypes (g_k) is described as a sequential addition of genotypes, so that the likelihood of an ordered sample is the product of the probabilities π that an additional genotype is g_i given previously added genotypes were $g_1, \ldots, g_{i-1}$. Approximations are then considered for these conditional probabilities. This was originally applied to inference of recombination rate in a panmictic population. Using de Iorio and Griffiths's approximations $\hat{\pi}$ in one-locus models, it turns out to perform well under stepwise mutation, where the expectation of the PAC-likelihood statistic was indistinguishable from the likelihood, but computation was much faster (Cornuet and Beaumont, 2007). In a linear stepping stone model, one can find small differences between the expectation of the PAC-likelihood statistic and the likelihood, but estimation of model parameters based on PAC-likelihood is essentially equivalent to maximum likelihood (ML) estimation while again requiring far less computation than likelihood computation via the importance sampling algorithm (F.R., unpublished data).

28.4 INFERENCE UNDER THE DIFFERENT MODELS

In this section I review the implementation and application of the different methodologies in specific cases. Published genetic and demographic data from Gainj- and Kalam-speaking people of New Guinea (Wood *et al.*, 1985; Long *et al.*, 1986) will conveniently illustrate several conclusions.

28.4.1 Migration-Matrix Models

For any migration matrix at stochastic equilibrium, the distribution of frequencies p_{ki} of allele k in each deme i follows some probability distribution with (parametric) covariances which can be written as

$$\mathrm{E}[(p_{ki} - \pi_k)(p_{ki'} - \pi_k)] = Q_{ii':k} - \pi_k^2, \tag{28.17}$$

where $Q_{ii':k}$ is the expected frequency of pairs of genes in demes i, i' that are of allelic type k. For mutation models assuming identical mutation rates between K alleles, this is also $(Q_{ii'} - \sum_{k=1}^{K} \pi_k^2)/K$, where $Q_{ii'}$ is the probability of IIS of pairs of genes in demes i, i'. Probabilities of IIS $Q_{ii':k}$ or $Q_{ii'}$ may be derived from the probabilities of IBD for various mutation models (Markov chain models, or stepwise mutation models; see e.g. Tachida, 1985; Rousset, 2004). For any migration matrix model, probabilities of IBD within and among demes can be computed as solutions of a linear system of equations (see e.g. Nagylaki, 1982 or Rousset, 1999a; 2004 for details and examples). In principle, the demographic parameters can be estimated by inverting such relationships. This approach has been taken seriously only in a few cases, in particular the island and isolation-by-distance model, as detailed below, and does not generate likelihood expressions in a straightforward way, although some heuristic likelihood formulas have been proposed (Tufto *et al.*, 1996) by using Gaussian approximations for the distribution of allele frequencies, with the covariances given above.

28.4.2 Island Model

In the island model, one has the well-known approximation $F_{\mathrm{ST}} \approx 1/(1 + 4Nm)$ (with $2N$ genes per deme, Wright, 1969). This has led to the usage of computing F_{ST}s and expressing

the results in terms of 'estimates of Nm', i.e. in terms of $(1/\hat{F} - 1)/4$. This usage is often problematic. The worst sin is to estimate an F_{ST} between a pair of samples far apart, to translate it into a nonzero 'Nm', and to conclude that the populations must have exchanged migrants in the recent past. In the context of the island model, an F_{ST} between a pair of subpopulations is not a function of the number of migrants exchanged specifically between these two subpopulations. More generally, in many models of population structure, it is expected that subpopulations that never exchange migrants will have nonzero 'Nm' values. It may be seen from the above definition of F_{ST} that its maximum value is the probability of identity within demes Q_2 (when $Q_3 = 0$), which results in a minimum possible value of 'Nm' of $(1/Q_2 - 1)/4$ which may well be > 1 even for demes that never directly exchange migrants. Likewise, the practice of equating $Nm > 1$ to panmixia and $Nm < 1$ to divergence is not useful.

Likelihood functions for allele frequency data may be derived relatively easily by diffusion techniques for the infinite island model (see (28B.5)) and can in principle be recovered by coalescent arguments (Balding and Nichols, 1994). These sampling formulas allow analytical insight, and may be used to define estimators of Nm as well as to discuss efficient estimation of F_{ST} by moment methods (See Appendix B). Kitada *et al.* (2000) have implemented likelihood estimation under this model. Approximation have also been considered such as the 'pseudo maximum likelihood estimator' (PMLE, Rannala and Hartigan, 1996) of the number of migrants in an island model (see (28B.6)). These authors found that this estimator of Nm generally (though not always) had lower mean square error than the moment estimator $(1/\hat{F} - 1)/4$, depending on sampling scheme and Nm values. The MLE is also biased when the number of sampled populations is small and some corrections have been proposed (Kitakado *et al.*, 2006). For the New Guinea population, application of pseudo maximum likelihood (PML) estimation yields an estimate of 10.2 migrants per generation. This is one-fourth of the average value, 41.87, that can be computed from maternal and paternal dispersal rates and total subpopulations sizes (Tables 1–3 in Wood *et al.* (1985)), but it is closer to one third if we take 'effective size' considerations into account following Storz *et al.* (2001).

In comparison with the simulation methods for likelihood computation, it should be noted that no mutation model has to be considered here. Wright's formula (28B.1) is an approximation for low mutation, common to the different 'Markov chain' models of mutation, and so is the likelihood formula (28B.5). In this respect it is analogous to the methods based on F-statistics.

28.4.3 Isolation by Distance

Here analytical insight is available only for the moment methods. The results reviewed here are not tied to a Gaussian model of dispersal. We consider

$$a(\mathbf{r}) \equiv \frac{Q_0 - Q_{\mathbf{r}}}{1 - Q_0}, \tag{28.18}$$

which is $F_{ST}/(1 - F_{ST})$ at (vectorial) distance $\mathbf{r}$. Approximations for F_{ST} immediately follow from those for $a(\mathbf{r})$. I will use a dot on a or Q to emphasize that the results given hold strictly only for IBD, but the differences with IIS do not affect the main practical conclusions drawn below (Rousset, 1997).

Two cases are usually considered, the one-dimensional model for populations in a linear habitat, and the two-dimensional model. In one dimension, at distance r,

$$\dot{a}(r) \approx \frac{A_1}{4N\sigma} + \frac{1 - e^{\frac{-(2u)^{1/2}r}{\sigma}}}{4N\sigma(2u)^{1/2}} \overset{r\text{small}}{\approx} \frac{A_1}{4N\sigma} + \frac{r}{4N\sigma^2} \approx \frac{A_1}{4D\sigma} + \frac{r}{4D\sigma^2}, \qquad (28.19)$$

where A_1 is a constant determined by the dispersal distribution, but not by N nor u. Its definition is given by Sawyer(1977, eq. 2.4).

In two dimensions, for genes at Euclidian distance r,

$$\dot{a}(\mathbf{r}) \approx \frac{-\ln((2u)^{1/2}) - K_0((2u)^{1/2}r/\sigma) + 2\pi A_2}{4N\pi\sigma^2} \overset{r\text{small}}{\approx} \frac{\ln(r/\sigma) - 0.116 + 2\pi A_2}{4N\pi\sigma^2}$$

$$\approx \frac{\ln(r/\sigma) - 0.116 + 2\pi A_2'}{4D\pi\sigma^2}, \qquad (28.20)$$

where K_0 is the modified Bessel function of second kind and zero order (e.g. Abramovitz and Stegun, 1972), and A_2 is of the same nature as A_1 above. Its definition is given by Sawyer, (1977, eq. 3.4); see also Rousset (1997 eq. A11).

In the last two equations the first expression is given for σ measured in the length unit of the model (i.e. one interdeme distance on the lattice), the second is the small distance/low mutation limit of the first, and the third is the second for any length unit. They are in terms of population density D per length or surface unit, and the A_2' constant depends on the length unit.

In the same equations, the second approximation shows a linear relationship between $\dot{a}(r)$ and geographical distance in one dimension, and between $\dot{a}(r)$ and the logarithm of geographical distance in two dimensions. In both cases, the slope of this relationship is a function of $D\sigma^2$.

These different expressions emphasize two points. First, differentiation is a function of the A constants, which are not simple functions of σ^2 but also of other features of the dispersal distribution. In fact, when the total migration rate is low, the differentiation between adjacent subpopulations is close to that expected under an island model with the same total number of migrants. This confirms that σ^2 is not the only relevant parameter of the dispersal distribution. Second, the value of A_2' depends on the spatial unit chosen to measure σ and D. A method of inference from A_2' values that would not take into account the discrepancy between the length unit used and the idealized interdeme distance would therefore be internally incoherent.

The above approximations allow a relatively simple description of the expected differentiation in these models as well as relatively simple estimation of $D\sigma^2$ from genetic data. Estimates of $a(r)$ at different distances may be obtained in some cases as estimates of $F_{ST}/(1 - F_{ST})$ for pairs of samples, and simply regressed to spatial distance (Rousset, 1997, as implemented in GENEPOP, Raymond and Rousset, 1995). An estimate of $1/(D\sigma^2)$ may be deduced from the slope of the regression. Two early applications of this method yielded estimates about twice the demographic estimate (Rousset, 1997). For the New Guinea population, the regression equation $F_{ST}/(1 - F_{ST}) \hat{=} 0.0191 + 0.0047 \ln(\text{distance in km.})$ provides an estimate of $D\sigma^2$ which is about twice the demographic estimate (after application of effective density correction following Storz et al. (2001), and after correction of clerical errors affecting the reported σ^2 of females and males in Rousset (1997), which should be 3.1 and 0.76 km^2 respectively).

When the migration rate is low, an estimate of the number of immigrants per generation may also be computed by the '$(1/\hat{F} - 1)/4$' method, taking the value of the estimated regression equation at the distance between the closest subpopulations as an estimate of $\hat{F}/(1 - \hat{F})$. The estimate of number of migrants in the New Guinea population is then 11.5, close to the pseudo maximum likelihood estimate (10.2, see above). This result illustrates the approximate convergence of estimates by different methods and under different dispersal models to Wright's classic result, even though the dispersal rate in this study is not precisely low (its average value being 0.43 from demographic data).

The regression of $F_{ST}/(1 - F_{ST})$ to distance is not always applicable, particularly when there are no recognizable demes of several individuals, as for 'continuous' populations. A variant based on the comparison of pairs of individuals has been designed to address this problem (Rousset, 2000). Simulations have shown that this method performs reasonably well when σ is small (a few times interindividual distance at most) and when most individuals are sampled within an area of about $20\sigma \times 20\sigma$ (Leblois *et al.*, 2003,2004). For higher dispersal, the variant considered by Vekemans and Hardy (2004) provides more accurate upper confidence bounds for $D\sigma^2$ (Watts *et al.*, 2007). Several comparisons have found agreement within a factor of two with independently derived demographic estimates (Rousset, 2000; Sumner *et al.*, 2001; Winters and Waser, 2003; Fenster *et al.*, 2003; Broquet *et al.*, 2006; Watts *et al.*, 2007). Whether this is considered an important discrepancy or not will depend on the accuracy expected from such analyses, but this is certainly much better than usually reported (see e.g. Slatkin, 1994; Koenig *et al.*, 1996). They actually go against an earlier long stream of reported discrepancies between genetic and demographic estimates, which needs explaining.

Part of the discrepancies hinge on misunderstandings of the models. For example, Wright assumed that the value of F-statistics under isolation by distance (Wright, 1946) was determined by the 'neighborhood size'. The value of this parameter would be $4D\pi\sigma^2$ under the assumption of two-dimensional Gaussian dispersal and its more general definition would be a function of 'the chance that two uniting gametes came from the same individual' (Wright, 1946). A third common 'definition' found in the literature is that the 'neighborhood size' would be the size of a subpopulation that would behave as a panmictic unit. It is not clear in which respect the subpopulation would behave as a panmictic unit nor whether there is a subpopulation that behaves as a panmictic unit in some useful sense. In any case Wright's measures do not correctly predict the value of unambiguously defined parameters in unambiguously defined models. In the analysis of Malécot's model, neither $D\sigma^2$ (because of the important A_2 term), nor the more generally defined neighborhood, determine the differentiation alone (Rousset, 1997). In one dimension, Wright proposed that $D\sigma$ was the important parameter, but the above results show that $D\sigma^2$ is important. One must give up the idea that $D\sigma^2$ equals neighborhood equals a number of individuals. In one dimension, $D\sigma^2$ scales as number of individuals times a length, not as a number of individuals, since density is a number of individuals per unit length.

The neighborhood concept was an attempt to account for different families of dispersal distributions. On the other hand, it has recurrently been assumed that differentiation is essentially a function of σ^2 and not of other features of the dispersal distribution. If so, it would be easy to seemingly improve on the regression method by considering only a family of dispersal distributions with a single parameter, completely determined by σ^2, for example a discretized Gaussian. In this case, F_{ST} or $a(r)$ values, not simply their increase

with distance, would contain information about σ^2. But such improvements would not be robust to misspecification of the dispersal distribution.

To explain reported discrepancies, it has often been argued that genetic patterns are highly sensitive to long-distance dispersal, which occurrence is easily missed in demographic studies. While some genetic patterns are indeed affected by long-distance dispersal (e.g. Austerlitz *et al.*, 2000), this is much less so for the patterns considered in the regression analyses, and this contributes to their concordance with demographic estimates. If a fraction m of immigrants come from an infinite distance (so that the 'true' σ^2 is infinite and does not predict any local pattern of differentiation), such migrants will be unrelated to their neighbors, and these migration events are analogous to mutation events. Hence we can deduce the effect of such immigrants from the effect of mutation, e.g.

$$\dot{a}(\mathbf{r}) \approx \frac{-\ln(\sqrt{2m}) - K_0(\sqrt{2m}r/\hat{\sigma})}{4D\pi\hat{\sigma}^2} + \text{constant}, \tag{28.21}$$

where $\hat{\sigma}$ is the parameter of the dispersal distribution for the fraction $1 - m$ of locally dispersing individuals. This result implies that there is an approximately linear increase of differentiation, determined by $D\hat{\sigma}^2$, roughly up to distance $0.56\hat{\sigma}/\sqrt{2m}$ (from Figure 3 in Rousset (1997)). For example if we ignore a 1 % (respectively, 0.1 %) tail of the distribution of dispersal distance in a demographic study which estimates $\hat{\sigma} = 10$ distance units, the prediction of increase of differentiation with distance will reach 20 % error at $0.56\hat{\sigma}/\sqrt{2m} = 39.6$ (respectively, 125) distance units. This is a wide overestimate of the error for any data set spread over such a distance, but more accurate predictors of bias will depend on the distribution of spatial distances in the sample.

Naive application of testing methodology has been another factor contributing to confusion. The absence of a pattern of isolation by distance (null slope of the regression, $D\sigma^2$ infinite) may be tested by the exact permutation procedure known as the *Mantel test* (Mantel, 1967; see Rousset and Raymond, 1997, for a simple description). In practice, nonsignificant test results have often been interpreted as evidence that dispersal is not localized. However, the Mantel test has often been applied in conditions of low power. In many populations with localized dispersal, the value of $D\sigma^2$ will be large, and thus expected patterns of isolation by distance (increase of differentiation with distance) will be weak (particularly in two-dimensional habitats), even though differentiation will be inferred by classical tests for differentiation.

Finally, variation in expected gene diversity due to spatial heterogeneity of demographic parameters may result in larger variation in expected differentiation than that due to isolation by distance. For example, if several demes with very small deme size and restricted dispersal are clustered in space, they will have low expected gene diversity and will show a larger differentiation between them than with more distant demes with higher expected diversity, and the above methods will obviously fail unless this heterogeneity is taken into account (Rousset, 1999b).

28.4.4 Likelihood Inferences

Maximum likelihood methods have not been developed and tested to a comparable level. The migration matrix models have been implemented for allele frequency and sequence data in the software MIGRATE. A more restricted set of models based on the 'isolation-with-migration' scenario has been implemented in the software IM (Nielsen and Wakeley,

2001; Hey, 2005). In the coalescent algorithms as in previous inferences based on per-locus information from samples taken at one time, it is not possible to estimate the deme sizes, mutation rates, and immigration rates separately: only products of deme size with migration probability and mutation probability, and (as a consequence) the ratio of migration and mutation probabilities, can be estimated. In the latest version of IM it is possible to analyze the divergence between two populations in terms of their deme sizes scaled relative to mutation rate ($N_1\mu$ and $N_2\mu$), of immigration rates in each of them m_1 and m_2, either scaled relative to deme size (e.g. N_1m_1) or relative to mutation rate (e.g. m_1/μ), of the scaled size of the ancestral population, and of relative fractions of this ancestral which contributed to the two descendant populations.

In a two-populations setting, one study has compared ML estimates with demographic estimates and with some other genetic estimation methods (Wilson *et al.*, 2004). Both ML and the two-locus method of Vitalis and Couvet (2001) produced estimates roughly in agreement with demographic data, while F_{ST} did not. It appears difficult to estimate all parameters of a four-demes migration-matrix model (Beerli, 2006). Likewise, it has not been possible to choose among different scenarios of colonization of the Americas using the IM software (Hey, 2005). The effect of unsampled demes on estimation of dispersal between two sampled demes has been investigated by Beerli (2004), for sequence data (100 000 bp per individual). As expected, the biases increase with immigration from the unsampled population(s), but it was found that estimates of immigration rate between the two populations were hardly affected by an equal total immigration rate from unsampled population(s) (estimates of scaled populations sizes were more affected). Abdo *et al.* (2004) argued that confidence intervals given by MIGRATE are not accurate, a fact attributed to too short run times of the algorithm (Beerli, 2006).

The above simulation scenarios are rather distinct from those considered in the previous section, and it would be hard to perform ML analyses of the New Guinea data using current software. With MIGRATE for example, estimation of a full migration matrix from the New Guinea data has been attempted (R. Leblois and F.R., unpublished results), yielding larger estimates of dispersal than inferred by the regression method described above and from demographic data. Attempts to estimate fewer parameters could be more successful. In addition, one can question the convergence of the Markov chain algorithm, a problem which remains with no easy solution (e.g. Brooks and Gelman, 1998). In this respect, the importance sampling algorithms of Griffiths and collaborators are more convenient as estimates are derived from independent runs of an absorbing Markov chain (28.15), so traditional techniques based on independent variables apply.

28.5   SEPARATION OF TIMESCALES

The properties of F-statistics illustrate the more general idea of separation of timescales, in which some events occur at a much faster rate than others. For example, in an island model, when the number of demes n_d increases indefinitely, the rate of coalescence of ancestral lineages of genes sampled in different demes decreases as $1/n_d$, while for genes sampled within the same deme, the probability of coalescence of ancestral lineages in some recent generation is nonvanishing in this limit. Thus the events in the genealogy of a sample can be described as a sum of two processes, a fast process by which lineages either coalesce within the demes they are sampled or separate in distinct demes, at rate $O(1)$ as

$n_d \to \infty$, and a slow process by which genes in different demes coalesce at rate $O(1/n_d)$ as $n_d \to \infty$ (e.g. Wakeley and Aliacar, 2001). This slow process is a rescaled version of the well-studied coalescence process for an unstructured population (the n-coalescent, Kingman, 1982; see **Chapter 25**).

Under a separation of timescales, the likelihood can be expressed in terms of distinct terms for fast and slow processes (see Nordborg, 1997 for a population with selfing), and in the above example, analytical results or simulation techniques for the n-coalescent can be used as soon as ancestral lineages have separated in different demes in a simulation of the ancestry of a sample. The traditional formula giving probability of identity conditional on allele frequency as $(1 - F)p^2 + Fp$ also results from such a separation of timescales, provided that allele frequency does not change in the total population until the completion of the fast process.

However, convergence to the n-coalescent may hold under sometimes restrictive conditions. No convergence to the n-coalescent has been found in one-dimensional models of isolation by distance (Cox, 1989; Wilkins, 2004). In two-dimensional models on a lattice of size $L \times L$, the genealogy of genes sampled far enough (at distance $O(L)$) from each other converges to that of an unstructured population (the n-coalescent) with coalescence events occurring at rate $O(1/[L^2 \ln(L)])$ as $L \to \infty$ (Cox, 1989; Zähle *et al.*, 2005). A separation of timescales may hold if the genealogy of genes sampled closer in space compounds a process of 'fast' coalescence (in less than $O[L^2 \ln(L)]$ generations) and the n-coalescent. The closest results are those of Zähle *et al.* according to which genes at distance $O(L^\beta)$ for $0 < \beta \le 1$ either coalesce in less than $L^2/2$ generations with probability $1 - \beta$, or follow the scaled n-coalescent with probability β. This differs qualitatively from the island model where the fast process becomes negligible in finite time, so defining a finite time span for a fast process from these results seems less than straightforward. Instead, they suggest applying n-coalescent approximations in a backward simulation algorithm when ancestral lineages are distant enough relative to the total size of the lattice, although the minimal distance to consider is itself not obvious since all results stand for $\beta > 0$ only, i.e. for spatial separation of genes increasing with L. As shown in Figure 28.2, a separation of timescales holds for fixed distance and fixed $N\sigma^2$, with the fast process vanishing in finite time, but the slower process is not an n-coalescent.

Some moment methods have attempted to extract more information from the data by taking into account allele size (for microsatellites) or DNA sequence divergence. However, in an island model, most of the information about Nm is in whether genes sampled within a deme have their most recent common ancestor within this deme (in which case they are identical, unless mutation rates are higher than migration rates). Otherwise, the ancestral lineages separate in different demes, and in this case the allelic divergence may contain little useful information about dispersal rates. This may explain why moment methods based on microsatellite allele size often yield estimates of demographic parameters with higher variance and mean square error than estimates derived from allelic identity statistics (Gaggiotti *et al.*, 1999; Balloux and Goudet, 2002; Leblois *et al.*, 2003) despite their lower asymptotic bias (Slatkin, 1995), except when mutation rates are larger than migration rates (Balloux *et al.*, 2000) or when genetic correlations are not determined by a distinctly fast process (Tsitrone *et al.*, 2001). A key issue in evaluating the performance of likelihood methods will be the extent to which they specifically capture the information from fast processes and how much better they are than pair-wise identity methods in this respect.

28.6   OTHER METHODS

As likelihood computations are often difficult, simulation methods based on other summary statistics have been developed. In essence, the likelihood of the data is substituted with the probability of observing in simulations values of a summary statistic S sufficiently close to its value observed in the data. These methods are reviewed by Beaumont (see **Chapter 30**) and have been applied to some subdivided population scenarios (Estoup *et al.*, 2004; Hamilton *et al.*, 2005). There is a huge collection of other methods of analysis of spatial patterns of genetic variation in the literature. Here again a small selection is presented on the basis of their impact in the field or of their perceived practical validity.

28.6.1   Assignment and Clustering

Of particular interest are assignment methods, which aim to assign individuals to their subpopulations of origin. For example, in an ecological perspective it may be useful to know whether immigrants differ from residents in some aspects of their behavior, and therefore to individually identify immigrants; further it might be of interest to identify their habitat of origin. Sometimes there will be independent information about the potential source populations. A more challenging problem (on which this section will focus) is to estimate dispersal rates by inferring the number of subpopulations and then assigning individuals to them (see **Chapter 30** for other aspects of these methods). In this perspective, the traditional problem of clustering becomes part of the assignment task.

Assignment methods stem from the idea that individuals are more likely to originate from populations with higher frequencies of the alleles they possess. Thus, considering an haploid organism for simplicity, for each individual one can compute a statistic such as

$$\prod_{k=1}^{K} \tilde{q}_{ki}^{x_k},\tag{28.22}$$

where $\tilde{q}_{ki}$ is the observed frequency for allele k in sample i (possibly with some correction such as excluding the focal individual itself), and $x_k = 1$ if the individual bears allele k and $x_k = 0$ otherwise. This is usually viewed as a likelihood statistic (e.g. Paetkau *et al.*, 1995). If each locus is considered independent of the others, multilocus statistics are the product of single locus statistics, and the individual is assigned to the sample i that maximizes the multilocus statistic.

It can be expected that, given some differentiation between different subpopulations, this method will preferentially assign an individual to its original subpopulation. It is also expected that such assignments will be more accurate when differentiation is higher. More generally, if the individuals are correctly assigned to their subpopulations of origin, then we could estimate from the same data and the same likelihood formulas the dispersal rates between each of them in the last round of dispersal, independently of past dispersal rates. Conversely if we cannot consistently estimate the dispersal rates in this way, this implies that there is no way to consistently assign individuals to their populations of origin. So we consider the question whether we can estimate the dispersal rate in order to address the question whether immigrants can be assigned to their population of origin.

Consider the allele frequencies in the subpopulations before the last round of dispersal (p_{ki} for allele k in subpopulation i), and the dispersal rate for this last round of dispersal

($m_{ii'}$ from subpopulation i' to i). For each locus and each individual in sample i the likelihood is approximately $\prod_{k=1}^{K} q_{ki}^{x_{ki}}$ where $q_{ki} = \sum_{i'} m_{ii'} p_{ki'}$ is the frequency of allele k in subpopulation i after the last round of dispersal. For the total sample from n_d demes the likelihood is therefore proportional to

$$\prod_{i=1}^{n_d} \prod_{k=1}^{K} q_{ki}^{n_{ki}}. \tag{28.23}$$

Since one can always explain the data by some model assuming that there was no dispersal in the last round, there is not enough information in the data to separately estimate the dispersal rates and the p_{ki}s, at least when each locus is considered independent of the others as done previously. This implies that it is not possible to consistently assign individuals to their populations of origin without introducing additional assumptions or additional knowledge.

The most obvious assumption is to assume no linkage disequilibrium within populations before dispersal, which is a good approximation for many organisms (though not highly selfing ones). In principle, one- and two-locus measures of genetic association contain information which allows to estimate separately dispersal rates and subpopulation sizes. A numerical method has been developed to use such information (Vitalis and Couvet, 2001). Assignment methods may be viewed as using multilocus associations, in that their current formulation relies on assuming no within-population disequilibria. Additional assumptions have been made, for example specifying a prior probability model for the distribution of allele frequencies, often a Dirichlet distribution as per an island model (see (28B.1); Rannala and Mountain, 1997; Falush *et al.*, 2003; Wilson and Rannala, 2003; Corander *et al.*, 2003). In this respect, although an asserted aim of such methods is to infer migration rates without the many assumptions of other methods, the stationary island model is still lurking in the background.

How do these methods perform in practice? Cornuet *et al.* (1999) reported >75 % correct assignment probabilities by Rannala and Mountain's method for populations diverged since several hundred generations. Évanno *et al.* (2005) found that the 'most likely' number of populations reported by the program STRUCTURE was a biased estimator of the actual number of populations. Waples and Gaggiotti (2006) reported that this program correctly identified the number of populations when the number of immigrants per population was less than 5, mutation rates were high, and for large sample size (20 loci genotype in 50 individuals from each population), but quickly degraded in other conditions. The latter study also reported similar problems with the methods implemented in the programs BAPS (Corander *et al.*, 2003) and IMMANC (Rannala and Mountain, 1997), and found that a method based on traditional contingency tests of spatial structure performed better than these different methods in identifying the number of populations. In a comparison with mark-recapture data, Berry *et al.* (2004) found good performance in estimating dispersal. In this study, the populations were known in advance and at most 4, the number of immigrants was known from mark-recapture data to be small, and the latter information was somehow used as prior information in STRUCTURE.

Most of the recently formulated methods have been presented as 'Bayesian', a label which in practice covers various compromises between subjective Bayesian (e.g. Lindley, 1990) and frequentist (Neyman, 1977) statistics, to the point of being uninformative. However, numerous accounts of supposed differences between Bayesian and frequentist

methods have drawn attention away from the real issue, which is the criterion by which to measure the performance of statistical methods. It is always possible to generate 'better than previous' methods by changing the measure of average performance. In the context of estimation of dispersal rates, the results of Beerli (2006) illustrate this, where performance averaged over a prior distribution for model parameters was better than that of maximum likelihood ignoring the prior distribution, a predictable result from textbook statistical theory (Cox and Hinkley, 1974, Chapter 11; Lehmann and Casella, 1998, Chapter 4). Likewise, the performance of an assignment method defined in terms of priors over allele frequencies may be evaluated in terms of its performance for any given allele frequency, or of averaged performance over the distribution of allele frequencies. The context of scientific inference should determine the appropriate measure of performance, but the averaged measure can balance misleading assignment inferences in some species with more efficient inferences in some other species, depending on the spectrum of allele frequencies in different species. One may need to know when this is the case, and comparing results for two choices of the prior distribution is not enough in this respect.

As seen above, a more prosaic problem with assignment methods is that it is difficult from available information to give simple bounds on the frequency of erroneous inferences, even averaged over prior distributions.

28.6.2   Inferences from Clines

An important class of models of spatial structure for selected genes are the cline models. Clines arise in two contexts: selection for two distinct alleles or genotypes in two adjacent habitats exchanging migrants, or selection against hybrid genotypes between two taxa ('tension zones'). Theoretical models predict the shape of clines, notably the steepness in the cline center, as a function of the σ parameter defined above and of one or several selection coefficients (for tension zones: Bazykin, 1969; Barton, 1979; for selection variable in space: Nagylaki, 1975). Additional information on dispersal and selection is given by linkage disequilibria between loci. Such methods, therefore, depend on assumptions specific to each case study (e.g. external information on recombination rates and epistasis between loci).

A typical expression for the shape of a cline, i.e. for allele frequency p at distance $x - x_0$ from the center of the cline, is $p = \left(1 + \exp((x_0 - x)(2s)^{1/2}/\sigma)\right)^{-1}$ (Bazykin, 1969; Barton, 1979). In the center of the cline, it holds approximately for different models of selection (Barton and Gale, 1993) and is relatively insensitive to the shape of the dispersal distribution, as a small rate of unaccounted long-distance immigration has little effect on the shape of the center of the cline (Rousset, 2001). Inferences about s and σ separately are possible by considering multilocus clines and by taking linkage disequilibria into account. Then, the expected shapes of clines must be computed numerically. Drift is neglected relative to selection: Allele frequencies in the different subpopulations are fixed values, functions of the parameters defining the demography and the selection regime, not random variables as in the neutral models.

The likelihood function is then given by multinomial sampling in each subpopulation. Thus, the statistical model is conceptually straightforward and relatively easily tailored to the specific demography and selection regime of different organisms, at least when only a few loci subject to selection need be considered and when the expected frequencies of each genotype are directly computed using recursion equations for specific values of the

parameters to be estimated (e.g. Lenormand *et al.*, 1999). Variants of the Metropolis algorithm (Metropolis *et al.*, 1953) including simulated annealing (Kirkpatrick *et al.*, 1983) have been used to find MLEs (a software ANALYSE is distributed by Barton and S. J. Baird, `http://helios.bto.ed.ac.uk/evolgen/Mac/Analyse/`). Sites *et al.* (1995) reported an agreement with demographic estimates of σ similar to that of isolation-by-distance analyses reported above.

28.7 INTEGRATING STATISTICAL TECHNIQUES INTO THE ANALYSIS OF BIOLOGICAL PROCESSES

Several of the methods reviewed in this chapter have been both widely used and widely criticized. Much of the criticisms rest on the difficulty of formulating precise quantitative models of population genetic processes. Theoretical studies of robustness are important, but may themselves overlook factors that turn out to be important in natural populations. For these reasons, this review has emphasized comparisons with demographic estimates. Independent demographic estimates may have their own problems, but it will be hard to detect misapplications of the genetic inferences if no such comparison is made. This last section discusses some of these problems and partial solutions to them.

We have seen that many discrepancies between 'models and data' inferred from empirical studies using F-statistics derive from various misunderstandings of the models. The more successful studies, in terms of comparisons with independent estimates, were conducted at a small spatial scale (between 1 and 20 σ). This is somehow unavoidable because of the need for good demographic data in these comparisons, but one may expect larger discrepancies over larger spatial scales, for many reasons: spatial variation of demographic parameters should be taken into account; the effects on genetic differentiation of some demographic events such as range expansions will be more likely to be observed (Slatkin, 1993); mutation will have measurable consequences (e.g. Estoup *et al.*, 1998); and selection variable in space may also affect the markers.

A frequently raised concern is the possible nonneutrality of the markers used. On the positive side a number of authors have realized that divergent selection would increase levels of differentiation between different subpopulations. Thus potentially selected loci may be detected in a first step by 'weak' statistical approaches such as classifying loci as showing structure or not by conventional significance tests (see Kreitman, 2000 for tests of selection at a molecular level not specifically using geographical information). Formal statistical evidence for selection may be obtained by other experiments in a second step. This approach has proven efficient (e.g. Feder *et al.*, 1997). Lewontin and Krakauer (1973) proposed a quantitative test of selection from the heterogeneity of F_{ST} estimates. This procedure was inadequate in several ways, but there have been more recent attempts to refine the detection of candidate selected loci (e.g. Beaumont and Nichols, 1996; Vitalis *et al.*, 2001; Beaumont and Balding, 2004).

Another often expressed criticism of the models and analyses reviewed above is that they assume equilibrium, while the populations are often not at demographic equilibrium, i.e. population sizes and migration rates fluctuate in time. If so, it is not clear what is estimated by such techniques: the present demography, an average over 'recent' times, a 'long-term' average, or none of them? If the fluctuations can be described as a fast process,

they may be described by some effective size correction. In other cases, such as the range expansion of a species, this cannot be so, and either modeling the demographic expansion, or using methods insensitive to it, are the only coherent alternatives. To understand this, it suffices to note that spatial patterns of pairs of genes approach equilibrium faster the smaller the spatial scale considered. Therefore, if the effect of a demographic event was captured by some effective size correction, the effective size would differ at different scales. Hence this effect cannot be described by a single effective size characterizing the total population.

One way to avoid some assumptions of equilibrium is to analyze sequential samples. All the above methods assume that samples have been taken at one point in time, yet sometimes temporal information is available. See e.g. Robledo-Arnuncio *et al.* (2006) for inferences of dispersal distributions from mother–offspring data, Wang and Whitlock (2003) for estimation of immigration rate and deme size from samples over wider time spans, and Ewing and Rodrigo (2006) for implementation of Markov chain algorithms for inference of changes in demographic parameters from sequential samples.

The idea of estimating dispersal parameters such as σ is also open to difficulties. By allowing an arbitrarily small fraction of immigrants to come from far enough, it is easy to design cases where the theoretical σ value would be arbitrarily large, and where long-distance immigrants would have arbitrarily small effect on the likelihood of samples. The question that one must address prior to statistical analysis is how important such long-distance immigrants are for population processes. For example, the speed of range expansions is known to be affected by the most extreme long-distance migrants in a way generally not characterized by the σ parameter (Mollison, 1977; Clark *et al.*, 2001), so if one is interested in characterizing such processes, not only it will be difficult to estimate σ but this may be irrelevant. One the other hand, some processes of local adaptation (as may lead to allele frequency clines, for example) are not very sensitive to long-distance migrants, and then approximations ignoring them are not only adequate but required to formulate good statistical inferences.

Current methods of estimation still have low range of applications, low efficiency, or both. In principle, this can be improved by the development of likelihood methods, yet this leaves room for different methodologies, and it is unclear how far research practices will be improved. One common theme is that genealogical structure is affected by events occurring at different timescales, and that inferences based on models of the faster processes could be relatively independent to uncontrolled historical processes, and therefore perhaps more reliable. It is not yet clear how much complexity we can add in the models, for given data, nor where will be the limit between reliable and unreliable inference; further, the answer will likely differ whether sequence data or allele frequency data are considered.

RELATED CHAPTERS

Chapter 25; **Chapter 26**; **Chapter 29**; and **Chapter 30**.

REFERENCES

Abdo, Z., Crandall, K.A. and Joyce, P. (2004). Evaluating the performance of likelihood methods for detecting population structure and migration. *Molecular Ecology* **13**, 837–851.

Abramovitz, M. and Stegun, I.A. (eds) (1972). *Handbook of Mathematical Functions*. Dover Publications, New York.

Austerlitz, F., Mariette, S., Machon, N., Gouyon, P.-H. and Godelle, B. (2000). Effects of colonization processes on genetic diversity: differences between annual plants and tree species. *Genetics* **154**, 1309–1321.

Balding, D.J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology* **63**, 221–230.

Balding, D.J. and Nichols, R.A. (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International* **64**, 125–140.

Balloux, F., Brünner, H., Lugon-Moulin, N., Hausser, J. and Goudet, J. (2000). Microsatellites can be misleading: an empirical and simulation study. *Evolution* **54**, 1414–1422.

Balloux, F. and Goudet, J. (2002). Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Molecular Ecology* **11**, 771–783.

Barton, N.H. (1979). The dynamics of hybrid zones. *Heredity* **43**, 341–359.

Barton, N.H. and Gale, K.S. (1993). Genetic analysis of hybrid zones. In *Hybrid Zones and the Evolutionary Process*, R.G. Harrison, ed. Oxford University Press, Oxford, pp. 13–45.

Bazykin, A.D. (1969). Hypothetical mechanism of speciation. *Evolution* **23**, 685–687.

Beaumont, M.A. and Balding, D.J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969–980.

Beaumont, M.A. and Nichols, R.A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London, Series B* **263**, 1619–1626.

Beerli, P. (2004). Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology* **13**, 827–836.

Beerli, P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**, 341–345.

Beerli, P. and Felsenstein, J. (1999). Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**, 763–773.

Berry, O., Tocher, M.D. and Sarre, S.D. (2004). Can assignment tests measure dispersal?. *Molecular Ecology* **13**, 551–561.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434–455.

Broquet, T., Johnson, C.A., Petit, E., Burel, F. and Fryxell, J.M. (2006). Dispersal kurtosis and genetic structure in the American marten, *Martes americana*. *Molecular Ecology* **15**, 1689–1697.

Chakraborty, R. (1992). Multiple alleles and estimation of genetic parameters: computational equations showing involvement of all alleles. *Genetics* **130**, 231–234.

Chuang, C. and Cox, C. (1985). Pseudo maximum likelihood estimation for the Dirichlet-multinomial distribution. *Communications in Statistics Part A: Theory and Methods* **14**, 2293–2311.

Clark, J.S., Lewis, M. and Horváth, L. (2001). Invasion by extremes: population spread with variation in dispersal and reproduction. *American Naturalist* **157**, 537–554.

Cockerham, C.C. (1973). Analyses of gene frequencies. *Genetics* **74**, 679–700.

Cockerham, C.C. and Weir, B.S. (1987). Correlations, descent measures: drift with migration and mutation. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 8512–8514.

Corander, J., Waldmann, P. and Sillanpää, M. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374.

Cornuet, J.M. and Beaumont, M.A. (2007). A note on the accuracy of PAC-likelihood inference with microsatellite data. *Theoretical Population Biology* **71**, 12–19.

Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A. and Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989–2000.

Cox, J.T. (1989). Coalescing random walks and voter model consensus times on the torus in $\mathbb{Z}^d$. *Annals of Probability* **17**, 1333–1366.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

Coyne, J.A., Barton, N.H. and Turelli, M. (1997). A critique of Sewall Wright's shifting balance theory of evolution. *Evolution* **51**, 643–671.

Crow, J.F. and Aoki, K. (1984). Group selection for a polygenic behavioural trait: estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 6073–6077.

DiCiccio, T.J. and Efron, B. (1996). Bootstrap confidence intervals (with discussion). *Statistical Science* **11**, 189–228.

Dobzhansky, T. and Wright, S. (1941). Genetics of natural populations. V. Relations between mutation rate and accumulation of lethals in populations of *Drosophila pseudoobscura*. *Genetics* **26**, 23–51.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, Boca Raton, FL.

Epperson, B.K. and Li, T.-Q. (1997). Gene dispersal and spatial genetic structure. *Evolution* **51**, 672–681.

Estoup, A., Beaumont, M., Sennedot, F., Moritz, C. and Cornuet, J.-M. (2004). Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution* **58**, 2021–2036.

Estoup, A., Rousset, F., Michalakis, Y., Cornuet, J.-M., Adriamanga, M. and Guyomard, R. (1998). Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). *Molecular Ecology* **7**, 339–353.

Évanno, G., Regnaut, S. and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611–2620.

Ewing, G. and Rodrigo, A. (2006). Coalescent-based estimation of population parameters when the number of demes changes over time. *Molecular Biology and Evolution* **23**, 988–996.

Falush, D., Stephens, M. and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.

Feder, J.L., Roethele, J.B., Wlazlo, B. and Berlocher, S.H. (1997). Selective maintenance of allozyme differences between sympatric host races of the apple maggot fly. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 11417–11421.

Fenster, C.B., Vekemans, X. and Hardy, O.J. (2003). Quantifying gene flow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculata* (Leguminosae). *Evolution* **57**, 995–1007.

Gaggiotti, O.E., Lange, O., Rassmann, K. and Gliddon, C. (1999). A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data. *Molecular Ecology* **8**, 1513–1520.

Goudet, J., Raymond, M., De Meeüs, T. and Rousset, F. (1996). Testing differentiation in diploid populations. *Genetics* **144**, 1931–1938.

Griffiths, R.C. and Tavaré, S. (1994). Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131–159.

Griffiths, R.C. and Tavaré, S. (1995). Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Mathematical Biosciences* **127**, 77–98.

Hamilton, G., Currat, M., Ray, N., Heckel, G., Beaumont, M. and Excoffier, L. (2005). Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* **170**, 409–417.

Hardy, O.J. and Vekemans, X. (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83**, 145–154.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hey, J. (2005). On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biology* **3**, e193.

Hudson, R.R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7**, 1–44.

de Iorio, M. and Griffiths, R.C. (2004a). Importance sampling on coalescent histories. *Advances in Applied Probability* **36**, 417–433.

de Iorio, M. and Griffiths, R.C. (2004b). Importance sampling on coalescent histories. II. Subdivided population models. *Advances in Applied Probability* **36**, 434–454.

de Iorio, M., Griffiths, R.C., Leblois, R. and Rousset, F. (2005). Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoretical Population Biology* **68**, 41–53.

Kingman, J.F.C. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19A**, 27–43.

Kirkpatrick, S., Gelatt, C. and Vecchi, M. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Kitada, S., Hayashi, T. and Kishino, H. (2000). Empirical Bayes procedure for estimating genetic distance between populations and effective population size. *Genetics* **156**, 2063–2079.

Kitakado, T., Kitada, S., Kishino, H. and Skaug, H.J. (2006). An integrated-likelihood method for estimating genetic differentiation between populations. *Genetics* **173**, 2073–2082.

Koenig, W.D., Van Vuren, D. and Hooge, P.N. (1996). Detectability, philopatry, and the distribution of dispersal distances in vertebrates. *Trends in Ecology and Evolution* **11**, 514–517.

Kreitman, M. (2000). Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics* **1**, 539–559.

Leblois, R., Estoup, A. and Rousset, F. (2003). Influence of mutational and sampling factors on the estimation of demographic parameters in a "continuous" population under isolation by distance. *Molecular Biology and Evolution* **20**, 491–502.

Leblois, R., Rousset, F. and Estoup, A. (2004). Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. *Genetics* **166**, 1081–1092.

Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, New York.

Lenormand, T., Bourguet, D., Guillemaud, T. and Raymond, M. (1999). Tracking the evolution of insecticide resistance in the mosquito *Culex pipiens*. *Nature* **400**, 861–864.

Lewontin, R.C. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175–195.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233. Correction: **167**, 1039.

Lindley, D.V. (1990). The present position in Bayesian statistics. *Statistical Science* **5**, 44–89.

Long, J.C. (1986). The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's *F*-statistics. *Genetics* **112**, 629–647.

Long, J.C., Naidu, J.M., Mohrenweiser, H.W., Gershowitz, H., Johnson, P.L. and Wood, J.W. (1986). Genetic characterization of Gainj- and Kalam-speaking peoples of Papua New Guinea. *American Journal of Physical Anthropology* **70**, 75–96.

Malécot, G. (1951). Un traitement stochastique des problèmes linéaires (mutation, linkage, migration) en génétique de population. *Annales de l'Université de Lyon A* **14**, 79–117.

Malécot, G. (1967). Identical loci and relationship. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, L.M. Le Cam and J. Neyman, eds. University of California Press, Berkeley, CA, pp. 317–332.

Malécot, G. (1975). Heterozygosity and relationship in regularly subdivided populations. *Theoretical Population Biology* **8**, 212–241.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–220.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Mollison, D. (1977). Spatial contact models for ecological and epidemic spread. *Journal of the Royal Statistical Society, Series B* **39**, 283–326.

Nagylaki, T. (1975). Conditions for the existence of clines. *Genetics* **80**, 595–615.

Nagylaki, T. (1976). The decay of genetic variability in geographically structured populations. II. *Theoretical Population Biology* **10**, 70–82.

Nagylaki, T. (1982). Geographical invariance in population genetics. *Journal of Theoretical Biology* **99**, 159–172.

Nagylaki, T. (1998). Fixation indices in subdivided populations. *Genetics* **148**, 1325–1332.

Nei, M. (1986). Definition and estimation of fixation indices. *Evolution* **40**, 643–645.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36**, 97–131.

Nielsen, R. and Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–896.

Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics* **146**, 1501–1514.

Paetkau, D., Calvert, W., Stirling, I. and Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology* **4**, 347–354.

Pannell, J.R. and Charlesworth, B. (1999). Neutral genetic diversity in a metapopulation with recurrent local extinction and recolonization. *Evolution* **53**, 664–676.

Rannala, B. and Hartigan, J.A. (1996). Estimating gene flow in island populations. *Genetical Research (Cambridge)* **67**, 147–158.

Rannala, B. and Mountain, J.L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 9197–9201.

Raufaste, N. and Bonhomme, F. (2000). Properties of bias and variance of two multiallelic estimators of F_{ST}. *Theoretical Population Biology* **57**, 285–296.

Raymond, M. and Rousset, F. (1995). GENEPOP version 1.2: population genetics software for exact tests and ecumenicism. *The Journal of Heredity* **86**, 248–249.

Robledo-Arnuncio, J.J., Austerlitz, F. and Smouse, P.E. (2006). A new method of estimating the pollen dispersal curve independently of effective density. *Genetics* **173**, 1033–1046.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219–1228.

Rousset, F. (1999a). Genetic differentiation in populations with different classes of individuals. *Theoretical Population Biology* **55**, 297–308.

Rousset, F. (1999b). Genetic differentiation within and between two habitats. *Genetics* **151**, 397–407.

Rousset, F. (2000). Genetic differentiation between individuals. *Journal of Evolutionary Biology* **13**, 58–62.

Rousset, F. (2001). Genetic approaches to the estimation of dispersal rates. In *Dispersal*, J. Clobert, E. Danchin, A.A. Dhondt and J.D. Nichols, eds. Oxford University Press, Oxford, pp. 18–28.

Rousset, F. (2002). Inbreeding and relatedness coefficients: what do they measure?. *Heredity* **88**, 371–380.

Rousset, F. (2004). *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton, NJ.

Rousset, F. (2006). Separation of time scales, fixation probabilities and convergence to evolutionarily stable states under isolation by distance. *Theoretical Population Biology* **69**, 165–179.

Rousset, F. and Raymond, M. (1997). Statistical analyses of population genetic data: old tools, new concepts. *Trends in Ecology and Evolution* **12**, 313–317.

Sawyer, S. (1977). Asymptotic properties of the equilibrium probability of identity in a geographically structured population. *Advances in Applied Probability* **9**, 268–282.

Searle, S.R. (1971). *Linear Models*. John Wiley & Sons, New York.

Sites, J.W. Jr., Barton, N.H. and Reed, K.M. (1995). The genetic structure of a hybrid zone between two chromosome races of the *Sceloporus grammicus* complex (Sauria, Phrynosomatidae) in central Mexico. *Evolution* **49**, 9–36.

Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research (Cambridge)* **58**, 167–175.

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**, 264–279.

Slatkin, M. (1994). Gene flow and population structure. In *Ecological Genetics*, L.A. Real, ed. Princeton University Press, Princeton, NJ, pp. 3–17.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462.

Smouse, P. and Williams, R.C. (1982). Multivariate analysis of HLA-disease associations. *Biometrics* **38**, 757–768.

Sokal, R. and Wartenberg, D.E. (1983). A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics* **105**, 219–237.

Storz, J.F., Ramakrishnan, U. and Alberts, S.C. (2001). Determinants of effective population size for loci with different modes of inheritance. *The Journal of Heredity* **92**, 497–502.

Sumner, J., Estoup, A., Rousset, F. and Moritz, C. (2001). 'Neighborhood' size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. *Molecular Ecology* **10**, 1917–1927.

Swofford, D. and Olsen, G. (1990). Phylogeny reconstruction. In *Molecular Systematics*, D. Hillis and C. Moritz, eds. Sinauer Associates, pp. 411–501.

Tachida, H. (1985). Joint frequencies of alleles determined by separate formulations for the mating and mutation systems. *Genetics* **111**, 963–974.

Tsitrone, A., Rousset, F. and David, P. (2001). Heterosis, marker mutational processes, and population inbreeding history. *Genetics* **159**, 1845–1859.

Tufto, J., Engen, S. and Hindar, K. (1996). Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics* **144**, 1911–1921.

Vekemans, X. and Hardy, O.J. (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* **13**, 921–934.

Vitalis, R. and Couvet, D. (2001). Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* **157**, 911–925.

Vitalis, R., Dawson, K. and Boursot, P. (2001). Interpretation of variation across marker loci as evidence of selection. *Genetics* **158**, 1811–1823.

Wakeley, J. and Aliacar, N. (2001). Gene genealogies in a metapopulation. *Genetics* **159**, 893–905. Correction in *Genetics* **160**, 1263.

Wang, J. and Whitlock, M.C. (2003). Estimating effective population size and migration rates from genetic samples over space and time. *Genetics* **163**, 429–446.

Waples, R.S. and Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology* **15**, 1419–1439.

Watts, P.C., Rousset, F., Saccheri, I.J., Leblois, R., Kemp, S.J. and Thompson, D.J. (2007). Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera) populations: analysis of 'neighbourhood' size using an improved estimator. *Molecular Ecology* **16**, 737–751.

Weir, B.S. (1996). *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Weir, B.S. and Cockerham, C.C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.

Wilkins, J.F. (2004). A separation-of-timescales approach to the coalescent in a continuous population. *Genetics* **168**, 2227–2244.

Wilson, A.J., Hutchings, J.A. and Ferguson, M.M. (2004). Dispersal in a stream dwelling salmonid: inferences from tagging and microsatellite studies. *Conservation Genetics* **5**, 25–37.

Wilson, G.A. and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**, 1177–1191.

Winters, J.B. and Waser, P.M. (2003). Gene dispersal and outbreeding in a philopatric mammal. *Molecular Ecology* **12**, 2251–2259.

Wood, J.W., Smouse, P.E. and Long, J.C. (1985). Sex-specific dispersal patterns in two human populations of highland New Guinea. *American Naturalist* **125**, 747–768.

Wright, S. (1931a). Evolution in Mendelian populations. *Genetics* **16**, 97–159. Reprinted in Wright (1986), pp. 98–160.

Wright, S. (1931b). Statistical theory of evolution. *Journal of the American Statistical Society* **26**(Suppl.), 201–208. Reprinted in Wright (1986), pp. 89–96.

Wright, S. (1946). Isolation by distance under diverse systems of mating. *Genetics* **31**, 39–59. Reprinted in Wright (1986), pp. 444–464.

Wright, S. (1949). Adaptation and selection. In *Genetics, Paleontology and Evolution*, G.L. Jepson, G.G. Simpson and E. Mayr, eds. Princeton University Press, pp. 365–389. Reprinted in Wright (1986), pp. 546–570.

Wright, S. (1969). *Evolution and the Genetics of Populations. II. The Theory of Gene Frequencies*. University of Chicago Press, Chicago, IL.

Wright, S. (1986). *Evolution: Selected Papers*. University of Chicago Press, Chicago, IL.

Zähle, I., Cox, J.T. and Durrett, R. (2005). The stepping stone model, II: genealogies and the infinite sites model. *Annals of Applied Probability* **15**, 671–699.

APPENDIX A: ANALYSIS OF VARIANCE AND PROBABILITIES OF IDENTITY

Here we detail the relationship between classical formulas for estimators of F-statistics and expressions in terms of frequency of identical pairs of genes. In the framework considered here, negative 'components of variance' (which are actually not variances) arise naturally.

We use the following notation: the total sample is made up of samples of n_i ($i = 1, \ldots, n_s$) individuals in n_s subpopulations; $X_{ij:k}$ is an indicator variable for gene j ($j = 1, \ldots, n_i$) in sample i being of allelic type k ($k = 1, \ldots, K$), i.e. $X_{ij:k} = 1$ if the sampled gene is of type k and $X_{ij:k} = 0$ otherwise; standard dot notation is used for sample averages: e.g. for weights w_j, $X_. \equiv (\sum_j w_j X_j)/(\sum_j w_j)$ is a weighted average of the Xs. Here the weighting for each individual will be simply 1. A discussion of optimal weighting with respect to allele frequencies or samples sizes will be given in Appendix B. The indicator variables are given a single index (X_j) if no reference is made to a specific sample. π_k is the expected frequency of allele k, the expectation of $X_{j:k}$ over independent replicates of some evolutionary process (typically a mutation–drift stationary equilibrium, but this is in no way required).

For haploid data the statistical model is generally described as

$$X_{ij:k} = \mu + \alpha_i + \varepsilon_{ij}, \tag{28A.1}$$

$\mu = \pi_k$ here, α_i is a random effect with zero mean and variance σ_a^2, and ε_{ij} is a random effect with zero mean and variance σ_e^2. It is also assumed that $E[\alpha_i \alpha_{i'}] = 0$ for $i \neq i'$ and that $E[\varepsilon_{ij}\varepsilon_{ij'}] = 0$ for $j \neq j'$ (e.g. Searle, 1971, p. 384), but this is precisely what we will not do here, in order to obtain the most general analogy. What remains more generally valid is a basic algebraic relationship of analysis of variance,

$$\sum_j w_j (X_j - \mu)^2 = \sum_j w_j (X_j - X_. + X_. - \mu)^2$$

$$= \sum_j w_j (X_j - X_.)^2 + \sum_j w_j (X_. - \mu)^2, \tag{28A.2}$$

for any variable X and weights w_j because we still have $\sum_j w_j (X_{j:k} - X_{:k}) = 0$ by definition of $X_{.:k}$.

The method of analysis of variance is then based on the computation of weighted sums of squares related as follows:

$$\sum_{\text{samples}}^{n_s} \sum_{\text{genes}}^{n_i} w_i (X_{ij:k} - X_{...k})^2 = \sum^{n_s} \sum^{n_i} w_i (X_{ij:k} - X_{i.:k})^2$$

$$+ \sum^{n_s} \sum^{n_i} w_i (X_{i.:k} - X_{...k})^2 \tag{28A.3}$$

$$\equiv SS_{\text{w[ithin]}} + SS_{\text{b[etween]}} \text{ for allele } k. \tag{28A.4}$$

For $w_i = 1$ these sums of squares will be expressed in terms of $S_1 \equiv \sum_i n_i$ and $S_2 \equiv \sum_i n_i^2$, of the observed frequency of pairs of different genes within samples which are both of type k, $\hat{Q}_{2:k} \equiv \sum_i \sum_{j \neq j'} X_{ij:k}X_{ij':k}/(S_2 - S_1)$, and of the observed frequency of pairs of different genes between samples which are both of type k, $\hat{Q}_{3:k} \equiv \sum_{i \neq i'} \sum_{j,j'} X_{ij:k}X_{i'j':k}/(S_1^2 - S_2)$. We first express the sums of squares using (28A.2), as follows:

$$SS_{\text{w}} = \sum_i \sum_j^{n_i} (X_{ij:k} - X_{i.:k})^2 = \sum_i \sum_j^{n_i} (X_{ij:k} - \pi_k)^2 - \sum_i n_i (X_{i.:k} - \pi_k)^2, \tag{28A.5}$$

and

$$SS_{\text{b}} = \sum_i \sum_j^{n_i} (X_{i.:k} - X_{...k})^2 = \sum_i n_i (X_{i.:k} - \pi_k)^2 - S_1 (X_{...k} - \pi_k)^2. \tag{28A.6}$$

The values of different variables that appear in these expressions, $(X_{j:k} - \pi_k)^2$ and $(X_{j:k} - \pi_k)(X_{j':k} - \pi_k)$, for pairs j, j' of different genes, are summarized in Table 28.1 where $Q_{.k}$ is the probability that both genes of a pair are of type k. Note that $E[(X_{j:k} - \pi_k)(X_{j':k} - \pi_k)]$ is the covariance between allele frequencies in the subpopulations from

Table 28.1 Values of variables in comparisons of pairs of genes.

Pair of gene	kk	k, not k	None k
Probability of each pair	$Q_{:k}$	$2(\pi_k - Q_{:k})$	$1 - 2\pi_k + Q_{:k}$
Frequency in total sample	$\hat{Q}_{:k}$	$2(\tilde{\pi}_k - \hat{Q}_{:k})$	$1 - 2\tilde{\pi}_k + \hat{Q}_{:k}$

Variable	Value of variable for each pair		
$(X_{j:k} - \pi_k)(X_{j':k} - \pi_k)$	$(1 - \pi_k)^2$	$-\pi_k(1 - \pi_k)$	π_k^2
$(X_{j:k} - \pi_k)^2 + (X_{j':k} - \pi_k)^2$	$2(1 - \pi_k)^2$	$(1 - \pi_k)^2 + \pi_k^2$	$2\pi_k^2$

which j and j' are sampled. From this table we see that

$$\sigma_a^2 + \sigma_e^2 = \mathrm{E}[(X_{j:k} - \pi_k)^2] = \pi_k(1 - \pi_k), \tag{28A.7}$$

$$\mathrm{E}[(X_{j:k} - \pi_k)(X_{j':k} - \pi_k)] = Q_{:k} - \pi_k^2. \tag{28A.8}$$

where $Q_{:k}$ is $Q_{w:k} = Q_{2:k}$ for two genes within a sample and $Q_{b:k} = Q_{3:k}$ for two genes between samples. In particular we have

$$Q_{2:k} - \pi_k^2 = \mathrm{Cov}[X_{ij} X_{ij'}] = \sigma_a^2 + \mathrm{E}[\varepsilon_{ij} \varepsilon_{ij'}], \tag{28A.9}$$

$$Q_{3:k} - \pi_k^2 = \mathrm{Cov}[X_{ij} X_{i'j'}] = \mathrm{E}[\alpha_i \alpha_{i'}]. \tag{28A.10}$$

The latter expression confirms that $\mathrm{E}[\alpha_i \alpha_{i'}] \neq 0$: in general two genes in different subpopulations are more likely to be identical than two independent genes, i.e. $Q_{3:k} - \pi_k^2 > 0$. In the present case one could consider a slightly different parameterization of the model, so that this positive component would appear as a variance (e.g. Cockerham and Weir, 1987), but more generally this would be confusing because we may also have to consider negative terms, as shown below for diploid data.

The table also shows that the sample averages of $(X_{j:k} - \pi_k)(X_{j':k} - \pi_k)$ and $(X_{j:k} - \pi_k)^2$ are $\hat{Q}_{:k} + \pi_k^2 - 2\pi_k \tilde{\pi}_k$ and $\hat{\pi}_k + \pi_k^2 - 2\pi_k \tilde{\pi}_k$, respectively. Here $\tilde{\pi}_k$ is the average allele frequency among all *pairs of genes* for which the average is written. This is the observed allele frequency by gene counting (denoted $\hat{\pi}_k$ or $\hat{\pi}_{i:k}$) when all pairs in the total sample or in sample i are considered. Among all pairs of genes sampled without replacement within each sample, this is $\hat{\pi}_{w:k} \equiv \sum_i n_i(n_i - 1)\hat{\pi}_{i:k}/(S_2 - S_1)$. Among all pairs of genes from two different samples, $\hat{\pi}$ has value $\hat{\pi}_{b:k}$ given by $(S_1^2 - S_2)\hat{\pi}_{b:k} + (S_2 - S_1)\hat{\pi}_{w:k} = (S_1^2 - S_1)\hat{\pi}$.

Then

$$\sum_i \sum_j^{n_j} (X_{ij:k} - \pi_k)^2 = S_1(X_{..:k} + \pi_k^2 - 2\pi_k X_{..:k}). \tag{28A.11}$$

Next

$$(X_{..:k} - \pi_k)^2 = \left(\frac{\sum_{i,j} X_{ij:k} - \pi_k}{S_1} \right)^2 \tag{28A.12}$$

$$= \frac{1}{S_1^2} \left(\sum_{i,j} (X_{ij:k} - \pi_k)^2 \right.$$

$$+ \sum_{i \neq i', j, j'} (X_{i'j':k} - \pi_k)(X_{i'j':k} - \pi_k)$$

$$+ \sum_{i, j \neq j'} (X_{ij:k} - \pi_k)(X_{ij':k} - \pi_k) \Bigg) \tag{28A.13}$$

$$= \frac{1}{S_1^2} \Bigg(S_1(\hat{\pi} + \pi_k^2 - 2\hat{\pi}\pi_k) + (S_1^2 - S_2)(\hat{Q}_{3:k} + \pi_k^2 - 2\hat{\pi}_{b:k}\pi_k)$$

$$+ (S_2 - S_1)(\hat{Q}_{2:k} + \pi_k^2 - 2\hat{\pi}_{w:k}\pi_k) \Bigg) \tag{28A.14}$$

$$= (\hat{\pi} + \pi_k^2 - 2\hat{\pi}\pi_k) + \frac{1}{S_1^2} \Big((S_1^2 - S_2)(\hat{Q}_{3:k} - \hat{\pi}) + (S_2 - S_1)(\hat{Q}_{2:k} - \hat{\pi}) \Big), \tag{28A.15}$$

by definition of $\hat{\pi}_{b:k}$. Likewise

$$(X_{i:k} - \pi_k)^2 = \left(\frac{\sum_{j=1}^{n_i} X_{ij:k} - \pi_k}{n_i} \right)^2 \tag{28A.16}$$

$$= \frac{1}{n_i^2} \left(\sum_{j=1}^{n_i} \left(X_{ij:k} - \pi_k \right)^2 + \sum_{j \neq j'} \left(X_{ij:k} - \pi_k \right) \left(X_{ij':k} - \pi_k \right) \right) \tag{28A.17}$$

$$= \frac{1}{n_i^2} \left(n_i(\hat{\pi}_{i:k} + \pi_k^2 - 2\hat{\pi}_{i:k}\pi_k) + n_i(n_i - 1)(\hat{Q}_{i2:k} + \pi_k^2 - 2\hat{\pi}_{i:k}\pi_k) \right) \tag{28A.18}$$

$$= (\hat{\pi}_{i:k} + \pi_k^2 - 2\hat{\pi}_{i:k}\pi_k) + \frac{n_i - 1}{n_i}(\hat{Q}_{i2:k} - \hat{\pi}_{i:k}), \tag{28A.19}$$

hence

$$\sum_{i=1}^{n_s} n_i(X_{i:k} - \pi_k)^2 = S_1(\hat{\pi}_k + \pi_k^2 - 2\hat{\pi}_k\pi_k) + \sum_i (n_i - 1)(\hat{Q}_{i2:k} - \hat{\pi}_{i:k}). \tag{28A.20}$$

Then from (28A.5), (28A.11) and (28A.20)

$$SS_w = (S_1 - n_s)(\hat{\pi}_k - \hat{Q}_{2:k}), \tag{28A.21}$$

and from (28A.6), (28A.15) and (28A.20)

$$SS_b = \sum_i (\hat{\pi}_{i:k} - \hat{\pi}_k) + \sum_i (n_i - 1)\hat{Q}_{i2:k} - (S_1 - S_2/S_1)\hat{Q}_{3:k} - (S_2/S_1 - 1)\hat{Q}_{2:k}. \tag{28A.22}$$

Note that as in (28A.20), allele frequencies terms do not reduce to a function of $\hat{\pi}$ only: the term $\sum_i (\hat{\pi}_{i:k} - \hat{\pi})$ will usually be nonzero when sample sizes are unequal. Taking

expectations, one has

$$
\begin{aligned}
\mathrm{E}\left[SS_{\mathrm{w}}\right] &= (S_1 - n_{\mathrm{s}})(\pi_k - Q_{2:k}), \\
\mathrm{E}\left[SS_{\mathrm{b}}\right] &= (S_1 - S_2/S_1)(Q_{2:k} - Q_{3:k}) + (n_{\mathrm{s}} - 1)(\pi_k - Q_{2:k}) \\
&= (S_1 - S_2/S_1)(\sigma_a^2 - \mathrm{E}[\alpha_i\alpha_{i'}] + \mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}]) + (n_{\mathrm{s}} - 1)(\sigma_e^2 - \mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}]).
\end{aligned}
$$
(28A.23)

These relationships hold whatever the model considered (fixed or random, etc.). They are formally equivalent to a standard analysis of variance (e.g. Searle, 1971) on the indicator variables $X_{ij:k}$, except that (1) $\mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}]$ and $\mathrm{E}[\alpha_i\alpha_{i'}]$ are not assumed null, and (2) the sums of squares are themselves summed over alleles. When we write these two modifications as '$\overset{1}{\to}$' and '$\overset{2}{\to}$' the equivalence of expectations in the standard formulas of analysis of variance with expressions in terms of probabilities of identity is as follows:

$$
\sigma_a^2 \overset{1}{\to} \sigma_a^2 - \mathrm{E}[\alpha_i\alpha_{i'}] + \mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}] \overset{2}{\to} Q_2 - Q_3 \equiv (1 - Q_3)F_{\mathrm{ST}}
$$

$$
\sigma_e^2 \overset{1}{\to} \sigma_e^2 - \mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}] \overset{2}{\to} 1 - Q_2 = (1 - Q_3)(1 - F_{\mathrm{ST}}). \quad (28A.24)
$$

Hence the 'intraclass covariance' $\sigma_a^2/(\sigma_a^2 + \sigma_e^2)$, often taken as a definition of F_{ST}, should be interpreted as

$$
F_{\mathrm{ST}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \overset{1}{\to} F_{\mathrm{ST}} = \frac{\sigma_a^2 + \mathrm{E}[\varepsilon_{ij}\varepsilon_{ij'}] - \mathrm{E}[\alpha_i\alpha_{i'}]}{\sigma_a^2 + \sigma_e^2 - \mathrm{E}[\alpha_i\alpha_{i'}]}, \quad (28A.25)
$$

where the latter expression may be considered more general.

For diploid data, the model is $X_{ijl:k} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijl}$ for gene l ($l = 1, 2$ for diploids) of individual j in population i. With $\sigma_a^2 \equiv \mathrm{E}[\alpha_i^2]$, $\sigma_b^2 \equiv \mathrm{E}[\beta_{ij}^2]$, $\sigma_e^2 \equiv \mathrm{E}[\varepsilon_{ijl}^2]$, we have

$$
\sigma_a^2 \overset{1}{\to} \sigma_a^2 - \mathrm{E}[\alpha_i\alpha_{i'}] + \mathrm{E}[\beta_{ij}\beta_{ij'}] \overset{2}{\to} Q_2 - Q_3 \equiv (1 - Q_3)F_{\mathrm{ST}}
$$

$$
\sigma_b^2 \overset{1}{\to} \sigma_b^2 - \mathrm{E}[\beta_{ij}\beta_{ij'}] + \mathrm{E}[\varepsilon_{ijl}\varepsilon_{ijl'}] \overset{2}{\to} Q_1 - Q_2 \equiv (1 - Q_3)F_{\mathrm{IS}}(1 - F_{\mathrm{ST}})
$$

$$
\sigma_e^2 \overset{1}{\to} \sigma_e^2 - \mathrm{E}[\varepsilon_{ijl}\varepsilon_{ijl'}] \overset{2}{\to} 1 - Q_1 = (1 - Q_3)(1 - F_{\mathrm{IS}})(1 - F_{\mathrm{ST}}),
$$
(28A.26)

where Q_1 is the probability of identity of genes within a diploid individual, Q_2 is for genes between individuals within subpopulations, and Q_3 between subpopulations. Thus in both formalisms we see that the 'components of variance' actually translate into more general expressions that can be negative. When there is an heterozygote excess within demes ($Q_1 < Q_2$), $\mathrm{E}[\varepsilon_{ijl}\varepsilon_{ijl'}]$ is negative, and $\sigma_b^2 - \mathrm{E}[\beta_{ij}\beta_{ij'}] + \mathrm{E}[\varepsilon_{ijl}\varepsilon_{ijl'}]$ is negative.

In the haploid case, (28A.23) implies that

$$
\frac{(S_1 - n_{\mathrm{s}})\mathrm{E}[SS_{\mathrm{b}}] - (n_{\mathrm{s}} - 1)\mathrm{E}[SS_{\mathrm{w}}]}{(S_1 - n_{\mathrm{s}})\mathrm{E}[SS_{\mathrm{b}}] + (W_c - 1)(n_{\mathrm{s}} - 1)\mathrm{E}[SS_{\mathrm{w}}]} = \frac{Q_2 - Q_3}{1 - Q_3}, \quad (28A.27)
$$

where SS_{w} and SS_{b} are now summed over alleles (e.g. $SS_{\mathrm{w}} \equiv \sum^{n_{\mathrm{s}}} \sum^{n_i} \sum_k (X_{ij:k} - X_{i.:k})^2$), and $W_c \equiv (S_1 - S_2/S_1)/(n_{\mathrm{s}} - 1)$. Although we have related the expectation of the

different terms to 'components of variance' in a model such as (28A.1), we note again that the last equality holds independently of such a model, because it is only based on the basic relationship (28A.2). Accordingly, an estimator of F_{ST} is the ratio of unbiased estimators

$$\frac{(S_1 - n_s)SS_b - (n_s - 1)SS_w}{(S_1 - n_s)SS_b + (W_c - 1)(n_s - 1)SS_w}. \tag{28A.28}$$

(see also Weir, 1996, p. 182). One could hope that this estimator is directly interpretable as $(\hat{Q}_2 - \hat{Q}_3)/(1 - \hat{Q}_3)$, where $\hat{Q}_j = \sum_k \hat{Q}_{j:k}$ are the frequencies of identical pairs of genes in the sample, computed by simple counting either within (for Q_2) or between (for Q_3) samples ($\hat{Q}_2$ being is an average over the different samples, weighted according to the number of pairs in each sample). But this is not so when sample sizes are unequal, because the term $\sum_i (\hat{\pi}_{i:k} - \hat{\pi})$ from (28A.22) remains in the above expression. The expression closest to $(\hat{Q}_2 - \hat{Q}_3)/(1 - \hat{Q}_3)$ that I have found for Weir and Cockerham's estimator is

$$\frac{\tilde{Q}_2 - \hat{Q}_3}{1 - \hat{Q}_3 + \sum_i (n_i - 1)(\hat{Q}_{i2} - \hat{Q}_2)\frac{S_2 - S_1}{(S_1^2 - S_2)(S_1 - n_s)}}, \tag{28A.29}$$

in terms of the weighted frequency

$$\tilde{Q}_2 = \frac{(S_1 - 1)\sum_i (n_i - 1)\hat{Q}_{i2} - (S_1 - n_s)(S_2/S_1 - 1)\hat{Q}_2}{(S_1 - n_s)(S_1 - S_2/S_1)}, \tag{28A.30}$$

and where $\hat{Q}_{i2}$ is the observed frequency of pairs of genes identical in state among all pairs taken without replacement within sample i. Compared to the analysis-of-variance estimator, the simple strategy of estimating any function of probabilities of identities by the equivalent function of frequencies of identical pairs of genes is equally 'unbiased', has no obvious drawback and is easily adaptable to different settings.

For multilocus data it is usual to compute the estimator as a sum of locus-specific numerators over a sum of locus-specific denominators; see e.g. Weir and Cockerham (1984) or Weir (1996) for details. Note that the sums are weighted differently in these two references. The numerator in Weir and Cockerham (1984), eqs. (2) and (10), is $\bar{n}/W_c$ times the one in Weir (1996), p.178–179. Parallel changes in the denominator ensure that the one-locus estimators are identical, but the multilocus estimators will be different if $\bar{n}/W_c$ varies between loci.

APPENDIX B: LIKELIHOOD ANALYSIS OF THE ISLAND MODEL

Sampling Formulas

Consider an infinite island model of haploid subpopulations where there are K alleles and n_s subpopulations are sampled. The following notation will be used: π_k is the frequency of allele k in the total population (which is not a random variable here) and p_{ki} is frequency of allele k in subpopulation i; n_{ki} is the number of genes of type k in the sample from subpopulation i; n_i is the size of sample i, $\bar{n}$ is the average n_i, $\tilde{\pi}_k \equiv \sum_i n_{ki}/\sum_i n_i$ and $\tilde{p}_{ki} \equiv n_{ki}/n_i$ are the observed frequencies of allele k in the total sample and in sample i, respectively.

The distribution of the p_{ki}s in population i follows a Dirichlet distribution,

$$L(p_{ki}, \ldots, p_{Ki}) = \Gamma(M) \prod_k \frac{p_{ki}^{M\pi_k - 1}}{\Gamma(M\pi_k)}. \qquad (28B.1)$$

This type of distribution arises as a diffusion approximation to the discrete generation Wright–Fisher model, where M is twice the number of migrant genes per generation (Wright, 1949), e.g. $2Nm$ or $4Nm$, and more generally in any scenario that can be approximated by the n-coalescent.

It should be noted that this equation is valid for each subpopulation with its own size, N_i, and its own immigration rate, m_i. Thus the likelihood of samples may be given for an infinite island model only characterized by the homogeneous dispersal of individuals to other demes: the N_is and m_is need not be identical in all subpopulations. This result implies that, with a large number of subpopulations, one can only estimate the products $N_i m_i$ for each deme.

Consider the vector of counts n_{ki} of allele k in subpopulation i, $\mathbf{n}_i \equiv (n_{1i}, \cdots, n_{Ki})$, and the corresponding multinomial coefficient $C(\mathbf{n}_i) \equiv n_i! / \prod_{k=1}^K n_{ki}!$. The conditional probability distribution of the ith sample, given the subpopulation frequencies $\mathbf{p}_i \equiv (p_{1i}, \ldots, p_{Ki})$, is multinomial:

$$C(\mathbf{n}_i) \prod_k^K p_{ki}^{n_{ki}}. \qquad (28B.2)$$

The probability distribution of a sample $\mathbf{n}_i$ in subpopulation i must be expressed as a function only of the parameters, M and of expected allele frequencies (expectations under stochastic model) $\boldsymbol{\pi} \equiv (\pi_1, \ldots, \pi_K)$, by combining (28B.1) and (28B.2) and summing over the set $\mathcal{S}$ of possible values of allele frequencies $\mathbf{p}_i$:

$$L(M, \boldsymbol{\pi}; n_{1i}, \ldots, n_{Ki}) = \int \ddot{\mathcal{S}} \int \Gamma(M) \prod_k^K \frac{p_{ki}^{M\pi_k - 1}}{\Gamma(M\pi_k)} C(\mathbf{n}_i) \prod_k^K p_{ki}^{n_{ki}} \, d\mathbf{p}_i \quad (28B.3)$$

$$= \frac{\Gamma(M)}{\Gamma(M + n_i)} C(\mathbf{n}_i) \prod_k^K \frac{\Gamma(M\pi_k + n_{ki})}{\Gamma(M\pi_k)}. \qquad (28B.4)$$

This distribution is the Dirichlet-multinomial distribution. In the infinite island model, subpopulation frequencies are independent from each other in each subpopulation, so that the likelihood of a total sample from n_s subpopulations is

$$L(M, \boldsymbol{\pi}) = \left(\frac{\Gamma(M)}{\Gamma(M + n_i)}\right)^{n_s} \prod_{i=1}^{n_s} C(\mathbf{n}_i) \prod_k^K \frac{\Gamma(M\pi_k + n_{ki})}{\Gamma(M\pi_k)}. \qquad (28B.5)$$

The pseudo–maximum likelihood estimator $\hat{M}_A$ of M had been previously defined by Chuang and Cox (1985) as the solution of $\partial \ln L / \partial M|_{\boldsymbol{\pi} = \tilde{\boldsymbol{\pi}}, M = \hat{M}_A} = 0$. From (28B.5), this is the solution of

$$0 = \left[\sum_k^K \sum_i^{n_s} \tilde{\pi}_k \left(\sum_{k=0}^{n_{ki}-1} \frac{1}{\tilde{\pi}_k M + k}\right) - \sum_i^{n_s} \left(\sum_{k=0}^{n_i-1} \frac{1}{M + k}\right)\right]. \qquad (28B.6)$$

Efficiency in the Island Model

When $M \to \infty$ (high migration rates) the Dirichlet-multinomial distribution converges to a multinomial with parameter $\boldsymbol{\pi}$, so the sampling distribution for the total sample is a product of multinomials with the same parameter $\boldsymbol{\pi}$: this corresponds to the case of no population differentiation. Thus we can construct asymptotically efficient statistics for detecting weak differentiation from a study of the properties of the likelihood when $M \to \infty$. To that aim it is simpler to express it as a function of $\psi \equiv 1/M$ and compute the Taylor expansion near $\psi = 0$. From (28B.6), it may be shown that

$$\frac{\partial \ln L}{\partial \psi} = \sum_{k}^{K} \sum_{i}^{n_s} \frac{n_{ki}(n_{ki} - 1)}{2\pi_k} - \sum_{i}^{n_s} \frac{n_i(n_i - 1)}{2} + O(\psi), \qquad (28B.7)$$

and a statistic of interest (effectively a score statistic, Cox and Hinkley, 1974, Chapter 9) may be constructed as

$$\tilde{U} \equiv \lim_{\psi \to 0} \left. \frac{\partial \ln L}{\partial \psi} \right|_{\boldsymbol{\pi} = \tilde{\boldsymbol{\pi}}} = \sum_{k}^{K} \sum_{i}^{n_s} \frac{n_{ki}(n_{ki} - 1)}{2\tilde{\pi}_k} - \sum_{i}^{n_s} \frac{n_i(n_i - 1)}{2}, \qquad (28B.8)$$

where the $\tilde{\pi}$s are the observed allele frequencies in the total sample, which are the MLEs of the πs in the case $\psi = 0$. This result draws a connection between the likelihood and the moment methods (see also Balding, 2003). Since the second sum in (28B.8) is fixed for given sample sizes n_i, the score statistic is essentially a sum of squares and can be considered in an analysis of variance framework. It shows that asymptotically efficient weights w_k of the sum of squares for the different alleles are proportional to $1/\tilde{\pi}_k$, and the weights of the sum of squares for the different samples are proportional to n_i^2 for the different samples, i.e. $w_i = n_i$ for each individual in (28A.3). The allele weighting is not new: it is implicit in the matrix formulations of Smouse and Williams (1982) and Long (1986) (see Weir and Cockerham, 1984; Chakraborty, 1992) and in standard test statistics such as the χ^2 or log-likelihood for multinomial models. Consistent with the above analysis assuming weak differentiation ($\psi \to 0$), it leads to estimators with efficient properties only for low differentiation (Raufaste and Bonhomme, 2000). By contrast the sample size weighting is odd and has not been previously considered in analysis of variance. But it may bring very little (F.R., unpublished data): the allele weighting is generally sufficient to turn moment statistics into efficient statistics when differentiation is low.

Weighting according to observed allele frequencies or to other measures of genetic diversity may have some drawbacks, particularly when one considers more general models than the island model. This weighting seems to imply that the ratio of expected sum of squares, conditional on genetic diversity, is independent of the value of the conditioning variable. As noted in the main Text, this is approximately so in the island model, and more generally under a separation of timescales when the slow process is an n-coalescent, but may not hold more generally. Then, the only consistent method would be to sum the $\hat{Q}$ terms directly in the numerator and denominator; any other method would introduce a bias. Further, selection of markers with specific levels of variability−as is often the case in practice−could also introduce an ascertainment bias.

Analysis of Population Subdivision

L. Excoffier

Zoological Institute, Department of Biology, University of Berne, Berne, Switzerland

This paper reviews the basic measures of intraspecific population subdivision, without assuming any particular spatial model. The structure of the paper follows from the historical development of the field, starting with the presentation of Wright's fixation indexes as correlations of gene frequencies, then Nei's extension of *F* statistics to multiple alleles as function of heterozygosities and his introduction of *G* statistics as function of gene diversities, to end with Cockerham's intraclass correlations obtained under an analysis of variance (ANOVA) framework. We also present an extension of the ANOVA approach to handle molecular data (DNA sequences, microsatellites) as well as dominant data. The relationships between different estimators, different approaches, and their potential limitations are discussed.

29.1 INTRODUCTION

29.1.1 Effects of Population Subdivision

Species or populations usually do not constitute a single panmictic unit where individuals breed at random over the whole species range. They are rather subdivided into smaller entities, which can be arranged in space, time, ecology, or otherwise. Different levels of population subdivisions may exist, which may be hierarchically arranged or not. A population can thus be divided into groups or regions, themselves divided into smaller units, and so on until we reach a basic unit of population that may be treated as a homogeneous group, a fundamental unit called a *deme* (Gilmour and Gregor, 1939) or a neighborhood (Wright, 1946) in the context of continuous habitats. Even within demes, individuals are often related to each other, due to some shared ancestry and finite deme sizes, which introduces some local levels of inbreeding often resulting in homozygote excess.

Population structure, if not properly recognized, can affect our interpretation of the observed pattern of genetic diversity, and can often mimic the effect of selection. Wahlund (1928) thus showed that hidden subdivision leads to an apparent excess of homozygotes (see also Nagylaki, 1985). When several loci are considered, population subdivision can

lead to observed patterns of linkage disequilibrium, even between physically unlinked loci (e.g. Ohta and Kimura, 1969; Nei and Li, 1973; Li and Nei, 1974; Ohta, 1982), and recently admixed populations showing high levels of linkage disequilibrium over long chromosomal segments can be interesting to locate disease genes (e.g. Chakraborty and Weiss, 1988; Chapman and Thompson, 2001). At the molecular level, the presence of hidden and unrecognized population subdivision can increase the number of segregating sites within samples relative to the average number of substitutions between pairs of sequences (Tajima, 1993; Wakeley, 1998), while local inbreeding due to selfing can result in a drastic reduction of genetic diversity (Fu, 1997; Nordborg and Donnelly, 1997). In a population made up of demes interconnected by migrations, the mean number of pairwise differences (or the number of heterozygous sites within individuals) only depends on the sum of deme sizes, irrespective of the migration pattern (e.g. Slatkin, 1987; Strobeck, 1987; Nagylaki, 1998a), but this property does not hold for the variance (Marjoram and Donnelly, 1994). It follows that statistical tests of selective neutrality can often be significant in absence of selection, when performed on samples drawn from subdivided populations (e.g. Tajima, 1993; Simonsen *et al.*, 1995; Nielsen, 2001). Also, hidden subdivision or population stratification can severely affect tests of genetic association and case control studies (e.g. Pritchard *et al.*, 2000b; Pritchard and Donnelly, 2001; Marchini *et al.*, 2004). When selection is present, population subdivision can allow the distinction between several forms of selection that have different effects on diversity within and between demes (Charlesworth *et al.*, 1997). The inference of population subdivision is also useful for its own sake. For instance, in conservation biology it may be crucial to understand phylogeographic patterns and to recognize subspecies or the effect of settlement history on genetic diversity (e.g. Taberlet *et al.*, 1998; Hewitt, 2001; Falush *et al.*, 2003b; Jaarola and Searle, 2004; Magri *et al.*, 2006). Moreover, the presence and the assessment of genetic structure is also essential in studies of species adaptation under Wright (1932; 1982) shifting balance theory.

In statistical terms, population structure introduces some correlation or covariance between genes taken from different subdivision levels. We are interested in describing here how we can detect, measure and test for the presence of internal subdivisions, often on the basis of the observed correlations, without assuming any particular migration model, which should be treated in the section on spatial genetics (**Chapter 28**). However, we sometimes have to assume that subdivisions are independent (implying no migration between demes).

Owing to space limitation, this chapter is not a comprehensive review of the huge literature on population subdivision. Additional and sometimes more exhaustive reviews can be found elsewhere (e.g. Jorde, 1980; Weir and Cockerham, 1984; Nei, 1986; Chakraborty and Danker-Hopfe, 1991; Weir and Hill, 2002; Rousset, 2004). Instead, we focus on introducing readers to the basic concepts and measures of population subdivision, trying to preserve the spirit in which they have been initially presented. We show the foundations and the relationships between different approaches, as well as their potential limitations. While the treatment presented here assumes that populations are monoecious, the results apply quite well to species with separate sexes, and interested readers could consult Wang (1997b) for a specific treatment of dioecious species.

29.2   THE FIXATION INDEX *F*

The fact that a population is subdivided means in genetic terms that individuals do not mate at random, and that the population is thus not panmictic. This departure from panmixia translates into various levels of apparent inbreeding when considering the total population. The concept of fixation index has been introduced by Sewall Wright to describe the effect of population structure on the amount of inbreeding at a given level of subdivision. Departure from panmixia indeed creates a correlation between homologous genes in uniting gametes relative to a pair of genes taken at random from the population. Wright (1921; 1922) proposed to use this correlation to describe the genetic structure and to quantify the effect of inbreeding.

As we shall see, the fixation index F is precisely a measure of that correlation. We shall briefly see how this correlation arises in the simple case of a single locus with two alleles A and a of respective and unknown frequencies p and $1 - p$ in a total population subdivided into arbitrary breeding units. Gametes unite to form diploid individuals with genotype proportions as shown in Table 29.1, where H_o denotes observed proportion of heterozygotes.

Let us define an indicator variable y, such that, in an elementary experiment consisting in drawing a gamete at random from the population, $y = 1$ if a gamete is of allelic type A and 0 otherwise. We see immediately that over replicates of this sampling process, $E(y) = p.1 + (1 - p).0 = p$, $E(y^2) = p$, and $\mathrm{var}(y) = E(y^2) - E(y)^2 = p(1 - p)$. The correlation ρ_{12} between two uniting gametes G_1 and G_2 can be written as

$$\rho_{12} = \frac{\mathrm{cov}(y_1, y_2)}{\sqrt{\mathrm{var}(y_1)\mathrm{var}(y_2)}}. \tag{29.1}$$

From Table 29.1, it is clear that $\mathrm{cov}(y_1, y_2) = E(y_1 y_2) - E(y_1)E(y_2) = p - H_o/2 - p^2 = p(1 - p) - H_o/2$, leading to

$$\rho_{12} = 1 - \frac{H_o}{2p(1 - p)}. \tag{29.2}$$

Note that the quantity $2p(1 - p)$ is the expected heterozygosity H_e under Hardy–Weinberg equilibrium (HWE) (panmixia), and therefore,

$$\rho_{12} = 1 - \frac{H_o}{H_e}. \tag{29.3}$$

Table 29.1   Pattern of union of gametes in the total population. H_o and p are the observed proportion of heterozygotes and the frequency of allele A in the total population respectively. The observed frequencies of the homozygotes depend only on these two random variables.

		Gamete 2		
		A	*a*	Total
Gamete 1	*A*	$p - H_o/2$	$H_o/2$	p
	a	$H_o/2$	$1 - p - H_o/2$	$1 - p$
	Total	p	$1 - p$	1

For a (total) polymorphic population at an arbitrary time considered to be the beginning of an evolutionary process, H_o should be equal to H_e in the initial generation if there is complete panmixia, and therefore $\rho_{12} = 0$. If after an arbitrary period of genetic drift, there is complete fixation of one or the other allele in every breeding unit (whatever it is), then $H_o = 0$ and $\rho_{12} = 1$. Therefore, Wright proposed to call this correlation of uniting gametes the 'fixation index' F because it is a relative measure of where the demes are on their course to fixation due to random genetic drift. Note that this index was first introduced in studies of inbreeding (Wright, 1921) under the notation f. Later, Wright (1922) proposed to use f in a narrower sense as a 'coefficient of inbreeding', as it is still known until today.

On the course to fixation, p and H_o are random variables that change at each generation because of genetic drift and other potential evolutionary forces. Thus, in practice, both H_o and H_e fluctuate at each generation in the total population and F, defined as a function of observed and expected heterozygosities as in (29.3), is also a random variable with a potentially complex distribution. However, let us consider the simple case where the number of demes is infinite and where gene frequencies within subdivisions only change because of genetic drift. In that case, p and therefore H_e remain constant even though the size of each deme, say N, is finite. The frequency of heterozygotes H_o in any deme should decrease *on average* by a factor $[1 - 1/(2N)]$ each generation, such that

$$F^{t+1} = 1 - \frac{H_o^{t+1}}{H_e} = 1 - \frac{\left(1 - \dfrac{1}{2N}\right) H_o^t}{H_e}.$$

Therefore, the fixation index F should approximately change over time as

$$F^{t+1} = 1 - \frac{H_o^0}{H_e}\left(1 - \frac{1}{2N}\right)^t = 1 - (1 - F^0)\left(1 - \frac{1}{2N}\right)^t, \qquad (29.4)$$

to asymptotically reach the value of 1.

Note that, in full generality, the observed proportion of heterozygotes H_o can be defined in terms of allele frequencies and fixation index by rearranging (29.2) as

$$H_o = 2p(1 - p)(1 - F). \qquad (29.5)$$

This relationship holds for any level of subdivision, as is shown in the next section.

29.3 WRIGHT'S F STATISTICS IN HIERARCHIC SUBDIVISIONS

Suppose now that the total diploid population is subdivided into a finite number of demes d. Since the derivation of the fixation index F in the previous section does not require that the gametes are uniting to produce zygotes, it can apply to any pair of gametes, and in fact to any pair of genes at a given locus. F can therefore be seen as a correlation between homologous genes taken from a given subdivision level (individuals, demes, or any other higher level) relative to any higher subdivision level (Wright, 1943; 1951; 1965; 1969). The correlation between genes within individuals (I) relative to the genes of the total population (T) is represented by F_{IT}. The correlation between genes within

a deme or subdivision (S) relative to the genes of the total population is represented by F_{ST}. Finally, the correlation between genes within individuals relative to those within a subdivision is Wright's local inbreeding coefficient f that is represented here by F_{IS}. The list of F statistics can be easily extended to further levels of subdivisions.

We first consider the correlation of genes within demes F_{ST}. In that case, (29.5) still holds, but H_o is not the observed heterozygosity anymore. Here, the probability is that two random genes *within a subdivision* are different, without need for these two genes to be on the homologous chromosomes of an individual. If we call this probability H_S, then we have

$$H_S = 2p(1 - p)(1 - F_{ST}). \tag{29.6}$$

Note that H_S can also be defined as an average probability over all d subdivisions as

$$H_S = \frac{1}{d} \sum_k^d H_{Sk}, \tag{29.7}$$

where H_{Sk} is the probability that two genes randomly chosen from the kth deme are different, which is equivalent to the expected heterozygosity in the kth deme under HWE and $H_{Sk} = 2p_k(1 - p_k)$. It should be clear however that since these two genes do not necessarily come from the same individual, the assumption of HWE does not need to be made here. Therefore,

$$H_S = \frac{1}{d} \sum_k^d 2p_k(1 - p_k) = 2\left(\bar{p} - \frac{1}{d} \sum_k^d p_k^2\right) = 2(\bar{p} - \overline{p^2})$$

$$H_S = 2[\bar{p} - \text{var}_S(p) - \overline{p}^2] = 2\bar{p}(1 - \bar{p}) - 2\text{var}_S(p), \tag{29.8}$$

where the S subscript stresses the fact that the variance of p is computed over subdivisions. Note that this equation also describes the 'Wahlund effect' (Wahlund, 1928), which represents the deficit of observed heterozygotes in a subdivided population made up of random mating demes. Rearranging (29.8) with (29.6) leads to the classical relationship

$$F_{ST} = \frac{\text{var}_S(p)}{\bar{p}(1 - \bar{p})}. \tag{29.9}$$

Therefore, the fixation index F_{ST} can also be defined as the standardized variance of allele frequencies. It is the ratio of the observed variance of gene frequencies over subdivisions divided by the maximum possible variance $\bar{p}(1 - \bar{p})$ that can be reached when alleles have gone to fixation in all subdivisions, that is when the allele frequencies within each deme are as divergent as they can be from the average population frequency. Here again, the fixation index F_{ST} is a relative measure of where the subdivisions are in the process of allelic fixation. Note that F_{ST} has not been derived by assuming random mating within demes, and therefore it should not be affected by local inbreeding. Its definition only depends on local allele frequencies. However, if there is random mating within demes, the correlation of pairs of genes within or between individuals is equivalent and $F_{ST} = F_{IT}$.

Let us now consider subdivisions where mating is not at random, so that $F_{IT} \neq F_{ST}$, and

$$H_o = 2\bar{p}(1 - \bar{p})(1 - F_{IT}). \tag{29.10}$$

The total proportion of heterozygotes H_o can also be expressed as an average over subdivisions, as for H_S in (29.7), but allowing for local inbreeding within subpopulations it becomes

$$H_o = \frac{1}{d} \sum_{k=1}^{d} 2p_k(1 - p_k)(1 - F_{ISk}).$$

Assuming that the level of local inbreeding is independent of the allele frequencies, which has been rightly criticized (see e.g. Barrai, 1971), Wright (1951; 1965) obtains

$$H_o = 2(1 - \bar{F}_{IS})(\bar{p} - \frac{1}{d} \sum_{k}^{d} p_k^2) = 2(1 - \bar{F}_{IS})(\bar{p} - \text{var}_S(p) - \bar{p}^2), \quad (29.11)$$

where $\bar{F}_{IS}$ is the unweighted average F_{IS} value over subdivisions. From (29.10), we have

$$\text{var}_S(p) = \frac{\bar{p}(1 - \bar{p})(F_{IT} - \bar{F}_{IS})}{(1 - \bar{F}_{IS})}. \quad (29.12)$$

Combining (29.12) with (29.9) leads to the classical relationship between fixation indexes as

$$(1 - F_{IT}) = (1 - \bar{F}_{IS})(1 - F_{ST}). \quad (29.13)$$

The bar over $\bar{F}_{IS}$ is often omitted in the literature, but we keep it to remind us that it is in fact the average inbreeding value over subpopulations. Note that F_{ST} has no meaning if the population is not subdivided and made up of only one deme. Note also that (29.11) and therefore (29.13) are valid even when local inbreeding depends on gene frequencies provided that $\bar{F}_{IS}$ is defined as a weighted average of the form (Wright, 1969, vol. 4, chap. 3)

$$\bar{F}_{IS_w} = \frac{\sum_{k}^{d} w_k p_k(1 - p_k)F_{ISk}}{\sum_{k}^{d} w_k p_k(1 - p_k)}, \quad (29.14)$$

where w_k is the weight given to subpopulation k. A common choice is to set the weights equal to the relative deme effective size, with $\sum_k w_k = 1$. Alternatively, if these sizes are unknown, one usually assumes that all w_k,s are equal to $1/d$, which has the effect of giving more weight to subpopulations where allele A is of intermediate frequency.

29.3.1 Multiple Alleles

The extension of the above derivations to multiple alleles is straightforward (Nei, 1977). Consider the case where we have r alleles at a given locus. Within each subpopulation, $r(r - 1)/2$ inbreeding coefficients are needed for a complete specification of the genotype frequencies $P_{kij}(i = 1 \ldots r, j = 1 \ldots i)$, as

$$P_{kij} = 2p_{ki} p_{kj}(1 - F_{IS_{kij}}),$$
$$P_{kii} = p_{ki}^2(1 - F_{IS_{kii}}) + F_{IS_{kii}} p_{ki}. \quad (29.15)$$

If we only consider the r homozygote genotypes (see Nagylaki, 1998b, for a complete treatment of all genotypes), a global inbreeding coefficient can be obtained as a weighted average of the form

$$F_{IS_k} = \frac{\sum_{i=1}^{r} p_{ki}(1-p_{ki})F_{IS_{kii}}}{\sum_{i=1}^{r} p_{ki}(1-p_{ki})},$$

which becomes

$$F_{IS_k} = \frac{\sum_{i=1}^{r}(P_{kii}-p_{ki}^2)}{\sum_{i=1}^{r} p_{ki}(1-p_{ki})} = \frac{H_{Sk}-H_{ok}}{H_{Sk}},$$

where $H_{Sk} = 1 - \sum_i p_{ki}^2$ and $H_{ok} = 1 - \sum_i P_{kii}$ are the expected and the observed heterozygosities in subpopulation k respectively. Similar to (29.14), we can define an average inbreeding coefficient over all subpopulations as

$$\bar{F}_{IS} = \frac{\sum_{k}^{d} w_k H_{Sk} F_{ISk}}{\sum_{k}^{d} w_k H_{Sk}} = \frac{H_S - H_o}{H_S}, \tag{29.16}$$

where $H_S = 1 - \sum_k \sum_i w_k p_{ki}^2$ and $H_o = 1 - \sum_k \sum_i w_k P_{kii}$ are respectively the global expected and observed heterozygosities. Similarly, the other fixation indexes can be obtained as ratios of heterozygosities

$$F_{IT} = \frac{H_T - H_o}{H_T}, \tag{29.17}$$

$$F_{ST} = \frac{H_T - H_S}{H_T}, \tag{29.18}$$

where $H_T = 1 - \sum_i \bar{p}_i^2$ and $\bar{p}_i = \sum_k w_k p_{ki}$.

With these definitions, fixation indexes can be viewed as ratios of expected or observed heterozygosities (Nei, 1977), which allows arbitrary numbers of alleles at a locus. This simplicity, however, hides the fact that subpopulations have to be given weights.

29.3.2 Sample Estimation of F Statistics

Up to this point, F statistics have been defined from population allele and genotype frequencies (which are random variables if the number of demes and deme sizes are finite or if there are other random evolutionary forces acting within demes). Some corrections need to be performed when allele frequencies are estimated from population samples. An estimation of these fixation indexes has been developed in Nei and Chesser (1983). The estimators differ from the above definitions by replacing expected and observed population heterozygosities (H_o, H_S, and H_T) by unbiased sample estimates ($\hat{H}_o$, $\hat{H}_S$,

and $\hat{H}_T$), taking into account sample sizes as (Nei and Chesser, 1983)

$$\hat{H}_o = 1 - \sum_{k=1}^{d} \sum_{i=1}^{r} w_k \hat{P}_{kii}$$

$$\hat{H}_S = \frac{\tilde{n}}{\tilde{n} - 1} \left(1 - \sum_{i=1}^{d} \sum_{k=1}^{r} w_k \hat{p}_{ki}^2 - \frac{\hat{H}_o}{2\tilde{n}} \right),$$

$$\hat{H}_T = 1 - \sum_{i=1}^{r} \left(\sum_{k=1}^{d} w_k \hat{p}_{ki} \right)^2 + \frac{\hat{H}_S}{\tilde{n}d} - \frac{\hat{H}_o}{2\tilde{n}d} \tag{29.19}$$

with $\tilde{n}$ being the harmonic mean of the sample sizes, and d the number of sampled demes. Note that the estimates of $\hat{H}_S$ and $\hat{H}_T$ are only unbiased if one assumes that there is no correlation between sample size and observed or expected heterozygosity levels (Nei and Chesser, 1983, p. 255). Sample estimators of F_{IS}, F_{ST}, and F_{IT} are obtained by substituting (29.19) in (29.16), (29.17), and (29.18) respectively.

However, the above estimation procedure only takes into account the sampling process within demes. It does not consider the sampling of demes within the population and therefore assumes that all demes have been sampled. Pons and Chaouche (1995) have developed estimators of F statistics that remove this assumption by extending the approach of Nei and Chesser (1983) to an additional level of sampling. They considered that d demes (supposed to be independent from each other) had been sampled and obtained

$$\hat{H}_o^* = \frac{1}{d} \sum_{i=1}^{d} \hat{H}_{ok} = 1 - \sum_{k=1}^{d} \sum_{i=1}^{r} \hat{P}_{kii}$$

$$\hat{H}_S^* = \frac{1}{d} \sum_{k=1}^{d} \frac{n_k}{n_k - 1} \left(1 - \sum_{i=1}^{r} \hat{p}_{ki}^2 - \frac{\hat{H}_{ok}}{2n_k} \right),$$

$$\hat{H}_T^* = \frac{1}{d(d-1)} \sum_{k=1}^{d} \sum_{k' \neq k}^{d} \left(1 - \sum_{i=1}^{r} \hat{p}_{ki} \hat{p}_{k'i} \right) \tag{29.20}$$

where n_k is the size of the kth sample. Again, sample estimators of F_{IS}, F_{ST}, and F_{IT} are obtained by substituting (29.20) into (29.16), (29.17), and (29.18) respectively. $\hat{H}_o^*$ is identical to $\hat{H}_o$ defined in (29.19), except that Pons and Chaouche (1995) have preferred to weight demes equally. The difference between $\hat{H}_S^*$ and $\hat{H}_S$ is in the weighting scheme of the sample sizes, but the main difference between the two above approaches is that between $\hat{H}_T^*$ and $\hat{H}_T$, where we note that $\hat{H}_T^*$ only depends on the comparison of allele frequencies between samples.

29.3.3 *G* Statistics

G_{ST} has been defined by Nei (1973) as an analog of F_{ST} by extending Nei's genetic distance between a pair of populations (Nei, 1972) to the case of a hierarchical structure of populations. He showed that gene frequency variation at loci with multiple alleles could be analyzed directly in terms of gene diversity, which is defined as the probability

that two randomly chosen genes from a population are different (Nei, 1973). For diploid organisms with random mating, this probability only depends on allele frequencies and should be equal to the average proportion of heterozygote individuals. However, there are many organisms or genetic systems where the concept of heterozygosity does not hold, but where gene diversity would always make sense (e.g. with mitochondrial DNA, haploid, or haplo-diploid organisms). By analogy with Wright's F_{ST}, Nei defined the coefficient of gene differentiation

$$G_{ST} = \frac{D_{ST}}{H_T},\tag{29.21}$$

expressing the magnitude of gene differentiation among subpopulations, where D_{ST} is the average gene diversity between subpopulations, including comparisons of subpopulations with themselves. However, for diploid random mating populations $D_{ST} = H_T - H_S$, and G_{ST} is identical to F_{ST} defined in (29.18) as function of heterozygosities.

Some confusion still prevails in the literature about the differences between F statistics and G statistics. Technically, G statistics are only function of allele frequencies (in gene diversities) and do not incorporate information on observed proportions of heterozygotes, unlike F statistics. They can thus only be analogs of F statistics that do not depend on population genotype frequencies, such as F_{ST}, or correlations of gene frequencies at higher levels of subdivisions.

The concept of gene identity is explicit in the definition of G statistics, where it has been introduced by Crow and Aoki (1984), but it is also implicit in the definition of F statistics as correlation coefficients. F statistics defined in (29.16–29.18) can also be defined in terms of identity coefficients (see **Chapter 28**) as

$$\bar{F}_{IS} \equiv \frac{Q_I - Q_S}{1 - Q_S}; \quad F_{IT} \equiv \frac{Q_I - Q_T}{1 - Q_T}; \quad F_{ST} \equiv \frac{Q_S - Q_T}{1 - Q_T},\tag{29.22}$$

where Q_I is the probability that two genes within individuals are identical, whereas Q_S and Q_T are the probabilities that two genes within subpopulations and within the total population are identical, respectively. Of course, the values taken by these probabilities will depend on a particular spatial and mutation model (see e.g. Crow and Aoki, 1984; Cockerham and Weir, 1993; Rousset, 2004).

Estimators of G statistics are obtained by estimating gene diversities similarly to heterozygosities in (29.19) and (29.20), except that they only depend on allele frequencies (Nei, 1987, p. 191; Pons and Chaouche, 1995).

29.4   ANALYSIS OF GENETIC SUBDIVISION UNDER AN ANALYSIS OF VARIANCE FRAMEWORK

Cockerham (1969; 1973) has shown how one could decompose the total variance of gene frequencies into variance components associated to different subdivision levels within the framework of an ANOVA of gene frequencies. He also showed that variance components, identity by descent measures (in absence of mutation and migration), and F statistics were just different ways to express correlations of genes (see Table 29.3). We present below his basic derivations in some detail, as it has become a method of choice for the analysis

of population subdivision. We shall use Cockerham's notations, but we will mention how to relate them to F statistics, as defined previously.

Cockerham's ANOVA framework preserves Wright's definition of F statistics in terms of correlation of gene frequencies. The notion of demes used in the ANOVA framework is close to the notion of population in the statistical sense, as demes are supposed to be independent replicates of the same evolutionary process (drawn from the same distribution). Here the unknown distribution is a stochastic evolutionary process essentially shaped by random genetic drift of gene frequencies between generations. The mutation process will be introduced later.

29.4.1 The Model

We assume that several demes of finite sizes have diverged simultaneously from an ancestral population that was at HWE. Since that time, the demes have remained separate and have been exposed to the same conditions. Samples from different demes are thus expected to differ from each other because of the sampling process of individuals within each deme (statistical sampling), and because of the stochasticity in the evolutionary process between generations (genetic sampling). We will consider a hierarchically structured diploid population, with genes in individuals, individuals in demes, and demes in the population.

Let x_{kij} be the jth allele ($j = 1, 2$) in the ith individual in the kth deme, and y_{kij} be the indicator variable for any given gene: y_{kij} is equal to 1 if x_{kij} is of type A and 0 otherwise. If p is the population frequency of the allele A, then $E(y_{kij}) = p$, $E(y_{kij}^2) = p$, and thus $\text{var}(y_{kij}) = p(1 - p)$. Cockerham (1969; 1973) considered a linear model of the form

$$y_{kij} = p + a_k + b_{ki} + w_{kij}. \tag{29.23}$$

The variable y_{kij} is assumed to differ from the mean p because of the additive, random, and uncorrelated effects due to demes (a), individuals within demes (b) and genes within individual (w).

We are interested in defining the relations (correlations, covariances) between genes found in different levels of the hierarchy. Formally, the expectations of the products of gene frequencies from different subdivision levels are

$$
\begin{aligned}
E(y_{kij} y_{k'i'j'}) &= \text{cov}(y_{kij}, y_{k'i'j'}) + E(y_{kij})E(y_{k'i'j'}) \\
&= \sigma_T^2 + p^2 && \text{if } k = k', i = i', j = j' \\
&= \text{cov}_{a \supset b} + p^2 && \text{if } k = k', i = i', j \neq j'. \\
&= \text{cov}_a + p^2 && \text{if } k = k', i \neq i' \\
&= \text{cov}_g + p^2 && \text{if } k \neq k' \tag{29.24}
\end{aligned}
$$

We have assumed that the demes are independent, implying that $\text{cov}_g = 0$, but this simplification could be removed, for instance, to accommodate migration (see e.g. Cockerham and Weir, 1987). The ANOVA layout corresponding to this hierarchy can be found in Table 29.2.

The covariance of gene frequencies within individuals within demes, $\text{cov}_{a \supset b}$, depends on the probability of genes being identical by descent $F_i = \text{Pr}(x_{ki1} \equiv x_{ki2})$, a notion

Table 29.2 Analysis of variance layout. The total sums of square (SS(T)) is decomposed in a standard fashion into sums of square due to different sources of variation. The expected mean squares are expressed here as functions of variance components. Comparisons with (29.27), (29.30), and (29.34) show that these variance components are a parameterization of the model equivalent to the correlations $\bar{\theta}$ and $\bar{F}$.

Source of variation	d.f.	Sums of squares*	Mean squares[†]	Expected mean squares
Among demes (AD)	$d-1$	$SS(AD) = \sum\limits_{k}^{d} 2n_k(y_{k..} - y_{...})^2$	$\dfrac{SS(AD)}{d-1}$	$\sigma_w^2 + 2\sigma_b^2 + 2n'\sigma_a^2$
Among individuals within demes (WD)	$n-d$	$SS(WD) = \sum\limits_{k}^{d} \sum\limits_{i}^{n_k} 2(y_{ki.} - y_{k..})^2$	$\dfrac{SS(WD)}{n-d}$	$\sigma_w^2 + 2\sigma_b^2$
Between genes within individuals (WI)	n	$SS(WI) = \sum\limits_{k}^{d} \sum\limits_{i}^{n_k} \sum\limits_{j}^{2} (y_{kij} - y_{ki.})^2$	$\dfrac{SS(WI)}{n}$	σ_w^2
Total	$2n-1$	$SS(T) = \sum\limits_{k}^{d} \sum\limits_{i}^{n_k} \sum\limits_{j}^{2} (y_{kij} - y_{...})^2$		

$$n = \sum_{k}^{d} n_k, \quad n' = \frac{n - \sum\limits_{k}^{d} \dfrac{n_k^2}{n}}{d-1},$$

* We use the conventional dot notation to specify different means of y (e.g. $y_{ki.} = \sum_j y_{kij}$).

[†] The mean squares among demes MS(AD), within demes MS(WD), and within individuals MS(WI) are obtained as usual by dividing the corresponding sum of squares by the appropriate degrees of freedom.

introduced by Malécot (1948), which is equivalent to Wright's total inbreeding coefficient F_{IT} here. The two alleles of an individual can be of type A because they are identical by descent with probability F_i, or not identical by descent with probability $(1 - F_i)p$. Therefore, $\Pr(x_{ki1} \equiv A, \; x_{ki2} \equiv A) = E(y_{ki1}y_{ki2}) = \Pr(x_{i2} \equiv A | x_{i1} \equiv A)$. $\Pr(x_{i1} \equiv A) = [F_i + (1 - F_i)p]p$. It follows that $\text{cov}(y_{ki1}, y_{ki2}) = F_i p(1-p)$. Taking the average F_i over all individuals, $\text{cov}_{a \supset b} = \bar{F}p(1-p)$, which shows that Malécot's identity by descent measure is also equivalent to the correlation of gene frequencies within individuals.

The covariance of gene frequencies between individuals depends on another identity measure called the *coancestry coefficient* $\theta_{ii'}$ by Cockerham, which is the probability that two random genes from different individuals, i and i', are identical by descent. Note that this coancestry coefficient is indeed equivalent for most practical purposes to Malécot (1948) *coefficient of kinship* $\Phi_{ii'}$. We now have $\Pr(x_{kij} = A, x_{ki'j'} = A) = p[\theta_{ii'} + (1 - \theta_{ii'})p]$ and $\text{cov}(y_{kij}, y_{ki'j'}) = \theta_{ii'}p(1-p)$, and again taking the average over all pairs of individuals within demes, $\text{cov}_a = \bar{\theta}p(1-p)$ showing that $\bar{\theta}$ is also the average correlation between two genes from two *different* individuals within the same deme. It is worth stressing that $\bar{\theta}$ can only be estimated if we have samples from several independent demes, because this correlation is relative to the correlation of the least related individuals in the population, which are assumed to be in different demes and therefore independent. In that sense, it differs from the coefficient of kinship, which is supposed to be an absolute measure of relatedness between individuals. In other words, the relatedness

Table 29.3 Relationship between intraclass correlations, variance components, Cockerham's gene correlations, and Wright's fixation indexes.

Intraclass correlations	Expressed as covariances	Expressed as variance components	Cockerham gene correlations	Wright's fixation indexes
ρ_a	$\dfrac{\mathrm{cov}_a}{p(1-p)}$	$\dfrac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma_w^2}$	$\bar{\theta}$	F_{ST}
$\rho_{a \supset b}$	$\dfrac{\mathrm{cov}_{a \supset b}}{p(1-p)}$	$\dfrac{\sigma_a^2 + \sigma_b^2}{\sigma_a^2 + \sigma_b^2 + \sigma_w^2}$	$\bar{F}$	F_{IT}
ρ_b	$\dfrac{\mathrm{cov}_b}{p(1-p)}$	$\dfrac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$	$\bar{f} = \dfrac{\bar{\theta} - \bar{F}}{1 - \bar{F}}$	$\bar{F}_{IS}$
ρ_w	$\dfrac{\mathrm{cov}_w}{p(1-p)}$	$\dfrac{\sigma_w^2}{\sigma_w^2} = 1$		

between individuals within demes cannot be measured without a reference point, which in this case is built by comparing individuals between demes. Note that for F or G statistics estimated from heterozygosities or gene diversities, the reference point is the comparison of a random pair of genes, either within or between demes, as is made clearer below. See also Weir (1996, see Chapter 5) for further discussions on this distinction. Because it is the variability of the genetic process among demes that introduces a correlation of genes between individuals within demes, a measure of the relatedness between individuals within demes is also a measure of the degree of differentiation between demes.

In Table 29.3, we have expressed each intraclass correlation ρ as a ratio of variance components, and made explicit the relationship between Wright's fixation indexes and Cockerham's correlations. We have also introduced the correlation $\bar{f}$, which is an average correlation of genes within individuals, like $\bar{F}$, but it is relative to the least related pair of genes within the same deme, instead of being relative to the least related pair of genes between demes. It is a composite measure that depends on the two parameters of the model $\bar{\theta}$ and $\bar{F} \cdot \bar{f}$ is the analog of Wright's $\bar{F}_{IS}$, the inbreeding coefficient within demes.

From Table 29.2, (29.27), (29.30), and (29.34) and Table 29.3, we see that

$$
\begin{aligned}
\sigma_a^2 &= p(1-p)\rho_a & &= \mathrm{cov}_a & &= p(1-p)\bar{\theta} \\
\sigma_b^2 &= p(1-p)(\rho_{a \supset b} - \rho_a) & &= \mathrm{cov}_{a \supset b} - \mathrm{cov}_a & &= p(1-p)(\bar{F} - \bar{\theta}). \\
\sigma_w^2 &= p(1-p)(1 - \rho_{a \supset b}) & &= \sigma_T^2 - \mathrm{cov}_{a \supset b} & &= p(1-p)(1 - \bar{F}) \quad (29.25)
\end{aligned}
$$

Thus, despite the conventional σ^2 notation, these variance components are indeed functions of gene covariances and can therefore take negative values. As explained by Rousset (2000), these variance components do not represent a decomposition of the total variance into a sum of variances as in conventional ANOVA.

29.4.2 Estimation Procedure

29.4.2.1 Expected Mean Squares

The correlation of genes at different levels of subdivision are estimated by the method of moments by equating the mean squares expressed in terms of allele and genotype frequencies to their expectation in terms of correlation parameters. The expected mean

squares are derived below in terms of the parameters of the model. From Table 29.2, we
have

$$E[SS(WI)] = E\left[\sum_k^d \sum_i^{n_k} \sum_j^2 (y_{kij} - y_{ki.})^2\right]$$

$$= \sum_k^d \sum_i^{n_k} \sum_j^2 \left[E(y_{kij})^2 + E(y_{ki.})^2 - 2E(y_{kij}y_{ki.})\right]$$

$$= 2n\left\{[p^2 + p(1-p)] + \left[p^2 + \frac{1}{2}p(1-p)(1+\bar{F})\right]\right.$$

$$\left. - 2\left[p^2 + \frac{1}{2}p(1-p)(1+\bar{F})\right]\right\}$$

$$= np(1-p)(1-\bar{F}) \tag{29.26}$$

and therefore

$$E[MS(WI)] = p(1-p)(1-\bar{F}). \tag{29.27}$$

The expectation of the sum of squares between individuals within demes is computed as

$$E[SS(WD)] = 2E\left[\sum_k^d \sum_i^{n_k} (y_{ki.} - y_{k..})^2\right]$$

$$= 2\sum_k^d \sum_i^{n_k} E(y_{ki.}^2) + E(y_{k..}^2) - 2E(y_{ki.}y_{k..}).$$

In order to derive this quantity, we need to compute the variance of gene frequencies
within deme $\mathrm{var}(y_{k..})$, which is obtained as

$$\mathrm{var}(y_{k..}) = \mathrm{var}\left(\frac{1}{2n}\sum_{i=1}^{n_k}\sum_{j=1}^2 y_{kij}\right)$$

$$= \frac{1}{4n_k^2}\left[\sum_{i=1}^{n_k}\sum_{j=1}^2 \mathrm{var}(y_{kij}) + 2\sum_{i=1}^{n_k} \mathrm{cov}(y_{ki1}, y_{ki2})\right.$$

$$\left. + 2\sum_{i=1}^{n_k}\sum_{i'<i}^{n_k}\sum_{j=1}^2\sum_{j'=1}^2 \mathrm{cov}(y_{kij}, y_{ki'j'})\right]$$

$$= \frac{1}{4n_k^2}\left[2n_k(p(1-p) + 2\sum_{i=1}^{n_k} F_i p(1-p) + 2\sum_{i=1}^{n_k}\sum_{i'<i}^{n_k} 4\theta_{ii'} p(1-p)\right]$$

$$= \frac{1}{4n_k^2}\left[2n_k p(1-p) + 2p(1-p)n_k\bar{F} + 4n_k(n_k-1)\bar{\theta}p(1-p)\right]$$

$$\mathrm{var}(y_{k..}) = \frac{p(1-p)}{2n_k}[1 + \bar{F} + 2(n_k-1)\bar{\theta}], \tag{29.28}$$

where $\bar{\theta}$ is the average correlation of genes between individuals (the $\theta_{ii'}$) or the average kinship coefficient between the individuals of the sample. The derivation of (29.28) is instructive, as it shows that the variance of gene frequencies within demes depends on the breeding structure of the population ($\bar{F}$) and the relatedness of the individuals within demes ($\bar{\theta}$). If the sample was noninbred and made up of unrelated individuals, the genetic variance would be removed and the variance would simply be equal to the binomial variance $p(1-p)/(2n_k)$.

The expected sum of squares within demes can now be computed as

$$E[SS(WD)] = 2 \sum_k^d \sum_i^{n_k} \left\{ \left[p^2 + \frac{p(1-p)}{2}(1+\bar{F}) \right] \right.$$

$$+ \left[p^2 + \frac{p(1-p)}{2n_k}(1+\bar{F}+(n_k-1)\bar{\theta} \right]$$

$$\left. -2 \left[p^2 + \frac{p(1-p)}{2n_k}(1+\bar{F}+(n_k-1)\bar{\theta} \right] \right\}$$

$$= (n-d)p(1-p)(1+\bar{F}-2\bar{\theta}), \tag{29.29}$$

and therefore

$$E[MS(WD)] = p(1-p)(1-\bar{F}+2(\bar{\theta}-\bar{F})). \tag{29.30}$$

Finally, the expectation of the sum of squares between demes is obtained as

$$E[SS(AD)] = E\left[\sum_k^d 2n_k(y_{k..}-y_{...})^2 \right] = \sum_k^d 2n_k[E(y_{k..}^2) + E(y_{...}^2) - 2E(y_{k..}y_{...})]. \tag{29.31}$$

In order to derive this quantity, we first need to get the total variance of gene frequencies:

$$\mathrm{var}(y_{...}) = \mathrm{var}\left(\frac{1}{2n} \sum_k^d \sum_i^{n_k} \sum_1^2 y_{kij} \right) = \frac{1}{4n^2} \left[\sum_k^d \mathrm{var}(2n_k y_{k..}) \right.$$

$$\left. +2 \sum_{k'<k}^d 16 n_k^2 n_{k'}^2 \mathrm{cov}(y_{k..}y_{k'..}) \right].$$

Because we have assumed that the demes were independent from each other, the covariance terms can be ignored, and the total variance reduces to

$$\mathrm{var}(y_{...}) = \left[\sum_k^d \frac{n_k^2}{n^2} \mathrm{var}(y_{k..}) \right] = \frac{p(1-p)}{2n} \left[1 + \bar{F} + 2\bar{\theta}\left(\frac{\sum_k^d n_k^2}{n} - 1 \right) \right]. \tag{29.32}$$

Then, the expectation of the product of the total mean $(y_{..})$ and the within population mean $(y_{k..})$ is obtained as

$$E(y_{k..}y_{...}) = E\left[y_{k..}\left(\frac{1}{2n}\sum_{k'=1}^{d}2n_k y_{k'..}\right)\right] = \frac{n_k}{n}\sum_{k'=1}^{d}E[y_{k'..}y_{k..}]$$

$$= \frac{n_k}{n}\left[E(y_{k..}^2) + \sum_{k'\neq k}^{d}E(y_{k'..}y_{k..})\right]$$

$$= \frac{n_k}{n}\left\{p^2 + \frac{p(1-p)}{2n_k}[1 + \bar{F} + 2(n_k-1)\bar{\theta}] + (d-1)p^2 + \sum_{k'\neq k}^{d}\text{cov}(y_{k'..}, y_{k..})\right\}$$

$$= p^2 + \frac{p(1-p)}{2n}[1 + \bar{F} + 2(n_k-1)\bar{\theta}]. \tag{29.33}$$

After some algebra, we obtain the expected mean squares between populations as

$$E[MS(AD)] = \frac{E[SS(AD)]}{d-1} = p(1-p)\left[1 - F + 2(\theta - F) + \frac{2\theta}{d-1}\left(n - \frac{\sum_{k}^{d}n_k^2}{n}\right)\right],$$
$$\tag{29.34}$$

which explains the weighting scheme of the variance components in Table 29.2.

29.4.2.2 Moment Estimators

Unbiased estimators of the variance components are obtained by the method of moments, equating the observed mean squares to their expectations in Table 29.2, and rearranging as

$$\hat{\sigma}_w^2 \overset{\triangle}{=} MS(WI)$$

$$\hat{\sigma}_b^2 \overset{\triangle}{=} \frac{1}{2}[MS(WD) - MS(WI)],$$

$$\hat{\sigma}_a^2 \overset{\triangle}{=} \frac{1}{2n'}[MS(AD) - MS(WD)] \tag{29.35}$$

from which the estimators of gene correlations are defined as

$$\hat{\bar{F}} = \frac{\hat{\sigma}_a^2 + \hat{\sigma}_b^2}{\hat{\sigma}_a^2 + \hat{\sigma}_b^2 + \hat{\sigma}_w^2}$$

$$\hat{\bar{\theta}} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_b^2 + \hat{\sigma}_w^2}.$$

$$\hat{\bar{f}} = \frac{\hat{\bar{F}} - \hat{\bar{\theta}}}{1 - \hat{\bar{\theta}}} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_w^2} \tag{29.36}$$

Note that these estimators do not depend on the unknown gene frequency p, and that they are not necessarily unbiased, as they are just ratios of unbiased quantities.

29.4.2.3 Weighted Averages for Several Alleles and Several Loci

The above derivations have been made for a single allele at a given locus. Similar computations could be made for each allele of a multiallelic (say r alleles) polymorphism, giving rise to $r-1$ independent estimates of the same parameters $\bar{f}$, $\bar{F}$, and $\bar{\theta}$. Noting that estimators of correlations are ratios of variance components, weighted estimators can be defined as

$$\hat{\bar{F}}_w = \frac{\sum\limits_{u=1}^{r} \hat{\sigma}_{b_u}^2 + \sum\limits_{u=1}^{r} \hat{\sigma}_{w_u}^2}{\sum\limits_{u=1}^{r} \hat{\sigma}_{T_u}^2} = \frac{\bar{\hat{\sigma}}_b^2 + \bar{\hat{\sigma}}_w^2}{\bar{\hat{\sigma}}_T^2}$$

$$\hat{\bar{\theta}}_w = \frac{\sum\limits_{u=1}^{r} \hat{\sigma}_{a_u}^2}{\sum\limits_{u=1}^{r} \hat{\sigma}_{T_u}^2} = \frac{\bar{\hat{\sigma}}_a^2}{\bar{\hat{\sigma}}_T^2}.$$

$$\hat{\bar{f}}_w = \frac{\sum\limits_{u=1}^{r} \hat{\sigma}_{b_u}^2}{\sum\limits_{u=1}^{r} \hat{\sigma}_{b_u}^2 + \sum\limits_{u=1}^{r} \hat{\sigma}_{w_u}^2} = \frac{\bar{\hat{\sigma}}_b^2}{\bar{\hat{\sigma}}_b^2 + \bar{\hat{\sigma}}_w^2} \qquad (29.37)$$

From (29.25), it is clear that these estimators are weighted averages of the intraclass correlations ρ, with weight for allele u equal to the total variance $\sigma_{Tu}^2 = p_u(1 - p_u)$. Similarly, if estimators are available from l loci, they can be combined as

$$\hat{\bar{F}}_w = \frac{\sum\limits_{t=1}^{l}\sum\limits_{u=1}^{r_t} \hat{\sigma}_{b_{tu}}^2 + \sum\limits_{t=1}^{l}\sum\limits_{u=1}^{r_t} \hat{\sigma}_{w_{tu}}^2}{\sum\limits_{t=1}^{l}\sum\limits_{u=1}^{r_t} \hat{\sigma}_{T_{tu}}^2}, \hat{\bar{\theta}}_w = \frac{\sum\limits_{t=1}^{l}\sum\limits_{u=1}^{r} \hat{\sigma}_{a_{tu}}^2}{\sum\limits_{t=1}^{l}\sum\limits_{u=1}^{r_t} \hat{\sigma}_{T_{tu}}^2}, \hat{\bar{f}}_w = \frac{\sum\limits_{t=1}^{l}\sum\limits_{u=1}^{r_t} \hat{\sigma}_{b_{tu}}^2}{\sum\limits_{t=1}^{l}\sum\limits_{u=1}^{r_t} \hat{\sigma}_{b_{tu}}^2 + \sum\limits_{t=1}^{l}\sum\limits_{u=1}^{r_t} \hat{\sigma}_{w_{tu}}^2}.$$

Alternative ways of building average estimators such as unweighted or least-squares estimators are discussed in Reynolds *et al.* (1983), while other weighting schemes are discussed in Weir and Cockerham (1984) and in Robertson and Hill (1984). It appears that weighted estimators as in (29.37) have a very low bias and are rather insensitive to unequal sample sizes, even though they may have a slightly larger variance than other combined estimators (Goudet *et al.*, 1996).

29.4.2.4 Likelihood-based Methods

A number of authors have derived likelihood-based estimators of F statistics by specifying a given parametric distribution of the unknown allele frequencies (p_i) in subpopulations (e.g. Rannala and Hartigan, 1996; Holsinger *et al.*, 2002; Nicholson

et al., 2002; Weir and Hill, 2002; Balding, 2003; Kitada and Kishino, 2004; Foll and Gaggiotti, 2006; Kitakado *et al.*, 2006). While some authors have assumed a Normal distribution of these allele frequencies (e.g. Nicholson *et al.*, 2002; Weir and Hill, 2002), most others have chosen to assume that the r allele frequencies in subpopulations followed a β or a Dirichlet distribution (Wright, 1951; Balding, 2003) with parameters $\alpha p_i (i = 1, \ldots, r)$, where p_i is the unknown frequency of the ith allele. An interesting discussion on the validity of a Dirichlet distribution for describing allele frequencies in subpopulations under different evolutionary models or for different markers can be found in Balding (2003). Under this framework, the variance of p_i over populations is given by $\text{var}(p_i) = p_i(1 - p_i)/(\alpha + 1)$, showing that $F_{ST} = 1/(\alpha + 1)$ from (29.9), and that $\alpha = F_{ST}^{-1} - 1$. It follows that the allelic counts (n_i) in each subpopulation follow a multinomial-Dirichlet distribution of the form

$$\Pr(n_i | \alpha, \mathbf{p}) = \frac{n! \Gamma(\alpha)}{\Gamma(n + a)} \prod_{i=1}^{r} \frac{\Gamma(n_i + \alpha p_i)}{n_i! \Gamma(\alpha p_i)}, \tag{29.38}$$

where $n = \sum_r n_i$, which provides a simple basis for maximum-likelihood or Bayesian estimation of F_{ST} (Balding, 2003). Note that (29.38) needs to be multiplied across loci and across subpopulations to get a global likelihood and a global F_{ST} estimate. This equation still depends on unknown allele frequencies, which can either be integrated out or coestimated with F_{ST} using an Markov chain Monte Carlo (MCMC) approach (Balding, 2003; Kitakado *et al.*, 2006). Interestingly, subpopulation specific or locus-specific F_{ST} can also be estimated, if one believes that some demes had distinct histories (e.g. Falush *et al.*, 2003a; Foll and Gaggiotti, 2006), or that selection has shaped genetic diversity at a given locus (Beaumont and Balding, 2004). Extensions of this approach have been proposed to take into account local inbreeding (Ayres and Overall, 1999) or linkage disequilibrium between loci (Kitada and Kishino, 2004).

29.4.3 Dealing with Mutation and Migration using Identity Coefficients

When mutations and migrations are considered, two major changes take place as compared to previous derivations: the global population allele frequencies cannot be considered as constant over time (due to mutations), and the subpopulations cannot be considered as independent (due to migration). Cockerham and Weir (1987; 1993) have overcome these difficulties by expressing correlation measures as functions of the identity coefficients Q. Note that a formalization of the effect of a population structure in terms of identity coefficients has led to the derivation of equilibrium values of the fixation indexes for different spatial models incorporating mutations (see e.g. Nagylaki, 1983; Crow and Aoki, 1984; Cockerham and Weir, 1987; Slatkin, 1991; 1993; Slatkin and Voelm, 1991; Rousset, 1996; 1997; 2004; Herbots, 1997; Wilkinson-Herbots, 1998; Weir and Hill, 2002; Wilkinson-Herbots and Ettridge, 2004).

The variance components now depend on migration and mutation parameters as well as on unknown allele frequencies, but an ANOVA can still be performed. In that case, however, Cockerham and Weir (1987) showed that the variance components corresponding

to the linear model described in (29.23) were equal to

$$\sigma_a^2 = Q_1 - Q_2$$
$$\sigma_b^2 = Q_0 - Q_1$$
$$\sigma_w^2 = 1 - Q_0, \tag{29.39}$$

where Q_0 is the probability that two genes from the same individual are identical in state, Q_1 is the same identity probability for two genes from different individuals within the same deme, and Q_2 is the identity probability for two genes in different demes. Therefore, the intraclass correlations defined in the same way as in (29.36) are now given by

$$\rho_0 = \frac{Q_0 - Q_1}{1 - Q_1}; \, \rho_1 = \frac{1 - Q_1}{1 - Q_2}; \, \rho_2 = \frac{Q_1 - Q_2}{1 - Q_2}, \tag{29.40}$$

where $\rho_0 \equiv \bar{F}_{IS}$, $\rho_1 \equiv F_{IT}$, and $\rho_2 \equiv F_{ST}$. Rousset (2002) more recently showed that such definitions of inbreeding coefficients as ratios of differences of probability of identity in state are very general and do not depend on a particular evolutionary model.

29.5 RELATIONSHIP BETWEEN DIFFERENT DEFINITIONS OF FIXATION INDEXES

The use of identity coefficients provides a convenient way to understand the relationship between the different indexes used to quantify population subdivision. It is interesting to compare the quantities defined in (29.40) with those defined in (29.22) as functions of similar but different identity coefficients to understand a major difference between the gene diversity and the intraclass correlation approach to estimate F statistics. In Nei's (1973; 1977) approach, the correlations are always relative to a random pair of genes at a given subdivision level (Q_S or Q_T in (29.22)), whereas in Cockerham's approach (1969; 1973) it is always relative to the least related pair of genes for a given subdivision level (Q_1 or Q_2 in (29.40)). Note that when the number of sampled diploid individuals within demes (n_k) is large then $\hat{Q}_S \approx \hat{Q}_1$. This is because $n_k(n_k - 1)$ pairs of genes are compared when computing $\hat{Q}_S$, whereas only $n_k(n_k - 2)$ pairs are considered when computing $\hat{Q}_1$. In other words, the n_k pairs of genes within individuals that are not taken into account when computing $\hat{Q}_1$ (as compared to $\hat{Q}_S$) become negligible relative to the number of between-individual gene pairs. For a similar reason, $\hat{Q}_T \approx \hat{Q}_2$ when the number of sampled demes is large, and therefore F statistics defined from gene diversities or heterozygosities are close to those defined in terms of variance components. We study below the relationships between different estimators of F_{ST} in more detail.

As usual now, let Q be the probability that two random genes are identical in state, but here we simply have Q_1 for genes within the same deme, and Q_2 for genes in different demes. The intraclass correlation within demes equivalent to ρ_2 in (29.40) is equal to (Cockerham and Weir, 1987)

$$\beta = \frac{Q_1 - Q_2}{1 - Q_2}, \tag{29.41}$$

while Crow and Aoki (1984) have redefined the G_{ST} statistic in (29.21) as a parameter corresponding to F_{ST} in (29.22). Following Cockerham and Weir (1993), we call this parameter G_{CA} (to differentiate it from the G_{ST} statistic defined in (29.21)). It is therefore equal to

$$G_{CA} = \frac{Q_1 - \bar{Q}}{1 - \bar{Q}}, \tag{29.42}$$

where $\bar{Q} = [Q_1 + (d-1)Q_2]/d$ is the probability corresponding to Q_T in (29.22) that two genes are alike when drawn randomly from the total population. Note that the two quantities are therefore related by the following relationships (Cockerham and Weir, 1987):

$$G_{CA} = \frac{(d-1)\beta}{d-\beta}; \quad \beta = \frac{dG_{CA}}{G_{CA} + d - 1}, \tag{29.43}$$

which corresponds to the relationship between F_{ST} and ρ_2 defined in (29.22) and (29.40) respectively. As explained above, $G_{CA} \approx \beta$ if the number of demes is large, but β should always be slightly larger than G_{CA} (Goudet, 1993).

If we now consider estimators based on these parameters, the relationship between $\hat{\beta}$ and $\hat{G}_{CA}$ would be identical to (29.43), but with d being here the number of sampled demes (Cockerham and Weir, 1993). For the estimator of G_{ST} as defined in (29.21) or F_{ST} as defined from (29.19), its relationship with $\hat{\beta}$ and $\hat{G}_{CA}$ depends on both the numbers of sampled demes and of sampled individuals per deme (see Cockerham and Weir, 1993, (29.4)). We note here that an estimator of F_{ST} based on the estimation of heterozygosities defined in (29.20) (Pons and Chaouche, 1995) should not depend on the number of demes.

Comparison between the actual values taken by different estimators has been achieved through simulations, or by comparing equilibrium values in different spatial models (e.g. Weir and Cockerham, 1984; Slatkin and Barton, 1989; Chakraborty and Danker-Hopfe, 1991; Rousset, 1997; Weir and Hill, 2002). The results of these comparisons vary in details, but the general trend is that estimators behave quite similarly and lead to nearly identical values provided that sample sizes and the number of sampled demes are large. For small samples (100 individuals or less), Chakraborty and Danker-Hopfe (1991) find that both heterozygosity and intraclass correlation approaches overestimate F_{ST} and underestimate $\bar{F}_{IS}$, while F_{IT} estimates are close to large samples values. From (29.43), it is clear that expectations and estimations based on heterozygosity are affected by the number of demes sampled, which is not the case for estimators based on intraclass correlation. In order to address this deficiency, Nei (1986; 1987, p. 163) has proposed to modify D_{ST} as $D'_{ST} = dD_{ST}/(d-1)$ to remove the comparison of samples with themselves. He also had to redefine H_T as $H'_T = D'_{ST} + H_S$, which results in a new definition of $G'_{ST} = D'_{ST}/H'_T$ similar to that of β in (29.41). Note that this modification is not entirely satisfactory, as the actual number of demes is usually unknown, and not necessarily equal to the number of demes sampled. Thus, even though different approaches lead to quite similar estimators in most practical situations, the ANOVA approach seems preferable for its clear statistical foundations (but see Nagylaki, 1998b for a statistically sound approach based on gene diversities). Another advantage is that the intraclass correlation approach can easily be applied to dominant data (Peakall *et al.*, 1995; Stewart and Excoffier, 1996), as well as to molecular markers (Weir and Basten, 1990; Excoffier *et al.*, 1992; Michalakis and Excoffier, 1996), while retaining exactly the same analytical

design. Moreover, its connections with identity by descent and kinship coefficients make it suitable for analyses of regular systems of mating. However, several features of the ANOVA approach have been criticized. First, estimates of variance components with the highest subdivision levels can be negative. This can arise by the way they are extracted from the mean squares using the method of moments (see (29.35)). When the true value of the variance component is positive but small, slightly negative estimates can be obtained due to random fluctuations with small samples, but negative values can also be obtained because the true variance components are negative. This latter situation is not embarrassing, because they are not obtained by a decomposition of the total variance into variances. Rather, they are functions of covariances (see Table 29.3 and (29.25)). As explained in **Chapter 28** (Appendix 1), they differ from the strictly positive variance components in conventional ANOVA because we do not assume that the effects due to deme (a), individuals (b), and genes (w) are uncorrelated between demes, between individuals within demes, or between genes within individuals. Negative coefficients can indeed arise in real populations if alleles are more alike between individuals than within individuals (if there is self-mating avoidance or separate sexes), or between demes than within demes (if there is predominant outbreeding) (see e.g. Wright, 1969; Cockerham, 1973 for a discussion on this subject). Note, however, that heterozygosity-based estimates of F statistics can also become negative in similar situations if random mating is not assumed (Goudet, 1993). In that case, it is clear that the definition of F statistics as correlations still makes sense, whereas definitions in terms of identity probabilities do not. A more justified criticism made to the approach of Weir and Cockerham is that they consider the different subpopulations as replicates, and therefore as having identical sizes and being subjected to identical evolutionary forces, which is usually implausible in biological situations except for experimental populations (Chakraborty and Danker-Hopfe, 1991). On the contrary, approaches based on gene diversity and heterozygosity do not make any such assumptions and can in principle handle unequal deme sizes if properly estimated, which may be difficult (Gaggiotti and Excoffier, 2000). It seems therefore that intraclass correlation approaches should be preferred when the number of demes is small, supposed so, or totally unknown. On the contrary, in the case of an exhaustive sampling, or when demes are supposed to be of highly unequal size, gene diversity and heterozygosity approaches could be preferred due to lower variance consideration (Cockerham and Weir, 1993). However, maximum-likelihood-based methods (e.g. Beerli and Felsenstein, 2001; Balding, 2003; Beerli, 2006) should be superior in the context of gene flow estimations.

29.6   *F* STATISTICS AND COALESCENCE TIMES

Fixation indexes and related quantities can be conveniently described in terms of average coalescence times within and between subdivision levels assuming small mutation rates (Slatkin, 1991; Slatkin and Voelm, 1991). Using this formalization, equilibrium properties of different spatial models of population subdivision can be easily obtained (see e.g. Slatkin, 1995; Wilkinson-Herbots, 1998; Rousset, 2004). This approach gives results essentially identical to classical approaches based on recursions (see e.g. Crow and Aoki, 1984; Cockerham and Weir, 1987; 1993; Rousset, 1996).

Following Slatkin (1991), let us consider a set of demes that have diverged from an ancestral population some T generations ago. The demes have remained separate without exchanging migrants ever since (models with migration are reviewed in **Chapter 28**). We assume that the demes are made up of haploid or diploid individuals that randomly unite their gametes (no inbreeding).

Two genes are of the same allelic type if no mutation occurred since their divergence from a most recent common ancestor (since their coalescence time), say some t generations ago. Conditional on t, this event has a probability $Q(t) = (1 - \mu)^{2t}$, where μ is the mutation rate per generation. Then the unconditional expectation of Q is

$$Q = \sum_{t=1}^{\infty} (1 - \mu)^{2t} \Pr(t) \tag{29.44}$$

where $\Pr(t)$ is the probability that there was a coalescence event t generations ago. For small μ, it becomes

$$Q \approx \sum_{t=1}^{\infty} (1 - 2\mu t) \Pr(t) = 1 - 2\mu\bar{t} \tag{29.45}$$

where $\bar{t}$ is simply the average coalescence time. Note that (29.45) is a good approximation if $\mu\bar{t} \ll 1$ (Slatkin and Voelm, 1991), but see Wilkinson-Herbots (1998) for the exact result with large mutation rates. It follows that we can use (29.45) directly to express β defined in (29.41) as a function of average coalescence times as (Slatkin, 1991)

$$\beta = \frac{\bar{t}_1 - \bar{t}_0}{\bar{t}_1} \tag{29.46}$$

where $\bar{t}_1$ is the average coalescence time of two genes from different demes and $\bar{t}_0$ is the average coalescence time of two genes drawn from the same deme. Similarly, Crow and Aoki's G_{CA} can be expressed as

$$G_{CA} = \frac{\bar{t} - \bar{t}_0}{\bar{t}}, \tag{29.47}$$

where $\bar{t}$ is the average coalescence time of two genes drawn from the whole population, either in the same or in different demes. This parameterization of the fixation index in terms of average coalescence times can be extended to higher subdivision levels such as equivalents of F_{SC} (correlation of a pair of genes taken within demes relative to a pair of genes taken at random from a group of demes) or F_{CT} (correlation of a pair of genes taken within a group of demes relative to a pair of genes taken at random from the population). They are good approximations if migration rates are much smaller between demes from different subdivision levels than from the same subdivision level (Slatkin and Voelm, 1991). As more generally stated by Rousset (2002; 2004, p. 58–62), inbreeding coefficients or F statistics can be seen as the increased probability of recent coalescence of one pair of genes relative to another pair, the choice of pairs of genes to be compared defining the inbreeding coefficient.

We can use these relationships to study the rate of approach of these statistics to equilibrium. The average coalescent time within diploid demes of size N is simply $\bar{t}_0 = 2N$, while the average coalescence time between demes is given by $\bar{t}_1 = T + 2N$

(Tajima, 1993), since two genes have to be in the same ancestral population in order to coalesce, then it takes again on average $2N$ more generations. From (29.46), it follows that $\beta = T/(T + 2N)$, suggesting that

$$\frac{\beta}{1 - \beta} = \frac{T}{2N} \tag{29.48}$$

can be used as an estimate of the divergence time (T) of the demes scaled by the population size. It is interesting to see that (29.48) is independent from the number of demes, which means that it can also be used as a genetic distance between populations (Slatkin, 1995) for loci with low mutation rates. Note that this is not the case for the same ratio of G_{CA} whose expectation is $G_{CA} = (d-1)T/[(d-1)T + 2Nd]$, implying

$$\frac{G_{CA}}{1 - G_{CA}} = \frac{T(d-1)}{2Nd}. \tag{29.49}$$

If the true (but usually unknown) number of demes is two, then the ratio of G_{CA} values is only one half that for β. As G_{CA} depends on the unknown number of demes d, it implies that the scaled divergence time between demes is poorly estimated from gene diversity measures like G_{CA}.

29.7 ANALYSIS OF MOLECULAR DATA: THE AMOVA FRAMEWORK

Because conventional ANOVA compares average gene frequencies among populations, it is not best suited to deal with molecular markers for which we have information not only on allele frequencies, but also on the amount of differences (mutations) between alleles. We show hereafter how ANOVA can be modified to directly incorporate this molecular information and thus to become an analysis of molecular variance (AMOVA) (Excoffier *et al.*, 1992). AMOVA can be considered as an extension of Cockerham's ANOVA of gene frequencies, where intraclass correlation coefficients do depend on the mutational process. In this sense, it is analogous to Cockerham and Weir's (1987; 1993) extension of their correlation measures to identity coefficients to include mutation and migration. The common denominator between the two approaches is to estimate F statistics through intraclass correlations defined as ratios of variance components.

29.7.1 Haplotypic Diversity

Molecular information at a given locus can be considered as a particular combination of alleles at different but completely linked loci, which is one way to define a haplotype. The nature of the molecular information may vary (DNA sequence, restriction fragment length polymorphism (RFLP), microsatellites), but the format of the analysis remains the same. For sake of simplicity, let us assume that each mutation occurs at a different site (the infinite-site model), and consider a haploid genetic system or assume that the gametic phase of S-linked polymorphic loci can be determined without ambiguity (e.g. mitochondrial DNA, Y chromosome, X chromosome in males). In that case, each haplotype can be coded as a vector of biallelic states of dimension S, where each element

Table 29.4 Analysis of Molecular Variance (AMOVA) layout for the linear model defined in (29.50). The sums of squares are expressed here as functions of squared Euclidean distances between haplotypes i and j (δ_{ij}^2's).

Source of variation	d.f.*	Sums of squares	Mean squares	Expected mean squares
Among groups	$G-1$	$SS(AG) = SS(T)$ $-\sum\limits_{g}^{G} \frac{1}{2n_g} \sum\limits_{k}^{d} \sum\limits_{k'}^{d} \sum\limits_{i}^{n_{gk}} \sum\limits_{j}^{n_{gk}} \delta_{ij}^2$	$\dfrac{SS(AG)}{G-1}$	$\sigma_w^2 + n''\sigma_b^2 + n'''\sigma_a^2$
Among demes within groups	$d-G$	$SS(AD) = \sum\limits_{g}^{G} \frac{1}{2n_g} \sum\limits_{k}^{d_g} \sum\limits_{k'}^{d_g} \sum\limits_{i}^{n_{gk}} \sum\limits_{j}^{n_{gk'}} \delta_{ij}^2$ $-SS(WD)$	$\dfrac{SS(AD)}{d-G}$	$\sigma_w^2 + n'\sigma_b^2$
Among haplotypes within demes	$n-d$	$SS(WD) = \sum\limits_{g}^{G} \sum\limits_{k}^{d_g} \frac{1}{2n_{gk}} \sum\limits_{i}^{n_{gk}} \sum\limits_{j}^{n_{gk}} \delta_{ij}^2$	$\dfrac{SS(WD)}{n-d}$	σ_w^2
Total	$n-1$	$SS(T) = \dfrac{1}{2n} \sum\limits_{i}^{n} \sum\limits_{j}^{n} \delta_{ij}^2$		

$d = \sum\limits_{g}^{G} d_g, n = \sum\limits_{g}^{G} \sum\limits_{k}^{d_g} n_{gk}, n_g = \sum\limits_{k}^{d_g} n_{gk}, S_G = \sum\limits_{g}^{G} \sum\limits_{k}^{d_g} \frac{n_{gk}^2}{n_g}, n' = \frac{n - S_G}{d - G}, n'' = \frac{S_G - \sum\limits_{k}^{d} \frac{n_k^2}{n}}{G - 1}, n''' = \frac{n - \sum\limits_{g}^{G} \frac{n_g^2}{n}}{G - 1},$

* G is the number of groups of demes.

y_s is an indicator variable that may have the value 0 or 1. We show later how to handle multiallelic markers.

We consider here a hierarchical model where genes are within demes, demes within groups of demes, and groups within the population. In that case, $\mathbf{y}_{gki}$ represents the ith haplotype of the kth deme in the gth group. We shall use a linear model similar to that seen in conventional ANOVA (29.23),

$$\mathbf{y}_{gki} = \mathbf{p} + \mathbf{a}_g + \mathbf{b}_{gk} + \mathbf{w}_{gki}, \tag{29.50}$$

where $\mathbf{p}$ is the vector of the expected allelic states. The effects are here $\mathbf{a}$ for the group, $\mathbf{b}$ for the deme, and $\mathbf{w}$ for the individual, and, as usual, are assumed random, additive, and independent, with associated variance components σ_a^2, σ_b^2, and σ_c^2.

The ANOVA layout for this partitioning of the total variance is shown in Table 29.4. We introduce here an alternative but completely equivalent way of computing the sums of squares. They are expressed as functions of square Euclidean distances instead of squared deviations from some mean. As a matter of fact, a sum of squared deviations from the mean can be written as a sum of squared differences between observations, barring a given constant, as

$$SS(z) = \sum_{i=1}^{N}(z_i - \bar{z})^2 = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N}(z_i - z_j)^2.$$

We can thus rewrite the sum of squares within demes

$$SS(WD) = \sum_g^G \sum_k^{d_g} \sum_i^{n_{gk}} (\mathbf{y}_{gki} - \mathbf{y}_{gk.})' \mathbf{W} (\mathbf{y}_{gki} - \mathbf{y}_{gk.}),$$

as

$$SS(WD) = \sum_g^G \sum_k^{d_g} \frac{1}{2n_{gk}} \sum_i^{n_{gk}} \sum_j^{n_{gk}} (\mathbf{y}_{gki} - \mathbf{y}_{gkj})' \mathbf{W} (\mathbf{y}_{gki} - \mathbf{y}_{gkj})$$

where $\mathbf{W}$ is a square weight matrix that can be used to specify the relationships between the different loci and/or their respective weights. The expression $(\mathbf{y}_{gki} - \mathbf{y}_{gkj})' \mathbf{W} (\mathbf{y}_{gki} - \mathbf{y}_{gkj})$ is nothing else but a squared Euclidean distance δ_{ij}^2 between haplotype i and j in a S-dimensional space. Note that if all loci are assumed independent and equally informative, then $\mathbf{W} = \mathbf{I}$, the identity matrix, and we can write the sum of squares in a simple way as

$$SS(WD) = \sum_g^G \sum_k^{d_g} \frac{1}{2n_{gk}} \sum_i^{n_{gk}} \sum_j^{n_{gk}} \delta_{ij}^2$$

$$= \sum_g^G \sum_k^{d_g} \frac{1}{2n_{gk}} \sum_i^{n_{gk}} \sum_j^{n_{gk}} \sum_s^S (y_{gki_s} - y_{gkj_s})^2. \tag{29.51}$$

In that case, the squared Euclidean distance is just the number of pairwise differences between two haplotypes, equivalent to a Hamming distance. The other sums of squares are given by

$$SS(AD) = SS(\text{Within Group}) - SS(WD)$$

$$= \sum_g^G \frac{1}{2n_g} \sum_k^{d_g} \sum_{k'}^{d_g} \sum_i^{n_{gk}} \sum_j^{n_{gk'}} \delta_{ij}^2 - \sum_g^G \sum_k^{d_g} \frac{1}{2n_{gk}} \sum_i^{n_{gk}} \sum_j^{n_{gk}} \delta_{ij}^2, \tag{29.52}$$

$$SS(AG) = SS(T) - SS(\text{Within Group})$$

$$= \frac{1}{2n} \sum_i^n \sum_j^n \delta_{ij}^2 - \sum_g^G \frac{1}{2n_g} \sum_k^{d_g} \sum_{k'}^{d_g} \sum_i^{n_{gk}} \sum_j^{n_{gk}} \delta_{ij}^2. \tag{29.53}$$

We can thus directly express the variance components and the intraclass correlations as a function of Euclidean distances between haplotypes (Table 29.4). By analogy with Wright's F statistics, we have called these intraclass correlations Φ statistics (Excoffier *et al.*, 1992). They are conventionally defined as

$$\Phi_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_T^2}; \quad \Phi_{SC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}; \quad \Phi_{CT} = \frac{\sigma_a^2}{\sigma_T^2}. \tag{29.54}$$

If $\mathbf{W} = \mathbf{I}$, the identity matrix, the estimated Φ statistics are equivalent to Weir and Cockerham (1984) weighted correlations of gene frequencies over all loci (Michalakis and Excoffier, 1996) as defined in (29.37) for another hierarchical model. For other weighting

schemes, the relationship between Φ statistics and Weir and Cockerham's statistics would be different, even though the underlying parameters (variance components) to estimate would remain identical.

29.7.2 Genotypic Data

Even though the AMOVA framework has been essentially developed for haplotypic data, data consisting of diploid genotypes can be accommodated to estimate correlations between genes in different individuals at any level of subdivision. This is possible because the correlation of genes between individuals (the coancestry coefficient $\bar{\theta}$ in Cockerham's terminology) does not depend on how genes are associated within individuals. Diploid molecular data can be transformed into pseudohaploid data by defining the arbitrary gametic phase in all individuals (Michalakis and Excoffier, 1996). If one assumes that the weighting matrix **W** has zero entries on off-diagonal elements (implying that the loci are independent), the Euclidean distances can be computed between pseudohaplotypes as if they were real haplotypes. Intraclass correlations are estimated as for the haploid case above, and the sums of squares within and between demes should be insensitive to the definition of the pseudohaplotypes. This procedure can thus be applied to loci with an arbitrary degree of linkage and even completely unlinked loci. This is because multilocus average estimators are constructed by summing up variance components on both the numerator and denominator, as in (29.37). Even though linkage information does not affect the way overall sums of squares and variance components are estimated, it should affect the variance of the estimates, which should be incorporated into the assessment of the significance of the inferred statistics, as discussed below.

29.7.3 Multiallelic Molecular Data

With the exception of RFLP and single nucleotide polymorphism (SNP) data, more than two allelic states are possible for most molecular data types like DNA sequences or microsatellites. Several options for the handling of such data are possible, depending on the underlying model of mutation we have in mind.

Under a strict infinite-site model, no more than two allelic states should coexist at any locus, even though the allelic repertoire can be much larger. In that case, the multiallelic data can be safely re-encoded into binary data.

29.7.3.1 DNA Sequences

Under a finite-site model, multiple allelic states can coexist at a given locus, and identity in state no longer means identity by descent. One can still re-encode the information into a binary vector (in an augmented space) if one can infer the phylogenetic tree underlying the observed data. Efficient techniques have been developed in the case of DNA sequence (e.g. Felsenstein, 1981; Swofford *et al.*, 1996; Piontkivska, 2004; Boussau and Gouy, 2006) as well as **Chapters 15** and **16**, often allowing the reconstruction of the actual number m of mutations having occurred in the ancestry of a sample of sequences. The binary vector used to compute the squared distances should then be of length m and encode the mutational status of each sequence instead of its nucleotide status. It should have an entry of 1 at a given position if the corresponding mutation has occurred on the lineage of the haplotype to the most recent common ancestor

of all haplotypes and an entry of zero otherwise. An unweighted squared Euclidean distance computed on such vectors is therefore equivalent to counting the number of mutations having happened on the branches of a phylogeny separating two haplotypes (a patristic distance). Although we can still partition total genetic diversity into variance components and perform an AMOVA analysis on such data, it should be clear that the resulting Φ statistics are not average correlations of nucleotide frequencies anymore, but are rather average correlations of mutation frequencies at different subdivision levels. The weight matrix **W** could be used in such an analysis to reflect the weight of different types of mutations, like transitions, transversions, or insertion/deletions. If no phylogenetic tree is available, the number of pairwise differences between sequences can be used as a squared Euclidean distance, which would be equivalent to a patristic distance computed on minimum spanning network relating sequences (Excoffier and Smouse, 1994).

Note that corrections for multiple substitutions per site (see e.g. Nei and Kumar, 2000) could be incorporated into a conventional ANOVA framework (see Weir and Basten, 1990). Because these corrections aim at uncovering the true number of mutations having occurred in the ancestry of a pair of sequences, they could also be incorporated into the AMOVA framework by using corrected genetic distances as approximations to squared Euclidean distances. Further work is needed to establish exactly how multiple substitutions affect the partitioning of genetic variance within and among populations.

A number of other studies have attempted to estimate F statistics from DNA sequences (Takahata and Palumbi, 1985; Lynch and Crease, 1990; Weir and Basten, 1990; Hudson *et al.*, 1992a; 1992b; Slatkin, 1993), some of them being reviewed in Charlesworth (1998). They all attempt in partitioning the total diversity into within and between components, and can all be related either to gene diversity or variance components approaches, implying they should lead to similar results for large number of large samples.

29.7.3.2 *Microsatellite Data*

Microsatellite data cannot be properly recoded as a binary vector because it is quite difficult to reconstruct a phylogenetic tree from it. Microsatellite data are supposed to predominantly follow a stepwise mutation model (SMM) (e.g. Weber and Wong, 1993, but see Leblois *et al.*, 2003), where identity in state does not mean identity by descent. However, the difference between allelic states (numbers of repeats of a motif of a few base pairs) is an indicator of the evolutionary proximity between alleles.

Slatkin (1995) has shown that the square of this difference in allele size increased approximately linearly with time. He considered a single microsatellite locus, where μ is the mutation rate per generation. The mutation can lead either to an increase or to a decrease in repeat number, and the increase size is a random variable with mean 0 and variance σ_{m}^2, assumed to be independent of allele size. Note that this mutation model is more general than the single-step model, for which $\sigma_{\mathrm{m}}^2 = 1$. Let y_i and y_j be as usual indicator variables, here equal to the number of repetitions of the microsatellite motif for the gene copies i and j respectively. If these gene copies had a most recent common ancestor t generations ago (i.e. their coalescence time is t) and we assume a Poisson distribution of mutations, then the expectation of $\delta_{ij}^2 = (y_i - y_j)^2$ increases linearly with coalescence time: $E(\delta_{ij}^2) = 2\mu t \sigma_{\mathrm{m}}^2$. Slatkin (1995) then defined the sum of squares of the

difference in allele size within d demes, from which samples of n diploid individuals are supposed to have been drawn, as

$$S_{\mathrm{W}} = \frac{1}{d} \sum_{k=1}^{d} \frac{2}{2n(n-1)} \sum_{i=1}^{2n} \sum_{j<i} \delta_{ij}^2, \qquad (29.55)$$

and the average number of square differences in repeat size between demes as

$$S_{\mathrm{B}} = \frac{2}{(2n)^2 d(d-1)} \sum_{k=1}^{d} \sum_{k'<k} \sum_{i=1}^{2n} \sum_{j=1}^{2n} \delta_{ij}^2. \qquad (29.56)$$

The average squared difference in repeat size in the total population is then obtained as

$$\bar{S} = \frac{2n-1}{2nd-1} S_{\mathrm{W}} + \frac{2n(d-1)}{2nd-1} S_{\mathrm{B}}. \qquad (29.57)$$

By analogy with Nei's G_{ST}, Slatkin (1995) defined an estimator of population subdivision

$$R_{ST} = \frac{\bar{S} - S_{\mathrm{W}}}{\bar{S}}. \qquad (29.58)$$

Noting that the δ_{ij}^2s are indeed squared Euclidean distances, the AMOVA framework is a natural alternative to Slatkin (1995) approach for the analysis of microsatellite data. It also implies that Slatkin's analysis is close to an ANOVA of allele size frequencies. However, it should be clear from the comparison of (29.41) with (29.42) that the analog of R_{ST} should be (Michalakis and Excoffier, 1996; Rousset, 1996),

$$\Phi_{ST} = \frac{S_{\mathrm{B}} - S_{\mathrm{W}}}{S_{\mathrm{B}}}. \qquad (29.59)$$

Therefore, assuming identical sample sizes in all demes, the correct expected relationships between R_{ST} and Φ_{ST} are

$$\Phi_{ST} = \frac{R_{ST}}{1 - c(1 - R_{ST})}, \qquad (29.60)$$

$$R_{ST} = \frac{(1-c)\Phi_{ST}}{1 - c\Phi_{ST}} \qquad (29.61)$$

where $c = (2n-1)/(2nd-1)$. (29.60) corrects a typo in (29.9) of Michalakis and Excoffier (1996). The equilibrium values of a statistic equivalent to Φ_{ST} under the SMM, but called ρ_{ST}, has been defined for an island model by Rousset (1996).

Even though the analysis of microsatellite data has exactly the same form as that of other data types, estimators based on such data have a much larger associated variance (Slatkin, 1995), which is essentially due to the SMM, and for which large samples do not necessarily produce more consistent estimates (Pritchard and Feldman, 1996). Several authors have underlined the need for a very large number of independent loci to get meaningful estimates of various genetic quantities from microsatellite data (Goldstein *et al.*, 1995; Zhivotovsky and Feldman, 1995; Bertorelle and Excoffier, 1998). Thus, even

though it remains to be investigated in further detail, F statistics derived under the SMM could have the same limitations and requirements. As has been shown empirically (e.g. Takezaki and Nei, 1996; Gaggiotti *et al.*, 1999; Balloux and Goudet, 2002), it may be adequate to compute F_{ST} estimates only from allele frequencies if a few microsatellite loci are available because, even though such estimates would be biased under an SMM, they would have a smaller mean square error.

29.7.4 Dominant Data

An important fraction of newly generated molecular data are dominant such as randomly amplified polymorphic DNA (RAPD) markers or amplified fragment length polymorphism (AFLP) markers. For these markers, diploid individuals typically show, for each locus, either a positive or a negative scoring. A negative scoring implies that the individual is homozygote for the recessive allele, while a positive scoring is obtained for individuals that are either homozygote or heterozygote for the dominant allele. Lynch and Milligan (1994) have been the first authors to propose an estimation procedure for quantifying the amount of genetic structure based on such markers. Using second order Taylor series expansions, they have developed an asymptotically unbiased estimator of G_{ST} as defined in (29.21), that depends on estimated gene diversities within ($\hat{H}_W$) and between ($\hat{H}_B$) demes, as (Lynch and Milligan, 1994)

$$\hat{G}_{ST} = \frac{\hat{H}_B}{\hat{H}_B + \hat{H}_W} \left(1 + \frac{\hat{H}_B \text{var}(\hat{H}_W) - \hat{H}_W \text{var}(\hat{H}_B) + (\hat{H}_B - \hat{H}_W)\text{cov}(\hat{H}_B, \hat{H}_W)}{\hat{H}_B(\hat{H}_B + \hat{H}_W)^2} \right)^{-1}.$$
(29.62)

Gene diversities are conventionally estimated from allele frequencies. However, instead of using the maximum-likelihood estimator of the recessive allele frequency ($\hat{p}$) equal to $\sqrt{\hat{P}}$, where $\hat{P}$ is the fraction of individuals that do not exhibit the marker allele, Lynch and Milligan proposed to use an asymptotically unbiased estimator, namely,

$$\hat{p} = \sqrt{\hat{P}} \left(1 - \frac{\text{var}(\hat{P})}{8\hat{P}^2} \right)^{-1},$$
(29.63)

where $\text{var}(\hat{P}) = \hat{P}(1 - \hat{P})/(2n_k)$ is the sampling variance of $\hat{P}$ in deme k. While this estimator is less biased than the maximum-likelihood estimator, it is still biased for small sample sizes and low frequencies of the recessive allele. In order to address this problem, Zhivotovsky (1999) has recently proposed an elegant Bayesian approach to estimate recessive allele frequencies. He empirically shows that his approach leads to a slightly improved estimation of recessive allele frequencies and to a much better estimation of the fixation index.

An alternative and economical way to test for the presence of genetic structure with dominant data is to consider differences in genotype frequencies between demes. Thus, Peakall *et al.* (1995) proposed to analyze multilocus RAPD data under the AMOVA framework. They used as a Euclidean square distance the number of pairwise difference between RAPD profile, and obtained in this case Φ statistics equivalent to the correlation of *phenotype* frequencies within or between demes. Because these Φ statistics were not comparable to conventional fixation indexes, we (Stewart and Excoffier, 1996) developed a methodology that allows the computation of correlations of *gene* frequencies from

dominant data under the AMOVA framework. This was done by realizing that the comparison of two diploid phenotypes corresponds to the comparisons of four pairs of genes. It is therefore possible to replace squared Euclidean distances between phenotypes by the sum of four expected square Euclidean distances between their constituent genes, and to adjust the degrees of freedom in order to get estimates of Φ statistics directly comparable to those defined from codominant markers. However, for each individual, the expected number of different genes given its phenotype depends on the frequency of the recessive allele, which thus needs to be estimated. It also depends quite strongly on a possible departure of HWE that is quite common in partially selfing plant communities where RAPD markers are most often used. If the selfing rate is S, then under inbreeding equilibrium, an asymptotically unbiased (and apparently unpublished) estimator of the recessive allele frequency equivalent to (29.63) can be obtained if $0 \leq S < 1$ as

$$\hat{p} = \frac{1}{4(1-S)} \left\{ (A\hat{P} + S^2)^{1/2} \left[1 - \frac{A^2 \text{var}(\hat{P})}{8(A\hat{P} + S^2)^2} \right]^{-1} - S \right\}, \qquad (29.64)$$

where $A = 8S^2 - 24S + 16$. Obviously, $\hat{p} = \hat{P}$ if $S = 1$. One can check that (29.64) reduces to (29.63) if $S = 0$. If departure from Hardy–Weinberg is not due to self-fertilization, but to other forms of inbreeding, (29.25) can still be used due to the relationship $S = 2\bar{F}_{IS}/(1 + \bar{F}_{IS})$. Note also that Zhivotovsky (1999) has proposed a Bayesian estimator of $\hat{p}$ that takes $\bar{F}_{IS}$ into account. More recently, Holsinger *et al.* (2002) have introduced a Bayesian approach to estimate F statistics from dominant markers allowing for unknown amount of local inbreeding. Under the assumption that, in populations related by migrations, the allele frequencies follow a β distribution (e.g. Crow and Kimura, 1970; Balding and Nichols, 1995), the parameters of which explicitly depend on F statistics (see (29.38)), Holsinger *et al.* (2002) are able to simultaneously estimate F_{ST} (as reported in (29.9)) and the local inbreeding coefficient F_{IS} (assumed to be identical in all subpopulations). Simulation results show that F_{ST} is usually well estimated but that a large number of loci and populations are necessary to get meaningful estimates of F_{IS} from dominant data. Further effect of selfing on the estimation of population subdivision can be found elsewhere (Maruyama and Tachida, 1992; Pollak and Sabran, 1992; Jarne, 1995; Nordborg, 1997; Wang, 1997a; 1997b).

29.7.5 Relation of AMOVA with other Approaches

The AMOVA framework is therefore fully equivalent to Cockerham's ANOVA framework. AMOVA is also closely related to Long *et al.* (1986) multivariate extension of Cockerham's ANOVA for loci with multiple alleles, which uses variance–covariance matrices of allele frequencies as weighting matrices similar to **W**. Explicit relationships between Long *et al.* (1986), Nei (1977; 1983), and Weir and Cockerham (1984) estimators can be found in Chakraborty and Danker-Hopfe (1991). For single-locus estimators, it is noted that when covariance of allele frequencies are neglected, Long's estimators are close to unweighted average F statistics $\hat{\theta}_U$ (Weir and Cockerham, 1984), while AMOVA produces estimates completely identical to Weir and Cockerham's weighted estimators $\hat{\theta}_W$ (Weir and Cockerham, 1984). For multilocus estimators, if some individuals are not typed at all loci, implying that there is some missing data, it is better to perform distinct analyses for each locus separately, and to combine variance components estimates for

each locus as in (29.37). In that case, the AMOVA haplotypic framework would wrongly assume that the squared distances are based on the same number of characters for each pairwise comparison, which could bias the estimators, particularly if missing data are not equally balanced among demes. The advantage of the AMOVA framework is therefore to highly increase the speed of the computations by carrying out a single analysis for arbitrary numbers of alleles and loci, and by allowing the covariance of genetic information (i.e. linkage disequilibrium) across loci to be taken into account.

29.8 SIGNIFICANCE TESTING

Event though the variance of some F statistics estimates has been derived (see e.g. Nei *et al.*, 1977; Cockerham and Weir, 1983; Weir and Cockerham, 1984), it seems of little use to build reliable significance tests for small samples. Instead, several traditional resampling techniques, like bootstrap, jackknife (Efron, 1982), or permutations are commonly used to assess the significance of the moment-estimated statistics. Likelihood-based estimations can lead to posterior distributions of the parameters of interest, from which credible interval can be constructed (e.g. Holsinger *et al.*, 2002; Balding, 2003; Foll and Gaggiotti, 2006), which is not detailed here.

29.8.1 Resampling Techniques

Typically, bootstrap and jackknife methods aim at defining confidence intervals around an estimated value, implicitly assuming that genetic structure exists. Therefore, appropriate bootstrapping methods should try to mimic the action of evolutionary forces that have created the genetic structure (Weir, 1996, p. 174). The use of nonparametric bootstrap of sampled loci is advocated because the genetic diversity observed at unlinked loci should be the result of the replication of independent evolutionary (coalescent and mutational) processes, giving us some idea of the genetic component of the variance. This nonparametric bootstrapping approach would then satisfactorily explore the stochasticity of the evolutionary process only if a large number of unlinked loci are available. However, bootstrapping over linked nucleotide sites of a DNA sequence would certainly not be appropriate, as it would only give us some idea of the stochasticity of the mutation process, which is much smaller than the stochasticity of the coalescent process. Even though nonparametric bootstrap procedures do not always allow one to estimate the complete stochasticity of the (often unknown) evolutionary process, it may be justified to appreciate the sensitivity of a given statistic to the choice of alleles, loci, or demes as is implemented in various computer packages (e.g. Goudet, 1995; Raymond and Rousset, 1995b; Excoffier *et al.*, 2005). Under an assumed spatial model, some parametric bootstrap procedure mimicking the coalescent and mutational process would be in order.

On the other hand, permutational approaches construct confidence intervals for a given statistic under the hypothesis of no genetic structure (e.g. Manly, 1991; Excoffier *et al.*, 1992; Hudson *et al.*, 1992a), and therefore do not require to specify the evolutionary forces at work that would lead to a genetic structure. However, note that mixed bootstrap and permutational approaches can be used to obtain power curves, for instance, as a function of the number of sampled loci (Excoffier *et al.*, 1992). Some care should however be taken

on how to perform permutations, depending on not only the statistic to test ($\bar{F}_{IS}$, F_{IT}, F_{ST}, or F_{CT}, as well as their corresponding variance components) but also on the nature of data (linked or independent loci, dominant data). For haplotypic data, with known gametic phase, all linked loci can be considered as a single locus, and it would be irrelevant to bootstrap loci. Differences between haplotypes within individuals within demes ($\bar{F}_{IS}$) can be tested by permuting haplotypes between individuals within demes. Differences between haplotypes within individuals within the total population (F_{IT}) can be tested by permuting haplotypes between individuals and between demes. Differences between demes (F_{ST}) are tested by permuting individuals between demes. If random mating is assumed or if no information is available on how haplotypes are grouped into individuals, F_{ST} can be tested by permuting haplotypes between demes. Under such permutation procedures, note that the total sums of squares and degrees of freedom remain unchanged in an ANOVA approach. If a higher level of subdivision is considered, such as groups of demes within populations, differences between groups (F_{CT}) are tested by permuting whole demes between groups, which changes the sums of square and the degrees of freedom.

With dominant data, $\bar{F}_{IS}$ or F_{IT} cannot be generally estimated (but see Holsinger *et al.*, 2002), and one has to estimate allele or haplotype frequencies (Lynch, 1991; Lynch and Milligan, 1994; Stewart and Excoffier, 1996; Zhivotovsky, 1999). Under the hypothesis of no population structure, genotypes can be permuted between demes only if frequencies have been estimated at the population level and not at the deme level because, otherwise, estimated frequencies would change with permutations. This feature of dominant data makes it much less powerful than codominant markers to study population subdivision (Lynch and Milligan, 1994; Stewart and Excoffier, 1996).

For data consisting of completely unlinked loci, permutations should be done as for haplotypic data, but independently for each locus. A problem that is not completely settled occurs for data consisting of incompletely linked loci. In that case, permutations cannot be done for each locus independently, as it would disrupt the potentially existing linkage disequilibrium pattern. Permuting all loci simultaneously as if they were completely linked seems, however, a conservative but not entirely satisfactory procedure (Michalakis and Excoffier, 1996).

29.8.2.9 Exact Tests

Exact tests of population subdivision have been proposed for testing differences in allele (Raymond and Rousset, 1995a) or genotype (Goudet *et al.*, 1996) frequencies among a set of demes. Contrary to previously described statistics, these tests do not aim at measuring directly correlations of genes within and among populations. They rather follow from the application of the classical Fisher exact test on R × C contingency tables. Here, rows are for demes, whereas columns are for alleles or genotypes. Under the null hypothesis of absence of genetic differentiation between samples, the conditional probability of the observed table Π_0 is computed as

$$\Pi_0 = \frac{\prod\limits_{k=1}^{d} n_{k.}! \prod\limits_{i=1}^{r} n_{.i}!}{n_{..}! \prod\limits_{k=1}^{d} \prod\limits_{i=1}^{r} n_{ki}!} \tag{29.65}$$

where, for allelic data, n_{ki} is the number of genes of allelic type i in the kth sample, $n_{.i}$ is the marginal count of the ith allele in the d samples, $n_{k.}$ is the marginal size of the kth sample, and $n_{..}$ is the total allelic counts over all samples. For genotype data, the n's are simply genotypic counts instead of allelic counts.

Instead of enumerating all possible contingency tables as in the Fisher exact test (Mehta and Patel, 1983), a Markov chain is used to efficiently explore the space of all possible tables by a random walk (see Guo and Thompson, 1992; Raymond and Rousset, 1995a). It is done in such a way that the probability to visit a particular table corresponds to its actual probability under the null hypothesis of no genetic structure. The P value of the test is then defined as the fraction of explored tables that are equally or less likely than the observed one. Simulations have confirmed that exact tests performed on genotype counts are less powerful than those performed on allelic counts (Goudet *et al.*, 1996), as they use less information. The incorporation of additional levels of subdivision could be envisioned by extending the analysis of two-dimensional tables to cubic ones, which could be quite challenging to implement in practice. Multilocus tests can be performed by combining probabilities of single-locus tests, assuming that all loci are independent.

29.9   RELATED AND REMAINING PROBLEMS

29.9.1   Testing Departure from Hardy–Weinberg Equilibrium

HWE can be disrupted by population subdivision or inbreeding and specific attempts have been made to detect discrepancy between observed and expected genotype frequencies. Recent attempts include procedures related to Fisher's exact test on R × C contingency tables, where alternative tables are generated either using permutations or by a random walk along a Markov chain (Guo and Thompson, 1992). Departure from HWE due to inbreeding can also be tested by the use of likelihood ratio statistics. Numerical methods to obtain minimum variance and maximum-likelihood estimators of $\bar{F}_{IS}$ have been proposed in Robertson and Hill (1984) and in Hill *et al.* (1995). Several Bayesian methods have been used to estimate inbreeding coefficients within demes in the case of biallelic loci (see Shoemaker *et al.*, 1998 for a review). For multiple alleles, a Bayesian MCMC method has been developed to test departure from HWE that is specifically due to inbreeding (Ayres and Balding, 1998). In that paper, a Bayesian procedure is described to estimate the posterior distribution of $\bar{F}_{IS}$ statistics associated with each genotype, or a global $\bar{F}_{IS}$ statistic, along with the distribution of allele frequencies. The use and effect of several priors are also discussed. Note that, in both maximum-likelihood and Bayesian procedures, care has to be taken to keep $\bar{F}_{IS}$ estimates within reasonable bounds (Robertson and Hill, 1984; Ayres and Balding, 1998). The power of different testing procedures is reviewed in Rousset and Raymond (1995). It is found that exact probability tests (Guo and Thompson, 1992) are more powerful than tests based on maximum likelihood (Hill *et al.*, 1995), minimum variance (Robertson and Hill, 1984), or variance component estimators of $\bar{F}_{IS}$ (Cockerham, 1969; 1973), when the departure from HWE is due to specific genotypes. Alternatively, Robertson and Hill's (1984) estimator is found to perform best when departure from HWE is due to regular inbreeding that can be accounted for by a single $\bar{F}_{IS}$ value for all genotypes.

29.9.2   Detecting Loci under Selection

It has since long been realized (Cavalli-Sforza, 1966; Lewontin and Krakauer, 1973) that gene frequencies at loci involved in adaptive events should show increased levels of differentiation among populations, while genes under balancing selection should present relatively uniform frequencies across populations, and thus lead to low F_{ST} values. Capitalizing on these ideas, Beaumont and Nichols (1996) proposed to detect loci under selection by comparing observed levels of genetic diversity within (heterozygosity) and between (F_{ST}) populations to those expected under a simple evolutionary scenario, like an infinite-island model. This approach does not seem to be sensitive to past population demography (Beaumont, 2005) and it seems applicable to a wide diversity of genetic markers assessed in several populations (e.g. Luikart et al., 2003, Table 29.2). It has been successfully applied to recent genome scans to detect outlier loci (e.g. Wilding et al., 2001; Storz et al., 2004; Murray and Hare, 2006), which can be considered as candidate loci for having had a history of selection. More recently, Beaumont and Balding (2004), proposed a Bayesian method to detect loci under selection based on (29.38), to identify signatures of selection, and compared its power to that of Beaumont and Nichols (1996) for loci under positive or balancing selection. They find that both methods can identify loci under positive selection if the selection coefficient is about five times larger than the migration rate between populations. However, for low levels of population differentiation ($F_{ST} = 0.1$), these approaches cannot really distinguish loci under balancing selection, even for quite strong selection coefficients. Because it is more and more widely recognized that our ability to evidence patterns of selection depends on a sound knowledge of past demographic history (e.g. Ometto et al., 2005; Williamson et al., 2005; Kelley et al., 2006), complex demographic inferences from observed patterns of genetic diversity will need to be developed to better understand selective processes.

29.9.3   What is the Underlying Genetic Structure of Populations?

The estimation procedures defined above all assumed that the (potentially hierarchical) genetic structure of the population was known and that the main issue was to estimate the degree of population subdivision. However, the history and therefore the true genetic structure of a population are very often unknown. Some attempts have been made to try to disentangle population history from population structure (e.g. Templeton et al., 1995; Templeton 2004), but these procedures have not been developed within a clear statistical framework. Moreover, individuals may not always be grouped into well-recognized sampling units, such that their assignment to a given subdivision may be a relevant issue. Quite recently, a number of studies have tried to address some of these problems. They attempted to partition sampled populations into distinguishable sets (e.g. Holsinger and Mason-Gamer, 1996; Dupanloup et al., 2002; Corander et al., 2003) to allocate individuals to specific or reconstructed populations (Pritchard et al., 2000a; Dawson and Belkhir, 2001; Falush et al., 2003a; Wilson and Rannala, 2003; Guillot et al., 2005; Corander and Marttinen, 2006), or to find the number of hidden subdivisions (Pritchard et al., 2000a; Guillot et al., 2005; Corander and Marttinen, 2006). Another complicating factor is that genetic structure estimates may differ depending on which chromosomal segments are studied (mitochondrial DNA, Y chromosome, or autosomes) because of sex-biased dispersal and sexual selection (Chesser, 1991; Chesser and Baker, 1996). In that case, the estimation of sex-specific F statistics has been proposed (e.g. Petit et al., 2001; Vitalis,

2002; Fontanillas *et al.*, 2004). It appears that the development of statistical procedures to uncover the demographic or selection history of a set of populations that best explains the observed genetic structure is certainly one of the most interesting challenges of population genetics.

Acknowledgments

REFERENCES

Ayres, K.L. and Balding, D.J. (1998). Measuring departures from Hardy-Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity* **80**, 769–777.

Ayres, K.L. and Overall, D.J. (1999). Allowing for within-subpopulation inbreeding in forensic match probabilities. *Forensic Science International* **103**, 207.

Balding, D.J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology* **63**, 221–230.

Balding, D.J. and Nichols, R.A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3–12.

Balloux, F. and Goudet, J. (2002). Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Molecular Ecology* **11**, 771–783.

Barrai, I. (1971). Subdivision and inbreeding. *American Journal of Human Genetics* **23**, 95–96.

Beaumont, M.A. (2005). Adaptation and speciation: what can F-st tell us? *Trends in Ecology and Evolution* **20**, 435–440.

Beaumont, M.A. and Balding, D.J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969–980.

Beaumont, M.A. and Nichols, R.A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society London, Series B* **263**, 1619–1626.

Beerli, P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**, 341–345.

Beerli, P. and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 4563–4568.

Bertorelle, G. and Excoffier, L. (1998). Inferring admixture proportions from molecular data. *Molecular Biology and Evolution* **15**, 1298–1311.

Boussau, B. and Gouy, M. (2006). Efficient likelihood computations with nonreversible models of evolution. *Systematic Biology* **55**, 756–768.

Cavalli-Sforza, L.L. (1966). Population structure and human evolution. *Proceedings of the Royal Society London, Series B* **164**, 362–379.

Chakraborty, R. and Danker-Hopfe, H. (1991). Analysis of population structure: a comparative study of different estimators of Wright's fixation indices. In *Handbook of Statistics*, C.R. Rao and R. Chakraborty, eds. Elsevier Science Publishers, Amsterdam, pp. 203–254.

Chakraborty, R. and Weiss, K.M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 9119–9123.

Chapman, N.H. and Thompson, E.A. (2001). Linkage disequilibrium mapping: the role of population history, size, and structure. *Advances in Statistics* **42**, 413–437.

Charlesworth, B. (1998). Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution* **15**, 538–543.

Charlesworth, B., Nordborg, M. and Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research* **70**, 155–174.

Chesser, R.K. (1991). Influence of gene flow and breeding tactics on gene diversity within populations. *Genetics* **129**, 573–583.

Chesser, R.K. and Baker, R.J. (1996). Effective sizes and dynamics of uniparentally and diparentally inherited genes. *Genetics* **144**, 1225–1235.

Cockerham, C.C. (1969). Variance of gene frequencies. *Evolution* **23**, 72–83.

Cockerham, C.C. (1973). Analysis of gene frequencies. *Genetics* **74**, 679–700.

Cockerham, C.C. and Weir, B.S. (1983). Variance of actual inbreeding. *Theoretical Population Biology* **23**, 85–109.

Cockerham, C.C. and Weir, B.S. (1987). Correlations, descent measures: drift with migration and mutation. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 8512–8514.

Cockerham, C.C. and Weir, B.S. (1993). Estimation of gene flow from F-statistics. *Evolution* **47**, 855–863.

Corander, J. and Marttinen, P. (2006). Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology* **15**, 2833–2843.

Corander, J., Waldmann, P. and Sillanpaa, M.J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374.

Crow, J. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper and Row, New York.

Crow, J.F. and Aoki, K. (1984). Group selection for a polygenic behavioral trait: estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 6073–6077.

Dawson, K.J. and Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research* **78**, 59–77.

Dupanloup, I., Schneider, S. and Excoffier, L. (2002). A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology* **11**, 2571–2581.

Efron, B. (1982). *The Jacknife, the Bootstrap and Other Resampling Plans. Regional Conference Series in Applied Mathematics*. Society for Industrial Applied Mathematics, Philadelphia, PA.

Excoffier, L., Laval, G. and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1**, 47–50.

Excoffier, L. and Smouse, P. (1994). Using allele frequencies and geographic subdivision to reconstruct gene genealogies within a species. Molecular variance parsimony. *Genetics* **136**, 343–359.

Excoffier, L., Smouse, P. and Quattro, J. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.

Falush, D., Stephens, M. and Pritchard, J.K. (2003a). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.

Falush, D., Wirth, T., Linz, B., Pritchard, J.K., Stephens, M., Kidd, M., Blaser, M.J., Graham, D.Y., Vacher, S., Perez-Perez, G.I., Yamaoka, Y., Megraud, F., Otto, K., Reichard, U., Katzowitsch, E., Wang, X., Achtman, M. and Suerbaum, S. (2003b). Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**, 1582–1585.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Foll, M. and Gaggiotti, O. (2006). Identifying the environmental factors that determine the genetic structure of populations. *Genetics* **174**, 875–891.

Fontanillas, P., Petit, E. and Perrin, N. (2004). Estimating sex-specific dispersal rates with autosomal markers in hierarchically structured populations. *Evolution* **58**, 886–894.

Fu, Y.X. (1997). Coalescent theory for a partially selfing population. *Genetics* **146**, 1489–1499.

Gaggiotti, O. and Excoffier, L. (2000). A simple method of removing the effect of a bottleneck and unequal population sizes on pairwise genetic distances. *Proceedings of the Royal Society London B* **267**, 81–87.

Gaggiotti, O.E., Lange, O., Rassmann, K. and Gliddon, C. (1999). A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data. *Molecular Ecology* **8**, 1513–1520.

Gilmour, J.S.L. and Gregor, J.W. (1939). Demes: a suggested new terminology. *Nature* **144**, 333.

Goldstein, D.B., Ruiz-Linares, A., Cavalli-Sforza, L.L. and Feldman, M.W. (1995). Microsatellite loci, genetic distances, and human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 6723–6727.

Goudet, J. (1993). The genetics of geographically structured populations. Ph.D., University of Wales, Bangor.

Goudet, J. (1995). Fstat version 1.2: a computer program to calculate F-statistics. *Journal of Heredity* **86**, 485–486.

Goudet, J., Raymond, M., De Meeüs, T. and Rousset, F. (1996). Testing differentiation in diploid populations. *Genetics* **144**, 1933–1940.

Guillot, G., Estoup, A., Mortier, F. and Cosson, J.F. (2005). A spatial statistical model for landscape genetics. *Genetics* **170**, 1261–1280.

Guo, S. and Thompson, E. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**, 361–372.

Herbots, H.M. (1997). The structured coalescent. In *Progress in Population Genetics and Human Evolution*, P. Donnely and S. Tavaré, eds. Springer-Verlag, New York, pp. 231–255.

Hewitt, G.M. (2001). Speciation, hybrid zones and phylogeography – or seeing genes in space and time. *Molecular Ecology* **10**, 537–549.

Hill, W.G., Babiker, H.A., Ranford-Cartwright, L.C. and Walliker, D. (1995). Estimation of inbreeding coefficients from genotypic data on multiple alleles, and application to estimation of clonality in malaria parasites. *Genetical Research* **65**, 53–61.

Holsinger, K.E., Lewis, P.O. and Dey, D.K. (2002). A Bayesian approach to inferring population structure from dominant markers. *Molecular Ecology* **11**, 1157–1164.

Holsinger, K.E. and Mason-Gamer, R.J. (1996). Hierarchical analysis of nucleotide diversity in geographically structured populations. *Genetics* **142**, 629–639.

Hudson, R.R., Boos, D.D. and Kaplan, N.L. (1992a). A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution* **9**, 138–151.

Hudson, R.R., Slatkin, M. and Maddison, W.P. (1992b). Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589.

Jaarola, M. and Searle, J.B. (2004). A highly divergent mitochondrial DNA lineage of *Microtus agrestis* in southern Europe. *Heredity* **92**, 228–234.

Jarne, P. (1995). Mating system, bottleneck, and genetic polymorphism in hermaphroditic animals. *Genetical Research* **65**, 193–207.

Jorde, L. (1980). The genetic structure of subdivided human populations. In *Current Developments in Anthropological Genetics*, J. Mielke and M. Crawford, eds. Plenum Press, New York, pp. 135–208.

Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W. and Akey, J.M. (2006). Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research* **16**, 980–989.

Kitada, S. and Kishino, H. (2004). Simultaneous detection of linkage disequilibrium and genetic differentiation of subdivided populations. *Genetics* **167**, 2003–2013.

Kitakado, T., Kitada, S., Kishino, H. and Skaug, H.J. (2006). An integrated-likelihood method for estimating genetic differentiation between populations. *Genetics* **173**, 2073–2082.

Leblois, R., Estoup, A. and Rousset, F. (2003). Influence of mutational and sampling factors on the estimation of demographic parameters in a "continuous" population under isolation by distance. *Molecular Biology and Evolution* **20**, 491–502.

Lewontin, R.C. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175–195.

Li, W.H. and Nei, M. (1974). Stable linkage disequilibrium without epistasis in subdivided populations. *Theoretical Population Biology* **6**, 173–183.

Long, J.C., Naidu, J.M., Mohrenweiser, H.W., Gershowitz, H., Johnson, P.L., Wood, J.W. and Smouse, P.E. (1986). Genetic characterization of Gainj- and Kalam-speaking peoples of Papua New Guinea. *American Journal of Physical Anthropology* **70**, 75–96.

Luikart, G., England, P.R., Tallmon, D., Jordan, S. and Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**, 981–994.

Lynch, M. (1991). Analysis of population genetic structure by DNA fingerprinting. *Experientia Supplementum* **58**, 113–126.

Lynch, M. and Crease, T.J. (1990). The analysis of population survey data on DNA sequence variation. *Molecular Biology and Evolution* **7**, 377–394.

Lynch, M. and Milligan, B.G. (1994). Analysis of population genetic structure with RAPD markers. *Molecular Ecology* **3**, 91–99.

Magri, D., Vendramin, G.G., Comps, B., Dupanloup, I., Geburek, T., Gomory, D., Latalowa, M., Litt, T., Paule, L., Roure, J.M., Tantau, I., van der Knaap, W.O., Petit, R.J. and de Beaulieu, J.L. (2006). A new scenario for the quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytologist* **171**, 199–221.

Malécot, G. (1948). *Les Mathématiques de l'Hérédité*. Masson, Paris.

Manly, B.F.J. (1991). *Randomization and Monte Carlo methods in Biology*. Chapman and Hall, London.

Marchini, J., Cardon, L.R., Phillips, M.S. and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics* **36**, 512–517.

Marjoram, P. and Donnelly, P. (1994). Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics* **136**, 673–683.

Maruyama, K. and Tachida, H. (1992). Genetic variability and geographic structure in partially selfing populations. *Japanese Journal of Genetics* **67**, 39–51.

Mehta, C.R. and Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in rxc contingency tables. *Journal of the American Statistical Association* **78**, 427–434.

Michalakis, Y. and Excoffier, L. (1996). A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* **142**, 1061–1064.

Murray, M.C. and Hare, M.P. (2006). A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, *Crassostrea virginica*. *Molecular Ecology* **15**, 4229–4242.

Nagylaki, T. (1983). The robustness of neutral models of geographical variation. *Theoretical Population Biology* **24**, 268–294.

Nagylaki, T. (1985). Homozygosity, effective number of alleles, and interdeme differentiation in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **82**, 8611–8613.

Nagylaki, T. (1998a). The expected number of heterozygous sites in a subdivided population. *Genetics* **149**, 1599–1604.

Nagylaki, T. (1998b). Fixation indices in subdivided populations. *Genetics* **148**, 1325–1332.

Nei, M. (1972). Genetic distance between populations. *American Naturalist* **106**, 283–292.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 3321–3323.

Nei, M. (1977). F-statistics and analysis of gene diversity in subdivided populations. *Annals of Human Genetics* **41**, 225–233.

Nei, M. (1986). Definition and estimation of fixation indices. *Evolution* **40**, 643–645.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., Chakravarti, A. and Tateno, Y. (1977). Mean and variance of FST in a finite number of incompletely isolated populations. *Theoretical Population Biology* **11**, 291–306.

Nei, M. and Chesser, R.K. (1983). Estimation of fixation indices and gene diversities. *Annals of Human Genetics* **47**, 253–259.

Nei, M. and Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford.

Nei, M. and Li, W.H. (1973). Linkage disequilibrium in subdivided populations. *Genetics* **75**, 213–219.

Nicholson, G., Smith, A.V., Jonsson, F., Gustafsson, O. and Stefansson, K. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistics Society, Series B* **64**, 695–715.

Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641–647.

Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics* **146**, 1501–1514.

Nordborg, M. and Donnelly, P. (1997). The coalescent process with selfing. *Genetics* **146**, 1185–1195.

Ohta, T. (1982). Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **79**, 1940–1944.

Ohta, T. and Kimura, M. (1969). Linkage disequilibrium due to random genetic drift. *Genetical Research* **13**, 47–55.

Ometto, L., Glinka, S., De Lorenzo, D. and Stephan, W. (2005). Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Molecular Biology and Evolution* **22**, 2119–2130.

Peakall, R., Smouse, P.E. and Huff, D.R. (1995). Evolutionary implications of allozyme and RAPD variation in diploid populations if dioecious buffalograss "Buchloe dactyloides". *Molecular Ecology* **4**, 135–147.

Petit, E., Balloux, F. and Goudet, J. (2001). Sex-biased dispersal in a migratory bat: a characterization using sex-specific demographic parameters. *Evolution* **55**, 635–640.

Piontkivska, H. (2004). Efficiencies of maximum likelihood methods of phylogenetic inferences when different substitution models are used. *Molecular Phylogenetics and Evolution* **31**, 865–873.

Pollak, E. and Sabran, M. (1992). On the theory of partially inbreeding finite populations. III. Fixation probabilities under partial selfing when heterozygotes are intermediate in viability. *Genetics* **131**, 979–985.

Pons, O. and Chaouche, K. (1995). Estimation, variance and optimal sampling of gene diversity. II. Diploid locus. *Theoretical and Applied Genetics* **91**, 122–130.

Pritchard, J.K. and Donnelly, P. (2001). Case-control studies of association in structured or admixed populations. *Theoretical Population Biology* **60**, 227–237.

Pritchard, J.K. and Feldman, M.W. (1996). Statistics for microsatellite variation based on coalescence. *Theoretical Population Biology* **50**, 325–344.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000a). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P. (2000b). Association mapping in structured populations. *American Journal of Human Genetics* **67**, 170–181.

Rannala, B. and Hartigan, J.A. (1996). Estimating gene flow in island populations. *Genetical Research* **67**, 147–158.

Raymond, M. and Rousset, F. (1995a). An exact test for population differentiation. *Evolution* **49**, 1280–1283.

Raymond, M. and Rousset, F. (1995b). GENEPOP Version 1.2: population genetics software for exat tests and ecumenicism. *Journal of Heredity* **86**, 248–249.

Reynolds, J., Weir, B.S. and Cockerham, C.C. (1983). Estimation for the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**, 767–779.

Robertson, A. and Hill, W.G. (1984). Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics* **107**, 703–718.

Rousset, F. (1996). Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics* **142**, 1357–1362.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219–1228.

Rousset, F. (2000). Inferences from spatial population genetics. In *Handbook of Statistical Genetics*, D. Balding, M. Bishop and C. Cannings, eds. John Wiley & Sons, Chichester.

Rousset, F. (2002). Inbreeding and relatedness coefficients: what do they measure? *Heredity* **88**, 371–380.

Rousset, F. (2004). *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton, NJ.

Rousset, F. and Raymond, M. (1995). Testing heterozygote excess and deficiency. *Genetics* **140**, 1413–1419.

Shoemaker, J., Painter, I. and Weir, B.S. (1998). A Bayesian characterization of Hardy-Weinberg disequilibrium. *Genetics* **149**, 2079–2088.

Simonsen, K.L., Churchill, G.A. and Aquadro, C.F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.

Slatkin, M. (1987). The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology* **32**, 42–49.

Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research* **58**, 167–175.

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**, 264–279.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462.

Slatkin, M. and Barton, N.H. (1989). A comparison of three indirect methods for estimating average levels of gene flow. *Evolution* **43**, 1349–1368.

Slatkin, M. and Voelm, L. (1991). FST in a hierarchical island model. *Genetics* **127**, 627–6629.

Stewart, N. and Excoffier, L. (1996). Assessing population genetic structure and variability with RAPD data: application to "vaccinium macrocarpon" (American cranberry). *Journal of Evolutionary Biology* **9**, 153–171.

Storz, J.F., Payseur, B.A. and Nachman, M.W. (2004). Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Molecular Biology and Evolution* **21**, 1800–1811.

Strobeck, K. (1987). Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* **117**, 149–153.

Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. (1996). Phylogenetic inference. In *Molecular Systematics*, D.M. Hillis, C. Moritz and B.K. Mable, eds. Sinauer Associates, Inc., Sunderland, MA, pp. 407–514.

Taberlet, P., Fumagalli, L., Wust-Saucy, A.G. and Cosson, J.F. (1998). Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology* **7**, 453–464.

Tajima, F. (1993). Statistical analysis of DNA polymorphism. *Japanese Journal of Genetics* **68**, 567–595.

Takahata, N. and Palumbi, S.R. (1985). Extranuclear differentiation and gene flow in the finite island model. *Genetics* **109**, 441–457.

Takezaki, N. and Nei, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* **144**, 389–399.

Templeton, A.R. (2004). Statistical phylogeography: methods of evaluating and minimizing inference errors. *Molecular Ecology* **13**, 789–809.

Templeton, A.R., Routman, E. and Phillips, C.A. (1995). Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, Ambystoma tigrinum. *Genetics* **140**, 767–782.

Vitalis, R. (2002). Sex-specific genetic differentiation and coalescence times: estimating sex-biased dispersal rates. *Molecular Ecology* **11**, 125–138.

Wahlund, S. (1928). Zusammensetzung von populationen und korrelationerscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas* **11**, 65–106.

Wakeley, J. (1998). Segregating sites in Wright's island model. *Theoretical Population Biology* **53**, 166–174.

Wang, J. (1997a). Effective size and F-statistics of subdivided populations. I. Monoecious species with partial selfing. *Genetics* **146**, 1453–1463.

Wang, J. (1997b). Effective size and F-statistics of subdivided populations. II. Dioecious species. *Genetics* **146**, 1465–1474.

Weber, J.L. and Wong, C. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics* **2**, 1123–1128.

Weir, B.S. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates, Inc., Sunderland, MA.

Weir, B.S. and Basten, C.J. (1990). Sampling strategies for distances between DNA sequences. *Biometrics* **46**, 551–582.

Weir, B.S. and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.

Weir, B.S. and Hill, W.G. (2002). Estimating F-statistics. *Annual Review of Genetics* **36**, 721–750.

Wilding, C.S., Butlin, R.K. and Grahame, J.W. (2001). Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology* **14**, 611–619.

Wilkinson-Herbots, H.M. (1998). Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biolology* **37**, 535–585.

Wilkinson-Herbots, H.M. and Ettridge, R. (2004). The effect of unequal migration rates on FST. *Theoretical Population Biology* **66**, 185–197.

Williamson, S.H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R. and Bustamante, C.D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 7882–7887.

Wilson, G.A. and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**, 1177–1191.

Wright, S. (1921). Systems of mating. *Genetics* **6**, 111–178.

Wright, S. (1922). Coefficients of inbreeding and relationship. *American Naturalist* **56**, 330–338.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution In *Proceedings of the 6th International Congress of Genetics*. Vol. 1. Brooklyn Botanic Garden, Menasha, pp. 356–366.

Wright, S. (1943). Isolation by distance. *Genetics* **28**, 114–138.

Wright, S. (1946). Isolation by distance under diverse systems of mating. *Genetics* **31**, 39–59.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics* **15**, 323–354.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* **19**, 395–420.

Wright, S. (1969). *The Theory of Gene Frequencies*. The University of Chicago Press, Chicago, IL.

Wright, S. (1982). The shifting balance theory and macroevolution. *Annual Review of Genetics* **16**, 1–19.

Zhivotovsky, L.A. (1999). Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology* **8**, 907–913.

Zhivotovsky, L.A. and Feldman, M.W. (1995). Microsatellite variability and genetic distances. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 11549–11552.

Conservation Genetics

M.A. Beaumont

School of Biological Sciences, University of Reading, Reading, UK

This chapter reviews some of the statistical methodologies used in the genetic analysis of endangered and managed populations. The topics covered include the estimation of effective population size, N_e, the detection of past changes in population size, the estimation of admixture proportions, and the analysis of local population structure, kinship, and relatedness through genotypic methods. The reasons why it may be useful to measure N_e are discussed. Using genetic information it is possible to estimate N_e from changes in gene frequencies between samples taken at different times, or from genotypic disequilibria in one sample, or from differences in gene frequencies between two recently descended populations. The different statistical methodologies that have been applied are discussed in some detail. The ability to detect past changes in population size may also be useful in conservation management, and some of the more recent approaches are described. Another area of concern in conservation biology is introgressive hybridisation. At the population level it is possible to infer admixture proportions by comparing the gene frequencies in the admixed and non-admixed populations, providing they are available, and the different statistical approaches to analysing this problem are reviewed. Statistical methods have also been developed to detect immigrant and hybrid individuals from their multi-locus genotype, and the main methods developed in this area are also discussed.

30.1 INTRODUCTION

The use of genetic analysis in conservation falls into two main areas (Hedrick, 2001; DeSalle and Amato, 2004): examination of the genetic consequences of small population size for mean fitness and the probability of extinction (Frankham, 1995; 2005); and the use of data from genetic surveys to infer aspects of the demographic history of populations (Avise, 1994). This latter may then impinge on the former, e.g. in the estimation of 'effective population size', but may also be independent – for example, in the detection of hybrids. In many ways the first aspect, although historically motivating the use of genetics in conservation, is the most challenging. The theories describing the effect of recurrent deleterious mutation in small populations (Lynch *et al.*, 1995) depend on parameters whose

values are the subject of some debate even in model organisms (Keightley *et al.*, 1998; Lynch *et al.*, 1999; Whitlock *et al.*, 2003). The relative importance of genetic factors on extinction risk, in comparison with the ecological consequences of small population size, have been questioned (Lande, 1998). However, an increasing number of studies suggest that the loss of genetic variation can indeed contribute to extinction (Saccheri *et al.*, 1998; Spielman *et al.*, 2004), and this remains an area of active investigation in conservation biology.

This review will concentrate on the problems of statistical estimation that arise when using genetic data to infer aspects of the demographic history of populations, which can then be used to make informed decisions in the management of populations (Luikart and England, 1999). The statistical methodology is derived from the general area of inference in population genetics covered in Part 2 of this volume. To avoid overlap I have chosen four topics that have direct relevance to conservation biology – estimation of effective population size and detecting historical population bottlenecks, the analysis of hybridisation, and genotypic analysis of local population structure. One area, that of subdivision and gene flow from different populations, is only tangentially covered in the sections on hybridisation and genotypic analysis, because it is the subject of chapters by Excoffier and Rousset (**Chapters 28** and **29**). Methods for inferring relatedness and recovering pedigrees from genotypic data are briefly covered at the end of this review.

30.2   ESTIMATING EFFECTIVE POPULATION SIZE

It is useful to estimate the current rate of inbreeding in populations, and recent changes in this rate, for a number of reasons. First, although the effect of small population size (large inbreeding rate) on population viability is controversial, it is potentially significant, and therefore quantification of inbreeding rate in natural populations whose viability can also be monitored is an important component of any study to resolve these questions. Second, the inbreeding rate can be useful for calibrating timescales of recent inbreeding. For example, information on census size is often available, and if the rate of inbreeding for a given census size can be calibrated for one population then predictions may be made about other (comparable) populations on other (comparable) timescales (e.g. O'Ryan *et al.*, 1998).

The idealized 'effective' population size that gives rise to the observed rate of inbreeding is denoted N_e (Wright, 1931). The effect of different mating systems on N_e has been the subject of many studies (reviewed in Caballero (1994)), and, unsurprisingly, given the complexity of population genetic phenomena, it turns out that there are subtly different definitions of effective population size (Crow and Kimura, 1970), depending on the phenomena studied. For example, for the same theoretical 'inbreeding effective size' (Crow and Kimura, 1970), different mating systems can lead to slightly different predictions in the degree of linkage disequilibrium observed (Weir and Hill, 1980). Since these phenomena (e.g. linkage disequilibrium) can be used to infer N_e, a tacit assumption is that, given the uncertainty in estimation, these differences do not matter very much, and an estimate of N_e made by one method will lead to reasonable predictions of other phenomena that depend on N_e. Although a careful consideration of the mating system can give some estimate of N_e from the census size (Caballero, 1994; Nunney, 1995), it is often more practicable to estimate N_e directly, using genetic information, and listed

below are some of the commoner approaches to this problem (see also Schwartz *et al.*, 1998). A final caveat is that implicit in the methods below, the estimate of N_e is made in a closed population over a short interval (with N_e the harmonic mean over this interval). On longer time scales, with mutation, immigration, and past population restructuring, estimates of (indeed, notions of) N_e obtained from genetic data become highly model dependent (Wakeley, 2001) and increasingly divorced from any possible measurement of local census size, mating systems or any factor with relevance to population management.

30.2.1 Estimating N_e Using Two Samples from the Same Population: The Temporal Method

Random genetic drift causes gene frequencies to change over an interval as a function of the population size and duration of the interval. Thus the gene frequencies in samples taken from a population at two time points can be compared, allowing N_e to be estimated. This approach is known as the *temporal method* and has a wide history of use with allozyme and microsatellite data (Krimbas and Tsakas, 1971; Nei and Tajima, 1981; Pollak, 1983). The bulk of the theory has been developed from the consideration of the Wright–Fisher model (see Jorde and Ryman, 1995, for a treatment with overlapping generations). Method-of-moments estimators were originally developed (Krimbas and Tsakas, 1971; Waples, 1989) but in recent years there has been a concentration on likelihood-based approaches. The general area has recently been reviewed in detail by Wang (2005), and a brief overview will be given here.

Assume a sample is taken at generation 0 from a diploid population of N individuals, comprising the total number of individuals available to be sampled. The frequency of a particular allele is p_0. The gene frequency in the next generation, p_1 is determined by taking a sample of size $2N_e$ with replacement from generation 0. With independent realisations of this sampling process, the mean, $E[p_1]$, is unchanged at p_0, and the variance, $E[(p_1 - p_0)^2]$, is $p_0(1 - p_0)/(2N_e)$ from binomial sampling. Extrapolating to t generations of such sampling, gives $E[p_t] = p_0$ and $E[(p_t - p_0)^2] = p_0(1 - p_0)(1 - (1 - 1/(2N_e))^t)$. Defining F to be $(p_t - p_0)^2/p_0(1 - p_0)$, $E[F] = 1 - (1 - 1/(2N_e))^t$ and hence, replacing $E[F]$ by an estimate $\hat{F}$ one can estimate $\hat{N}_e$ as $t/(2\hat{F})$. In general p_t and p_0 are unknown and are estimated by x_0 and x_t, with sample sizes S_0 and S_t, using some sampling scheme. The nature of the sampling scheme affects aspects of the estimation (Nei and Tajima, 1981; Pollak, 1983; Waples, 1989): e.g. whether the samples are taken with or without replacement. A commonly used method of moments estimator that takes into account uncertainty in the gene frequencies is

$$\hat{N}_e = \frac{t}{2[\hat{F} - 1/(2S_0) - 1/(2S_t)]}, \tag{30.1}$$

(Krimbas and Tsakas, 1971; Waples, 1989). A number of different estimators $\hat{F}$ have been proposed (Nei and Tajima, 1981; Pollak, 1983).

The likelihood-based method of Williamson and Slatkin (1999) uses the Wright–Fisher model. They wish to estimate $p(\mathbf{n}_0, \mathbf{n}_t|N)$ where $\mathbf{n}_0$ and $\mathbf{n}_t$ are vectors of counts of different allelic types scored at a locus at generation 0 and generation t, and N is the number of (diploid) individuals. At time 0 there is a vector $\mathbf{p}_0$ of the (unknown) allele frequencies in the entire population from which $\mathbf{n}_0$ is sampled with replacement. In this model the population allele frequencies are discrete and could be obtained as counts divided by $2N$. A tacit assumption is that N is always N_e in this model. Williamson and

Slatkin's method sums over the uncertainty in $\mathbf{p}_0$ in the following way. In the simplest case of diallelic loci, there are $2N + 1$ possible states of $\mathbf{p}_0$. Denote by $\mathbf{x}_0$ the vector of length $2N + 1$ giving the initial (prior) probability distribution of the ordered set of states of $\mathbf{p}_0$. Then

$$p(\mathbf{n}_0, \mathbf{n}_t | N) = (\mathbf{x}_0^{\mathrm{T}} \mathbf{s}_0) \mathbf{x}_0^{\mathrm{T}} \mathbf{M}^t \mathbf{s}_t, \tag{30.2}$$

where $\mathbf{s}_0$ and $\mathbf{s}_t$ are vectors of multinomial probabilities of sampling $\mathbf{n}_0$ and $\mathbf{n}_t$ from each of the possible population frequencies in generation 0 and t, and $\mathbf{M}$ is the standard Wright–Fisher transition matrix (see Ewens, 2004) where the element m_{ij} gives the probability of moving from state i to j by multinomial sampling. Equation (30.2) is evaluated for a range of possible value of N to obtain the maximum likelihood estimate. As Williamson and Slatkin (1999) point out, the likelihood can be evaluated straightforwardly only when the number of possible states is small. With diallelic loci there are only $2N + 1$ possible states and therefore the size of the computational problem will scale as $(2N)^2$. However with larger numbers of alleles estimation rapidly becomes unmanageable. An importance-sampling strategy has been proposed by Anderson *et al.* (2000) to overcome this.

An advantage of Williamson and Slatkin's method is that it is straightforward to make a composite estimate of N_e from samples taken at different time points, and also to make estimates of population growth rates using genetic samples. In this latter case, population size changes exponentially from initial size N_0 with rate r as $N_i = N_0 r^t$, rounded down to the nearest integer, and $\mathbf{M}^t$ in (30.2) is replaced by the multiplied sequence of matrices M_i with columns equal to the number of possible allelic states in the ith generation and rows equal to the number of possible states in the $(i - 1)$th generation. Using simulations, they demonstrate that with a large number of loci (150) it is possible to obtain reasonably accurate joint estimates of N_0 and r.

Wang (2001) has made two modifications to the approach of Williamson and Slatkin (1999), resulting in the ability to estimate likelihoods very rapidly, even for multi-allelic loci. These improvements have enabled detailed simulation testing to be carried out. As in the study of Williamson and Slatkin (1999), Wang is also able to extend the method to work with multiple samples and to estimate changes in N_e. One of the changes proposed by Wang (2001) is to suggest that the likelihoods for multi-allelic loci with K alleles can be well approximated by converting them into K biallelic loci, whose frequencies $(\mathbf{n}_{0k}, \mathbf{n}_{tk})$ consist of the frequencies of the kth allele and the sum of the frequencies of the others. The likelihood is then calculated as

$$p(\mathbf{n}_0, \mathbf{n}_t | N) = \left(\prod_{k=1}^{K} p(\mathbf{n}_{0k}, \mathbf{n}_{tk} | N) \right)^{\frac{k-1}{k}}.$$

There is no theory to say why and under what circumstances this approximation will work. However Wang (2001) extends the transition matrix approach to the three-allele case, which is feasible for small sample sizes and small N, and is able to demonstrate that the approximation and the exact method are comparable in precision and accuracy. The second modification is to use a variety of computational 'tricks' to greatly speed up evaluation of (30.2), the most important of which is to note that only the diagonal and adjacent off-diagonal elements of $\mathbf{M}$ contribute significantly to the calculation of the likelihood.

Like Williamson and Slatkin (1999) and Wang (2001) demonstrated modest gains in accuracy and precision when using the likelihood method. The gains are most substantial when initial allele frequencies are low. For alleles at intermediate frequencies, however, as noted by Williamson and Slatkin, the improvement on moment-based methods through the use of likelihood methods appears to be small. The accuracy and precision are approximately linearly proportional to sample size, number of generations between samples, and the number of independent alleles used (i.e. the sum of $K - 1$ over loci), and inversely proportion to N_e. The method of Wang (2001) has been further extended in Wang and Whitlock (2003) to allow for the estimation of immigration rates.

An alternative likelihood-based approach suitable for multiple alleles has been developed using coalescent theory (Berthier *et al.*, 2002; Beaumont, 2003a; Anderson, 2005). The use of a coalescent model assumes that the changes in gene frequencies can be accurately described by the diffusion approximation, and will therefore lead to discrepancies with those based on the Wright–Fisher model when the population size is low.

30.2.2   Estimating N_e from Two Derived Populations

It is commonly the case that a once large population occupying a contiguous area becomes fragmented into isolated populations. Also, managed populations are often divided into separate groups. In these cases, when there is knowledge of the time of splitting, it is possible to infer the N_e for each sub-population from the amount of genetic divergence between the two populations. Most of the statistical methods devised for this case appear to be based on the likelihood method (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1981). These authors used a Brownian-motion approximation for genetic drift based on 2-allele models and extensions, discussed in a little more detail in a later section. More recently, methods based on coalescent theory have been developed which are directly applicable to multi-allelic models. This approach will be described here.

A method based on coalescent theory for estimating N_e in recently isolated, genetically diverging populations, was described in O'Ryan *et al.* (1998) and also by Nielsen *et al.* (1998) (see also Saccheri *et al.*, 1999). For a review of coalescent theory see Nordborg (**Chapter 25**), Hudson (1991) and Donnelly and Tavaré (1995). In the method described in O'Ryan *et al.* (1998), the genetic samples are taken from s isolated populations that are derived from some common ancestral population at time T in the past. In the discussion below N_e refers to the number of (diploid) individuals in the population, and sample sizes refer to the number of chromosomes. The populations are assumed to be have been isolated sufficiently recently that mutations arising since the time of isolation can be ignored. The genealogies for a particular locus of the s populations can be traced back to a set of founder lineages that have descendants in the samples. The founders are assumed to be sampled from an ancestral gene frequency vector $\mathbf{x}$ that is common to all the populations. The gene frequency configuration in the founder lineages is given by $\mathbf{a_f}$, and that in the sample is given by $\mathbf{a_s}$. The number of founder lineages, n_f, is a random variable given by

$$n_f = n_s - n_c,$$

where n_s is the sample size and n_c is the number of coalescences over the period T in the genealogy of the sample. The probability function $p(n_c|T/(2N_e), n_s)$ is given in Tavaré (1984; see also O'Ryan *et al.*, 1998). Thus the probability of a sample configuration for

a particular population can be calculated as

$$p(\mathbf{a_s}|T/(2N_e), n_s, k, \mathbf{x}) = \sum_{\mathbf{a_f}} p(\mathbf{a_s}|\mathbf{a_f}, n_s)\, p(\mathbf{a_f}|\mathbf{x}, n_f)\, p(n_c|T/(2N_e), n_s), \qquad (30.3)$$

where the summation is over all possible $\mathbf{a_f}$. In the case of populations that split from a common ancestor, the likelihoods can be multiplied over populations, each with their own $T/(2N_e)$, but assuming a common $\mathbf{x}$. The first term in (30.3) is the probability of obtaining the sample configuration from the founder configuration by successively choosing a lineage at random and duplicating it n_c times, and is given by

$$p(\mathbf{a_s}|\mathbf{a_f}, n_s, n_f) = \frac{\prod_{i=1}^{k} \binom{a_{si}-1}{a_{fi}-1}}{\binom{n_s-1}{n_f-1}}.$$

This can be derived by considering it as an urn problem (Slatkin, 1996; Nielsen *et al.*, 1998): the denominator is the number of ways of allocating the n_s lineages in the sample to the n_f founder lineages, and the numerator is the number of different ways of allocating the sample lineages within each allelic class to the corresponding class in the founders. The second term in (30.3) is the multinomial probability of drawing $\mathbf{a_f}$ from $\mathbf{x}$ given n_f. Clearly the number of possible configurations $\mathbf{a_f}$ is very large, and evaluation of (30.3) can be problematic. An additional problem is that the evaluation of $p(n_c|T/(2N_e), n_s)$ can be numerically unstable. O'Ryan *et al.* (1998) use importance sampling (Griffiths and Tavaré, 1994) as an alternative method of evaluation, and this is embedded within a Markov chain Monte Carlo (MCMC) sampler to infer marginal posterior densities for individual parameters (see Beaumont, 2003a, for a detailed description of this general approach).

The general scheme for two populations is illustrated in Figure 30.1. Using the MCMC scheme, marginal posterior distributions for $T/(2N_e)$ can be obtained. In the study of O'Ryan *et al.* (1998) a number of isolated populations were modelled in this way and the effective population size could be estimated because the splitting times were known. As an example, one population, (St. Lucia), was seeded from another umfolozi-hluhluwe complex (UHC) 16 years prior to sampling. Cull data suggested that the generation time was 7.5 years in these park populations. Seven microsatellite loci were analysed in samples of around 20–30 individuals (the sample size varied among loci due to missing data) sampled from the two parks. Figure 30.2(a) shows the posterior distribution for N_e in the St. Lucia population. There was extensive census data available and the harmonic mean census size was estimated as 58 for the St. Lucia population, giving a modal estimate of effective to census size of 13 %. In the UHC population, which was estimated to have a harmonic mean census size over this period of over 2000, the posterior density appeared to rise to an asymptote (a rectangular prior for N_e was used to make estimation practicable), and most values from $N_e > 500$ had similar posterior density (Figure 30.2b). Essentially, this means that an infinite population size has similar likelihood to most values of N_e greater than 500.

Obviously this modelling approach is only applicable when it can be assumed that the time scale is short enough that the effect of mutation can be ignored. Specifically it is assumed that all the descendent copies of alleles originally present in the founders are identical.
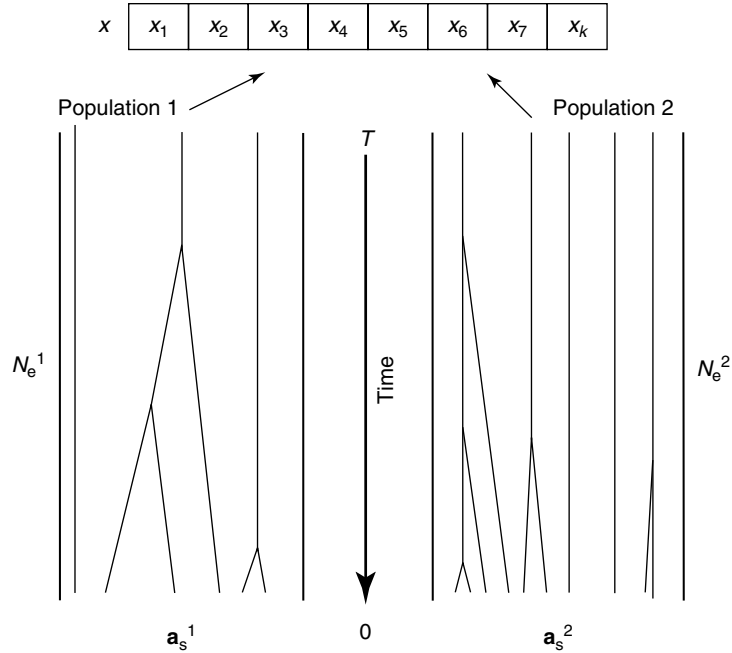
Figure 30.1 Illustration of the genealogy of samples taken from two different populations. The two populations are assumed to have been derived from a common ancestral population with allele frequencies $x_1, \ldots, x_k$. The gene frequencies in the two populations diverge over the time period of duration T. From a genealogical perspective this is illustrated as coalescences of lineages derived from the present day genetic samples with frequencies $\mathbf{a_s}^1$ in population 1 and $\mathbf{a_s}^2$ in population 2. The rates of coalescence depend on the population sizes N_e^1 and N_e^2.

It is possible to extend the methodology described here to test whether a model of population splitting and divergence through drift explains the data better than a model of gene flow. As discussed in (**Chapter 28**), the probability of the sample configuration $\mathbf{a_s}$ in a model of gene flow has a very simple description if an infinite island or continent-island model is assumed. In this case it is given by the multinomial Dirichlet

$$p(\mathbf{a_s}|M, n_s, k, \mathbf{x}) = \frac{\prod_{i=1}^{k} \binom{a_{si}+Mx_i-1}{a_{si}}}{\binom{n_s+M-1}{n_s}},$$

(Rannala and Hartigan, 1996; Balding and Nichols, 1995; 1997). As with the drift model, the assumption here is that the difference between the gene flow model and the isolation model is that in the latter case all lineages are available to coalesce until the time the population is founded whereupon the founder lineages have genotypes drawn at random from $\mathbf{x}$. By contrast in the gene flow model, looking backward in time, lineages emigrate out of the population (their genotypes being then drawn at random from $\mathbf{x}$) and are no longer available to coalesce. This corresponds to the 'scattering phase' of Wakeley (1999). Thus an assumption in this model is that the descendent lineages of each immigrant that
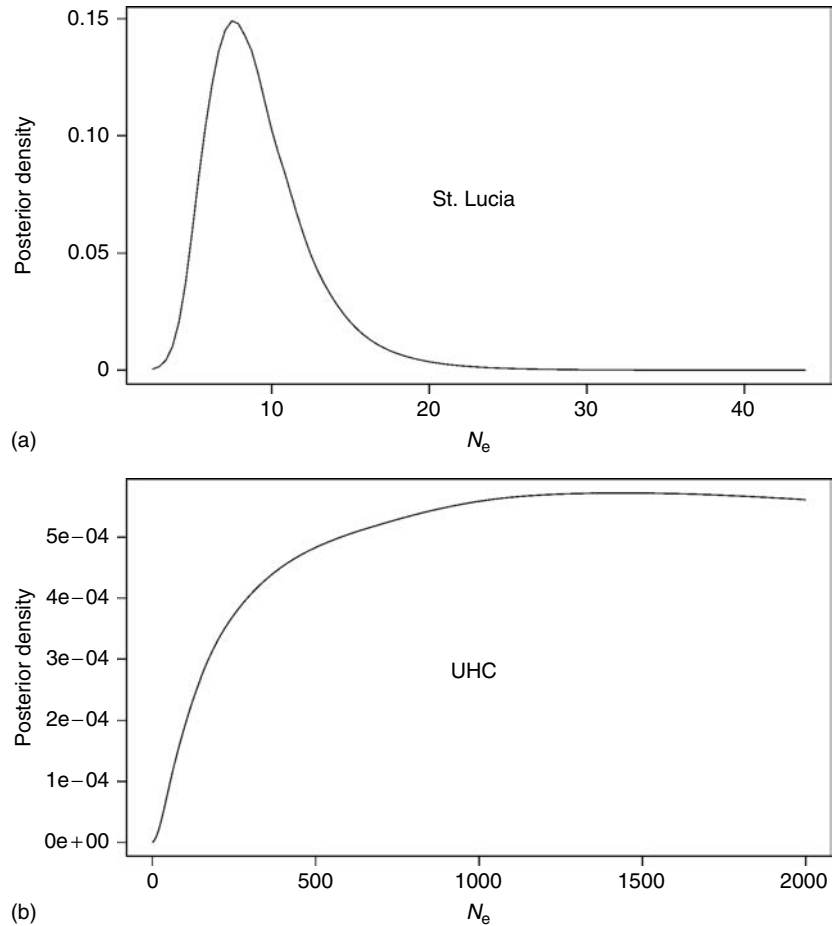
Figure 30.2 The marginal posterior distributions for N_e over a 16-year period, estimated from buffaloes sampled from the St. Lucia (a) and UHC (b) national parks in South Africa. Uniform priors are assumed for N_e. A rectangular prior was used for the UHC population, with an upper limit on N_e of 2000.

are present in the sample are all identical by descent (IBD), and that the population size of the external metapopulation, or continent is sufficiently large that **x** does not change between the times of the first and last immigration event. The differences between the two models are illustrated in Figure 30.3.

As described in Ciofi *et al.* (1999) it is possible to reparameterize both the drift model and gene flow model in terms of F, the probability that two lineages coalesce before the type of either lineage is drawn from **x** as $(1 - F)/F$ for M and $-\log(1 - F)$ for T/N_e. An indicator variable can then be introduced such that when $I = 0$ the probability of the sample configuration is calculated from the isolation-drift model, or when $I = 1$ the probability is calculated from the gene flow model. Thus it is possible to use $p(\mathbf{a_s}|I, F, n_s, k, \mathbf{x})$ to drive an MCMC scheme as described in Ciofi *et al.* to estimate the marginal posterior density $p(I|n, k, \mathbf{a_s})$. Simulations suggest that with 25 individuals,
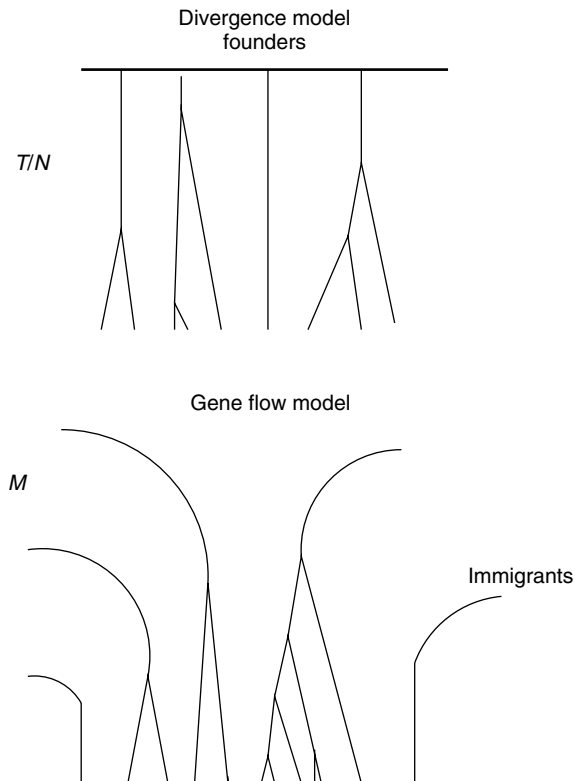
Figure 30.3   The genealogies of samples taken under a model of divergence (as in Figure 30.1) compared with those in a gene flow model. The symbols are defined in the text.

5 loci, 10 alleles per locus, and 5 populations, the two models can be distinguished relatively well (contribution by Beaumont in Balding *et al.* (2002)). Applications of the approach, implemented in the program *2mod*, are described in Ciofi *et al.* (1999), Goodman *et al.* (2001), Palo *et al.* (2001), and Hanfling and Weetman (2006). A method with related aims, but allowing for mutation as well as drift, is described in Nielsen and Wakeley (2001) and Hey and Nielsen (2004). In this case the model consists of a pair of populations that split from an ancestral population and there is gene flow between them. The data are assumed to consist of homologous copies of a single sequence (i.e. this is a one-locus model), and follow an infinite sites mutation model without recombination. Immigration and time of splitting can be jointly inferred using MCMC, thereby enabling the relative importance of each to be assessed.

Alternative methods for inference in these drift-based models are those by Wang (2003), using the Wright–Fisher model and the approximations discussed above, and also another approximation described by Nicholson *et al.* (2002), which is somewhat similar in spirit to that of Cavalli-Sforza and Edwards (1967). In the method of Nicholson *et al.* (2002), it is assumed that the data are frequencies of biallelic (SNP) markers, and the sample frequencies are a random draw from some unknown sub-population gene frequency, which is in turn a random draw from a distribution with mean corresponding to the baseline gene

frequency, p (shared between populations), and variance $cp(1 - p)$. The distribution is assumed to be Gaussian, and probability in the tails that extend outside (0,1) are given as atoms of probability on 0 and 1. The parameter c approximates T/N_e in the standard Wright–Fisher diffusion model, where T is the time of the split and N_e is the harmonic mean effective size over the interval. Using MCMC, Nicholson *et al.* (2002) are able to integrate out the unknown gene frequencies and obtain marginal posterior distributions for individual c_j for the jth population.

30.2.3 Estimating N_e Using One Sample

Often temporal information is not available for estimating short-term N_e and there have been attempts to estimate it using genotypic information from single samples. The essential idea here is that the gametes are sampled by a finite number of zygotes, which leads to small departures from the expected genotype frequencies both within and between loci. Thus, there are two possible approaches using genotypic data: to detect departures from Hardy–Weinberg equilibrium or to detect departures from linkage equilibrium. Both approaches have been studied and these will be briefly described here. A fundamental concern is that many other processes can cause departures from equilibrium. For example, admixture between two partially separated subpopulations can lead to departure from Hardy–Weinberg expectation (Wahlund effect) and cause substantial linkage disequilibrium. Furthermore, genotyping errors, particularly prevalent in microsatellites, can cause apparent substantial deviations from equilibrium proportions of genotypes, at least for particular loci, which can bias the results.

For a two-allele system with zygotes sampled from a base population with frequencies p and $1 - p$, the expected proportion of heterozygotes in a population of size N is $p(1 - p)/N$ in excess of that expected from an infinite population (Robertson, 1965; see also Cannings and Edwards, 1969). This result can be obtained by noting that N maternal gametes fuse with N paternal gametes to produce the zygote. The gene frequencies in these two classes of gamete will be slightly different, with frequencies p_f and p_m (having expectation p). Thus the expected proportion of heterozygotes is $p_f(1 + p_m) + p_m(1 + p_f)$. The expected proportion, based on the combined gene frequencies, $p_o = (p_f + p_m)/2$, is $2p_o(1 - p_o)$. Thus there is a heterozygote excess of $(p_f - p_m)^2/2$. Assuming that $E[p_f - p_m] = 0$, the expected value of the squared difference is given by the sum of binomial sampling variances, giving an expected heterozygote excess of $p(1 - p)/N$ (i.e. the expected frequency of heterozygotes is $2p(1 - p)(1 + 1/2N)$). Note that Crow and Kimura (1970), following Hogben (1946) and Levene (1949), suggest $2p(1 - p)(1 + 1/(2N - 1))$, instead, which can be obtained by assuming that all gametes contributing to the observed zygotes can potentially be joined together in a zygote.

Based on the Robertson (1965) result, Pudovkin *et al.* (1996) proposed an estimator that appeared to be unbiased. This has been investigated further by Luikart and Cornuet (1999) through simulation testing in which they aimed to identify the effect of sampling error and different mating systems on the robustness of the method. They concluded that in general the method slightly overestimated N_e, but this was generally less than 10 % for sample sizes of 30, and became smaller with larger sample sizes. The bias was largest (25 %) for extreme polygyny (1 male mating with 99 females), which is perhaps unsurprising, given that the method assumes equal male and female contributions. Overall, they note that relatively large numbers of loci are required because otherwise the confidence intervals

are typically very wide. The method has been developed further by Balloux (2004), who extended its use to subdivided population, and corrected some errors in the original treatment by Pudovkin *et al.* (1996).

The use of linkage disequilibrium to estimate N_e has been proposed by Langley *et al.* (1978), Laurie-Ahlberg and Weir (1979), and Hill (1981). There is an extensive and technical literature deriving approximate expectations of linkage disequilibrium statistics in terms of demographic parameters, stemming from the work of Hill and Robertson (1968) and Ohta and Kimura (1969). These approximation are discussed by Hudson (1985) and McVean (**Chapter 27**). The squared correlation in gene frequencies between two loci is defined to be

$$r^2 = \frac{D^2}{(1 - F_A)(1 - F_B)},$$

where $D^2 = \sum \sum (f_{ij} - p^A_i p^B_j)^2$ for loci A and B with alleles i, j, and $F_A = \sum (p^A_i)^2$, $F_B = \sum (p^B_j)^2$ (Hudson, 1985; see also McVean, **Chapter 27**). These definitions are based on known population gene frequencies. For samples (where usually the gametic phase of heterozygotes is unknown) there are several estimators available (see Hill, 1981; Weir, 1996).

In a finite population of N_e diploid individuals at equilibrium $E[r^2]$ tends to a steady state value that is approximately independent of the mutation model and mutation rate when $4N_e c$ is large, for recombination rate c. An approximation for this value, $E[r^2] \approx 1/(4N_e c)$, was first obtained by Hill and Robertson (1968) based on observations of simulations, and analytically by Ohta and Kimura (1969). Following these authors, most analytical approaches have tended to solve $E[r^2] \approx E[D^2]/E[(1 - F_A)(1 - F_B)]$. A number of other approximations have been obtained (see also McVean, **Chapter 27**), and the most widely used approximation is

$$E[\tilde{r}^2] = \frac{(1 - c)^2 + c^2}{2N_e c(2 - c)} + \frac{1}{n},$$

where $\tilde{r}^2$ is a sample estimate (Laurie-Ahlberg and Weir, 1979; Hill, 1981, suggest using the Burrows method in Cockerham and Weir (1977); see Weir and Hill, 1980, for other approximations for different estimators and mating systems).

Laurie-Ahlberg and Weir (1979) used the equation above to provide an estimator for N_e. Their analysis was based on 17 enzyme loci surveyed in nine laboratory populations of *Drosophila melanogaster*, from which they selected unlinked pairs of loci. Some of these pairs involved the same loci, and were thus not independent. In total they analysed 4–18 pairs among eight of the nine populations. The sample sizes were generally between 50 and 100. The census sizes fluctuated in these populations, but never exceeded 500–700, and could reach values as low as 10–12. With $c = 0.5$, Laurie-Ahlberg and Weir estimated $\hat{N}_e = m/3 \sum_i [(\tilde{r}^2_i - 1/n_i)]$. In four of the populations the denominator was negative, implying an estimated N_e of infinity. Of the remaining four populations $\hat{N}_e$ varied from 3 to 27. Laurie-Ahlberg and Weir made no attempt to give standard errors on these estimates.

Laurie-Ahlberg and Weir suggested that it is necessary for $n \gg \hat{N}_e$ to be able to estimate population size with some precision. For most taxa of conservation interest, linkage information will be unavailable, and, on the assumption that the markers are

randomly scattered through the genome (as with microsatellites identified through cloning, for example) the assumption of $c = 0.5$ would generally be reasonable.

Hill (1981) performed a similar analysis on other data sets, including linked pairs of markers. Hill suggested using information on variances to provide a weighted estimator across pairs of loci. He suggested that pairs of loci could be treated as uncorrelated. Simulation studies indicated that $\text{var}[\tilde{R}_i^2] \approx 2(\text{E}[\tilde{R}_i^2])^2$ (i.e. approximating χ^2 with 1 df), and Hill therefore suggested a weighted estimator

$$\frac{1}{\hat{N}} = \frac{\sum_i \alpha_i/\text{var}[\alpha_i]}{\sum_i 1/\text{var}[\alpha_i]} = \frac{\sum_i \gamma_i(\tilde{r}_i^2 - 1/n_i) \big/ \left(\frac{\gamma_i}{N} + \frac{1}{n_i}\right)^2}{\sum_i 1 \big/ \left(\frac{1}{N} + \frac{1}{\gamma_i n_i}\right)^2},$$

where $\alpha_i = (\tilde{r}_i^2 - 1/n_i)/\gamma_i$ and $\gamma_i = ((1 - c_i)^2 + c_i^2)/(2c_i(2 - c_i))$. The variance of $\hat{N}$ is approximated as $\hat{N} = \text{var}[\hat{N}] = 2N^2/\sum_i 1/(1 + N/(\gamma_i n_i))^2$. It is necessary to substitute estimates of N into these equations.

Hill analysed two *Drosophila* data sets: one based on chromosomes extracted from a sample of 198 flies isolated from a wild population, and another using a sample of around 700 flies from a caged population. In the first sample, 11 enzyme loci were scored, all of which were linked with most $c_i < 0.1$, and $\tilde{r}_i^2$ was measured in 25 pair-wise comparisons. The estimate of $\hat{N}_e$ was negative for these data. In the caged populations Hill analysed data from seven loci. The estimated census figure was around 1000. He used 9 pair-wise comparisons of linked markers (with all c_i less than 0.15) and 12 pair-wise comparisons of unlinked loci. The estimate of $\hat{N}_e$ was 363 with standard deviation 170.

Thus these studies appear to indicate that information on linkage disequilibrium has limited power to estimate effective population size. Even with large sample sizes and linked loci the standard errors on the estimates are fairly large, and negative estimates of N_e are common. Furthermore, as Hill (1981) points out, with very tightly linked loci, estimates of population size will be influenced by earlier demographic events, including gene flow. Also linked loci are only readily available in model organisms.

Notwithstanding these problems, Waples (1991) pointed out that if many pairs of loci are used, estimates of $\text{E}[\tilde{r}^2]$ can be made with reasonable precision. His own simulations confirmed the observation of Hill (1981) that pair-wise comparisons among k loci appeared to behave as independent data points with $\sum_{i=1}^{m} \tilde{r}_i^2$, for $m = k(k - 1)/2$, distributed as approximately χ^2 with m degrees of freedom. Recently it has been noted that the method of Hill (1981) appears to be quite biased when the sample sizes are low (England *et al.*, 2006) and Waples (2006) has developed a correction for this effect. Additional simulation-based tests by England *et al.* (2006) and Waples (2006) provide some encouragement that this general approach may be an effective method for estimating current population sizes.

The method has been used to estimate effective population sizes in managed or endangered populations (e.g. Bartley *et al.*, 1992; Bucci *et al.*, 1997). In the latter example, as part of a more detailed genetic survey, N_e was estimated in 5 Italian populations of the endangered pine *Pinus leucodermis* using 23 RAPD markers, assumed to be unlinked. The sample sizes were around 20–30 within each population. They estimated N_e using the method of Laurie-Ahlberg and Weir (1979), and estimated confidence limits for r^2 using the χ^2 approximation above, as suggested by Waples (1991), which were then transformed to give approximate limits on $\hat{N}_e$. Two of the 5 populations gave estimates

of infinity for N_e and the remaining three estimates were: 17 (11–31); 17 (11–30); 31 (14–650).

Thus, in conclusion, there appears to be scope for using multi-locus genotype information for estimating N_e with reasonable precision, and it may well be worthwhile investigating this approach further, perhaps using likelihood-based approaches to extract as much information as possible from the genotypic information. The development of methods to calculate approximate likelihoods and posterior distributions based on summary statistics (Beaumont *et al.*, 2002; discussed further below) may be helpful in this regard. If unlinked markers are used, low levels of gene flow may not affect the estimates too greatly, or, alternatively, can be included in an enhanced model.

30.2.4   Inferring Past Changes in Population Size: Population Bottlenecks

In order to obtain a better understanding of the genetic consequences of small size for population viability, it is useful to be able to identify whether a population has undergone a recent reduction in population size. For example inbreeding depression is assumed to be a transient phenomenon. A reduction in population size can inflate the frequency of rare deleterious recessives at some loci, leading to a reduction in population fitness, which is then 'purged' by natural selection. Thus populations that have had small sizes for a long time are unlikely to have high frequencies for highly deleterious genes (although mildly deleterious genes may be at high frequencies because of the strength of drift in small populations; Lynch *et al.*, 1995). Furthermore, if it is possible to model the time scale of the reduction in population size, one may be able to distinguish among different ecological explanations for the reduction. For these reasons the development of statistical methods to infer past changes in population size is useful in conservation biology. Although the focus in this section is on past reductions in size, the same methodology can also be used to detect population expansions, and this case will also be addressed.

A number of methods for detecting historical changes in population size have relied on discrepancies between different summary statistics that are used to estimate $\theta = 4N_e\mu$ assuming a stable population model, where N_e is the diploid population size and μ is the mutation rate. When the data do not fit a stable population model, the different
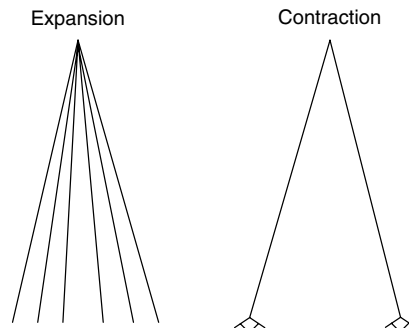
Figure 30.4   Two extreme examples illustrating how changes in population size affect the shape of genealogies. On the left is a genealogy that could only occur in the idealized case of a very small population that instantaneously increases to a size that is sufficiently large for no coalescences to occur during this interval. On the right is a genealogy that might be found in samples from a population undergoing a recent population contraction.

summary statistics give differing estimates of θ. This is best explained from a genealogical perspective (Figure 30.4) (see also **Chapter 25**). If populations have recently expanded, coalescences have a higher probability of occurring nearer the time of the most recent common ancestor than would be expected in a stable population. This is because the ancestral population size is smaller than the current population size, and therefore the coalescence rate is relatively higher. It should be noted that the schematic in Figure 30.4 is exaggerated: such a genealogy could only arise if there was a large change in population size over a very short period. For most reasonable demographic changes there are still many recent coalescences because the coalescence rate is almost quadratic in the number of lineages at any time (**Chapter 25**), and this tends to have a dominating effect on the shape of genealogies. This tendency for short recent coalescences is intensified when there is a population contraction. In this case there is a possibility that the most recent common ancestor occurs within the timescale of the contraction, or alternatively that some lineages 'escape' back into the ancestral population, which has longer coalescence times. In the example in Figure 30.4, two lineages are present in the ancestral population. Mutations occur along the lineages and thus for a sample from an expanding population most mutations have only a few descendants (a single descendent in the extreme example in Figure 30.4). Each gene in a pair differs by unique mutations. By contrast in a contracting population the majority of mutations have many descendants, and pairs of genes are either identical because of recent shared ancestry or differ by the same mutations. The differing distribution of mutations between these two scenarios alters the relationship, e.g. between summary statistics that are influenced by the total number of mutations in the genealogy and those that measure pair-wise differences between genes. Thus, conditioning on, say, measures of pair-wise divergence, there will be too many/too few alleles or segregating sites if the populations have expanded/contracted. Or, conversely, conditioning on the number of alleles or segregating sites, there will be too few/too many pair-wise differences if the populations have expanded/contracted.

There is an asymmetry in the detectable effects of population expansions and contractions. Contractions, by accelerating drift, produce instantaneous changes in the gene frequencies, whereas expansions do not. The only information about an expansion comes from mutations or recombination events that occur in the genealogical history over the time of expansion. Thus if the expansion has occurred very recently there will be little information in the data with which to detect it. Similarly if the population contraction is recent, although it should be easy to detect, without mutations there is no information on the shape of the contraction and the change in gene frequencies will depend only on the harmonic mean N_e over the interval. Thus the effects of a recent 'bottleneck' which traditionally means a contraction followed by an expansion, will be difficult to distinguish from any other model of recent population decline, and are considered together here.

For sequence data there are a large number summary statistics that have been devised to detect departures from the standard neutral model. Often these have been ascribed to the effects of selection rather than demographic history. This aspect is discussed in the chapters by Neuheuser (**Chapter 22**) and McVean (**Chapter 27**). A number of the tests are carried out by the Arlequin program (`http://lgb.unige.ch/arlequin/software/`) and the methodological section of the Arlequin manual provides a good description of the statistics, tests, and underlying assumptions.

One of the earliest tests for departures from neutrality (and also as a test for past changes in population size), is the test of Watterson (1978) based on the sampling theory

of Ewens (1972). This test generates the distribution of sample homozygosity (average pair-wise similarity), $\sum x_i^2$, from sample gene frequencies x_i, conditional on the observed number of alleles, assuming the standard neutral model. In a survey of many loci the observation of homozygosities that are extreme under the theoretical sampling distribution would then imply a departure from the model assumptions. If only one of the loci studied was strongly affected, it might be reasonable to infer that departure from neutrality at that locus was a likely cause. Alternatively, if many loci appeared to show the same extreme heterozygosities, then it may be more likely that an historic change in population size is the explanation. If the homozygosity calculated from the real data is too high in comparison with the theoretical distribution this would suggest a population expansion, and if too low it would suggest a contraction.

For DNA sequences one of the most commonly used statistics for detecting past changes in population size is Tajima's D (Tajima, 1989), which is based on the difference in estimates of θ, assuming an infinite allele model (IAM), from average pair-wise differences between sequences and from the number of segregating sites, scaled by an estimate of the standard deviation of the difference:

$$D = \frac{\pi - \frac{S}{a_1}}{\sqrt{\widehat{\mathrm{var}}\left(\pi - \frac{S}{a_1}\right)}},$$

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}.$$

π is the average number of pair-wise differences between sequences, which directly gives an estimate of θ (Tajima, 1983). S is the number of segregating sites in the sample, which is identical to the number of mutations in the genealogical history of the sample under the infinite sites assumption, and $\sum_{i=1}^{n-1} \frac{1}{i}$ is equal to half the expected total length of lineages in the genealogy, given the sample size, n, so that $E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$ (Watterson, 1975). The variance of the difference between the two estimators is given in Tajima (1989) as:

$$\widehat{\mathrm{Var}}\left(\pi - \frac{S}{a_1}\right) = \frac{S}{a_1}\left(\frac{n+1}{3(n-1)} - \frac{1}{a_1}\right) + S(S-1)\left(\frac{\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}}{a_1^2 + a_2}\right),$$

where

$$a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}.$$

Negative values of Tajima's D indicate population growth (too many segregating sites/too few pair-wise differences) whereas positive values of Tajima's D indicate population contraction (too few segregating sites/too many pair-wise differences). A significance test can be developed by simulating data sets under the coalescent with the same sample size and number of segregating sites as in the real data. An obvious caveat is that any test based on summary statistics such as Tajima's D measures departures from the standard

neutral model, assuming a particular mutation model, and not growth or decline *per se*. There are many ancestral processes, including variability in mutation rates among sites and selection, that can affect the value of the summary statistics.

Another summary statistic, similar in spirit to Tajima's D, is Fu's F_S (Fu, 1997). This is based on the discrepancy between the number of alleles observed in the sample and that predicted from Ewens' sampling theory for the IAM (Ewens, 1972), given $\hat{\theta}$ estimated as π, discussed above.

$$F_S = \log\left(\frac{S'}{1 - S'}\right),$$

where (from Ewens, 1972)

$$S' = p(k \geq k_O | \theta = \pi) = \sum_{k=k_O}^{n} \frac{S_k \pi^k}{S_n(\pi)},$$

and

$$S_n(\pi) = \pi(\pi + 1)\ldots(\pi + n - 1),$$

and S_k, an unsigned Stirling number of the first kind, is the coefficient of π^k in $S_n(\pi)$. If there has been population growth, there will be an excess of alleles (in particular, an excess of singleton haplotypes) given π, therefore S' will be small and F_S negative.

The statistical power of tests for population growth based on these summary statistics has been evaluated under a number of different scenarios using simulations by Ramos-Onsins and Rozas (2002). They found that tests based on F_S of Fu (1997), and another similar test developed by themselves, had a greatest power to detect population growth. Interestingly they found that tests based on the distribution of pair-wise differences (Rogers and Harpending, 1992) tended to be very conservative. The power of the tests to detect population decline has not been investigated.

A key point to reiterate is that these tests cannot in general be used to distinguish between departures from the standard neutral model caused by selection or by past changes in population size. A reasonable way to tackle this problem is to use many loci, on the assumption that selection will only affect a small proportion of loci, and if all loci are supporting the same tendency to depart from the standard neutral model it would be reasonable to ascribe this to the effect of past population history rather than selection. Until recently the markers that have been best suited for this are microsatellites because of their polymorphism and the relative ease of obtaining large numbers of loci. In conservation biology most of the tests for past population bottlenecks have been designed for microsatellites.

Based on these considerations Cornuet and Luikart (1996) proposed a test for bottlenecks using many loci in which coalescent simulations are performed to derive distributions of heterozygosities conditional on the observed number of alleles. Cornuet and Luikart consider two mutation models: the IAM, generally considered to be more appropriate for allozyme data; and the stepwise mutation model (SMM), often assumed for microsatellites (the distributed program, BOTTLENECK, Piry *et al.*, (1999), also incorporates the two-phase model (TPM) of Di Rienzo *et al.* (1994)). In the case of the IAM they used the method of Watterson (1978) described above, but looking at the distribution of heterozygosity (calculated as $\frac{n}{n-1}(1 - \sum x_i^2)$), conditional on the observed number of alleles. This is independent of θ (Watterson, 1978) and can be simulated by first creating a

sample genealogy and then adding mutations at a frequency proportional to branch length until the required number of alleles are obtained. In the case of the SMM, where θ does affect the distribution of heterozygosities, a uniform prior distribution of θ is assumed and the conditional heterozygosity is simulated by sampling θ uniformly and then simulating a genealogy with stepwise mutations, rejecting those that do not give rise to samples with the required number of alleles (see Cornuet and Luikart, 1996, for specific details of how to sample θ efficiently).

Cornuet and Luikart propose two methods for using the simulated heterozygosities to test for significant departures of the observed data from the null hypothesis. Firstly they propose a type of sign test whereby each locus is scored as having a heterozygosity above or below its expected value. For each locus it is possible to estimate by simulation the probability of obtaining a heterozygosity greater than the expected value. From this one can obtain a probability distribution of observing $l = 0 \ldots L$ loci with heterozygosities greater than the expected value, and thereby estimate the probability of observing at least l_O loci with greater than expected heterozygosity. The other test Cornuet and Luikart propose is to calculate deviations of the observed heterozygosity from the simulated mean scaled by the simulated standard deviation in heterozygosity. Under the null hypothesis, the sum of these should approximate normal deviates with mean 0 and variance equal to the number of loci. Cornuet and Luikart (1996) carry out a number of tests of their method using simulated data sets. They suggest that their test is most powerful with loci evolving according to the IAM compared to the SMM, and find that the second test is somewhat more powerful than the first, although the difference is not large. If the test is applied to microsatellites, where there is abundant evidence from pedigree information that the IAM does not apply, it would seem reasonable never to cite results assuming the IAM because of the tendency to type-I error, and a case could be made for only citing results for the TPM model, using parameters that are reasonable, given current pedigree information.

Other summary statistics can also be monitored in microsatellites. For example, Luikart *et al.* (1998) suggest that the shape of the frequency spectrum can provide a useful graphical test of a bottleneck. In the case of microsatellites the distribution of lengths holds useful information on the past demographic history (Reich and Goldstein, 1998; Reich *et al.*, 1999). For example, as noted by Cornuet and Luikart (1996) the distribution of lengths becomes more uneven with gaps when the population has been subject to a bottleneck. This is exploited in the test of Garza and Williamson (2001), who monitor the distribution of the statistic $M = k/r$, where k is the number of alleles, and r is the difference in length between the longest and shortest microsatellite allele in the sample. In a population that has a history of recently reduced population size, or bottleneck, there are expected to be relatively few alleles for a given value of r, giving a small M. The mean value, $\bar{M}_O$, is calculated across loci and compared to relevant quantiles of simulated values of $\bar{M}$ obtained from coalescent simulations corresponding to the same number of loci, and assuming a TPM model. The critical values depend on the parameters chosen for the TPM model. If the mutation process allows larger than single-step changes in length then these can give rise to allele frequency distributions that are similar to those that arise from population contraction.

An additional summary statistic that has been used to detect population expansions from microsatellite is the expansion index of Kimmel *et al.* (1998). This has been shown by King *et al.* (2000) to be more sensitive to population growth than summary statistics based on the kurtosis of the length distribution or the variance among loci of the variances

in microsatellite length (Di Rienzo *et al.*, 1998; Reich and Goldstein, 1998; Reich *et al.*, 1999). As with the summary statistics discussed earlier it is based on the discrepancy of two different estimators of θ in a standard neutral model. For microsatellites evolving according to a strict stepwise model one estimate of θ is $\theta_V = 2V$, where V is the sample variance in allele length. Another estimate is $\theta_f = (1/F^2 - 1)/2$ (Ohta and Kimura, 1973), where F is the sample homozygosity. The expansion index of Kimmel *et al.* (1998) is then

$$\log(\beta_1) = \log(\hat{\theta}_{\bar{V}}) - \log(\hat{\theta}_{\bar{F}}).$$

King *et al.* (2000) also suggest

$$\log(\beta_2) = \frac{1}{L} \sum_{i=1}^{L} \left((\log(\hat{\theta}_V))_i - (\log(\hat{\theta}_F))_i \right).$$

In the first case the two estimates of θ are calculated from the mean homozygosity and variance across or loci, and hence $\log(\beta_1)$ depends only on the marginal distribution of these statistics. In the second case the two estimates of θ are calculated for each locus separately, and $\log(\beta_2)$ depends on the joint distribution of homozygosity and variance across loci. King *et al.* (2000) show that the latter statistic is the most sensitive. Although devised to detect population growth, the imbalance index should also be useful for detecting population contractions or bottlenecks.

In addition to performing tests to detect departures from the null model of historically constant population size, it is possible to estimate parameters in specific models of population growth and/or decline. Primarily this has been done by the method of moments (e.g. Rogers, 1995; Reich and Goldstein, 1998; Reich *et al.*, 1999; Schneider and Excoffier, 1999). Increasingly, however, likelihood-based methods are being used. It is beyond the scope of this chapter to consider these in detail. Beaumont (2003b) gives a description of some of these approaches from the point of view of human demographic history, and general likelihood-based approaches are described in **Chapter 26**.

An example from an analysis of microsatellite data is given in Figure 30.5, which comes from Beaumont (1999), using a similar approach to that of Wilson and Balding (1998) and Wilson *et al.* (2003). The general idea is that the probability of the data and a genealogy consistent with the data, as a function of some unknown demographic and genetic parameters (such as mutation rate), can be derived from coalescent theory: $P(G, S|\Phi)$. MCMC is then used to obtain $P(\Phi, G|S)$ by sampling over genealogies and parameter values in the demographic model. In the example illustrated, the method has been applied to a sample from the last remaining population of the Northern Hairy Nosed Wombat, studied by Taylor *et al.* (1994). The species is believed to have undergone a sharp decline in the last 120 years. Sixteen loci were genotyped in a sample of 28 individuals. Seven loci were monomorphic and one locus had length differences that were not consistent multiples of two, leaving eight polymorphic loci. All the loci were polymorphic when samples from the Southern Hairy Nosed Wombat were included, suggesting that none of the monomorphic loci had unusually low mutation rates. Two analyses were performed with the data, one using the polymorphic data only and the other using the combined set of 15 loci. The results illustrated in Figure 30.5 strongly support a substantial population bottleneck. The population size was around 70 when the samples were taken, and even if one assumes that there were only 10 breeding adults, the data would suggest the
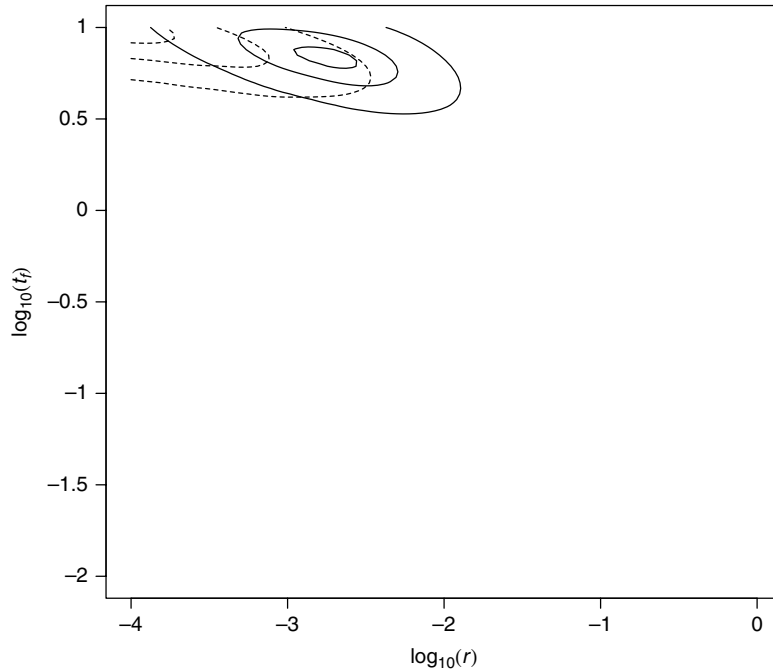
Figure 30.5 The joint posterior density for a parameter measuring the duration over which a population of wombats has been changing in size and a parameter measuring the degree to which it has changed. Specifically on the y axis is the logarithm to base 10 of the number of generations since the population started to change in size divided by the current population size, and on the x axis is the logarithm to base 10 of the ratio of current population size to ancestral population size. Contours giving the 10, 50, and 50 % highest posterior density intervals are shown. The dotted lines show the density for the polymorphic loci only.

decline began at least 1000 years ago, and that there has been a thousand-fold to ten thousand-fold decline in numbers. These conclusions will be highly model dependent, as discussed earlier, because of the difficulty in estimating the trajectory of a bottleneck without mutational information. An additional caveat is that the assumption of the SMM is restrictive, and inferences about reductions in population size will be affected by the presence of mutations that cause greater than single-step changes in allele length. Recently Calmet (2003) has modified the likelihood-based method of Beaumont (1999) to include the TPM, which corrects for the confounding effects of the mutation process.

It can be seen that there is a significant ascertainment bias associated with the use of the polymorphic markers alone. The effect is counter-intuitive in that the polymorphic markers suggest a more severe contraction over a longer time scale. Essentially, as discussed in Beaumont (1999), the polymorphic markers, while supporting a bottleneck effect imply a larger number of mutations in the ancestral population than the monomorphic loci. The number of mutations depends on $2N_1\mu = \theta/r$, and thus, for a given prior on θ, smaller values of r imply larger numbers of mutations in the pre-bottlenecked population. The ascertainment effect is much stronger for SNP markers, as discussed in Wakeley *et al.* (2001).

In conclusion, there appears to be considerable potential for the use of genetic data to make inferences about parameters in demographic models. However, a key caveat is whether one can have confidence in the models. In particular, as emphasised by Wakeley (1999), models with population structure can mimic many effects of past changes in population size without any change in census size. It is quite possible that most inferences about changes in population size are strongly confounded with phylogeographic effects.

30.2.5 Approximate Bayesian Computation

A problem with genealogical inference is that the parameter space is potentially very large and is not necessarily well explored by Monte Carlo methods, leading, e.g. to problems of convergence with MCMC. This will be exacerbated by the large volumes of genetic data that may soon be generated, even for non-model organisms. Motivated by difficulties in analysing human data a number of approximate approaches have been developed (**Chapter 26**). One of these, approximate Bayesian computation (ABC), has been used extensively for genetic analysis problems in conservation and management because it is relatively straightforward to model the complex scenarios that frequently arise in applied problems (as in, e.g. Miller *et al.* (2005)).

A number of related techniques have been developed (reviewed in Beaumont *et al.* (2002)). In particular Weiss and von Haeseler (1998) developed a method whereby a vector S_O of length d summary statistics are measured from the data and then n coalescent simulations are performed for a fixed parameter value Φ. For each simulation we measure summary statistics $S_{1,...,n}$, and the proportion of simulations that give rise to summary statistics sufficiently close to those measured from the data are recorded. This provides an approximate Monte Carlo estimate of the likelihood of $P(S = S_O|\Phi)$. If the parameter values are explored on a grid a likelihood surface can be estimated. Weiss and von Haeseler (1998) used this approach to detect patterns of population growth and also decline in different human populations from mtDNA sequence data.

A Bayesian approach taken by Pritchard *et al.* (1999) is to simulate the parameter values from the prior, $\Phi_i \sim P(\Phi)$, and then simulate data sets with these parameter values, and measure the summary statistics, $S_i \sim P(S|\Phi)$, to obtain samples from the joint distribution $P(S, \Phi)$. The posterior distribution $P(\Phi|S = S_O)$ is given by the conditional density

$$p(\Phi|S = S_O) = \frac{P(S_O, \Phi)}{P(S = S_O)}.$$

As with Weiss and von Haeseler (1998) this is estimated by choosing to accept simulations that give values of S_i that are close to S_O within some tolerance. Pritchard *et al.* (1999) apply the method to infer population growth from haplotype frequencies of linked microsatellite loci in the human Y chromosome. They used three summary statistics: the number of distinct haplotypes, mean heterozygosity, and mean variance in repeat length. The latter two were motivated by the study of Kimmel *et al.* (1998), discussed above. A problem for these methods is that only a limited number of summary statistics can be used for accurate inference because fewer points will be accepted for a given level of tolerance as the number of summary statistics is increased. The method of Pritchard *et al.* (1999) has been used by Estoup *et al.* (2001) to explore a relatively complex demographic and phylogeographic model of the history of introductions of cane toads to islands in the Caribbean and Pacific.

Beaumont *et al.* (2002) have pointed out that the method can be made more efficient by treating it as a problem of conditional density estimation, and using non-parametric regression methods to substantially increase the number of points used in the estimation of the posterior density, thereby allowing many more summary statistics to be used. Here, S_O and S are scaled so that each summary statistic in S has unit variance. The aim is to estimate the posterior mean $\alpha = \mathrm{E}(\Phi|S = S_O)$ (see, e.g. Ruppert and Wand, 1994, for background to the approach). Beaumont *et al.* (2002) use least squares to minimize

$$\sum_{i=1}^{n} \left\{ \Phi_i - \alpha - \beta^{\mathrm{T}}(S_i - S_O) \right\}^2 K_\delta(\| S_i - S_0 \|),$$

where $\| x \| = \sqrt{\sum_{i=1}^{d} x_i^2}$, and $K_\delta()$ is a weight given by an Epanechnikov kernel. This is a quadratic function that has a maximum at $\| S_i - S_0 \| = 0$, and is zero for $\| S_i - S_0 \| \geq \delta$ Standard weighted regression gives the estimates $\hat{\alpha}$ and $\hat{\beta}$. In order to estimate posterior densities, Beaumont *et al.* (2002), took a heuristic approach, in which they make an assumption that the errors are constant in the interval and adjust the parameter values as

$$\Phi_i^* = \Phi_i - (S_i - S_O)^{\mathrm{T}}\hat{\beta}.$$

Scripts written in R are available to infer parameters using this method (`http://www.rubic.rdg.ac.uk/~mab/stuff/ABC_distrib.zip`).

An alternative ABC approach that does not require simulation from the prior, which will typically be rather inefficient, is that of Marjoram *et al.* (2003), which uses a rejection approach as in Pritchard *et al.* (1999) embedded within an MCMC sampler in which the likelihood ratio calculation is replaced by an accept/reject step depending on whether simulated summary statistics lie within a tolerance range (see also **Chapter 26**). The method of Beaumont *et al.* (2002) has been applied to a number of problems in population genetics, including some in conservation and population management (Estoup *et al.*, 2004; Chan *et al.*, 2006; Miller *et al.*, 2005).

30.3 ADMIXTURE

Often, closely related taxa that were previously isolated from each other, have been brought into contact as a consequence of recent habitat disturbance. Hybridisation occurs when individuals from different taxa mate together and produce offspring, and this may be followed by introgression when there is back-crossing into either population. Introgressive hybridisation, and its detection is a significant area of concern in conservation genetics (Haig, 1998), and can be studied at three levels. At the highest level one can ask questions based on gene frequencies in separate populations, the subject of this section. At the next level one can identify hybrid individuals through genotypic modelling, discussed in the following section. Lastly, at the lowest level, hybrid individuals can be identified by inferring degrees of relatedness between individuals and the reconstruction of pedigrees, described in the final section.

The early development of statistical methods to infer admixture proportions was stimulated by an interest in human populations (e.g. Glass and Li, 1953). Two populations

diverge from a common ancestor and then hybridize in a single event (Figure 30.6). The current samples are taken some time after this event from the known descendants of the two parental populations and the hybrid population. The main aim of the statistical analyses is to estimate the admixture proportion μ. This has been called an *intermixture model* by Long (1991) to distinguish it from a gene flow model where the admixture occurs on a longer time scale. The various statistical methods that have been developed to estimate μ have made different simplifying assumptions about the parameters in the model. In general, other than the approach of Bertorelle and Excoffier (1998), the relationship between the parental populations has been ignored. Most approaches have modelled the gene frequency, x, of a particular allele in the hybrid population as $x_h = \mu x_1 + (1 - \mu)x_2$, where the indices refer to the hybrid, parental 1, and parental 2 populations respectively. Obviously, if there are many alleles at many loci, no single value of μ will explain all the allele frequencies and a maximum likelihood or least-squares estimate of μ is made where the variance in x_h from its expected value is modelled according to different assumptions. In the approaches of Glass and Li (1953) and Elston (1971) genetic drift is assumed absent, sampling error is assumed absent in the parentals, and all the variation is due to sampling error in the hybrids. The approach of Thompson (1973) takes into account drift in all populations as well as sampling error. That of Long (1991) separately estimates both drift and sampling error in the hybrid population, but assumes that the parental frequencies are known without drift or sampling error. Long suggests an approach to take into account sampling (but not drift) error in the parental frequencies, but this is not used in the most widely implemented version of Long's method, that of Chakraborty *et al.* (1992).

Thompson (1973) provides the framework for an analysis of the admixture model including drift in all populations. The gene frequencies are first arcsine square root transformed, $x' = \sin^{-1}(\sqrt{x})$. Under this transformation drift can be modelled as a Brownian-motion diffusion with variance $1/(8N_e)$ per generation, where N_e is the effective number of individuals, assumed to be diploid. In the case of k alleles at a locus, rather more complex transformations and projection to a $k - 1$ dimensional space can be used (see discussion in Thompson, 1975a). At the time of the admixture event the vector of transformed gene frequencies in the hybrid population is given as $\mathbf{x}'_h = \mu \mathbf{x}'_1 + (1 - \mu)\mathbf{x}'_2$. Following drift, the observed sample vectors are $\mathbf{a}_h$, $\mathbf{a}_1$, $\mathbf{a}_2$, the component elements of which are normal random variables with mean x_i and variance $T/8N_e + 1/8n$ for sample size n. Thus it is possible to estimate $p(\mathbf{a}_h, \mathbf{a}_1, \mathbf{a}_2 | \mathbf{x}'_h, \mu, \mathbf{x}'_1, \mathbf{x}'_2)$. Thompson (1973) uses this to obtain conditional likelihood functions for μ. This model has also been studied using a coalescent-based method similar to that discussed in Section 30.2 (Chikhi *et al.*, 2001), and is described in slightly more detail later.

The method of Long is based on a weighted least squares solution of $\mathbf{a}_h = \mu \mathbf{x}_1 + (1 - \mu)\mathbf{x}_2 + \varepsilon$, where ε has zero mean and separate components of variance due to sampling and drift. The weighting accounts for the inhomogeneity of the variances, and is solved iteratively in Long (1991). Chakraborty *et al.* (1992) provide a closed form expression:

$$\hat{\mu} = \frac{\sum_{l=1}^{L} \sum_{i=1}^{r_l} (a_{1li} - a_{2li})(a_{hli} - a_{2li})/a_{hli}}{\sum_{l=1}^{L} \sum_{i=1}^{r_l} (a_{1li} - a_{2li})^2/a_{hli}},$$

for L loci with r_l alleles at each locus. This has mean square error

$$\text{MSE} = \frac{\sum_{l=1}^{L} \sum_{i=1}^{r_l} [(a_{hli} - a_{2li}) - (\hat{\mu} a_{1li} - a_{2li})]^2 / a_{hli}}{\sum_{l=1}^{L} r_l - 1}.$$

Long (1991) suggests that since the sample size is known, drift in the hybrid population can be separately estimated as $\hat{F} = (\text{MSE} - 1/2n)/(1 - 1/2n)$ where F has the usual definition (drift variance)/$(x_h(1 - x_h))$. An estimate of T/N_e can be made using the Wright's formula $F = 1 - \exp(-T/(2N_e))$.

Another approach to inferring admixture proportions, with relevance to conservation, is that by Bertorelle and Excoffier (1998). They consider a model where two parental populations, each of size N chromosomes, themselves diverged from a common population over a time period τ, hybridize in one event at time t_A from the present to produce a hybrid population of size N, with μN chromosomes from one population and $(1 - \mu)N$ chromosomes from the other.

The essential approach is to define the expected admixture proportion as a function of the ratios of expected coalescence times between groups of genes. Genetic distances defined for the genetic marker that is being studied have expectations that depend only on the product of coalescence time and mutation rate. Assuming a constant mutation rate, the ratios of coalescence times can be replaced by ratios of expected genetic distances. An estimator can be obtained by substituting the expected distances by the estimated distances. Bertorelle and Excoffier define two estimators, one of which is described here. They derive the expected coalescence times for a pair of genes, one sampled from the hybrid population and one sampled from a parental. In this case, there can be no coalescences since the time of admixture. There are two possibilities: with probability μ the hybrid lineage is derived from the same parental populations as that being compared, and therefore the expected coalescence time is $t_A + t_1$, where t_1 is the expected coalescent time in population 1; alternatively, with probability $1 - \mu$, the hybrid lineage comes from the other parental, in which case the expected coalescence time is t_{12}. Two equations can be obtained, one for each of the parental populations in the pair, and a least-squares estimate for μ can be obtained. The expected times, t_1, t_2 and t_{12} are replaced by the estimated genetic distances d_1, d_2 and d_{12}, on the assumption that for both microsatellites and sequences the typical genetic distances have the form $d = 2\nu t$, where ν is the mutation rate and t is the expected coalescence time. Bertorelle and Excoffier denote this estimator m_Y.

It should be noted that this involves an estimate of t_A which is in general unknown. Bertorelle and Excoffier suggest replacing the estimate of $(2\nu)t_A$ by the minimum observed distance in pair-wise comparisons of genes between either parental and the hybrid population. However, in practice this is almost always equivalent to assuming $t_A = 0$. For multi-locus data, Bertorelle and Excoffier suggest that the distances should be calculated as averages over loci, rather than separately calculating estimates of μ for each locus before averaging.

Using molecular distances, they make comparisons between m_Y and the estimator of Roberts and Hiorns (1965), which they denote m_R, and that of Chakraborty et al. (1992; based on Long, 1991, discussed earlier), denoted m_C. Their overall finding is that in general m_Y is much less biased than either m_R or m_C, but tends to have a higher variance. When the parental populations have been separated for a substantial period and when the mutation rate is high the variance in m_Y tends to approach that of the other estimators.

However, with single microsatellite loci, the variance is still substantially larger than that of the other estimators. With many microsatellite loci (the situation commonly found in practice), the variance becomes close to that of the other estimators with negligible bias, while the other estimators remain appreciably biased towards admixture proportions of 0.5.

Bertorelle and Excoffier apply their method to microsatellite data from grey wolf, coyote, and red wolf populations obtained by Roy *et al.* (1994). They use the data from the red wolf, and use data from populations believed to be respectively pure wolf, pure coyote, wolf hybrid, coyote hybrid. They then infer admixture proportions in the wolf hybrid, coyote hybrid, and red wolf populations. They suggest that estimates of sampling error in m_Y, m_R and m_C can be made by bootstrap resampling chromosomes independently across loci. This is equivalent to adding an extra inbreeding step into each population, and it is more appropriate to resample among loci, conditioning on the observed frequencies at each locus. This procedure will take into account both the genealogical and sampling variance. For reasonable sample sizes the bulk of the variance in the estimators will be between loci.

For the coyote and wolf hybrid populations they obtain estimates of admixture that are consistent across estimators – ~ 0.5 in the case of the wolf hybrid populations and ~ 0.1–0.15 in the case of coyote hybrid populations. The bootstrap standard error in the case of the coyote hybrid population is sufficiently large that the estimate is consistent with the true admixture proportion of zero. Thus, for these known hybrid populations the results support the conclusions of Roy *et al.* (1994).

The method of Bertorelle and Excoffier (1998) has been extended to include multiple parental populations (Dupanloup and Bertorelle, 2001). Essentially the estimation procedure described above for m_Y, is generalized to allow estimates of multiple μ_i, where $\sum_{i=1}^{d} \mu_i = 1$. Standard errors for the m_{Yi} are estimated using the same bootstrapping procedure as described above. They evaluate the performance of the m_{Yi} using simulations and conclude that despite the increased number of parameters the level of precision in the estimates remains approximately the same as in the original study with two parental populations. A general improvement to the method of Bertorelle and Excoffier (1998) for sequence data, assuming an infinite sites model, has been described by Wang (2006). He derives formulae assuming an infinite sites model for the expected number of segregating sites in each population, in each pair of populations pooled together, and in all 3 populations pooled together. The actual numbers are counted in the data and the method of least squares is used to estimate parameters. This method appears to provide more accurate point estimates for the parameters in the general admixture model of Figure 30.6.

Chikhi *et al.* (2001) analyse the same model as that of Thompson (1973) using the coalescent method described in Section 30.2.2. Importance sampling can be used to estimate the likelihood $p(\text{data}|\mathbf{x}_1, \mathbf{x}_2, \mu, T/N_1, T/N_2, T/N_h)$ and then marginal posterior densities (e.g. $p(\mu|\text{data})$, $p(T/N_h|\text{data})$, etc.) are estimated with MCMC. Although the use of the likelihood method should allow for increased accuracy and precision the method is highly computer intensive and no simulation-based study of precision and accuracy has been published to allow comparison with earlier studies. On an example data set of gene frequencies obtained from Jamaican people, who are assumed to be of mixed European and African descent (Parra *et al.*, 1998) to the point estimates for μ were similar to those obtained by the method of Long (1991), but with estimates of error that were more than twice as great (assuming the posterior distribution is sufficiently close to normal

so that Bayesian and frequentist estimates are comparable). By estimating the degree of drift subsequent to admixture for all subpopulations it does provide extra detail in the analysis, unlike commonly used methods. The posterior distributions for T/N are, however, generally broad.

As discussed in Section 30.2.1, an alternative way of computing the likelihoods is to use the Wright–Fisher model. Wang (2003) has inferred the parameters discussed above using the same approach as in Wang (2001). In addition, like Bertorelle and Excoffier (1998), and unlike Chikhi *et al.* (2001), who used uniform Dirichlet priors for the parental gene frequencies, Wang (2003) explicitly models the genetic divergence between the parental populations by introducing additional scaled parameters T_A, N_{A1}, and N_{A2}, for the ancestral divergence time, and the ancestral population sizes, using a uniform prior for the common ancestral gene frequencies. Wang demonstrates that the assumption of independent uniform priors for the parental gene frequencies is quite problematic for most scenarios in which there is a high degree of divergence between the ancestral populations (the ideal case for detecting admixture). In this case there is tendency for estimates of the admixture proportion to be biased away from 0.5 towards 0 or 1. In a detailed comparison of the performance of different methods for inferring admixture, Choisy *et al.* (2004) noted a similar phenomenon, which could be partially corrected by using Dirichlet $(1/K, \ldots, 1/K)$ priors for the parental gene frequencies, as in Rannala and Mountain (1997), discussed further below. The prior of Wang (2003) is computationally intensive, and an alternative would be to use a Dirichlet distribution, parameterized by F_{ST} as discussed in Sections 30.2.2 and 30.4–i.e. Dirichlet $(\theta_j x_1, \ldots, \theta_j x_K)$, where $\theta_j = 1/F_{ST_j}$ for the jth parental population and $x_1, \ldots, x_K$ are the ancestral allele frequencies for alleles $1, \ldots, K$. The method of Wang (2003) is substantially faster than that of Chikhi *et al.* (2001), and, because of this, extensive simulation testing is possible, and Wang demonstrates that the likelihood-based approach has superior performance, generally, than moment-based estimators.

The likelihood-based models described thus far are models of drift without mutation. They assume the ancestral variation is partitioned among the descendent populations solely through drift and admixture. This is due to the convenience of modelling–e.g. it allows relatively straightforward analytical expressions for the likelihood in Wang (2003). In order to allow for mutations over the period since the common ancestral population, full genealogical models would have to be considered. The most straightforward implementation would be to use MCMC, or to compute likelihoods using importance sampling (**Chapter 26**). To date, no such methods have been developed for the specific class of intermixture models considered here. A Bayesian approach has been introduced for microsatellite data, however, using the ABC technique discussed above. In this case, 15 summary statistics are computed from the data and compared with those obtained from simulations. These comprise: for all 3 populations, numbers of alleles, heterozygosity, a modified version of the M statistic of Garza and Williamson, and pair-wise F_{ST} for all pairs; the $(\delta\mu)^2$ statistic of Goldstein *et al.* (1995) for the two parental populations; the linkage disequilibrium estimator, D' for the hybrid population; and the m_Y statistic (above). They use the SIMCOAL2 program (Laval and Excoffier, 2004) to infer the parameters in the full model of Figure 30.6. When the times since admixture are short (i.e. when the assumptions of drift-based models are best met) their method has only marginally weaker performance than that of Wang (2003), and substantially improved performance when they are longer.
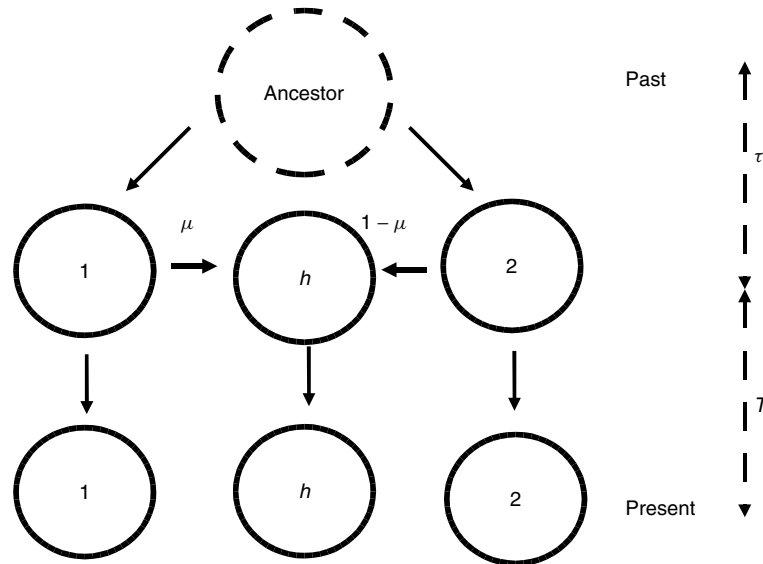
Figure 30.6 A figure illustrating the model of admixture described in the text. The two parental populations (1 and 2), diverge from a common ancestor over a period τ and then hybridize to produce a population h, with a proportion μ of alleles coming from population 1. These populations then remain distinct and diverge from each other over a period T without intervening migration.

30.4   GENOTYPIC MODELLING

In contrast to the approach described in the previous section, where gene frequencies are used to infer admixture, a relatively new area of modelling is to use multi-locus genotypic information to detect individuals that are immigrants or have immigrants in their recent ancestry. Thus, in effect, the method will only work if there is appreciable gametic disequilibrium within a population caused by gene flow. An influential paper in the context of conservation was that of Paetkau *et al.* (1995; see also Waser and Strobeck, 1998). They studied populations of polar bears in Canada and were interested in knowing whether populations managed as independent units were in fact connected through dispersal of individuals. The approaches that have been taken to address questions such as this have led to several overlapping areas of research that are outlined in the following sections.

30.4.1   Assignment Testing

In order to identify migrant individuals Paetkau *et al.* (1995) made an estimate of the gene frequencies at each locus in each of the populations from which they had samples. For each individual they calculated the joint probability of obtaining its multi-locus genotype from each of the populations by multinomial sampling, $\prod p(\mathbf{a}_j | \hat{\mathbf{x}}_j, n_j)$, with observed genotype at the jth locus, $\mathbf{a}_j$, and estimate of gene frequencies $\hat{\mathbf{x}}_j$. The individual was then 'assigned' to the population in which this probability was highest. The rationale behind this approach is that, in the absence of immigration, the genotype of each individual is a random draw from its own population gene frequencies and therefore, providing gene

frequencies vary sufficiently among populations, individuals should be assigned to their own populations. Migrants can be detected because they have a higher probability of being drawn from some other population. In the study of Paetkau *et al.* (1995), 4 populations of polar bears were sampled, using eight microsatellite loci. They found that 60 % of individuals were assigned to the population in which they were sampled, 33 % to the nearest population and 7 % to more distant populations. In a later study on brown bears from seven populations in NW Canada using the same loci (Paetkau *et al.*, 1998), 92 % were assigned to the population from which they were sampled, and those assigned to other populations illustrated biologically plausible patterns of dispersal.

Estimation of the population gene frequencies is an important part of the procedure, and in the method of Paetkau *et al.* in order to avoid zero elements in $\hat{\mathbf{x}}$, the genotype of each individual is incorporated into the estimate of $\hat{\mathbf{x}}$ for each population, when the assignment test is made. Alternative methods may be preferable, and, in particular, it would be useful to estimate $\hat{\mathbf{x}}$ jointly with the assignment parameters. In the initial study no attempt was made to estimate confidence intervals, but resampling schemes have latterly been considered (see `http://www2.biology.ualberta.ca/jbrzusto/Doh.php` for more details).

Some of these points are addressed in a study by Rannala and Mountain (1997). Their approach differs from that of Paetkau *et al.* in that they include Bayesian estimation of the allele frequencies in each population, and a likelihood ratio test to compare different hypotheses. They consider a number of populations with many loci sampled. For ease of exposition, one locus and two populations are considered here. They condition on the number of alleles, k, observed at a locus among all populations. The observed allele frequency counts are given by the vectors $\mathbf{a}_0$ and $\mathbf{a}_1$ for the two populations. The posterior distribution for the unknown gene frequency vector $\mathbf{x}$ is given by $p(\mathbf{x}|\mathbf{a}_0)$ and $p(\mathbf{x}|\mathbf{a}_1)$. To estimate this a Dirichlet prior is assumed with parameters all equal to $1/k$ and written $D(1/k, \dots 1/k)$, giving a posterior $p(\mathbf{x}|\mathbf{a}_0) = D(\mathbf{a}_0 + 1/k)$ and $p(\mathbf{x}|\mathbf{a}_1) = D(\mathbf{a}_1 + 1/k)$. Rannala and Mountain describe their prior as assigning equal probability density to the frequencies of the alleles, although this only occurs with a Dirichlet $D(1, \dots 1)$. Clearly the choice of prior may depend on the genetic model assumed.

Under a model where none of the individuals are immigrants or have immigrant ancestry, their genotypes are assumed to be multinomial samples of size 2 taken, e.g. from the posterior $p(\mathbf{x}|\mathbf{a}_0)$, defined above. Integrating out $\mathbf{x}$, the marginal probability of an individual having genotype $\mathbf{X} = (X_i, X_j)$ is then multinomial-Dirichlet,

$$p(\mathbf{X}|\mathbf{a}_0) = \int p(X|\mathbf{x}) p(\mathbf{x}|\mathbf{a}_0) \, \mathrm{d}\mathbf{x}.$$

The probabilities for each locus can then be multiplied together, under the assumption that they are independent.

Other hypotheses can be considered, e.g. whether the individual has ancestry from one immigrant d generations earlier. The probability of observing genotype $\mathbf{X} = (X_i, X_j)$ in an individual where one allele is drawn at random from one population and one is drawn at random from the other population, is $p(\mathbf{X}|\mathbf{a}_0, \mathbf{a}_1) = 1/2(p(X_i|\mathbf{a}_0) p(X_j|\mathbf{a}_1) + p(X_i|\mathbf{a}_1) p(X_j|\mathbf{a}_0))$, where the probabilities are calculated as single draws from the multinomial Dirichlet outlined above.

Rannala and Mountain (1997) extend the argument above to consider the probability of observing the genotype, $p(\mathbf{X}|\mathbf{a}_0, \mathbf{a}_1, d)$, given that an individual might have one ancestor,

d generations ago, from a different population (with all other ancestors coming from the same population). There is a probability of $1/2^{d-1}$ that the parent which is immigrant or has immigrant ancestry contributes an immigrant gene, in which case the probability is as given above, and there is a probability $1 - 1/2^{d-1}$ that both genes come from the same population in which case the probability is calculated as a sample of size 2 from the multinomial Dirichlet, $p(\mathbf{X}|\mathbf{a}_0)$ or $p(\mathbf{X}|\mathbf{a}_1)$.

Rannala and Mountain suggest that for particular individuals, hypotheses can be tested using likelihood ratio tests of the form

$$\Lambda = \frac{p(\mathbf{X}|\mathbf{a}_0)}{p(\mathbf{X}|\mathbf{a}_0, \mathbf{a}_1, d)}.$$

Critical regions of a given size can be estimated by parametric bootstrapping–i.e. Monte Carlo simulations of genotypes $\mathbf{X}$ with probability $p(X|\mathbf{a}_0)$ under the null hypothesis. The power of the tests can be estimated by additionally simulating under the alternative hypothesis–genotypes $\mathbf{X}$ with probability $p(\mathbf{X}|\mathbf{a}_0, \mathbf{a}_1, d)$.

A comparative analysis of the performance of the two approaches on test data sets has been carried out by Cornuet *et al.* (1999). They introduce the use of a distance-based method for assigning individuals to populations. The method is analogous to that of Paetkau *et al.* but individuals are assigned to populations with the smallest genetic distance. Cornuet *et al.* study the most commonly used genetic distances for microsatellite or allozyme data (**Chapter 29**), modified to take into account that individuals are compared with populations. In addition to the question of assigning individuals to particular populations, Cornuet *et al.* also consider the question whether an individual is likely to have come from the population in which it resides. In order to test the latter, which they term *testing for exclusions*, Cornuet *et al.* suggest simulating genotypes at random from an estimate of the sample population gene frequencies and comparing the likelihood (or genetic distance) of the individual's genotype with the distribution of likelihoods or distances from random sampling.

Cornuet *et al.* tested the methods using data simulated from a model of diverging populations with mutations according to either an IAM or SMM. Their overall conclusion was that the Rannala and Mountain method outperformed all other methods, and the genetic distance methods, in general, performed less well than the methods using likelihood. Since Rannala and Mountain's calculation of likelihoods differs from that of Paetkau *et al.* only in the estimation of population frequencies $\mathbf{x}$, this implies that, taking a Bayesian approach, the choice of a suitable prior distribution for $\mathbf{x}$ is an important consideration. Tests for exclusions also appear to work well using the Rannala and Mountain method to calculate likelihoods. However, Cornuet *et al.* noted that the method used for simulating null distributions was an important component of testing–with increasing numbers of loci there was a tendency to incorrectly exclude individuals from all populations.

30.4.2 Genetic Mixture Modelling and Clustering

One aspect of the method of Rannala and Mountain, discussed above, is the necessity for a large number of hypothesis tests, requiring some care in specifying the critical values for assessing significance. Also it would be better if the estimates of the population frequencies $\mathbf{x}$ could take into account the possibility that some individuals are immigrant

rather than resident. An additional feature, common to these approaches (but examined by Cornuet *et al.*), is that hypothesis testing is limited to the populations that are actually included in the survey, whereas immigrants may come from other, unsurveyed, populations.

These aspects are addressed in a study by Pritchard *et al.* (2000), which has resulted in the very widely used program, STRUCTURE. They initially consider the problem in terms of one sample composed of individuals of possibly heterogenous origin, where the genotype of each individual is a mixture of alleles drawn from an unknown number of contributing populations with unknown gene frequencies in each population. The aim is to infer jointly: (1) the number of populations contributing to the genotypes of individuals; (2) the allele frequencies in each of these populations; (3) the proportions from each of these populations contributed to the genotype of each individual. This can be regarded as a standard mixture problem in statistics (Robert, 1996), often analysed using Bayesian MCMC methods, and Pritchard *et al.* use similar approaches to solve it.

In this approach, it is assumed that there are K populations contributing to the gene pool of the sample population. Pritchard *et al.* consider two main models: in one case each individual is drawn from one of the K populations; whereas in the more general case a proportion of the individual's alleles are drawn from each of the K populations. The former situation is a special case of the latter, and this will be described here. The K potential source populations each have an unknown gene frequency distribution at each locus, p_{kl} for the $k = 1 \ldots K$ populations and $l = 1 \ldots L$ loci. The p_{kl} are elements of some multidimensional vector $\mathbf{P}$. For the ith individual it is assumed that the proportion of its genotype that is drawn from population k is q_{ik}, an element of the vector $\mathbf{Q}$. For the purpose of analysis they introduce an indicator vector $\mathbf{Z}$ where the element z_{ail} gives the population from which allele copy $a = 1, 2$ at locus l in individual i originates, and can take any value from $1 \ldots K$. This use of indicator variables (missing data) is widely used in Bayesian mixture modelling with Gibbs sampling (e.g. Robert, 1996). The genotype of each individual is given by vector $\mathbf{X}$ with elements x_{ail}. It is assumed that the alleles at each locus in each individual are drawn independently. With this formulation, the likelihood $p(\mathbf{X}|\mathbf{Z}, \mathbf{P}, \mathbf{Q})$ can be calculated straightforwardly.

The interest is in estimating the posterior distribution $p(\mathbf{Z}, \mathbf{P}, \mathbf{Q}|\mathbf{X})$ for the parameters of interest. In order to estimate this, priors are required for $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{Z}$. They assume a Dirichlet prior for $\mathbf{P}$, as in Rannala and Mountain (1997), but a uniform $D(1, \ldots 1)$ rather than (for J alleles at a locus) the $D(1/J, \ldots 1/J)$ prior of the latter authors. They assume a prior of $p(z_{ail} = k) = 1/K$ for $\mathbf{Z}$. For $\mathbf{Q}$ they model the prior hierarchically as a Dirichlet with form $D(\alpha, \ldots \alpha)$. When α is small the genotype of an individual is drawn almost completely from one population, and when α is large the genotype is evenly mixed from all the populations. The prior for α is uniform on [0,10]. The posterior distribution of α gives evidence of the degree to which there is heterogeneity among individuals in their ancestry.

Generally they wish to estimate $p(\mathbf{P}, \mathbf{Q}, \alpha|\mathbf{X})$ (marginal to $\mathbf{Z}$, which is introduced to make the Gibbs sampling tractable, see Robert, 1996). This is performed by successively sampling from the full conditional distributions $p(\mathbf{P}, \mathbf{Q}|\mathbf{X}, \mathbf{Z})$, and $p(\mathbf{Z}|\mathbf{P}, \mathbf{X}, \mathbf{Q})$ and updating α using Metropolis–Hastings sampling. The conditional distributions and sampling methods are given in Pritchard *et al.* (2000).

The estimation of K, the number of populations contributing to genotype distributions among the samples, is more problematic. Pritchard *et al.* develop an *ad hoc* procedure

for estimating $P(\mathbf{X}|K)$ using the sample mean and variance of the Bayesian Deviance $-2 \log p(\mathbf{X}|\mathbf{Z}, \mathbf{P}, \mathbf{Q})$ over values of $\mathbf{Z}, \mathbf{P}, \mathbf{Q}$ generated by the Gibbs sampler. An assumption of the method is that the sampled values of $p(\mathbf{X}|\mathbf{Z}, \mathbf{P}, \mathbf{Q})$ are lognormal, and Pritchard *et al.* argue that it should be viewed as an approximate guide, which appears to give satisfactory answers with test data. With a suitable prior for K it is then possible to estimate $P(K|\mathbf{X})$.

An additional point that Pritchard *et al.* note is that since the population labels are arbitrary, they can be switched in $\mathbf{Z}, \mathbf{P}, \mathbf{Q}$ without affecting the likelihood. Thus $p(\mathbf{Z}, \mathbf{P}, \mathbf{Q}|\mathbf{X})$ has $K!$ symmetric modes. If the Gibbs sampler explores these fully then, for example, a summary statistic such as the posterior mean of the q_{ik} will have expectation $1/K$ irrespective of the amount of substructure in the data. In their analyses the Gibbs sampler appears to explore only one mode, and thus they find that posterior mean values and 95 % credible intervals are satisfactory summaries of the posterior distribution. It is reasonable to assume that when there is less information on population structure label-switching will be more frequent, and this will in turn make detection of population structure more difficult if based on such summaries.

The method of Pritchard *et al.* can also be used to answer questions about the ancestry of individuals in a similar way to the method of Rannala and Mountain (1997). In order to avoid restructuring of the model, the approach they have taken is to specify that individuals have resident ancestry with fixed probability v and non-resident ancestry with probability $1 - v$. If they have non-resident ancestry, they can be themselves immigrants or have one immigrant parent, or one immigrant grandparent, etc. This is equivalent to using geographic information as a prior to restrict the possible values the q_{ik} can take. The Gibbs sampler can then be run as before, but the q_{ik} can only take a restricted set of values that can be mapped back to each of the above possibilities. In this way, by interpreting the q_{ik}, it is possible to estimate the posterior probability of each type of ancestry, whether immigrant or non-immigrant, and from which population. The method, however, relies on the user specifying in advance a particular value of v. The method offers an improvement on the approach of Rannala and Mountain in that it avoids the problems of multiple hypothesis testing.

Dawson and Belkhir (2001) have developed a method to partition a sample into subgroups within which there is random-mating, similar to the model without admixture analysed by Pritchard *et al.* (2000). There are, however, a number of differences between the two approaches. Following the notation of Pritchard *et al.* (2000), used above, Dawson and Belkhir calculate the likelihood function $p(\mathbf{X}|\mathbf{Z}, \mathbf{P}, K)$, and use Metropolis–Hastings sampling to estimate $p(\mathbf{Z}, \mathbf{P}, K|\mathbf{X})$. The prior for $\mathbf{Z}$ is such that all partitions have equal prior probability, and the prior for $\mathbf{P}$ is a Poisson–Dirichlet distribution, which is appropriate for the infinite alleles model, and is parameterized by θ, which is equivalent to the scaled mutation rate. A point prior value of θ was used in their examples. Unlike the method of Pritchard *et al.* (2000) the number of partitions, K, is allowed to vary from 1 to some upper value, and within this interval the prior for K is given by a power function $p(K) = Au^K$ for $0 < u \leq 1$. The main concern in Dawson and Belkhir (2001) is the problem of label-switching discussed above. Although Pritchard *et al.* (2000) are careful to highlight this, it still remains the case that in their method the MCMC 'works' only because it does not converge. This seems somewhat unsatisfactory and Dawson and Belkhir attempt to avoid this by recording the proportion of times during the MCMC simulation in which groups of individuals co-occur together within

the same partition. The posterior probability of co-assignment of d individuals to the same partition forms a measure of similarity among the d individuals, which can then be used in an agglomerative clustering algorithm to maximize the minimum probability of coassignment within clusters. In practice, Dawson and Belkhir use $d = 2$, which corresponds to the 'furthest-neighbour' or 'complete-linkage' clustering algorithm. They find that the resulting tree is a more useful graphical summary to identify cryptic population structure than the posterior distribution for K, which tends to overestimate the number of partitions, and appears to be affected by the value chosen for θ. They suggest that a future improvement would be to put a more diffuse prior on θ.

30.4.3 Hybridisation and the Use of Partially Linked Markers

The method of Pritchard *et al.* has been used to analyse hybridisation of wildcats with domestic cats in Scotland (Beaumont *et al.*, 2001). Applying the method, implemented in the program STRUCTURE, to wild-living cats in Scotland there is very strong evidence of two groups. In this analysis samples from known domestic cats were also included, but these were set to have $q_i = 0$ (where the subscript k has been dropped and q_i is the proportion of non-domestic ancestry of individual i). The point estimates of the q_i and their 95 % credible intervals are illustrated in Figure 30.7 for the wild-living cats. Museum specimens are shown as dotted lines in the figure. Interestingly, when logit transformed, these point estimates correlate strongly with the results from the first axis of classical metrical scaling of genetic distances between cats. As discussed in Beaumont *et al.* (2001), given that the linkage disequilibrium caused by admixture will decay rapidly for unlinked genes, the existence of a second group of cats does not necessarily mean that a 'pure' wildcat population with no introgression has been identified. It only suggests that this group of cats shows little evidence of very recent domestic cat ancestry.
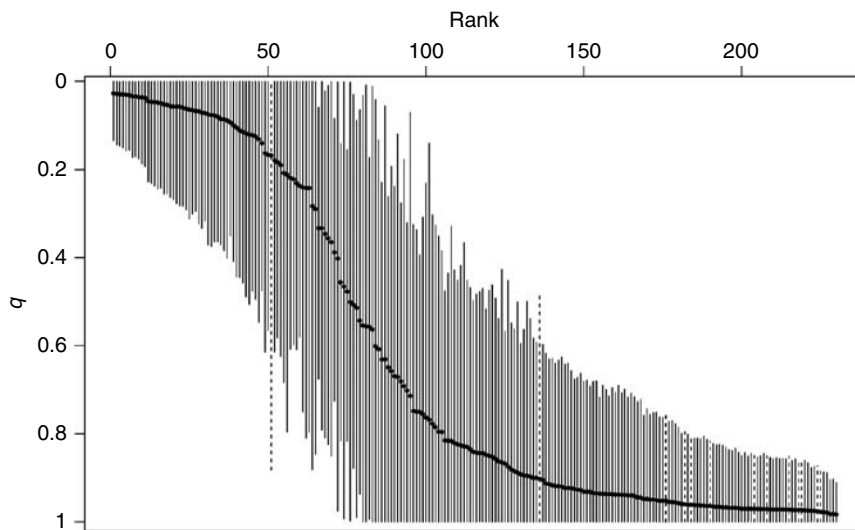
Figure 30.7 The figure shows the means and 95 % credible intervals for estimates of the probability that individual wild-living cats have purely non-domestic ancestry. These are plotted against the rank of the mean estimate. Vertical dashed lines refer to museum specimens.

Motivated by considerations such as this, Falush *et al.* (2003) have extended the basic approach in the program STRUCTURE to allow the use of closely linked (but recombining) markers. A further improvement lies in the priors used for the parental gene frequencies, which follow the Dirichlet approximation for modelling immigration-drift equilibrium. This allows additionally for F_{ST} to be inferred in each population, as in, e.g. the *2mod* program of Ciofi *et al.* (1999) discussed above. To allow for recombination, in their extension of the method, they modify **Z** so that the z_{ail} are no longer independent at each locus, but form a Markov chain, with neighbouring loci tending to originate from the same population with probability inversely related to the map distance between markers and the total rate of recombination events, r, per unit distance. In this model r can be loosely equated to the time since admixture. The idea is that the indicator for the first locus is drawn according to **Q** above and then the indicator for the second locus is either the same as this if there is no recombination event, or is drawn randomly from **Q** if there is an event. Although this approach has not been used to analyse the Scottish wildcats it has been used to look at wildcats in Hungary (Lecis *et al.*, 2006). In this case the results suggested that the admixture was probably ancient, but could not be dated accurately. More definite results come from an analysis of wolf populations in Italy (Verardi *et al.*, 2006). In wolves there is the prospect of historical admixture with domestic dogs, and the aim of this study was to elucidate how much had occurred, and over what period. In this case they were able to obtain clear estimates of admixture time of $70(\pm 20)$ generations (around $140-200$ years), although the overall amount of admixture was quite low. This study illustrates the potential power of genotypic methods to enhance frequency based analysis of population structure and hybridisation.

The use of multilocus genotypic information to make inferences about recent migration has great potential, and a number of papers have used methods similar to those of Pritchard *et al.* (2000) to study different aspects of population structure. For example, Anderson and Thompson (2002) have developed a method for inferring the proportion of hybrids in a sample. In their method they consider hybridisation over n generations. It is then possible to partition genotype frequencies into classes corresponding to the pedigree of the individuals–e.g. when $n = 2$, whether they are purely of one species or the other, or F1 hybrids, or F2 hybrids, or back-crosses. For each individual the posterior probability of belonging to one of these genotype frequency classes can be computed using MCMC, and hence non-admixed individuals can be detected (at least, given the n generations of admixture). There are similarities between these methods and those used for genetic stock identification (GSI), where often the population affinities of individuals is of less interest than some estimate of the degree of admixture. A Bayesian method specifically for GSI, using MCMC, has also been developed by Pella and Masuda (2001). Corander and Marttinen (2006) discuss the general problem of quantifying the extent of hybridisation and admixture in populations, and propose an alternative approach.

30.4.4   Inferring Current Migration Rates

As noted above, the method of Pritchard *et al.* (2000) also allows calculation of the posterior probability that an individual has a specified degree of immigrant ancestry. In order to calculate this a point prior is specified for v, the probability that an individual has resident ancestry. Taking this approach further, it is possible to use genotypic methods to infer parameters in a demographic model, such as the current levels of gene flow between populations. The model of Wilson and Rannala (2003) assumes that all populations in

the system have been sampled, and that the individuals each have at most one immigrant ancestor. The data consist of the genotypes (X) and sampling locations of individuals (S), and the parameters are the population source of the immigrant ancestor for each individual (M), the number of generations back in the past that the immigrant ancestor occurred (t), the probability that two alleles in an individual are IBD from a recent ancestor (F), leading to a departure from Hardy−Weinberg equilibrium, and the population frequencies (p). The prior for M and t for each individual can then be written as a function of the immigration rate, m_{lq}, the proportion of individuals in population q that are immigrant from population l. It is this hierarchical parameter that is of most interest. MCMC is then used to obtain the posterior distribution of the parameters. Thus the method can identify the immigrant ancestry of particular individuals, but it also, for example, constructs a matrix with point estimates of current migration rates between populations. An example where this approach has been used is in the study of movement between Orang-utan populations in Borneo (Goossens *et al.*, 2006). In this case the model of Wilson and Rannala (2003) was used to demonstrate that there was very little current gene flow between populations on two sides of a river. Goosens et al note that in more general comparisons the method appeared to converge well when the migration rate is low, but convergence was more problematic on data sets with higher migration rates.

The method of Wilson and Rannala (2003) need not be closely correlated with, e.g. pair-wise F_{ST}, both because the latter takes some time to equilibrate and also because in a migration model F_{ST} is a function of the product of the effective size, N, and migration rate m. Thus it should be possible to obtain estimates of N given information on the genetic divergence between populations. Interestingly it is not necessary to use a likelihood framework to make such estimates. Vitalis and Couvet (2001) have developed method-of-moments estimators to infer effective size and immigration rate using single locus and two locus measures of identity by state.

30.4.5 Spatial Modelling

A common application of some of these genotypic methods has been to discover clusters of individuals, with a view to better understand population structure (see Excoffier and Heckel, 2006, for a general review of software for performing these analyses). Typically the data are based on individuals of known origin, and the idea is to then apply, e.g. the program STRUCTURE of Pritchard *et al.* (2000) to see if it identifies interesting groupings (see Corander *et al.*, 2003, for an alternative approach to this problem). Thus inference of parameter K, the number of clusters, becomes a focus of interest, and this is rather problematic, since it is only estimated through an approximation that may not always be valid, although further modifications have been suggested (Evanno *et al.*, 2005). These considerations have motivated the development of methods that identify the number of groups through model-selection and also incorporate explicit spatial information into the genotypic models. These spatial models tend to borrow heavily from the techniques of geostatistics.

One example of this is the study by Guillot *et al.* (2005), who propose a model similar to that of Pritchard *et al.* (2000), with K populations, and with the aim of inferring K. However the prior for K is structured around a Voronoi tessellation. Here, a series of points, the number, m, of which is drawn from a Poisson distribution with parameter λ are located uniformly at random over a rectangle, covering the geographic area of interest. Around each point it is possible to draw a convex polygon that contains the

region of the rectangle that is closer to that point than to any other point. The rectangle can be divided into m such non-overlapping regions. This is a Voronoi tessellation, and, given K populations, each tile of the tessellation is assumed to be assigned uniformly at random to one of the K populations. One implication of this model is that, although the tessellation provides a spatial element, it only specifies loosely, via the interaction of K and λ, how the space is subdivided. Thus, e.g. if K is 2 and λ is high the two populations are distributed as a mosaic through the region, whereas if λ is low the region is likely to be dominated by two main blocks. In contrast to the method of Pritchard *et al.* (2000), reversible jump MCMC (Green, 1995) is used to infer the posterior distribution of K. The method appears able to be able to recover accurately spatial genetic discontinuities in simulated data sets. In an application of this method (implemented in the program GENELAND) to roe deer populations in France (Coulon *et al.*, 2006) the method could find weak evidence of spatial discontinuity into two populations, north and south of a region including a highway, canal, and river running close together. In contrast, analysis with *Structure* suggested that there was only one population.

An alternative spatial model, specifically for assigning individuals to populations (also possible with *Geneland*), is presented in Wasser *et al.* (2004), which includes an application of their method to the problem of locating the geographic origin in Africa of elephants from their DNA samples. Theirs is primarily an assignment method, similar to that of Rannala and Mountain (1997). However, they incorporate a spatial model for the population frequencies. Importantly also they explicitly allow for genotyping errors in the microsatellite data that they analyse. The frequencies f_{jlk} for allele j at locus l in population k are modelled logistically $f_{jlk}(\theta) = \exp(\theta_{jlk})/(\sum_{j'} \exp(\theta_{j'lk}))$, and the θ is modelled by a Gaussian process. A Gaussian process is one that generates random variables, any sample n of which are drawn from a n dimensional multivariate Gaussian distribution (and hence any linear combination of these random variables is also Gaussian). The mean vector and covariance matrix of this distribution are determined by the problem under consideration. For spatial problems, as here, this is more generally called a *Gaussian random field*. In the case of the spatial model here, for the same allele at the same locus in two different locations k and k' the covariance depends on d the distance between the two locations and they use the function given by $(1/\alpha_0) \exp[-(\alpha_1 d)^{\alpha_2}]$. Wasser *et al.* (2004) further extend the model by allowing a location W to be specified so that genetic samples can be assigned to locations that are not specified in advance. This is implemented using MCMC, allowing the posterior distribution of W to be computed. With this method Wasser *et al.* analysed samples from 399 individuals, and demonstrated that they were able to make accurate spatial assignments. The implication of this study is that in future, DNA from ivory samples could be used to determine their provenance, and thereby help control the trade in ivory.

30.5 RELATEDNESS AND PEDIGREE ESTIMATION

It is tempting to predict that the genotypic methods discussed above will in future become less distinct from those based on the kin structure of populations. The latter aim to estimate the relatedness of pairs of individual, test hypotheses about degrees of relationships between individuals, and more generally uncover the recent pedigree of a sample. If such methods could be embedded into a population model, then it would be

feasible to unite frequency-based methods for uncovering demographic history with the genotypic approaches. For example, recently methods have been developed for partitioning a sample of individuals into groups according to their degrees of relationship (e.g. sib, half-sib, or non-sib groups), and in this regard the methods are similar to the multilocus techniques described above. Currently, analysis of relatedness has been used to estimate genetic components of variance of phenotypic trait in outbred populations (e.g. Mousseau *et al.*, 1998), and thus can potentially be used in conservation to quantify the amount of phenotypically important genetic variation in endangered populations (Storfer, 1996; Carvajal-Rodríguez *et al.*, 2005; Thomas, 2005). Furthermore information about the kin structure of populations can help inform conservation decisions (Regnaut *et al.*, 2006).

There have been a number of recent reviews in this area (Blouin, 2003; Thomas, 2005; Oliehoek *et al.*, 2006; see also **Chapters 23** and **24**), and further relevant information is available in the chapters on pedigree analysis in the current volume. The aim of this section is to briefly introduce the general concepts and mainly follows the treatment in Thomas (2005) and Oliehoek *et al.* (2006).

An early interest has been to estimate the coefficient of relatedness, r. In a large random mating population

$$r = 2\theta = \phi/2 + \Delta, \tag{30.4}$$

where θ is the coefficient of coancestry or kinship between two individuals, and is the probability that an allele taken at random from a locus in one individual is IBD with an allele taken at random in another individual, ϕ is the probability that at a particular locus the two individuals have exactly one allele each i.e. IBD, and Δ is the probability that the two individuals have exactly two alleles that are IBD. The only other possibility is that no alleles are IBD. With inbreeding, it is of course possible that the alleles within individuals are IBD, leading to more complex possibilities, but those are typically not considered, by making the assumption of a large random mating population. Different degrees of relationship have different expected values for r, ϕ and Δ. The coefficient of relatedness can be viewed as the proportion of alleles in one individual that are IBD with those in another.

There are a number of different ways to estimate r. For example a simple method of moments estimator can be constructed from the same considerations that are used in the estimation of F_{ST} (Ritland, 1996): a pair of randomly chosen alleles, one from individual x and the other from individual y, has a probability of being identical in state (IIS), s_{xy}, given by $s_{xy} = \theta + (1 - \theta)s$, where s is the probability of choosing two identical alleles at random from the population (i.e. the sample homozygosity). This can be rewritten as

$$\theta = \frac{s_{xy} - s}{1 - s}. \tag{30.5}$$

As with estimators of F_{ST}, this invites the substitution of s_{xy} and s by estimates from the data. Various versions of this type of estimator, with different weighting schemes, are discussed in Ritland (1996), Lynch and Ritland (1999), and Oliehoek *et al.* (2006). It should be noted that Lynch and Ritland (1999) suggest that the specific estimator derived above (their equation 10) has poor sampling variance in comparison with others. It is given here solely because of its simplicity of derivation. As with F_{ST}, less biased estimates will be obtained by averaging the numerator and denominator across loci before evaluating the fraction. A more sophisticated class of moment estimators uses a similar approach to

that above to estimate the parameters in (30.4) (Lynch and Ritland, 1999; Wang, 2002). A widely-used method of moments estimator, derived using rather different considerations, is that of Queller and Goodnight (1989), which has often been used to estimate the relatedness of groups of individuals. It can, however, be used to estimate r for a pair of individuals, but in this case it is necessary to use multi-allelic loci because the estimator is undefined for heterozygotes when there are biallelic loci. A recent comparison of methods based on empirical data with known pedigrees and also on simulated data suggests that the method of Lynch and Ritland has generally superior performance among these estimators (Csillery *et al.*, 2006). This is supported, at least for panmictic populations, by the study of Oliehoek *et al.* (2006), but they also observed that for structured populations a modified version of (30.5) was more accurate.

Likelihood-based methods have also been developed (Thompson, 1975b). Maximum likelihood estimates tend to have marginally improved mean square error performance but can be biased with small sample sizes (Milligan, 2003; Thomas, 2005). The basic form of the likelihood calculation that has been used (following Thomas, 2005) is:

$$P(G|\phi, \Delta) = \prod_l \left\{ (1 - \phi - \Delta) P_{[0]}(g_l|x) + \phi P_{[1]}(g_l|x) + \Delta P_{[2]}(g_l|x) \right\},$$

where g_l is the vector of the four allele types at a particular locus (in this example, which assumes independence across loci, it is assumed that the alleles in an individual are unordered with respect to which parent they were inherited from), G is the genotype across loci, x is the baseline frequency. Note that this will typically be estimated from the data in a separate computation, although in principle it could be imputed by MCMC, as in the program STRUCTURE. The indices [0], [1], [2] refer to the number of alleles that are IBD, giving, e.g. $P_{[1]}((\{a, b\}, \{a, c\})|x) = 2 p_a p_b p_c$. Point estimates can then be obtained by maximising the likelihood (Milligan, 2003). One use of the likelihood approach has been to estimate heritability in outbred populations (Mousseau *et al.* 1998). For each pair of individuals the likelihood of the joint genotypes and phenotypes can be calculated as a function of the degree of relationship and parameters in a quantitative genetic model, which can then be inferred. Typically, since the computations are based on all pairs of individuals, these give composite likelihood estimates and the confidence intervals have to be computed through bootstrapping.

Recently, moving beyond these pair-wise analyses, there has been an increase in likelihood-based approaches for uncovering the kin-structure of populations. Methods have been developed for calculating the likelihood for groups of individuals that share two parents or none (Painter, 1997). In this case it is equivalent to finding a simple partition of the sample into full sib families and single individuals. Alternatively it is possible to identify groups that share two parents (full-sibs), one parent (half-sibs) or none (Thomas and Hill, 2002; Wang, 2004). The *Parentage* program of Wilson (in Emery *et al.* (2001)) directly assigns individuals to parents. It is possible to specify the genotypes of known potential parents, and impute the genotypes of missing parents. The methods of Wang (2004) and Emery *et al.* take care to deal with the possibility of genotyping error, which can have serious consequences for such analyses. Hadfield *et al.* (2006) have developed a method, extending that of Emery *et al.* that allows for assignment of parents but also the estimation of a number of parameters relating to social structure, such as the size of male territories.

An ideal goal is to be able to reconstruct the pedigree of a sample from genotypic information (Steel and Hein, 2006). Parentage assignment (Emery *et al.*, 2001; Hadfield *et al.*, 2006) can be regarded as construction of a pedigree back one generation in a discrete-generation model, with a population size given by the sum of the number of observed and missing parents, and with a prior distribution for the genotypes of the missing parents, as well as for family sizes and mating structure as in Hadfield *et al.* (2006). Such a system could then in principle be taken back an arbitrary number of generations. The pedigree can be regarded as a more detailed genealogical model for a diploid sexual organism than the coalescent, and, as with the coalescent, it may often be a nuisance parameter to be integrated out while inferring parameters related to demography and life-history. The most complete approach to date in this regard is that of Gasbarra *et al.* (2007), which is based on a population-based model of pedigrees constructed backwards in time (Gasbarra *et al.*, 2005). The method of Gasbarra *et al.* (2007) uses a highly complex MCMC updating scheme, and can be applied to both unlinked and partially linked genetic data. The pedigree is constructed back to some pre-specified point in time. Pedigrees are sampled from their posterior distribution, given the genotypic data. It is also possible to use the method to infer the phase of multi-locus data. Such a method has enormous potential in conservation genetics, both in being able to infer social structure as well as demographic parameters, such as immigration rates.

Acknowledgments

Related Chapters

Chapter 22; **Chapter 23**; **Chapter 24**; **Chapter 25**; **Chapter 26**; **Chapter 27**; **Chapter 28**; and **Chapter 29**.

Websites

Below is a list of websites containing programs that perform some of the analyses described in this chapter.

- Estimation of effective size and migration using temporally spaced samples; admixture; relatedness:
  `http://www.zoo.cam.ac.uk/ioz/software.htm`
- Estimation of effective size; changes in population size; migration and isolation; ABC methods:
  `http://www.rubic.rdg.ac.uk/~mab/`
- Estimation of effective size; hybridization:
  `http://swfsc.noaa.gov//staff.aspx?Division=FED&id=740`
- Assignment testing; bottleneck detection:
  `http://www.montpellier.inra.fr/URLB/`
- General population genetics software (ARLEQUIN); general-purpose programs for simulating gene-frequency data under a wide variety of conditions; programs for performing ABC analysis:
  `http://cmpg.unibe.ch/software.htm`

- Bottleneck detection:
  `http://swfsc.noaa.gov/textblock.aspx?Division=FED&id=3298`
- Estimation of migration rates:
  `http://www.rannala.org`
- Analysis of population structure; population assignment; hybridization:
  `http://pritch.bsd.uchicago.edu/software.html`
- Analysis of population structure; population assignment: hybridization:
  `http://www.rni.helsinki.fi/~jic/bapspage.html`

REFERENCES

Anderson, E.C. (2005). An efficient Monte Carlo method for estimating N_E from temporally spaced samples using a coalescent-based likelihood. *Genetics* **170**, 955–967.

Anderson, E.C. and Thompson, E.A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**, 1217–1229.

Anderson, E.C., Williamson, E.G. and Thompson, E.A. (2000). Monte Carlo evaluation of the likelihood for N_e from temporally spaced samples. *Genetics* **156**, 2109–2118.

Avise, J.C. (1994). *Molecular Markers, Natural History and Evolution*. Chapman & Hall, London.

Balding, D.J., Carothers, A.D., Marchini, J.L., Cardon, L.R., Vetta, A., Griffiths, B., Weir, B.S., Hill, W.G., Goldstein, D., Strimmer, K., Myers, S., Beaumont, M.A., Glasbey, C.A., Mayer, C.D., Richardson, S., Marshall, C., Durrett, R., Nielsen, R., Visscher, P.M., Knott, S.A., Haley, C.S., Ball, R.D., Hackett, C.A., Holmes, S., Husmeier, D., Jansen, R.C., ter Braak, C.J.F., Maliepaard, C.A., Boer, M.P., Joyce, P., Li, N., Stephens, M., Marcoulides, G.A., Drezner, Z., Mardia, K., McVean, G., Meng, X.L., Ochs, M.F., Pagel, M., Sha, N., Vannucci, M., Sillanpaa, M.J., Sisson, S., Yandell, B.S., Jin, C.F., Satagopan, J.M., Gaffney, P.J., Zeng, Z.B., Broman, K.W., Speed, T.P., Fearnhead, P., Donnelly, P., Larget, B., Simon, D.L., Kadane, J.B., Nicholson, G., Smith, A.V., Jonsson, F., Gustafsson, O., Stefansson, K., Donnelly, P., Parmigiani, G., Garrett, E.S., Anbazhagan, R. and Gabrielson, E. (2002). Discussion on the meeting on 'statistical modelling and analysis of genetic data'. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **64**, 737–775.

Balding, D.J. and Nichols, R.A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3–12.

Balding, D.J. and Nichols, R.A. (1997). Significant genetic correlations among Caucasians at forensic DNA loci. *Heredity* **78**, 583–589.

Balloux, F. (2004). Heterozygote excess in small populations and the heterozygote-excess effective population size. *Evolution* **58**, 1891–1900.

Bartley, D., Bagley, M., Gall, G. and Bentley, B. (1992). Use of linkage disequilibrium data to estimate effective size of hatchery and natural fish populations. *Conservation Biology* **6**, 365–375.

Beaumont, M.A. (1999). Detecting population expansion and decline using microsatellites. *Genetics* **153**, 2013–2029.

Beaumont, M.A. (2003a). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–1160.

Beaumont, M.A. (2003b). Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity* **92**, 365–379.

Beaumont, M.A., Barratt, E.M., Gottelli, D., Kitchener, A.C., Daniels, M.J., Pritchard, J.K. and Bruford, M.W. (2001). Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology* **10**, 319–336.

Beaumont, M.A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

Berthier, P., Beaumont, M.A., Cornuet, J.M. and Luikart, G. (2002). Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetics* **160**, 741–751.

Bertorelle, G. and Excoffier, L. (1998). Inferring admixture proportions from molecular data. *Molecular Biology and Evolution* **15**, 1298–1311.

Blouin, M.S. (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology and Evolution* **18**, 503–511.

Bucci, G., Vendramin, G.G., Lelli, L. and Vicario, F. (1997). Assessing the genetic divergence of *Pinus leucodermis* Ant. endangered populations: use of molecular markers for conservation purposes. *Theoretical and Applied Genetics* **95**, 1138–1146.

Caballero, A. (1994). Developments in the prediction of effective population size. *Heredity* **73**, 657–679.

Calmet, C. (2003). Inférences sur l'histoire des populations à partir de leur diversité genetique: étude de séquences démographiques de type fondation-explosion. Ph.D. thesis, University of Paris.

Cannings, C. and Edwards, A.W.F. (1969). Expected genotypic frequencies in a small sample: deviation from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **21**, 245–247.

Carvajal-Rodríguez, A., Rolan-Alvarez, E. and Caballero, A. (2005). Quantitative variation as a tool for detecting human-induced impacts on genetic diversity. *Biological Conservation* **124**, 1–13.

Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution* **32**, 550–570.

Chakraborty, R., Kamboh, M.I., Nwankwo, M. and Ferrell, R.E. (1992). Caucasian genes in American blacks: new data. *American Journal of Human Genetics* **50**, 145–155.

Chan, Y.L., Anderson, C.N.K. and Hadly, E.A. (2006). Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA. *PLoS Genetics* **2**, 451–460.

Chikhi, L., Bruford, M.W. and Beaumont, M.A. (2001). Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**, 1347–1362.

Choisy, M., Franck, P. and Cornuet, J.M. (2004). Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Molecular Ecology* **13**, 955–968.

Ciofi, C., Beaumont, M.A., Swingland, I.R. and Bruford, M.W. (1999). Genetic divergence and units for conservation in the Komodo Dragaon *Varanus komodoensis*. *Proceedings of the Royal Society of London, Series B* **266**, 2269–2274.

Cockerham, C.C. and Weir, B.S. (1977). Digenic descent measures for finite populations. *Genetical Research* **30**, 121–147.

Corander, J. and Marttinen, P. (2006). Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology* **15**, 2833–2843.

Corander, J., Waldmann, P. and Sillanpaa, M.J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374.

Cornuet, J.M. and Luikart, G. (1996). Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**, 2001–2014.

Cornuet, J.M., Piry, S., Luikart, G., Estoup, A. and Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989–2000.

Coulon, A., Guillot, G., Cosson, J.F., Angibault, J.M.A., Aulagnier, S., Cargnelutti, B., Galan, M. and Hewison, A.J.M. (2006). Genetic structure is influenced by landscape features: empirical evidence from a roe deer population. *Molecular Ecology* **15**, 1669–1679.

Crow, J.F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row, New York.

Csillery, K., Johnson, T., Beraldi, D., Clutton-Brock, T., Coltman, D., Hansson, B., Spong, G. and Pemberton, J.M. (2006). Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. *Genetics* **173**, 2091–2101.

Dawson, K.J. and Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research, Cambridge* **78**, 59–77.

DeSalle, R. and Amato, G. (2004). The expansion of conservation genetics. *Nature Reviews Genetics* **5**, 702–712.

Di Rienzo, A., Donnelly, P., Toomajian, C., Sisk, B., Hill, A., Petzl-Erler, M.L., Haines, G.K. and Barch, D.H. (1998). Heterogeneity of microsatellite mutations within and between loci and implications for human demographic histories. *Genetics* **148**, 1269–1284.

Di Rienzo, A., Peterson, A.C., Garza, J.C., Valdes, A.M., Slatkin, M. and Freimer, N.B. (1994). Mutational processes of simple sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 3166–3170.

Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29**, 410–421.

Dupanloup, I. and Bertorelle, G. (2001).. Inferring admixture proportions from molecular data: extension to any number of parental populations. *Molecular Biology and Evolution* **18**, 672–675.

Elston, R.C. (1971). The estimation of admixture in racial hybrids. *Annals of Human Genetics* **35**, 9–17.

Emery, A.M., Wilson, I.J., Craig, S., Boyle, P.R. and Noble, L.R. (2001). Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. *Molecular Ecology* **10**, 1265–1278.

England, P.R., Cornuet, J.M., Berthier, P., Tallmon, D.A. and Luikart, G. (2006). Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conservation Genetics* **7**, 303–308.

Estoup, A., Beaumont, M.A., Sennedot, F., Moritz, C. and Cornuet, J.M. (2004). Genetic analysis of complex demographic scenarios: the case of spatially expanding populations in the cane toad, *Bufo marinus*. *Evolution* **58**, 2021–2036.

Estoup, A., Wilson, I.J., Sullivan, C., Cornuet, J.-M. and Moritz, C. (2001). Inferring population history from microsatellite and enzyme data in serially introduced cane toads *Bufo marinus*. *Genetics* **159**, 1671–1687.

Evanno, G., Regnaut, S. and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611–2620.

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Ewens, W.J. (2004). *Mathematical Population Genetics: Theoretical Introduction*, vol. 1. Springer-Verlag, New York.

Excoffier, L. and Heckel, G. (2006). Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics* **7**, 745–758.

Falush, D., Stephens, M. and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies author(s). *Genetics* **164**, 1567–1587.

Felsenstein, J. (1981). Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution* **35**, 1229–1242.

Frankham, R. (2005). Genetics and extinction. *Biological Conservation* **126**, 131–140.

Frankham, R. (1995). Conservation genetics. *Annual Review of Genetics* **29**, 305–327.

Fu, Y.-X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925.

Garza, J.C. and Williamson, E.G. (2001). Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology* **10**, 305–318.

Gasbarra, D., Sillanpaa, M.J. and Arjas, E. (2005).. Backward simulation of ancestors of sampled individuals. *Theoretical Population Biology* **67**, 75–83.

Gasbarra, D., Pirinen, M., Sillanpää, M.J., Salmela, E. and Arjas, E. (2007). Estimating genealogies from unlinked marker data: a Bayesian approach. *Theoretical Population Biology* (Submitted: preprint at `http://www.rni.helsinki.fi/dag/genealogy.tar`).

Glass, B. and Li, C.C. (1953). The dynamics of racial intermixture – an analysis based on the American Negro. *American Journal of Human Genetics* **5**, 1–20.

Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L. and Feldman, M.W. (1995).. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 6723–6727.

Goodman, S.J., Tamate, H.B., Wilson, R., Nagata, J., Tatsuzawa, S., Swanson, G.M., Pemberton, J.M. and McCullough, D.R. (2001). Bottlenecks, drift and differentiation: the population structure and demographic history of sika deer (Cervus nippon) in the Japanese archipelago. *Molecular Ecology* **10**, 1357–1370.

Goossens, B., Chikhi, L., Ancrenaz, M., Lackman-Ancrenaz, I., Andau, P. and Bruford, M.W. (2006). Genetic signature of anthropogenic population collapse in orang-utans. *PLoS Biology* **4**, 285–291.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo and Bayesian model determination. *Biometrika* **82**, 711–7732.

Griffiths, R.C. and Tavaré, S. (1994). Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131–159.

Guillot, G., Estoup, A., Mortier, F. and Cosson, J.F. (2005). A spatial statistical model for landscape genetics. *Genetics* **170**, 1261–1280.

Hadfield, J.D., Richardson, D.S. and Burke, T. (2006). Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Molecular Ecology* **15**, 3715–3730.

Haig, S.M. (1998). Molecular contributions to conservation. *Ecology* **79**, 413–425.

Hanfling, B. and Weetman, D. (2006). Concordant genetic estimators of migration reveal anthropogenically enhanced source-sink population structure in the River Sculpin, *Cottus gobio*. *Genetics* **173**, 1487–1501.

Hedrick, P.W. (2001). Conservation genetics: where are we now? *Trends in Ecology and Evolution* **16**, 629–636.

Hey, J. and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**, 747–760.

Hill, W.G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetical Research* **38**, 209–216.

Hill, W.G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226–231.

Hogben, L. (1946). *An Introduction to Mathematical Genetics*. W. W. Norton, New York.

Hudson, R.R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631.

Hudson, R.R. (1991). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, K.J. Futuyama and J. Antonovics, eds. Oxford University Press, Oxford, pp. 1–44.

Jorde, P.E. and Ryman, N. (1995). Temporal allele frequency change and estimation of effective size in populations with overlapping generations. *Genetics* **139**, 1077–1090.

Kimmel, M., Chakraborty, R., King, J.P., Bamshad, M., Watkins, W.S. and Jorde, L.B. (1998). Signatures of population expansion in microsatellite repeat data. *Genetics* **148**, 1921–1930.

King, J.P., Kimmel, M. and Chakraborty, R. (2000). A power analysis of microsatellite-based statistics for inferring past population growth. *Molecular Biology and Evolution* **17**, 1859–1868.

Keightley, P.D., Caballero, A. and GarciaDorado, A. (1998). Population genetics: surviving under mutation pressure. *Current Biology* **8**, R235–R239.

Krimbas, C.B. and Tsakas, S. (1971). The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control – selection or drift? *Evolution* **25**, 454–460.

Lande, R. (1998). Anthropogenic, ecological and genetic factors in extinction and conservation. *Research in Population Economics* **40**, 259–269.

Langley, C.H., Smith, D.B. and Johnson, F.M. (1978). Analysis of linkage disequilibrium between allozyme loci in natural populations of Drosophila melanogaster. *Genetical Research* **32**, 215–229.

Laurie-Ahlberg, C.C. and Weir, B.S. (1979). Allozyme variation and linkage disequilibrium in some laboratory populations of *Drosophila melanogaster*. *Genetics* **92**, 1295–1314.

Laval, G. and Excoffier, L. (2004).. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* **20**, 2485–2487.

Lecis, R., Pierpaoli, M., Biro, Z.S., Szemethy, L., Ragni, B., Vercillo, F. and Randi, E. (2006). Bayesian analyses of admixture in wild and domestic cats (*Felis silvestris*) using linked microsatellite loci. *Molecular Ecology* **15**, 119–131.

Levene, H. (1949). On a matching problem arising in genetics. *Annals of Mathematical Statistics* **20**, 91–94.

Long, J.C. (1991). The genetic structure of admixed populations. *Genetics* **127**, 417–428.

Luikart, G., Allendorf, F.W., Cornuet, J.M. and Sherwin, W.B. (1998). Distortion of allele frequency distributions provides a test for recent population bottlenecks. *The Journal of Heredity* **89**, 238–247.

Luikart, G. and Cornuet, J.M. (1999). Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics* **151**, 1211–1216.

Luikart, G. and England, P.R. (1999). Statistical analysis of microsatellite DNA data. *Trends in Ecology and Evolution* **14**, 253–256.

Lynch, M., Blanchard, J., Houle, D., Kibota, T., Schultz, S., Vassivlieva, L. and Willis, J. (1999). Perspective: spontaneous deleterious mutation. *Evolution* **53**, 645–663.

Lynch, M., Conery, J. and Burger, R. (1995). Mutation accumulation and the extinction of small populations. *The American Naturalist* **146**, 489–518.

Lynch, M. and Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152**, 1753–1766.

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003).. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15324–15328.

Miller, N., Estoup, A., Toepfer, S., Bourguet, D., Lapchin, L., Derridj, S., Kim, K.S., Reynaud, P., Furlan, F. and Guillemaud, T. (2005). Multiple transatlantic introductions of the western corn rootworm. *Science* **310**, 992.

Milligan, B.G. (2003). Maximum-likelihood estimation of relatedness. *Genetics* **163**, 1153–11167.

Mousseau, T.A., Ritland, K. and Heath, D.D. (1998). A novel method for estimating heritability using molecular markers. *Heredity* **80**, 218–224.

Nei, M. and Tajima, F. (1981). Genetic drift and estimation of effective population size. *Genetics* **98**, 625–640.

Nicholson, G., Smith, A.V., Jonsson, F., Gustafsson, O., Stefansson, K. and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **64**, 695–715.

Nielsen, R., Mountain, J.L., Huelsenbeck, J.P. and Slatkin, M. (1998). Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution* **52**, 669–677.

Nielsen, R. and Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–896.

Nunney, L. (1995). Measuring the ratio of effective population size to adult numbers using genetic and ecological data. *Evolution* **49**, 389–392.

Ohta, T. and Kimura, M. (1969). Linkage disequilibrium due to random genetic drift. *Genetical Research* **13**, 47–55.

Ohta, T. and Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research, Cambridge* **22**, 201–204.

Oliehoek, P.A., Windig, J.J., van Arendonk, J.A.M. and Bijma, P. (2006). Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics* **173**, 483–496.

O'Ryan, C., Harley, E.H., Bruford, M.W., Beaumont, M., Wayne, R.K. and Cherry, M.I. (1998). Microsatellite analysis of genetic diversity in fragmented South African buffalo populations. *Animal Conservation* **1**, 85–94.

Paetkau, D., Calvert, W., Stirling, I. and Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology* **4**, 347–354.

Paetkau, D., Shields, G.F. and Strobeck, C. (1998). Gene flow between insular, coastal and interior populations of brown bears in Alaska. *Molecular Ecology* **7**, 1283–1292.

Painter, I. (1997). Sibship reconstruction without parental information. *Journal of Agricultural Biological and Environmental Statistics* **2**, 212–229.

Palo, J.U., Makinen, H.S., Helle, E., Stenman, O. and Vainola, R. (2001). Microsatellite variation in ringed seals (*Phoca hispida*): genetic structure and history of the Baltic Sea population. *Heredity* **86**, 609–617.

Parra, E.J., Marcini, A., Akey, J., Martinson, J., Batzer, M.A., Cooper, R., Forrester, T., Allison, D.B., Deka, R., Ferrell, R.E. and Shriver, M.D. (1998). Estimating African American admixture proportions by use of population-specific alleles. *American Journal of Human Genetics* **63**, 1839–1851.

Pella, J. and Masuda, M. (2001). Bayesian methods for analysis of stock mixtures from genetic characters. *Fishery Bulletin* **99**, 151–167.

Piry, S., Luikart, G. and Cornuet, J.M. (1999). BOTTLENECK: a computer program for detecting recent reductions in the effective population size using allele frequency data. *The Journal of Heredity* **90**, 502–503.

Pollak, E. (1983). A new method for estimating the effective population size from allele frequency changes. *Genetics* **104**, 531–548.

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A. and Feldman, M.W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**, 1791–1798.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Pudovkin, A.I., Zaykin, D.V. and Hedgecock, D. (1996). On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics* **144**, 383–387.

Queller, D.C. and Goodnight, K.F. (1989). Estimating relatedness using genetic markers. *Evolution* **43**, 258–275.

Ramos-Onsins, S.E. and Rozas, J. (2002). Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution* **19**, 2092–2100.

Rannala, B. and Hartigan, J.A. (1996). Estimating gene flow in island populations. *Genetical Research, Cambridge* **67**, 147–158.

Rannala, B. and Mountain, J.L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 9197–9201.

Regnaut, S., Christe, P., Chapuisat, M. and Fumagalli, L. (2006). Genotyping faeces reveals facultative kin association on capercaillie's leks. *Conservation Genetics* **7**, 665–674.

Reich, D.E., Feldman, M.W. and Goldstein, D.B. (1999). Statistical properties of two tests that use multilocus data sets to detect population expansions. *Molecular Biology and Evolution* **16**, 453–466.

Reich, D.E. and Goldstein, D.B. (1998). Genetic evidence for a Paleolithic human population expansion in Africa. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 8119–8123.

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research* **67**, 175–185.

Robert, C.P. (1996). Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richarson and D.J. Spiegelhalter, eds. Chapman & Hall, London, pp. 441–464.

Roberts, D. and Hiorns, R. (1965). Methods of analysis of the genetic composition of a hybrid population. *Human Biology* **37**, 38–43.

Robertson, A. (1965). The interpretation of genotypic ratios in domestic animal populations. *Animal Production* **7**, 319–324.

Rogers, A.R. (1995). Genetic evidence for a Pleistocene population explosion. *Evolution* **49**, 608–615.

Rogers, A.R. and Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* **9**, 552–569.

Roy, M.S., Geffen, E., Smith, D., Ostrander, E.A. and Wayne, R.K. (1994). Patterns of differentiation and hybridization in North American wolf-like canids, revealed by analysis of microsatellite loci. *Molecular Biology and Evolution* **11**, 553–570.

Ruppert, D. and Wand M.P., (1994).. Multivariate locally weighted least squares regression. *Annals of Statistics*, **22**, 1346–1370.

Saccheri, I., Kuussaari, M., Kankare, M., Vikman, P., Fortelius, W. and Hanski, I. (1998). Inbreeding and extinction in a butterfly metapopulation. *Nature* **392**, 491–494.

Saccheri, I.J., Wilson, I.J., Nichols, R.A., Bruford, M.W. and Brakefield, P.M. (1999). Inbreeding of bottlenecked butterfly populations: estimation using the likelihood of changes in marker allele frequencies. *Genetics* **151**, 1053–1063.

Schneider, S. and Excoffier, L. (1999). Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* **152**, 1079–1089.

Schwartz, M.K., Tallmon, D.A. and Luikart, G. (1998). Review of DNA-based census and effective population size estimators. *Animal Conservation* **1**, 293–299.

Slatkin, M. (1996). Gene genealogies within mutant allelic classes. *Genetics* **145**, 579–587.

Spielman, D., Brook, B.W. and Frankham, R. (2004). Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 15261–15264.

Steel, M. and Hein, J. (2006). Reconstructing pedigrees: a combinatorial perspective. *Journal of Theoretical Biology* **240**, 360–367.

Storfer, A. (1996). Quantitative genetics: a promising approach for the assessment of genetic variation in endangered species. *Trends in Ecology and Evolution* **11**, 343–348.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Tavaré, S. (1984). Lines-of-descent and genealogical processes, and their application in population genetics models. *Theoretical Population Biology* **26**, 119–164.

Taylor, A.C., Sherwin, W.B. and Wayne, R.K. (1994). Genetic variation of microsatellite loci in a bottlenecked species: the northern hairy-nosed wombat. *Molecular Ecology* **3**, 277–290.

Thomas, S.C. (2005). The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. *Philosophical Transactions of the Royal Society of London, Series B* **360**, 1457–1467.

Thomas, S.C. and Hill, W.G. (2002). Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genetical Research* **79**, 227–234.

Thompson, E.A. (1973). The Icelandic admixture problem. *Annals of Human Genetics* **37**, 69–80.

Thompson, E.A. (1975a). *Human Evolutionary Trees*. Cambridge University Press, Cambridge.

Thompson, E.A. (1975b). The estimation of pairwise relationships. *Annals of Human Genetics* **39**, 173–188.

Verardi, A., Lucchini, V. and Randi, E. (2006). Detecting introgressive hybridization between free-ranging domestic dogs and wild wolves (*Canis lupus*) by admixture linkage disequilibrium analysis. *Molecular Ecology* **15**, 2845–2855.

Vitalis, R. and Couvet, D. (2001). Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* **157**, 911–925.

Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics* **153**, 1863–1871.

Wakeley, J. (2001). The coalescent in an island model of population subdivision with variation among demes. *Theoretical Population Biology* **59**, 133–144.

Wakeley, J., Nielsen, R., Liu-Cordero, S.N. and Ardlie, K. (2001). The discovery of single-nucleotide polymorphisms – and inferences about human demographic history. *American Journal of Human Genetics* **69**, 1332–1347.

Wang, J. (2001). A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical Research, Cambridge* **78**, 243–257.

Wang, J. (2002). An estimator for pairwise relatedness using molecular markers. *Genetics* **160**, 1203–1215.

Wang, J.L. (2003). Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**, 747–765.

Wang, J.L. (2004). Sibship reconstruction from genetic data with typing errors. *Genetics* **166**, 1963–1979.

Wang, J.L. (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society of London, Series B* **360**, 1395–1409.

Wang, J.L. (2006). A coalescent-based estimator of admixture from DNA sequences. *Genetics* **173**, 1679–1692.

Wang, J.L. and Whitlock, M.C. (2003). Estimating effective population size and migration rates from genetic samples over space and time. *Genetics* **163**, 429–446.

Waples, R.S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* **121**, 379–392.

Waples, R.S. (1991). Genetic methods for estimating the effective size of Cetacean populations. *Genetic Ecology of Whales and Dolphins*, (special issue 13). *Report of the International Whaling Commission*, pp. 279–300.

Waples, R.S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics* **7**, 167–184.

Waser, P.M. and Strobeck, C. (1998). Genetic signatures of interpopulation dispersal. *Trends in Ecology and Evolution* **13**, 43–44.

Wasser, S.K., Shedlock, A.M., Comstock, K., Ostrander, E.A., Mutayoba, B. and Stephens, M. (2004). Assigning African elephant DNA to geographic region of origin: applications to the ivory trade. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 14847–14852.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276.

Watterson, G.A. (1978). The homozygosity test of neutrality. *Genetics* **88**, 405–417.

Weir, B.S. (1996). *Genetic Data Analysis*, 2nd edition. Sinauer Associates, Sunderland.

Weir, B.S. and Hill, W.G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**, 477–488.

Weiss, G. and von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics* **149**, 1539–1546.

Whitlock, M.C., Griswold, C.K. and Peters, A.D. (2003). Compensating for the meltdown: the critical effective size of a population with deleterious and compensatory mutations. *Annales Zoologici Fennici* **40**, 169–183.

Williamson, E.G. and Slatkin, M. (1999). Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics* **152**, 755–761.

Wilson, I.J. and Balding, D.J. (1998). Genealogical inference from microsatellite data. *Genetics* **150**, 499–510.

Wilson, G.A. and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**, 1177–1191.

Wilson, I.J., Weale, M.E. and Balding, D.J. (2003). Inferences from DNA data: population histories, evolutionary processes, and forensic match probabilities. *Journal of the Royal Statistical Society, Series A* **166**, 155–188.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Human Genetic Diversity and its History

G. Barbujani

Dipartimento di Biologia ed Evoluzione, Università di Ferrara, Ferrara, Italia

and

L. Chikhi

Laboratoire Evolution et Diversité Biobgique, Université Paul Sabatier,
Toulouse, France

Aspects of the evolutionary and historical processes that shaped present-day human diversity can be inferred from patterns of genetic variation within and between populations. The main questions currently being addressed include the evolutionary relationships among contemporary humans and the different human forms documented in the fossil record, the extent and causes of the genetic differences among modern populations, and the implications of such differences for applied research in fields such as medical genetics, pharmacogenetics, and forensic science. In this chapter, we outline the main current models of human evolution and review the available ancient and modern DNA evidence in the light of these models. We suggest that, despite ongoing controversy, most data are easier to reconcile with a model in which the ancestors of modern populations dispersed recently (<200 000 years ago) from Africa and essentially or completely replaced previously settled human forms. However, it has become increasingly evident that models assuming the expansion of a small African group are oversimplified, and more complex scenarios need be envisaged, including genetic substructuring in the expanding population. How much substructure is needed to explain existing data is unclear, and whether this apparent substructure could be partly due to admixture between modern and archaic humans is debated. The available studies among modern populations show that the differences tend to be patterned in geographical space, but that variation tends to be continuous and strong genetic boundaries rarely occur, suggesting a major role of isolation by distance in shaping human diversity. As a consequence, global human diversity is not well described by a small set of well-defined races. The extent to which racial classifications might be useful in some applications remains to be demonstrated. Throughout the chapter, we emphasise the large amount of uncertainty surrounding the interpretation of genetic data, when they are used either to make inferences on our origins or to predict disease risk.

31.1   INTRODUCTION

Individuals of all species form phenotypic and genotypic clusters that are generally structured in space and represent the outcome of the species' evolutionary history. Identifying the geographical patterns of human genetic diversity, and understanding their evolutionary and historical causes are the goals of anthropological genetics. Until recently, to address questions in this field, scientists had only measures of anatomical traits available and, from the mid twentieth century, a handful of polymorphic genetic markers. With the development of fast and inexpensive methods for DNA typing, the amount of relevant information available has increased dramatically, and so has the possibility to test hypotheses about the evolutionary factors underlying human diversity. In this way, some traditional questions found an apparently satisfactory answer, but many others emerged, so that the field seems no less controversial now than it used to. The main controversies concern the degree of genetic differentiation among human populations, whether the existing differences can be regarded as discontinuous and, if so, whether it is possible to agree on a non-subjective list of distinct human groups, each of which could potentially be associated with specific disease risks. A lesser, but by no means less debated problem, is whether such groups can or cannot be legitimately called *races*.

Interpreting the data bearing on human diversity and on its history is not straightforward for a number of reasons, including the obvious difficulty for all, including the authors of this paper, to study our species with the same emotional detachment with which we study *Escherichia coli* or *Drosophila melanogaster*. Words such as diversity and race do not immediately correspond to obvious biological concepts, and take different meanings in science from their non-scientific usage. Even in the technical literature, a high degree of ambiguity is present, so that according to Templeton (1999) the genetic differences among human populations are among the lowest in mammals, whereas according to Sarich (2000) they are highest of all mammals. Therefore, it seems useful to briefly review the available evidence, to outline the main open questions, and to consider the potentials and the limitations of the approaches chosen to address them.

In this chapter, we shall deal with some of the many studies leading to our current understanding of the patterns of human biological diversity and the evolutionary and demographic processes that generated them. Given the increasing role of genetic data in the debate, we describe in some detail the main methods used to interpret them. We also outline, the way some concepts took shape in the course of time. In the next section, we deal with historical inference, outlining the main lines of evidence that suggest a recent African origin of humankind. We include a short description of the palaeontological and archaeological data and of the models that provide the framework necessary for interpreting the biological evidence. In Section 31.3, we move to the studies describing how human genetic diversity is structured in space. This leads to a final section in which we deal with a number of recently emerged practical issues, in fields such as medicine, pharmacogenetics, and forensics, for which assumptions on the structure of human diversity appear to be important.

31.2 HUMAN GENETIC DIVERSITY: HISTORICAL INFERENCES

31.2.1 Some Data on Fossil Evidence

It is currently accepted that hominins, i.e. bipedal primates (Lewin and Foley, 2004), separated from the other African apes (*Gorilla* and *Pan* genera) some 5–8 million years (MY) ago. On the basis of fossil evidence, the first small-brained clearly bipedal primates are documented in Africa from at least 6 MY ago in what are now Chad and Kenya. Between 6 and 4 MY ago, the data are scanty, and it is only from approximately 4 MY ago that the hominin fossil record starts to be wider, with the appearance of the australopithecines (genus *Australopithecus*, meaning the 'southern ape man'), the most famous being 'Lucy' (*A. afarensis* 3.2 MY old, Figure 31.1). There is no consensus on the number of australopithecine species that existed between 4 and 1 MY ago (e.g. Klein, 1999), but many believe that there may have been as many as seven, some of them living at the same time (Toth and Schick, 2005; Klein, 2005; Scarre, 2005).

Australopithecines are usually divided into robust (sometimes referred to as *Paranthropus*) and gracile (genus *Australopithecus*) forms, and it is from the latter that the genus *Homo* (more precisely, *H. habilis* and *H. rudolfensis*) probably evolved, perhaps from east African *A. garhi* (Figure 31.1). This split probably took place around 2.5 or 2.0 MY ago, when the first stone tools appear in the African archaeological record. The early *Homo* forms include *H. habilis, H. rudolfensis*, and *H. ergaster*, the latter appearing sometime around 1.8 MY. It is from *H. ergaster* that *H. heidelbergensis* and *H. erectus* are thought to have derived, with *H. heidelbergensis* being regarded as the most likely ancestor of Neandertals and modern humans. *H. erectus* separated from *H. ergaster* less than 2 MY ago (Figure 31.1). With *H. ergaster, H. heidelbergensis, H. neandertalensis*, and *H. erectus*, we enter the heart of the controversies surrounding the models of modern human emergence, in particular, the contribution of so-called archaic forms to the ancestry of modern *H. sapiens*.
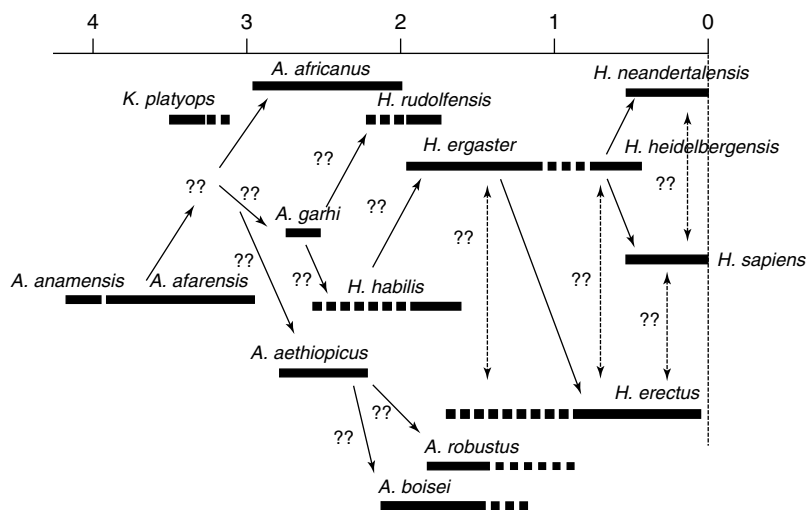
Figure 31.1 A schematic representation of hominin evolution. *X* axis: million years ago.

Until recently, *H. erectus* was thought to be the first *Homo* species to have left Africa and reached eastern Asia more than 1.6 MY ago (Antón and Swisher, 2004). The recent Dmanisi finds in the Caucasus, attributed to *H. ergaster* or to a closely related species (Vekua *et al.*, 2002; Toth and Schick, 2005) and dated about 1.77–1.95 MY ago, suggested that other *Homo* species also left Africa, which would imply that the three species potentially involved in the *H. sapiens* ancestry developed from *H. ergaster* ancestors. It thus appears that between 1 and 2 MY ago these potential ancestors were already geographically separated, namely *H. erectus* in eastern Asia, *H. neandertalensis* in western Asia and Europe, and *H. sapiens* in Africa. The first European settlers (ancestors of *H. neandertalensis*) are often pooled with their African contemporaries (ancestors of *H. sapiens*) under the name *H. heidelbergensis*, mentioned above, and may have lived around 1.1–1.3 MY ago.

Many details of this history are unknown and, among those that are known, just a few can be properly discussed here. But, before moving further to the models of modern human origins, we note that many fossils are still not assigned with certainty to a genus. As a consequence, there is still incomplete agreement on how many species of humans (or australopithecines, for that matter) have lived in the past. In fact, simply deciding which fossils were part of the human lineage and which were not depends largely on what are considered the key human features (Tattersall, 1995; Toth and Schick, 2005; Klein, 1999). We have only a vague idea about variation in time and space within fossil species, and hence the attribution to species and the reconstruction of their evolutionary relationships contain elements of arbitrariness (see Figure 31.1). For example, some authors do not consider the differences between *H. erectus* and *H. ergaster* large enough to warrant the need for a separate *H. ergaster* species. Moreover, proposed hominid species increased in numbers during the early twentieth century, decreased when the tendency prevailed to group together individuals in a limited number of species, and then increased again. Many early controversies originated from Darwin's intuition that bipedal locomotion, tool making, and a large brain case size developed simultaneously. We now know that these features appeared at different times, respectively around circa 6, 2.5, and 2 MY ago (Klein, 1999; Toth and Schick, 2005). Dating can also be problematic, as has been suggested, for instance, for the Dmanisi specimen, who might be as recent as 0.79 MY ago, or as old as 2.4 MY due to stratigraphic uncertainty (Klein, 2005). Finally, technologies are also not always easily linked to one species, or even one genus. In brief, there is currently no uncontroversial *Homo* fossil older than 1.8 MY and no uncontroversial data suggesting sustained occupation of temperate areas before 1 MY ago (Klein, 2005). The first dispersal of *Homo* species outside Africa probably started as early as ∼2 MY, but more recent dates (∼1 MY) cannot be ruled out (Antón and Swisher, 2004; Klein, 2005).

31.2.2 Models of Modern Human Origins

Despite the many uncertainties, most palaeoanthropologists would agree that *H. ergaster* is the ancestor from which Asian *H. erectus* and African *H. heidelbergensis* evolved, and that *H. neandertalensis* and *H. sapiens* then evolved from *H. heidelbergensis*. However, the relative contributions of these three forms, which may (Tattersall, 1997) or may not (Wolpoff, 1999) be regarded as separate species, to modern human morphologies remains controversial, as well as the exact timing (Grine *et al.*, 2007) and route (Anikovich *et al.*, 2007) of their dispersal in the Near East and Eurasia.

In recent decades, the debate has focused on two main models of human evolution. Under the Out of Africa (OOA), or Recent African Origin (Howells, 1976; Stringer and Andrews, 1988; Stringer, 1989) or Replacement (e.g. Relethford, 2003) model, present-day populations are regarded as deriving from a major spatial and demographic expansion of African populations of *H. sapiens*, starting less than 200 KY ago. These populations colonised the whole planet, replacing, essentially with no admixture, previously settled *H. erectus* and *H. neandertalensis* populations. On the other hand, the Multi-Regional (MR) model posits that Eurasian populations of *H. erectus* or *H. neandertalensis* are part of the genealogy of present-day humans (Wolpoff *et al.*, 1988). This means that only one human species existed in the last MY or so, encompassing all these different morphologies, which should then be simply classified as archaic and modern forms of a polytypic *H. sapiens* species (Relethford, 2003). These basic models really represent the extremes of a range of possible models, whose classification, in turn, is not completely clear (but see Aiello, 1993). For instance, Bräuer (1984) and Smith (1985) proposed that both admixture and replacement occurred between expanding African and previously settled Eurasian populations. Despite their similarities, these two hypotheses are considered, respectively, as versions of the OOA and of the MR models because of the greater weight they give to either admixture (Bräuer, 1984) or continuous gene flow (Smith, 1985) over long periods.

The MR model stems from work by Weidenreich (1943; 1947) and others, who described morphological similarities between Asian people and the fossil *Sinanthropus* (Peking Man), and between modern Australians and *Pithecanthropus* (Java Man) fossils, the so-called regional continuity traits (both Peking and Java man are now classified as *H. erectus*). These findings led to the proposal that anatomically modern Africans, Asians, Australians, and Europeans had descended from anatomically archaic ancestors who largely occupied the same regions. On the contrary, a comparatively recent origin of our species was proposed to explain the similarities between the first anatomically modern Europeans of the Cro-Magnon type and 130-KY-old fossils from the Omo Kibish region of Ethiopia (Stringer, 1978). In large-scale comparisons of modern and ancient bone specimens, the data appeared to be in significantly better agreement with the OOA rather than the MR model (Waddle, 1994; Lahr, 1994).

Further support for the OOA model comes from studies showing little or no evidence for hybrid populations displaying both archaic and modern traits, which suggests that the fossil record exhibits a discontinuous process of demographic replacement rather than continuity (Stringer and Andrews, 1988; Lahr and Foley, 1998). The controversy is still going on (Zilhão, 2006), and different authors appear to interpret the apparent lack of continuity in very different ways regarding the admixture process itself. For instance, it has been claimed that, even with significant admixture, complete continuity is not expected for morphological traits (Relethford, 1999; 2001), and hence lack of continuity may not necessarily play against the MR model. If all models involving any contribution of archaic populations to the modern gene pool are to be considered as versions of the MR model, rejecting it becomes in effect impossible.

31.2.3 Methods for Inferring Past Demography

Genetic data have the advantages over other types of data that their transmission mode is well known and that theory describing the evolution of genes in populations is well developed. In the following, we discuss from a biologist's point of view the manner in

which genetic data have been used in recent decades to infer the demographic history of populations. For more statistical and mathematical treatments, see **Chapters 26**, **28**, **29** and **30**.

31.2.3.1 *Summary Statistics*

Population genetics inference is possible only if past demographic events leave specific genetic signatures in present-day populations. In particular, the effect of changes in population size on different summary statistics has been extensively studied. For diploids, it is customary to define the scaled mutation rate $\theta = 4N_e\mu$, where N_e is the effective population size of a stationary (or demographically stable) population and μ the locus mutation rate.

For neutral markers, evolving under an infinite-allele model in a Wright–Fisher population, Kimura and Crow (1964) showed that at mutation-drift equilibrium $E[H] = 1/(\theta + 1)$, where $E[.]$ represents expectation and H is the probability that two alleles drawn at random from the population are the same, which can be estimated by $\Sigma(n_i/n)^2$, where n_i is the observed count of the ith allele, and n is the sample size. Ewens (1972) derived from what is now known as *Ewens' sampling formula*, an estimator of H given n_A, the number of distinct alleles in the sample, which depends on and n and θ but not the observed n_i values. Significant differences between these two H estimates can be interpreted in terms of departure from any of the neutrality, size constancy, or mutation model assumptions Watterson (1978). For instance, when large populations go through a bottleneck (Nei *et al.*, 1975), rare alleles are lost first, which only mildly influences $\Sigma(n_i/n)^2$, whereas n_A is substantially reduced, and so the former estimate of H is significantly lower than that expected given n_A until a new equilibrium is approached. Conversely, expanding populations tend to accumulate new (and hence rare) alleles, and $\Sigma(n_i/n)^2$ is initially significantly higher than the estimate of H based on n_A.

The rationale behind this approach, and Watterson's (1978) related approach, has since been extended to different genetic markers by assuming different mutation models and using different summaries of the data (Tajima, 1989a; 1989b; Fu and Li, 1993). For DNA sequence data, the mean number of nucleotide differences between sequences, π, and the number of segregating sites, S (i.e. single nucleotide polymorphisms (SNPs)), provide two estimators, θ_S and θ_π, of the scaled mutation parameter, θ, at mutation-drift equilibrium that are differently affected by demographic events and selection (Tajima, 1989a; 1989b). Tajima (1989a) hence suggested the use of $D = (\theta_\pi - \theta_S)/[\mathrm{var}(\theta_\pi - \theta_S)]^{1/2}$, now known as *Tajima's D* (see also **Chapter 26**), as a measure of departure from equilibrium conditions. Demographic bottlenecks and balancing selection tend to reduce S without much affecting π, and hence lead to positive D values, whereas population expansions and positive selection (selective sweeps) lead to negative values. For microsatellites, equilibrium equations have not been derived analytically, but simulation-based approaches have been developed (Cornuet and Luikart, 1996).

Beyond the use of summary statistics, theoretical work has also focused on other properties of allele-frequency distributions, or, for sequence data, on the properties of the so-called mismatch distributions. The mismatch distributions are obtained by plotting a histogram of all pairwise comparisons between sequences in a sample, and have been shown to be sensitive to demographic changes (Slatkin and Hudson, 1991; Rogers and Harpending, 1992; Excoffier and Schneider, 1999; Ray *et al.*, 2004). For other markers,

Marth *et al.* (2004) studied the properties of the full allele-frequency spectrum under different demographic scenarios, and found computationally quick ways to estimate these frequency spectra. Garza and Williamson (2001) have shown that bottlenecked populations have 'gappy' allele-frequency distributions at microsatellite loci because the number of alleles is strongly reduced, whereas the allelic range is only marginally affected.

   Summary statistics based on genetic data may be easy to compute, but may discard most of the genealogical information in the data (Felsenstein, 1992). As a consequence, different sets of evolutionary factors can exert similar effects on summary statistics and cannot be separated unless detailed archaeological or historical information is available. Furthermore, these methods can *detect* demographic events, but usually provide poor *quantification* or *dating* of such events.

   Indeed, the magnitude of a departure from equilibrium may not be a monotonic function of the time since the demographic event. Whereas mismatch distributions appear to follow simple temporal dynamics, this is not the case of Tajima's *D* (Fay and Wu, 1999). After a population expansion from a small population size, a unimodal mismatch distribution is expected whose mode will move from low to high values as mutations accumulate, whereas its height will decrease (Rogers and Harpending, 1992). On the contrary, simulations have shown that, after a bottleneck, Tajima's *D* is first positive, and then becomes negative when the population recovers from the bottleneck, before tending towards equilibrium (Fay and Wu, 1999). Thus, Tajima's *D*, one of the most popular statistics used to detect population size changes, may be transiently positive or negative, depending on the severity of the change and on the number of generations since the original demographic event, and hence it can behave differently for haploid and diploid genomes. The mutation model assumed can also generate significant departures from equilibrium (Aris-Brosou and Excoffier, 1996), but in many studies rejection of equilibrium is directly taken as evidence of demographic changes or selection. Future work will probably need to separate the effects of the mutation model from those of population structure and selection. Because crucial evolutionary information is lost as summary statistics are estimated, such an endeavour will require methods that are more efficient at retrieving information on demographic parameters from genetic data.

31.2.3.2   *Likelihood-based and Bayesian Approaches*

Following Felsenstein (1992), several authors have developed statistical methods that aim at extracting information from the full allelic distribution (Griffiths and Tavaré, 1994; Kuhner *et al.*, 1995; Wilson and Balding, 1998; Beaumont and Rannala, 2004). The aim of likelihood-based approaches is to compute or approximate the probability $P_M(D|\theta)$ of the observed data D under some demographical model M, defined by a set of parameters $\theta = (\theta_1, \ldots, \theta_k)$. This probability is the likelihood $L_M(\theta|D)$, or simply $L_M(\theta)$. Some methods focus on point estimators, e.g. maximum likelihood estimates (MLE), others take a Bayesian perspective and compute a posterior probability distribution for θ. Using Bayes formula, we can write the following:

$$P_M(\theta|D) = P_M(\theta) \times P_M(D|\theta)/P_M(D) = P_M(\theta) \times L_M(\theta|D)/P_M(D)$$

$$= P_M(\theta) \times L_M(\theta)/P_M(D). \tag{31.1}$$

   Since the denominator is constant, *given* the data, $P_M(\theta|D)$ is proportional to $P_M(\theta) \times L_M(\theta)$. In the Bayesian framework, $P_M(\theta)$ summarises knowledge (or lack thereof) about

θ *before* the data are observed and is referred to as the *prior*. $P_M(\theta|D)$ is the posterior and represents updated knowledge about θ *after* the data have been observed. Equation (31.1) asserts that the posterior is obtained by weighting the prior with the likelihood function. The use of priors involves subjectivity, but this is introduced via clear and explicit assumptions about parameters. It also allows one to explicitly account for uncertainty in all parameters, rather than assuming point values for, say, mutation rates or generation times (Goldstein *et al.*, 1995; Chikhi *et al.*, 1998).

Computing the likelihood can present a substantial difficulty for these approaches, even for simple demographic models involving one population (Wilson and Balding, 1998; Beaumont, 1999; Storz and Beaumont, 2002; Wilson *et al.*, 2003). For complex models, there is often no solution (but see Kuhner *et al.*, 1995; Chikhi *et al.*, 2001; Nielsen and Wakeley, 2001; Beerli and Felsenstein, 2001; Rannala and Yang, 2003; Hey and Nielsen, 2004). Whereas population data can be rapidly simulated using coalescent methods (see **Chapters 25**, **26**, **29** or **30**), the probabilities involved are too low for classical Monte Carlo (MC) integration even for reasonable sample sizes and multilocus data sets.

In order to improve the efficiency of MC integration and significantly reduce the computation time, importance sampling (IS) schemes have been introduced (Griffiths and Tavaré, 1994; Stephens and Donnelly, 2000). The idea is to sample coalescent trees from distributions that are as close as possible to the conditional distribution, given the data $P(T|D)$, where T represents the trees. One way of doing that is to construct coalescent trees starting from the data so that trees that cannot possibly produce the data are not explored (see Griffiths and Tavaré, 1994; Stephens and Donnelly, 2000). Under the classical frequentist approach, some optimisation algorithm is used to find the MLE. Alternatively, the likelihood profile is obtained using a grid across the parameter space. On the contrary, most Bayesian methods of the last decade use Markov chain Monte Carlo (MCMC) algorithms to obtain samples from the required posterior distribution. This approach has proven useful to analyse various demographic models, including migration between splitting populations of variable size (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004; an application in Hey, 2005) and admixture (Chikhi *et al.*, 2001; 2002).

The application of full-likelihood methods to real data is often difficult, and limited to specific demographic models. Analyses may run for weeks with no result because the MCMC algorithm fails to reach equilibrium, due to the high dimensionality of the problems addressed. This has favoured a renewed interest in approximate methods (Tavaré *et al.*, 1997; Weiss and von Haeseler, 1998; Excoffier *et al.*, 2005). One line of research consists in simplifying the calculation of the full likelihood into the product of simpler components that are treated as if they were independent, sometimes called a *composite likelihood approach* (Wang, 2003; McVean *et al.*, 2004). The use of product of approximate conditionals (PAC) likelihood by Li and Stephens (2003) follows a similar rationale and was used to simplify inference with recombining sequence data (see also **Chapter 26**).

Another possibility is to simulate a parameter value θ under the prior, but accept it as an approximate simulation from the posterior $P_M(\theta|D)$ if a dataset D^* simulated under the model given this value of θ is sufficiently close to the real data. How close D^* must be from the real data and how to measure 'closeness' are the focus of ongoing research. Different authors suggest either the use of summary statistics (e.g. Pritchard *et al.*, 1999; Beaumont *et al.*, 2002; Marjoram *et al.*, 2003), or the data directly (Marjoram *et al.*, 2003) to obtain a distance measure $d(D^*, D)$ between the observed

and simulated data. Not all methods using this approach are Bayesian (e.g. Weiss and von Haeseler, 1998), but most are, and these have been dubbed Approximate Bayesian Computation (ABC) methods (Beaumont *et al.*, 2002). Their main advantages are that they are extremely flexible, suitable for any demographic model under which data can be simulated, and, contrary to MCMC-based methods, computations do not need to reach an equilibrium. Non-genetic information, such as that deriving from archaeology, can also be incorporated into the priors. Figure 31.2 shows a simple example where posterior distribution obtained using a full-likelihood method is compared to posteriors obtained using an ABC approach with different summary statistics or combinations thereof (see also **Chapters 26** and **30**).

Summary statistics could also be used in simulation-based approaches not requiring likelihoods or posterior distributions to be estimated. Akey *et al.* (2004) analysed the properties of four summary statistics, namely Tajima's D, Fu and Li's D^* and F^*, and Fay and Wu's H, both under the standard neutral model and four alternative demographic scenarios (an exponential expansion, a bottleneck, a two-island model, and an ancient population split). They then compared average summary statistics estimated from the data and from the simulations, choosing parameter values that minimised the difference. By analysing variation at 132 genes sequenced in African-American and European-American samples in this way, the authors could (1) reject implausible evolutionary hypotheses, (2) estimate parameter values under the likely hypotheses, and (3) detect loci potentially under selection (outliers). Eight loci appeared as outliers under all the simulated scenarios and were thus termed *demographically robust selection genes*.

Detection of loci under selection has again come to be a major focus of research (see **Chapters 22**, **27** and **30**, Beaumont and Balding, 2004; Beaumont, 2005). Many methods analysing genomic data derive from the suggestion of Cavalli-Sforza (1966) that all demographic parameters being equal, very high or very low values of the standardised allele-frequency variance, F_{ST}, can only be due, respectively, to diversifying or stabilising selection. The main problem with this approach is that the null distribution of F_{ST} is unknown (Nei *et al.*, 1977). Possible solutions include simulating the data under simple models such as Wright's island model (Beaumont and Nichols, 1996), incorporating different mutation models (Flint *et al.*, 1999), or avoiding assumptions on the demographic history by using many random loci (Goldstein and Chikhi, 2002; Akey *et al.*, 2002). Akey *et al.* (2002) and Storz *et al.* (2004) chose this approach to study variation of, respectively, 26 000 SNPs in Africans, Europeans, and Asians, and microsatellite loci in Europeans and Africans. In both cases, all outlier loci showed reduced variability in non-African populations. Assuming the OOA model, these results might point to adaptive selective sweeps following expansion from Africa. However, the choice of the populations sampled could also generate spurious signals of selection. For instance, complex demographic processes, such as the wave of advance and admixture proposed by Eswaran (2002) and Eswaran *et al.* (2005) could also produce unusual patterns in genomic data, with some neutral loci showing drastically extreme behaviours. Recent work indicates that the fate of mutations during spatial expansions, whether to increase in frequency or disappear, is not easily predicted (Edmonds *et al.*, 2004; Klopfstein *et al.*, 2006).

31.2.4 Reconstructing Past Human Migration and Demography

In the last 25 years, genetic data have been increasingly used to investigate aspects of the human past that were previously studied only by paleontologists and archaeologists.
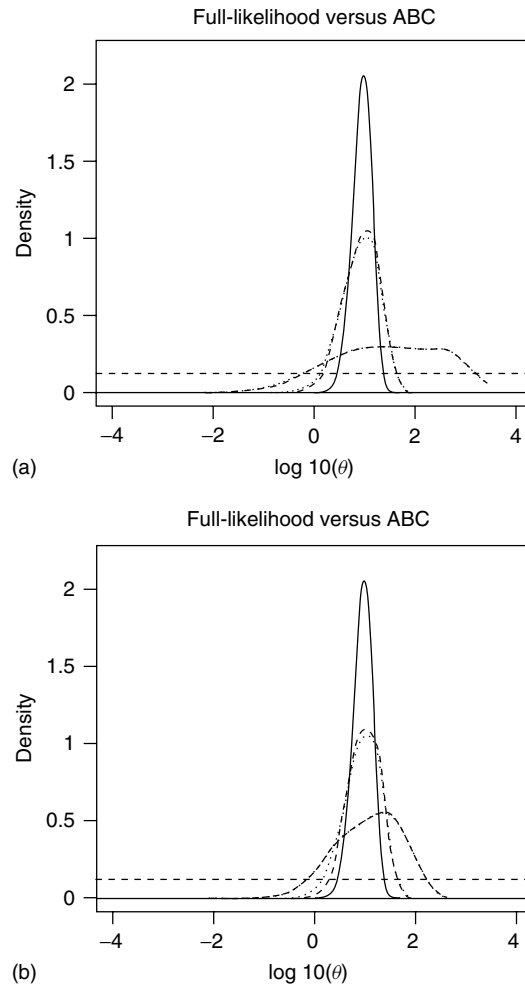
Figure 31.2 A comparison of full-likelihood and ABC approaches. The posterior densities for $\log(\theta)$ are compared, where $\theta = 2Nu$ and u is the mutation rate. The horizontal dashed lines represent the flat prior distributions for $\log(\theta)$. The data analysed were simulated using a value of $\theta = 10$. The solid lines in each panel represent the posterior obtained using Beaumont's (1999) full-likelihood method as implemented in the msvar software. Three summary statistics were used in the ABC approaches: expected heterozygosity (H_e), variance in allele size (va), and number of alleles (n_A). The three curves in each panel were obtained by using H_e only (dash-dotted line), H_e and va (dotted) or H_e, va and n_A (dashed). The upper and lower panels were obtained by changing the tolerance level. In the upper and lower panel, the summary statistics were allowed to differ from the real data by 50 and 30 % respectively. Altogether 10 000 coalescent simulations were performed using an R code originally written by M. Beaumont.

The first large-scale analyses of human DNA diversity were descriptions of mitochondrial DNA and showed that, contrary to what was expected at that time, geographical structuring is weak in humans. The estimated evolutionary trees are rather shallow, with long terminal branches and an excess of low-frequency polymorphisms (Cann *et al.*,

1987; Vigilant *et al.*, 1991; Ingman *et al.*, 2000), all features expected in populations that expanded rapidly or were subjected to diversifying selection. Because this signal was also detected in several autosomal regions (Goldstein *et al.*, 1995; Marth *et al.*, 2003; Voight *et al.*, 2005), it was interpreted as resulting from a demographic process, thus supporting a comparatively recent origin of our species from a small group of founders. Various independent lines of evidence suggest that these founders lived in Africa (Reich and Goldstein, 1998; Takahata *et al.*, 2001). The deepest branches of the mitochondrial trees consistently separated African sequences from heterogeneous clusters containing both African and non-African sequences, a pattern also observed in successive Y-chromosome studies (Underhill *et al.*, 2000; Ke *et al.*, 2001). Thus, both mitochondrial and Y-chromosome gene genealogies contained a significant signal of a recent (<200 KY ago) population bottleneck followed by a demographic (e.g. Slatkin and Hudson, 1991; Rogers and Harpending, 1992) and spatial (Ray *et al.*, 2004) expansion.

These results are not easy to reconcile with the predictions of the MR model. In addition, the OOA model seems to better account not only for highest genetic variation in Africa but also for two other findings, namely the limited human genetic diversity in general and the limited diversity among continental populations, both greater in all species of great apes despite their smaller geographical range and census sizes (Gagneux *et al.*, 1999; Kitano *et al.*, 2003; Fischer *et al.*, 2006, but see Yu *et al.*, 2003). Therefore, in the last decade of the twentieth century, a relatively clear picture of human demographic history seemed to be emerging. In synthesis, human genomic diversity appears to result from the recent expansion of a single African population. This population replaced the pre-existing human forms encountered in its dispersal across the other continents, in the course of a process of which the spatial details and the exact timing needed to be worked out.

In the last 5–10 years, however, the resequencing of autosomal regions provided evidence that it is necessary to envisage more complex sets of evolutionary pressures, not all of them fully understood yet, to account for the current patterns and levels of DNA diversity. Indeed, studies of autosomal and X-chromosome variation showed no general excess of rare polymorphisms (Hey, 1997; Harding *et al.*, 1997; Harris and Hey, 1999; Hammer *et al.*, 2004). That points to diversifying selection, rather than to a sudden demographic growth, as the cause of the excess polymorphism observed in non-recombining DNA regions (Przeworski *et al.*, 2000). In addition, Tajima's *D* shows different values in Africans and non-Africans (reviewed in Garrigan and Hammer (2006)), which may be explained by different selective pressures in the different continents, and/or different demographic histories in African versus Eurasian populations.

At present, the conclusion that the human gene pool is entirely derived from a single and small ancestral population that expanded from East Africa in the last 200 KY is being questioned anew. Opinions also differ on whether data support a single expansion, or rather a more complex set of demographic changes, during which distinct archaic African human forms came in contact and mixed, before or after migrating to other continents (Garrigan and Hammer, 2006). The latter view is supported by the observation of deep haplotype divergence and long-range linkage disequilibrium in the X chromosome.

On the other hand, explicit modelling of evolutionary processes in the geographical space, incorporating population structure, gives results in good agreement with the hypothesis that it was a single group of African hunter–gatherers who expanded dramatically in numbers. The most recent of such studies concluded that humans started expanding very recently (<60 KY ago) from a population whose effective size N_e was close to 1000

individuals (Liu *et al.*, 2006). The likely origin of the expansion was located in East Africa (Liu *et al.*, 2006), i.e. the region where the oldest modern *Homo* fossils have been found (Trinkaus, 2005). An analysis of the correlation between genetic and geographic distances concluded that current human diversity reflects a series of founder effects starting from central or western Africa, but paucity of samples from East Africa may have influenced this result (Ramachandran *et al.*, 2005). Ray *et al.* (2005) also explicitly modelled the spatial population structure and found a likely expansion centre in north Africa, but argued that this result is probably an artefact due to ascertainment bias; once the bias is removed, the most likely origin of modern humans returns to be East Africa.

A certain degree of inconsistency across studies is not surprising. The contrasting conclusions drawn from analyses of variation at, respectively, autosomal loci and uniparentally transmitted markers, reflect, at least in part, the different effective population sizes for these classes of markers, which in turn have an impact on the temporal change of various statistical indexes, including Tajima's *D* (Fay and Wu, 1999). And, more generally, current genetic diversity is the result of a history spanning hundreds of thousand years, of which we can infer from genetic data only some trends through time and place, with no guarantee that these trends affected in the same way all geographical regions and time periods. However, at present, the patterns observed in studies of different markers and different population samples are not easily combined in a coherent and comprehensive picture (Harris and Hey, 1999; Wall and Przeworski, 2000; Chikhi and Beaumont, 2005). All factors considered, most authors currently maintain that the OOA model is in clearer agreement with genetic data, but more elaborate versions of the model seem necessary to account for the way different genome regions seem to vary (Excoffier 2002; Goldstein and Chikhi, 2002). Efforts are currently directed at better understanding the effects of ancient population structuring (Sherry *et al.*, 1998; Goldstein and Chikhi, 2002; Excoffier, 2002; Harding and McVean, 2004), which may have generated genetic patterns of ambiguous interpretation. Indeed, in the absence of detailed knowledge of ancient gene flow, isolation, and extinctions, models must consider a broad range of possibilities, and then the data may not contain sufficient information for us to discriminate between alternative hypotheses.

Two examples, among the many possible, illustrate this problem. High coalescence times were estimated for some nuclear genes (Harding *et al.*, 1997; Templeton, 2002; Garrigan *et al.*, 2005; Garrigan and Hammer, 2006), and interpreted as resulting from admixture between modern and archaic humans, and hence in favour of the MR model. But with an effective size (N_e) of around 10 000 and a generation time of 20 years, the expected time to the most recent common ancestor (T_{MRCA}) is close to 800 KY (Goldstein and Chikhi, 2002), i.e. close to the values estimated by Harding *et al.* (1997). Even higher coalescence times such as those found by Garrigan *et al.* (2005) can be accounted for by assuming a larger N_e, as has been found by some authors (e.g. Hammer *et al.*, 2004) and a larger generation length. However, subdivision of the ancestral human population in genetically differentiated subpopulations or, in other words, ancient population structure, seems to more simply explain the available data. On the other hand, population structure and extinctions can lead to low N_e estimates, even in populations whose census size was actually large. Thus, the low N_e value, around 10 000, often estimated and used for humans, could have been caused by extinctions of early communities that were spread across Africa and Eurasia, rather than by a really small founding human population (Eller, 2002).

The second example comes from the study of the geographic distribution of derived haplotypes (the 'D' haplogroups) at microcephalin and abnormal spindle-like microcephaly (ASPM), two loci involved in the determination of brain size, and hence potential determinants of some cognitive abilities. The D haplogroups are recent (their estimated ages are 37 and 5.8 KY, with large standard errors) and widespread outside Africa. These two findings led to the conclusion that the D haplogroups carried a selective advantage, possibly related with the emergence of modern human cognition, and hence increased in frequency through a selective sweep (Evans *et al.*, 2005; Mekel-Bobrov *et al.*, 2005). This conclusion was supported by neutrality tests and by simulations, in which neither a population expansion from Africa nor presence of ancient population subdivision in Africa could generate similar patterns in short evolutionary times. However, simulations considering *at the same time* the presence of ancestral differences among African subpopulations and expansion from one subpopulation, did indeed reproduce under neutral conditions the pattern observed (Currat *et al.*, 2006). Studies currently in progress suggest that non-pathologic variation at the microcephalin and ASPM loci is not associated with variation in IQ tests (Mekel-Bobrov *et al.*, 2007); to establish whether or not a selective sweep played a role in the diffusion of the D haplogroups, more robust demographic models are still needed.

Data from other fields, such as archaeology and linguistics, provide potentially useful information to refine demographic models. However, language change may or may not be associated with population changes, and, similarly, changes in the material culture documented in the archaeological record are not necessarily caused by demographic changes (Collard and Shennan, 2000). Therefore, predicting evolutionary change from linguistic and archaeological information remains a complicated task. Multidisciplinary, comparative studies have enormous potential, but a shared methodology that everybody might be satisfied with has not yet emerged (cf Cavalli-Sforza *et al.*, 1988; Bateman *et al.*, 1990).

A substantial leap forward in our ability to reconstruct at least aspects of our evolutionary past is represented by the technical possibility to directly study fragments of DNA coming from fossil bones. The main drawback of the polymerase chain reaction (PCR)-based methods for the study of ancient DNA is the risk of amplifying contaminating DNA (Serre *et al.*, 2004). In the study of non-human species, it is simple to tell contaminating, generally human, from endogenous DNA. Much greater problems of authenticity arise in the study of human fossil DNA, which cannot differ much from those of potential contaminators (Pääbo *et al.*, 2004).

The risk to mistake modern human sequences for the sequences endogenous to the ancient sample has not affected the study of the mitochondrial relationships between anatomically modern and archaic Europeans, the Neandertals. At present, seven sequences of the Neandertal mtDNA covering >300 bp have been published, spanning a time range between 50 and 29 KY ago. All these Neandertal sequences differ from those of modern individuals by at least 16 fixed nucleotide substitutions (Caramelli *et al.*, 2006), and hence cannot possibly result from contamination by modern people who manipulated the specimens. In addition, the Neandertal sequences bear no special resemblance to those of the Europeans, who, under the MR but not the OOA model, should be their direct descendants (Krings *et al.*, 2000).

Comparison of the first Neandertal sequence with modern European sequences under reasonable assumptions on population sizes, and hence on the impact of genetic drift, led Nordborg (1998) to conclude that a direct genealogical link between them is unlikely,

but still cannot be ruled out with statistical confidence. However, later work on the increasing number of sequences available shows that the Neandertal contribution to the modern European mtDNA gene pool is very limited (Serre *et al.*, 2004) and probably does not exceed 0.1 % (Currat and Excoffier, 2004). Conversely, Plagnol and Wall (2006), analysing modern patterns of linkage disequilibrium, concluded that even a value greater than 5 % is compatible with the data, and suggested that admixture took place outside Europe, namely in west Africa, a place, however, for which no fossils are currently available. Admixture between archaic and modern Europeans has also been proposed to account for the results of a study of the microcephalin locus (Evans *et al.*, 2006). Evidence against a significant Neandertal contribution to the mtDNA of modern Europeans comes from the analysis of two sequences from 24-KY old anatomically modern humans of the Cro-Magnoid type. Despite the limited chronological distance from Neandertals, these sequences fall in the cluster of modern European sequences, and are clearly distinct from the cluster formed by the Neandertal sequences, thus suggesting that Neandertals are unlikely to be the genetic ancestors of anatomically modern Palaeolithic Europeans (Caramelli *et al.*, 2003).

Because DNA is generally degraded in ancient samples, so far applications have been essentially limited to the study of mtDNA, which is present in multiple copies in each cell, and hence has a higher probability to be retrieved in reasonably good, amplifiable, conditions. Hopefully, technological advances may allow this to change in the near future (Binladen *et al.*, 2006; Green *et al.*, 2006). However, at present, there is no way to look directly at past nuclear DNA variation in humans.

The main feature of modern nuclear DNA variation at the global level are the continent-wide gradients of allele frequencies, spanning much of Eurasia and extending into the Americas and Oceania (Cavalli-Sforza *et al.*, 1994; Barbujani *et al.*, 1994; Chikhi *et al.*, 1998; Serre and Pääbo, 2004; Mulligan *et al.*, 2004; Ray *et al.*, 2005; Ramachandran *et al.*, 2005; Liu *et al.*, 2006; Amos and Manica, 2006), with a more complicated pattern in Africa, probably reflecting complex gene-flow patterns within the continent, and selection (Reed and Tishkoff, 2006). Also, in good agreement with processes of gene flow originating in Africa are measures of genetic diversity estimated from microsatellite data, with a clear trend from minimal values in the Americas to maximum values in Africa, and populations known to be of particularly large or small size behaving as outliers (Conrad *et al.*, 2006).

Even in the absence of data convincing everybody that all modern human genealogies can be traced back to recent African ancestors, it is clear that Africa played a special role in human evolution, as shown by patterns of genetic diversity (Watkins *et al.*, 2001) and of linkage disequilibrium (Tishkoff *et al.*, 1996). Two main migration routes have been proposed on the basis of the available genetic data. An exit and dispersion through the Near East into Eurasia is commonly accepted and in good agreement with fossil evidence (Lahr and Foley, 1998) and nuclear diversity (Ramachandran *et al.*, 2005; Liu *et al.*, 2006) but another exit through the so-called southern route is currently receiving increasing attention (Luis *et al.*, 2004; Mellars, 2006). Indeed, mitochondrial DNA diversity in Asian genetic isolates has been interpreted as suggesting an early (65 KY ago) expansion through the Red Sea and India, directly into southeast Asia and Oceania (Macaulay *et al.*, 2005). Levels of linkage disequilibrium are maximal in the Americas and Oceania, lower in Eurasia and minimal in southern and central Africa, consistent with repeated founder

effects in the course of an expansion from Africa (Tishkoff and Verrelli, 2003; Tishkoff *et al.*, 1998; 2000; Sawyer *et al.*, 2005; Montpetit *et al.*, 2006).

31.3   HUMAN GENETIC DIVERSITY: GEOGRAPHICAL STRUCTURE

Humans have been described for centuries as subdivided in discrete and easily identifiable genetic subgroups or races. During much of the twentieth century, most geneticists believed that these races differed profoundly in their physical and mental characteristics and that these differences have a biological basis and are hereditary (Provine, 1986). At the end of the century, anthropologists and population geneticists had mostly abandoned this idea, which was perceived to conflict with the available genetic evidence. However, genetic epidemiologists brought it back to the centre of the attention with the beginning of the twenty-first century.

Indeed, our probability to develop a disease reflects the interplay between two main classes of risk factors, genetic and environmental. Both are at present impossible to exactly measure, either because they are unknown (e.g. all the loci involved in the onset of a complex disease, and the way they interact) or because they are difficult to quantify (e.g. the effects on health of our diet, of the exposure to chemicals, smoke, alcohol, and of our lifestyle in general). As a consequence, epidemiologists are forced to use approximations in their effort to prevent disease occurrence, and it has been proposed that a way to approximate genetic disease risks is to classify people in traditional racial categories (Risch *et al.*, 2002; Burchard *et al.*, 2003; Mountain and Risch, 2004; Tang *et al.*, 2005; Risch, 2006). The task may not be easy because in the course of history no consensus has emerged on the number and definition of human races, with catalogues including from two to hundreds of proposed races.

31.3.1   Catalogues of Humankind

Only in the last few centuries could questions about human diversity be addressed scientifically, but in fact these questions are as old as humankind. Differences among humans were evident to the earliest naturalists; descriptions and classifications of human types can be found in many ancient texts. A central question was whether we belong to one or to several species. Such polygenistic theories flourished up to the nineteenth century (Cohen, 1991), but the development of evolutionary studies, and the demonstration that there is no reduction of fertility in crosses between members of very distant human populations (Chung *et al.*, 1967) led to their dismissal.

Science has long left creationism and polygenism behind, but the idea that a scientific study of humans necessarily starts from their racial classification was long unchallenged (Cohen, 1991) and still remains attractive to some. People differ in their aspect; all of us recognise variation in facial traits, hair textures, height, body structure, and skin colour. The typological approach consists in identifying some basic human types, defined on the basis of such traits, and then attributing individuals to one of those types, or races. Accordingly, starting with Linnaeus and for at least two centuries, analyses of human biological diversity were essentially race catalogues, differing from each other for the number and definition of the various items (Bernasconi, 2001). In turn, racial types were thought to be associated with other characteristics such as temperament, behaviour, and intellectual abilities (Cole, 1965). The main difficulty was represented by the fact that

it is fairly simple to list typical anatomical features of a region or a population, but each human group, however defined, includes variable proportions of people who do not resemble much the typical individual. Even in the early years, it was evident that single morphological traits do not allow a stable classification of humans into discrete groups. Therefore, races were defined by a combination of several traits, often both biological and non-biological, the latter including language, house-building and tool-making techniques, mating systems, and other cultural variables (see Cohen, 1991, where reference to the original eighteenth and nineteenth century sources can be found).

The consequences of these difficulties are evident in the lists of races described in Table 31.1. Starting from Linnaeus' six races and going through Buffon's, Blumenbach's, and Cuvier's systems into the twentieth century, the number of races increased. In his *Systema naturae,* Linnaeus first defined the species *Homo sapiens* within the order primates and divided it in six varieties, four corresponding to continents (Australia was

Table 31.1 An incomplete scheme of proposed catalogs of human races. Reference to the original works can be found in Cohen (1991) and Barbujani (2005).

Author	Number of races	Races proposed
Linnaeus (1735)	6	Europaeus, Asiaticus, Afer, Americanus, Ferus, Monstruosus
Buffon (1749)	6	European, Tartar, Laplander, south Asian, Ethiopian, American
Blumenbach (1795)	5	Caucasian, Mongolian, Ethiopian, American, Malay
Cuvier (1828)	3	Caucasoid, Negroid, Mongoloid
Deniker (1900)	29	–
Weinert (1937)	17	–
Von Eickstedt (1937)	38	–
Biasutti (1959)	53	–
Coon (1962)	5	Congoid, Capoid, Caucasoid, Mongoloid, Australoid
Garn (1965)	9	African, European, Asian, Indian, Amerind, Melanesian, Polynesian, Micronesian, Australian
US Office of Management and Budget (1997)	5	African-American, White, American Indian or Alaska Native, Asian, Native Hawaiian or Pacific Islander
Metropolitan Police Service, London, UK (2005) (`http://en.wikipedia.org/wiki/Race`)	16	White-British, White-Irish, any other white background, Asian-Indian, Asian-Pakistani, Asian-Bangladeshi, any other Asian background, Black Caribbean, Black African, any other black background, Chinese, White and Black Caribbean, White and Black African, White and Asian, any other mixed background, any other race
Risch *et al.* (2002), Figure 1	5	African, Caucasian, Pacific islanders, east Asian, Native American
Risch *et al.* (2002), Table 3	5	African Americans, Caucasians, Hispanic Americans, east Asians, Native Americans

missing) and two, *H.s. ferus* and *H.s. monstruosus* (savage and monstrous), designating respectively wild human forms whose existence was not confirmed by successive work, and carriers of congenital malformations. These two races were eliminated from the list and Australia was added by Buffon, who considered laplanders as an additional, separate race. At the end of the eighteenth century, it was the German anatomist Blumenbach who, while refusing a relationship between humans and the other primates, proposed that humankind be composed of five races, broadly corresponding to the inhabitants of the five continents, four of them regarded as more or less serious degenerations from the European race, which he first termed *Caucasian*.

In the nineteenth and early twentieth centuries, as anthropologists came in contact with more and more populations, fitting all of them into pre-existing races proved difficult. The catalogues became broader, crossing in some instances the 100-mark (Molnar, 1975), and the borders between races therein ambiguous. In this way, the need emerged to explicitly define what a race is. Although published definitions differ, many converge around the idea that races are subspecies, namely collections of individuals associated with a geographical region, and separated by biological boundaries from other groups that differ from them in some measurable features (Mayr, 1947; Coon *et al.*, 1950). Classical population-genetics theory and empirical data show that genetic variation is shaped by a combination of factors affecting the genome (such as mutation, recombination, and selection) and acting at the population level (such as subdivision, admixture, migration, and population-size changes). Under reproductive isolation, genetic drift affects independently each reproductive group. This tends to reduce the group's internal variation, while groups diverge from each other, both phenomena causing the onset of zones of rapid genetic change, i.e. boundaries (Barbujani *et al.*, 1989). Conversely, when groups or populations exchange migrants, the effects of drift are opposed by those of gene flow, and genetic variation between populations and groups is continuous (Wright, 1943). Therefore, genetically differentiated groups that one can legitimately regard as races are surrounded by biological boundaries, which are the product of genetic drift affecting populations connected by little (or no) migratory exchanges.

By the second half of the twentieth century, a growing number of investigators had become unhappy with the idea that the discontinuous entities thus far defined give a faithful representation of human variation. This assumption appeared to conflict with the continuous change observed for most morphological (and, later, genetic) traits. An additional problem was the fact that different traits yield discordant human taxonomies (King and Wilson, 1975; Brown and Armelagos, 2001). Individuals can be clustered on the basis of, say, skin colour, but the clusters are not the same if a different trait, say skull shape, β-globin, or lactose intolerance, is considered (Relethford and Harding, 2001). With the introduction of the concept of cline (Huxley, 1938), namely a gradient of morphological measures or allele frequencies in the geographical space, continuous models of human population structure became conceivable, which ultimately led to the proposal that the concept of race is misleading for describing human biological diversity (Montagu, 1941; Livingstone, 1962). Dobzhansky (1967) maintained that human races could nevertheless be defined, if not by fixed allelic differences, at least as open genetic systems, each differing to some extent from its neighbours for some allele frequencies. However, it has been objected that according to this definition any human population would be a distinct race, which does not correspond to the general use of the concept in

evolutionary biology. This debate is still open, with different authors siding with either Dobzhansky or Livingstone.

In what can be regarded as the last attempt to base a racial classification on anatomy, Coon (1963) proposed that the apparent complexity of the human population structure disappears if one disregards the effects of recent migration. In that way, humans can be subdivided into five major races, two in Africa, and one in Europe, Australia, and Asia, the latter including native American populations. However, by the early 1960s, the typological approach had shown its drawbacks, and the number and definition of human races had become extremely uncertain. In the meantime, genetic data had begun to accumulate, as well as quantitative methods for their analysis (Edwards and Cavalli-Sforza, 1965; Cavalli-Sforza, 1966; Sokal *et al.*, 1989). Most studies of the last decades of the twentieth century focused, then, on the levels and patterns of genetic variation in the geographical space, summarised in the atlas of Cavalli-Sforza *et al.* (1994).

31.3.2 Methods for Describing Population Structure

In this section, we shall focus on two main sets of methods for assessing the strength of population structure from multilocus genotypes given a tentative classification based on some non-genetic criteria such as physical morphology or geography. Under the first approach, similar to an analysis of variance, genetic diversity is compared within and among the pre-defined groups. The alternative approach is to define a criterion of similarity among genotypes and hence classify them into groups ignoring the available group labels. Later, the group labels are revealed in order to quantify the accuracy of the assignments. We now discuss these two approaches in more detail.

31.3.2.1 *Genetic Diversity within and among Populations*

Classical analysis of variance is unsuitable for assessing genetic variance within and between groups because the distribution of allele frequencies is not expected to be normal, but is better approximated by the beta or Dirichlet distributions (Wright, 1931). An ideal statistic for comparing variation within and between groups should have the following properties: (1) take value 0 when there is no polymorphism; (2) increase with the number of alleles; (3) reach a maximum when the frequencies of all alleles are equal, for any number of alleles; and (4) be convex, i.e. it should increase when groups are pooled, unless these averages are identical (Lewontin, 1972). Wright's (1943) F_{st} is an example of a convex statistics. These criteria led Lewontin (1972) to apportion the global species diversity at three hierarchical levels using the Shannon information measure, defined as

$$H = -\Sigma p_i \ln_2 p_i,$$

where p_i is the frequency of the ith allele at a locus, and summation is over all alleles. H is calculated independently for each locus and each population, and its average is calculated respectively within populations (H_{pop}), among populations attributed to one race (H_{race}), and in the entire species ($H_{species}$). The global genetic diversity was then partitioned at the three levels as follows:

Variance within populations: $V_{wp} = H_{pop}/H_{species}$

Variance between populations, within races: $V_{\mathrm{bp}} = (H_{\mathrm{race}} - H_{\mathrm{pop}})/H_{\mathrm{species}}$

Variance between races: $V_{\mathrm{br}} = (H_{\mathrm{species}} - H_{\mathrm{race}})/H_{\mathrm{species}}$.

Although it is customary to refer to these V indexes as variances, this is not strictly correct. In particular, since V_{bp} and V_{br} are estimated by subtraction, when there is little geographical structure these statistics can take negative, if small, values (see e.g. Jorde *et al.*, 2000).

With the advent of techniques for the direct study of DNA, measures of molecular differentiation between alleles were incorporated into the statistics. Much like in Lewontin's approach, analysis of molecular variance (AMOVA) is a non-parametric method for the analysis of variance that subdivides genetic diversity estimated from molecular data into hierarchical components (Excoffier *et al.*, 1992; see also **Chapter 29**).

AMOVA tests by a permutational procedure the significance of each variance component, whether estimated from allele-frequency differences or considering molecular information also. For this purpose, the null hypothesis is that all samples come from an unstructured population, without differences among the groups (populations or races) defined within it, so that all variation is due to the random sampling of individuals or populations. The three variances (which one can refer to as *pseudovariances*) are recalculated by assigning individuals and populations to random geographic locations, according to three independent resampling schemes; the procedure is repeated many times, so as to obtain empirical null distributions for the three pseudovariances. The observed variances are eventually compared with these distributions, and observed values falling in the upper percentiles are judged significant at the appropriate level.

31.3.2.2 *Classification Methods*

In discriminant analysis (Ripley, 1996), also known as *supervised classification*, each individual's group allocation is ignored in sequence, and their genotype is assigned to the most likely source population using the pre-defined allocations of all other individuals as the reference data set. The similarity between genotypes is usually measured as the minimum (and most likely) number of mutational events separating them. Both parametric and non-parametric methods of discriminant analysis are available. However, the standard parametric methods, namely linear, logistic, and quadratic discriminant analysis, assume that the variables are at least approximately normal. Rannala and Mountain (1997) proposed an assignment method specifically designed for genetic data, which has also been used for discriminant analysis (Romualdi *et al.*, 2002). Under the assumptions of Hardy–Weinberg and linkage equilibria, this method computes the probability of multilocus genotypes originating from different potential reference/source populations and allocates them to the most likely group of origin. Among non-parametric methods, the simplest approach is to assign each genotype to the group to which the majority of the k nearest genotypes belong; k is chosen arbitrarily in a range of reasonable values. Under more complex approaches, such as kernel methods, the density of the genotype frequencies is initially estimated in each group, and then a new observation (genotype) is assigned to the group for which its estimated density is maximal.

Unsupervised clustering methods ignore any pre-defined group assignments and infer from the data groups of genetically similar individuals. The underlying idea is to use a

model-based approach to define the probability of generating the data assuming K hidden partitions. Typically, the basic model assumes that, within the partitions, alleles are independent within and between loci. Once this probability can be calculated or approximated, it is possible to construct an MCMC algorithm that will explore the parameter space in such a way that the parameter values sampled during an MCMC run are visited in proportion to their posterior probability of generating the data. In the general problem of uncovering hidden structure, parameters may include the number of partitions (Dawson and Belkhir, 2001) or the proportion of genes (in a population or in an individual) coming from any of K partitions conditional on K (Pritchard *et al.*, 2000; Falush *et al.*, 2003).

The most popular such approach is the one originally developed by Pritchard *et al.* (2000), later extended by Falush *et al.* (2003) and implemented in the STRUCTURE software (hereafter named the Pritchard and Falush method; see also **Chapter 30**). This method allows data analysis under two ancestry models, with or without admixture. In the simplest case of no admixture, the relevant parameters are (1) the population of origin of each individual and (2) the allele frequencies of each population. When admixture is assumed, an extra parameter is considered, (3) the proportion of each individual's ancestry from each population. The number of partitions K is not estimated and so must be specified, but Pritchard *et al.* (2000) proposed a criterion for choosing from multiple runs of the algorithm that providing the optimal K value. Recent simulation work on complex population structure where patterns of gene flow are not homogeneous has shown that this *ad hoc* method can be misleading (Evanno *et al.*, 2005).

Another approach is the Dawson and Belkhir (2001) method, based on an MCMC algorithm implemented in the PARTITION software, in which the parameter space is defined by (1) K, the number of possible partitions; (2) the distribution of individuals in the K partitions; and (3) the allele frequencies within partitions. During the exploration of the parameter space, the likelihood of the data is estimated for the different K values and for possible assignments of individuals to the K partitions. Thus, when the chain reaches equilibrium, the different values of K have been sampled in proportion to their probability of generating the data, and so it is possible to estimate the posterior probability distribution of K. The Dawson and Belkhir method also estimates the probability that sets of individuals lie in the same partition. Only the maximum number of partitions allowed has to be specified beforehand, whereas in the Pritchard and Falush method a pre-specified number of populations, K, is given and the algorithm uses a Gibbs sampler to obtain the posterior distribution of the parameters, i.e. the values of each parameter conditional on the observed individual's genotypes and on the number of populations K considered.

The Pritchard and Falush approach has some drawbacks, but it is also the most flexible. Besides assigning individuals to their population of origin and identifying possible migrants or admixed individuals, STRUCTURE can also incorporate information about the geographic origin of individuals, which is treated as a prior in the cluster analysis. The algorithm can also allow for allele frequencies to be correlated across populations (Falush *et al.*, 2003). This can be a crucial point when the groups being studied are not genetically distinct.

Other recent methods are suitable to detect and quantify population structure (Corander *et al.*, 2004; 2007; Corander and Marttinen, 2006). Corander and colleagues have developed Bayesian model–based clustering methods implemented in the different versions of the BAPS software. In the most recent version, Corander *et al.* (2007) proposed a method analogous to Dawson and Belkhir's in the sense that it tries to find the partition

that best fits the data, treating as random variables the allele frequencies and the number of genetically divergent populations. However, contrary to the alternative approaches, BAPS is based on *a priori* information provided by the geographic location of sampled individuals. Given a maximum value of partitions, the algorithm uses a stochastic optimisation procedure (rather than an MCMC approach) to find the clustering solution with the highest marginal likelihood of K (i.e. the most probable number of differentiated populations conditional on observed data). For more extensive, recent reviews, the interested reader should refer to Beaumont (2004) and Chikhi and Bruford (2005).

31.3.3 Identifying the Main Human Groups

In the first global analysis of human genetic diversity, Lewontin (1972) quantified variation within and among seven groups, namely Caucasians (including western Asians and north

Table 31.2 Estimated percentages of the global human diversity at three hierarchical levels of population subdivision.

Polymorphism, number of loci	References	Within population	Between populations, within race or continent	Between races or continents
Protein, 17	Lewontin (1972)	85.4	8.3	6.3
Protein, 18	Latter (1980)	84.0	5.6	10.4
Protein, 18	Latter (1980)	87.0	5.5	7.5
Protein, 18	Latter (1980)	83.8	6.6	9.6
Protein, 25	Ryman *et al.* (1983)	86.0	2.8	11.2
mtDNA	Excoffier *et al.* (1992)	75.4	3.5	21.1
Autosomal DNA, 109	Barbujani *et al.* (1997)	84.4	4.7	10.8
MtDNA (non-coding)	Seielstad *et al.* (1998)	81.4	6.1	12.5
Y chromosome, 10	Seielstad *et al.* (1998)	35.5	11.8	52.7
Autosomal DNA, 90	Jorde *et al.* (2000)	84.8	1.6	13.6
MtDNA (non-coding)	Jorde *et al.* (2000)	71.5	6.1	23.4
Y chromosome, 10	Jorde *et al.* (2000)	83.3	18.5	−1.8
Autosomal DNA, 21	Romualdi *et al.* (2002)	82.9	8.2	8.9
Y chromosome, 10	Romualdi *et al.* (2002)	42.6	17.3	40.1
β-Globin	Romualdi *et al.* (2002)	79.4	2.8	17.8
Autosomal DNA, 377	Rosenberg *et al.* (2002)	93.2	2.5	4.3
Autosomal DNA, 377	Excoffier and Hamilton (2003)	87.6	3.1	9.2
Y chromosome, 13	Wilder *et al.* (2004)	64.3	17.0	18.7
MtDNA (coding)	Wilder *et al.* (2004)	59.9	21.1	18.9
X chromosome, 17	Ramachandran *et al.* (2004)	90.4	4.6	4.9
Autosomal indels, 40	Bastos-Rodrigues *et al.* (2006)	85.7	2.3	12.1
Median, all loci[a]		82.8	6.1	11.1
Median, autosomal[a]		85.7	4.6	9.7

[a] These values were obtained by treating all studies equally. The medians at the three levels of population subdivision do not sum up to 1, and hence they were normalised by dividing them by 100.6 for all loci, 99.0 for autosomal loci.

Africans), Black Africans, Mongoloids, south Asian aborigines (dark-skinned populations of India and the Asian southeast), Amerinds, Oceanians, and Australians (Table 31.2). The estimated variances differed across the 17 loci, as well as the sample sizes but, on average, 85.4 % of the total was attributed to differences between members of the same population, 8.3 % to differences between populations of the same group, and 6.3 % to differences between groups. Lewontin concluded that racial classification has no genetic or taxonomic significance.

Nei and Roychoudhury (1972) compared three groups, white, black, and Japanese, and found that allele frequencies in the three groups are remarkably similar, despite the conspicuous phenotypic differences. They suggested that the genes controlling pigmentation and facial structures evolved under stronger natural selection than 'average genes' (Nei and Roychoudhury, 1972), a notion that gained broad acceptance (see, e.g. Rees, 2003). On the other hand, Lewontin's bold conclusions also prompted a number of criticisms for more than three decades (see Edwards, 2003) discussed below. However, reanalyses of Lewontin's data by three alternative methods, based on the proportion of alleles that two random multilocus genotypes have in common, gave only slightly different results, with variances within populations estimated around 84.0 % of the total (Latter, 1980). Genetic variances largely overlapping with Lewontin's and Latter's estimates were also inferred from enzyme and serum protein allele frequencies, and from blood groups, in three populations from Africa, Europe, and Asia (Ryman et al., 1983).

The first apportionment of human genetic variances at the DNA level was an illustration of the AMOVA algorithm. Excoffier et al. (1992) found that the differences between continents in mitochondrial DNA RFLP diversity were larger than estimated in protein studies, and populations contained 75 % of the global mtDNA diversity. Conversely, in large-scale analyses of autosomal DNA, genetic variances were very close to those inferred from protein polymorphisms, with differences among individuals of the same population accounting for 84.5 % of the overall variance for both microsatellites and RFLPs (Barbujani et al., 1997). Genetic differences among continental populations were small in absolute terms, between 8 and 11.7 % of the total, but yet significantly greater than zero at most loci.

Similar figures have been inferred from studies of other autosomal polymorphisms (Rosenberg et al., 2002), including Alu insertions (Romualdi et al., 2002). Actually, the variances among continents estimated in the former study, by far the largest so far for the number of screened loci, are less than 5 %, although a reanalysis of the same data based on a more appropriate mutational model suggests a twice-as-large value (Excoffier and Hamilton, 2003). Despite these inconsistencies, the impression one receives from studies of autosomal polymorphisms (Table 31.2) is that a large share of the global genetic diversity, very close to Lewontin's 85 %, is present in each human population, with small differences between non-coding DNA regions (e.g. Excoffier and Hamilton, 2003) and coding regions subjected to selection, such as the β-globin locus (Romualdi et al., 2002).

Greater differences between populations and continents are usually inferred from markers transmitted by one parent only. In studies of biallelic Y-chromosome markers, between 40 % (Romualdi et al., 2002) and more than half of the total variance (Seielstad et al., 1998) was found to differentiate continents, but that component was zero in a study of short tandem repeats (STRs) (Jorde et al., 2000). The absence of detectable differences among continents in the study of STRs may depend, at least in part, on the high mutation rate for STRs and on constraints on the number of repeats, both phenomena leading, in

the long run, populations of all continents to converge towards similar allelic distributions (Jorde *et al.*, 2000).

In general, given the smaller population sizes for uniparentally transmitted loci (in principle, one-fourth of the autosomal population size), higher diversity between populations and groups thereof should be expected for these markers, under a simple model of population differentiation by genetic drift. Therefore, the high variances between populations and groups inferred from the Y chromosome are easier to reconcile with theory than the lower variances inferred from mtDNA. This finding may reflect either an increased average impact of gene flow on females, or of drift on males, or both. Increased female migration is a consequence of patrilocality, i.e. a higher tendency to move to the spouse's place of origin for women than for men (Seielstad *et al.*, 1998); increased drift in males may mean that their effective population sizes are reduced by their higher variance in reproductive success, associated with the practice of polygyny (Dupanloup *et al.*, 2003).

All the above factors may have, or have had, an impact on the patterns of human diversity, but a part of the differences in the apportionment of genetic variances inferred from different markers may be due to biases in the choice of the markers. After identifying areas of high polymorphism on the Y chromosome, Wilder *et al.* (2004) sequenced more than 6 Kb in individuals from four continents, compared the observed sequence variation with variation in mtDNA sequences, and actually observed a slight excess of between-continent diversity for mtDNA than for the Y chromosome. Therefore, when sequences of random Y-chromosome regions are compared with mtDNA sequences, (1) similar distributions of variances emerge, and (2) the differences between populations (of the same or of different continents) is larger than inferred from autosomal DNA, which makes sense in the light of the different effective population sizes, fourfold larger for autosomal than for mitochondrial and Y-chromosome polymorphisms.

In all these studies, the global species diversity is probably overestimated because of (1) the unavoidable selection of populations that are far removed in the geographical space, to the exclusion of populations that may be of ambiguous classification, such as those at the boundaries between continents and (2) the over-representation of populations of anthropological interest, often reproductively isolated and hence presumably highly differentiated, rather than admixed urban populations. Because the internal diversity also tends to be low in genetic isolates, one can also expect the fraction of within-population variance to be underestimated. A random sample of humankind, say based on individuals collected at the nodes of a regular grid superimposed on the world map and taking population densities into account, would contain a greater proportion of admixed individuals and recent immigrants, presumably resulting in even higher diversity within populations.

On the other hand, genetic differences between putative races or continents, albeit small, are significantly greater than zero. This prompts the question whether these differences are large enough to allow one to distinguish discrete racial group, as suggested, among others, by Edwards (2003).

In evolutionary trees inferred from microsatellites (Bowcock *et al.*, 1994), or combining information from microsatellites, insertion/deletion and restriction polymorphisms (Jorde and Wooding, 2004), people of similar geographical origin tend to form clusters, and plausible evolutionary relationships among populations are apparent. As is commonly observed in humans (starting from Cann *et al.*, 1987), these trees have little deep structure and long terminal branches.

The question whether the limited genetic variances among populations and groups thereof are large enough for individuals to be accurately classified in continental groups was addressed considering Alu insertion/deletion polymorphisms and Y-chromosome SNPs. Using several discriminant analysis methods, Romualdi *et al.* (2002) could correctly assign to the right continent up to 73 % of their multilocus genotypes, with Y-chromosome data allowing a more accurate assignment. Accuracy decreased drastically as the number of partitions increased, i.e. when genotypes were assigned to subcontinental regions, and the misclassification rate decreased only slightly by adding new markers, once some 20 polymorphisms were considered (Romualdi *et al.*, 2002).

The choice of the markers and of the populations deeply affects the results of the analysis, suggesting that discordant variation is not only typical of morphological traits but also of genetic polymorphisms. Analysis of two sets of Alu polymorphisms by STRUCTURE showed that, depending on the loci considered, either three clusters (two worldwide distributed and one Eurasian) or four clusters (two in Eurasia and two corresponding respectively to Africa with Oceania and Asia with the Americas) were identified. These clusters did not overlap (Romualdi *et al.*, 2002), neither did they correspond to the clusters obtained by STRUCTURE in studies of X-chromosome STR markers, loci of pharmacogenomic relevance (Wilson *et al.*, 2001), and autosomal STR markers (Rosenberg *et al.*, 2002). Rosenberg *et al.* (2002) inferred the existence of six clusters from the analysis of a 377 STR markers from the Centre pour l'Etude des Polymorphismes Humains (CEPH) Diversity Panel, five of them corresponding to the main continents (with central Asia clustering with Europe rather than with east Asia), plus the Kalash population of Pakistan, presumably a genetic isolate. In a successive analysis of a dataset of almost 1000 STR markers by the same authors, the Kalash no longer seemed to be a genetic outlier, and appeared to be part of the European-Western-Central Asian cluster, but things became more complicated in the Americas, where two new clusters emerged (Rosenberg *et al.*, 2005).

The 377-marker CEPH dataset has been reanalysed by several authors. Using a Bayesian MCMC approach, Corander *et al.* (2003) found that more than six groups are needed to represent global human genetic diversity, with evidence for subdivision in South America). Barbujani and Belle (2006) used a method for recognising genetic boundaries in maps of genomic variation (Manni *et al.*, 2004), finding evidence for several distinct clusters in the Americas, and for three clusters separated by zones of rapid genomic change in Africa, a result consistent with extensive diversity known to exist in that continent (Kaessmann *et al.*, 1999; Yu *et al.*, 2003; Watkins *et al.*, 2003). Evidence for substructuring in Africa had also been detected by STRUCTURE in one analysis where the inferred clusters were four (Bamshad *et al.*, 2003). On the other hand, Serre and Pääbo (2004) investigated how study design can influence the inferred clustering. By resampling the CEPH dataset through a scheme devised to reduce the effects of the necessary discontinuous geographical collection of samples, they came to the conclusion that only genetic gradients, and not boundaries, are apparent when individuals are sampled homogeneously from all continents, and concluded that there is no evidence for major genetic discontinuities in the human species.

Is there a way to reconcile all these results? And which is the real structure of the human population? Some inconsistencies may be due to differences in the distribution of the studied populations. However, several of the above analyses were based on the same 377-STR CEPH diversity dataset, and yet their results are discordant as for the number

and geographical span of the genetic clusters identified. Studies of different dataset yield an even more complicated picture (Wilson *et al.*, 2001; Romualdi *et al.*, 2002; Bamshad *et al.*, 2003; Tang *et al.*, 2005; Barbujani, 2005), and it is legitimate to say that the main common element one can recognise is that each study is inconsistent with all the others. One plausible conclusion is then that this incongruence represents a basic feature of human diversity, and that racial categories are imposed on data rather than inferred from them (Cooper, 2005). Different genetic polymorphisms show discordant geographic patterns, i.e. are differently distributed over the planet, and their distributions are generally weakly correlated or uncorrelated. Genetic clusters can be inferred from the data, but the fact that two populations fall in the same cluster (or in different clusters) when described at loci A, B, C does not imply that they will fall in the same cluster (or in different clusters) based on loci X, Y, Z.

31.3.4   Continuous versus Discontinuous Models of Human Variation

Continuous genetic variation in the geographic space, such as that commonly observed in humans, may result from three classes of evolutionary phenomena, one of which, selection along a gradient, is expected to affect single loci, but not the genome as a whole (Cavalli-Sforza, 1966; Kayser *et al.*, 2003). When clines are observed over a number of loci, they likely reflect population dispersal, or isolation by distance, or both. Isolation by distance can be loosely defined as the process whereby the effects of drift are independent in each population, but gene flow rates are higher between close than between distant populations, and so genetic similarity tends to decline with the spatial distances (Wright, 1943). Population dispersal differs from isolation by distance in that it entails directional migration, and can generate clinal genetic change either by demic diffusion or by founder effects. Demic diffusion (Menozzi *et al.*, 1978) means that the geographical expansion is accompanied by demographic growth. If the expanding population mixes and forms hybrids with previously settled populations, the fraction of members of the expanding population that will contribute to the hybrids decreases with the distance from the place of origin, and so does their contribution to the gene pool of the hybrid populations. Repeated founder effects may also result in clines because, as small numbers of individuals disperse, alleles are lost from their gene pool, and occasionally reintroduced by local gene flow.

Application of a formal model of isolation by distance to the analysis of worldwide patterns of protein and DNA polymorphisms, and of craniometric measures, showed an excellent fit of the model for both genetic and anatomical data, suggesting that patterns of human diversity can largely be accounted for by the simple interaction of drift and geographically structured gene flow (Relethford, 2004). These results suggest that, as a rule, migrational networks left a deep mark in the observed patterns of human biodiversity. The exceptions, namely zones of rapid genetic change or genetic boundaries, point to migrational or reproductive barriers. In the few studies so far attempting to map them at the DNA level, genetic boundaries have been shown to occur mostly between small genetic isolates, sometimes separated by just a few kilometres, such as the Suruì and Karitiana populations of Brazil (Barbujani and Belle, 2006). This suggests that, at the small geographical scale, chance and possibly adaptation to local conditions may lead to even sharp population differentiation, thus causing local departures from the general clinal pattern.

31.4 FINAL REMARKS

As we have seen, the interpretation of the data on human genetic diversity is controversial in several areas. However, it is at least well established that the main fraction of the global human genetic diversity is found between individuals of the same population. We also know that the remaining fraction, representing differences among populations and population groups, is small but not zero. Indeed, it is large enough for most of us to form an opinion, if rough, on the likely origin of the people we meet just by looking at them, and for population geneticists to reconstruct aspects of population history. However, different loci show uncorrelated, and sometimes very different, patterns of variation, a phenomenon termed *discordant variation* (Brown and Armelagos, 2001). Therefore, predictions of epidemiological relevance based on the patient's physical aspect or supposed racial affiliation may be severely inaccurate (Wilson *et al.*, 2001; Cooper *et al.*, 2003; Cooper, 2005; Orduñez *et al.*, 2005; Bamshad, 2005). Countries or regions where several subpopulations coexist in reproductive isolation, with little or no genetic exchange, may represent exceptions. Examples include India, where genetic differences are apparent among castes (Cordaux *et al.*, 2004; Watkins *et al.*, 2005); the Yanomama (Neel, 1978) and other Amazonian tribes (Crawford, 1998), in which small population sizes combined with the tendency of tribes to split along family lines led to extreme genetic drift effects; and various urban areas of the United States (Shriver *et al.*, 2004).

DNA variation has probably been studied among US population groups more extensively than in any other area of the world, and US ethnic groups show different allele frequencies at many loci (Shriver *et al.*, 2005; Redd *et al.*, 2006). The problem is to what extent and for what purposes the results of these studies can be generalised to other populations. Characteristic alleles, present at high frequencies in one US population while rare or absent in the others, were initially defined population-specific alleles (PSA), and proposed for studies of admixture and DNA-based personal identification (Shriver *et al.*, 1997; Parra *et al.*, 1998). Later, the set of markers was extended and refined, but also relabelled as AIM (ancestry informative markers), thus suggesting that by using these markers it is possible to make one more step, namely to correlate skin colour with biological ancestry (Shriver *et al.*, 2003). This view was echoed by authors who concluded that human genomic diversity is discontinuously distributed according to groups that correspond well to 'common concepts of race' (Rosenberg *et al.*, 2002; 2005). However, in a study of Brazilian individuals, Parra *et al.* (2003) found no correlation between skin colour and frequency of 10 AIM, selected as those that had shown the maximum differentiation between Africans and Europeans. Thus, genetic markers that are population specific, and perhaps even ancestry informative in certain populations, are neither population specific nor ancestry informative elsewhere. Local genetic differentiation may be sharp for certain loci and useful for certain purposes, but studies of single countries are unlikely to give us a sensible, all-purpose description of human diversity in general, especially if these countries are inhabited by people whose ancestors evolved elsewhere for millennia.

Overall, the global patterns of human diversity are the likely product of a comparatively short evolutionary history. During less than 200 KY, a population of six billion individuals has quickly developed, largely or exclusively, from a presumably small number of recent African founders. In their dispersal across the planet, these ancestors are unlikely to have mixed with the descendants of people who had previously left Africa. The main evolutionary processes affecting them were probably repeated bottlenecks, dispersals, and

isolation by distance, all processes accounting for the broad gradients of genetic diversity observed (Cavalli-Sforza *et al.*, 1994). However, migration and genetic drift were not the only forces shaping human genetic diversity. Recent admixture episodes and selection have left a mark in current genetic diversity. Genomic regions of intense, and often environment specific, selection have been identified analysing both low-resolution and high-resolution single nucleotide polymorphism data (Kayser *et al.*, 2003; Akey *et al.*, 2002; 2004; Storz *et al.*, 2004; Voight *et al.*, 2006), and the different selection regimes, along with chance, may explain why patterns of genetic variation are discordant across loci.

Discordant variation has practical implications. Calculating how many polymorphic markers would be necessary to confidently assign individuals to racial groups, Risch *et al.* (2002) estimated that, with average allele-frequency differences between racial groups $= 0.6$, 17 markers would be enough, with an error rate as low as 10^{-5}. More markers would be needed if average allele-frequency differences are $= 0.1$, but with 474 loci the error rate would still be a comfortable 10^{-3}, and with 901 markers an excellent 10^{-5}. Because allele-frequency differences ≥ 0.4 are not uncommon between US ethnic groups, Risch *et al.* (2002) concluded that statistically robust racial clusters can be obtained by typing a few hundred markers. A similar rationale underlies a study suggesting that the limited genetic differences among putative races are sufficient to define well-distinct genetic clusters, provided one is willing to combine information at many loci (Edwards, 2003). The problem is that, for the calculations proposed in both such studies, genetic diversity is assumed to be concordant across loci; the more the loci in the analysis, the stronger the signal of differentiation. As we have seen, in humans that is not the case.

An additional problem is that, to test whether genetic differences confirm common concepts of race (as stated by Risch, 2006), one needs a list of biological races that most people are reasonably happy with, but so far there has been no agreement on such a list. This can only mean that, in practice, the task of compiling it is either impossible, or so difficult that nobody has been able to succeed. In particular, classical attempts to infer racial subdivision from skin colour have been remarkably unsuccessful (Cohen, 1991), and now we know why (Jablonski, 2006). Tens of loci affect pigmentation in humans, and they have epistatic effects. Most likely, these loci evolved under contrasting selective pressures, so that their current diversity represents a trade-off between protection against UV solar radiation (greater with darker skins) and the synthesis of vitamin D (more efficient with lighter skins) (Makova and Norton, 2005). There is evidence of relaxation of selection in non-African populations (Harding *et al.*, 2000), and different loci were probably selected at different times and places, so that some candidate genes show evidence of selection in the ancestral African population(s), others during the exit from Africa, and others in response to more recent local pressures (McEvoy *et al.*, 2006). Women have, on average, lighter skins than men of the same population, and this is probably accounted for by the higher vitamin D requirements during pregnancy and lactation (Madrigal and Kelly, 2006).

The importance of selection in shaping human pigmentation diversity had been already pointed out in the seventies (Nei and Roychoudhury, 1972), and is indirectly confirmed by the fact that skin-colour differences between populations and continents are fourfold greater than genetic differences estimated from autosomal loci (Relethford, 2002). Analogous cases can be made for traits such as lactase persistence and thalassemia. Because of decreasing levels of the enzyme lactase in the intestine, most people of the world lose the ability to digest lactose after weaning. However, lactase persists

through lifetime at high frequency in populations from Europe and Africa, and remains at substantial frequencies in other populations sharing the habit to drink fresh milk (Campbell *et al.*, 2005). The alleles for lactose tolerance are different in Africa and Europe (Tishkoff *et al.*, 2007), and in Europe there is a clear geographic parallelism among variation in lactose tolerance, cattle milk genes, and locations of the Neolithic farming sites, implying a gene-culture coevolution between cattle and humans (Beja-Pereira *et al.*, 2003). Similarly, various pathologic variants of the haemoglobins including those responsible for thalassemias are common in areas of Africa, Asia, and Europe, where malaria is endemic. These alleles are maintained by balancing selection, because the heterozygotes are protected against the malarial parasite (Weatherall, 2001). In all these cases, common selective pressures have led to evolutionary convergence of populations that had only comparatively weak evolutionary relationships, and so should not be expected to resemble each other as much in other genome districts.

Recent analyses of genomic variation have confirmed this broad picture, at the same time adding intriguing details. Within the context of the International HapMap project, 51 autosomal regions, spanning 13 MB of the human genome, were resequenced (Gabriel *et al.*, 2002). The study of African, European, and Asian individuals shows that half of the blocks are cosmopolitan, another quarter is observed only in Africa, and the rest are essentially shared between continents, with only 4 % of them being restricted to either Europe or Asia. Similar results were obtained in another study of the HapMap genotype data, estimating allele sharing for 1536 randomly selected individual SNPs (Montpetit *et al.*, 2006). In brief, the study of a sizable fraction of the genome in geographically distant populations shows that both single SNPs and large genomic regions in linkage disequilibrium are, with few exceptions, either specifically African, or generically human. Considering the expression level of various genes as variable phenotypes, Spielman *et al.* (2007) found substantial population differences, suggesting that variation at a few regulatory loci may account for even extensive population differences in liability to diseases.

This distribution of human genetic diversity suggests that the best way to predict whether certain individuals will have certain health risks or will benefit from pharmaceutical treatment is to study their genes, rather than relying on ill-defined racial stereotypes. This seems particularly true in the study of individual response to drug treatment, i.e. pharmacogenomics. Drug response is determined by many genetic and non-genetic factors, but variation at loci coding for drug-metabolizing enzymes accounts for a large fraction of the hereditary differences among normal, fast, and slow metabolisers, i.e. respectively those who benefit from a certain drug at the standard dosage, and those who do not, or even show negative side effects. Different populations show different allele frequencies at loci of pharmacogenomic interest (Meyer, 2004). However, most populations contain the whole spectrum of genotypes, which leaves little hope to develop different drugs, or drug dosages, specific for the Asian, African, or European market. Rather, it is now technically conceivable to concentrate the efforts on the genetic dissection of individual drug metabolism pathways, which in time may lead to tailoring-specific pharmacological treatment for different classes of metabolisers (Weber, 2001), no matter what their skin colour is and where in the world they live. Similarly, a serious understanding of the causes of disease requires an analysis of both their genetic and social causes, the latter being poorly approximated by ethnic and racial labels (Collins, 2004).

In the last 10 years, human genetic variation has also been actively studied for forensic purposes. The police authorities of many countries currently identify crime suspects by means of DNA profiling, and use the word race to summarise the general appearance of people. Association between race and patterns of DNA variation has gained a broad acceptance in courts (Cho and Sankar, 2004) and claims have been made that all scientific issues associated with DNA-based personal identification are resolved (Lander and Budowle, 1994). However, comparison of the racial categories used by the American federal bureau of investigations (FBI) and by the British police (Table 31.1) shows that these catalogs differ in the number and definition of races. People from the Indian subcontinent are either classified as whites or blacks in the United States, whereas they are attributed to three different Indian races under the UK system; the Irish and the British enjoy a special status in the UK system; and the possibility of mixed origins is contemplated in the United Kingdom, but not in the US system, where, instead, we find two races, 'Hispanics' and 'American Indian or Alaska natives', matching only with the 'Any other' race in the UK system. As a matter of fact, these differences do not matter much; Alaskan natives hardly cause criminal problems in the United Kingdom. However, these examples show that such race lists include categories that are considered useful for the practical needs of the local police, and are not *all-purpose* scientific descriptions of human variation.

The differences between UK and US forensic race catalogues reflect that common concepts of race are culture specific. In particular, there is no reason to believe that people know the race they belong to, and hence that self-assessed racial classification is a reasonable basis for categorizing genotypes. Under the one-drop rule that has been used in the United States to define the border between whites and blacks (everybody with one drop of black blood is black), many are considered (and consider themselves) as African Americans, because they have even a small documented fraction of black ancestors. Similarly, the hispanics or latinos (Risch, 2006), namely the Spanish-speaking community of the United States, include individuals whose origins represent very diverse mixtures, including native American, African, and European. They are defined as a group based on two non-biological factors, language and immigration, and would not refer to themselves as Hispanics in their country of origin. It may make sense to estimate disease prevalence and perhaps even allele frequencies among Hispanics, but only in very specific and local contexts. Moreover, disease prevalence also differs according to diverse social factors such as education level, often more strongly than is reflected in 'racial' labels such as Hispanic (Collins, 2004).

Genomic data are accumulating rapidly, and some or many of the views expressed in this chapter may have to change in the future. While we wait for the future to come, however, we have to conclude that the available studies of human biological diversity have not made it possible yet (1) to agree on a race list; (2) to place races on the world's map; and (3) to associate each race with diagnostic alleles or haplotypes. Humans differ genetically, but their differences do not seem to come in well-defined and consistent racial packages. After more than a century, it is hard to disagree with Darwin (1871) who (using the words race and species as synonymous) wrote: 'But the most weighty of all the arguments against treating the races of man as distinct species, is that they graduate into each other, independently in many cases, as far as we can judge, of their having intercrossed. Man has been studied more carefully than any other animal, and yet there is the greatest possible diversity amongst capable judges whether he should be classed

as a single species or race, or as two (Virey), as three (Jacquinot), as four (Kant), five (Blumenbach), six (Buffon), seven (Hunter), eight (Agassiz), eleven (Pickering), fifteen (Bory de St-Vincent), sixteen (Desmoulins), twenty-two (Morton), sixty (Crawfurd), or as sixty-three, according to Burke.'

Acknowledgments

REFERENCES

Aiello, L.C. (1993). The fossil evidence for modern human origins in Africa: a revised view. *American Anthropologist* **95**, 73–96.

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A. and Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology* **2**, e286.

Akey, J.M., Zhang, G., Zhang, K., Jin, L. and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Research* **12**, 1805–1814.

Amos, W. and Manica, A. (2006). Global genetic positioning: evidence for early human population centers in coastal habitats. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 820–824.

Anikovich, M.V., Sinitsyn, A.A., Hoffecker, J.F., Holliday, V.T., Popov, V.V., Lisitsyn, S.N., Forman, S.L., Levkovskaya, G.M., Pospelova, G.A., Kuz'mina, I.E., Burova, N.D., Goldberg, P., Macphail, R.I., Giaccio, B. and Praslov, N.D. (2007). Early upper paleolithic in eastern Europe and implications for the dispersal of modern humans. *Science* **315**, 223–226.

Antón S.C., and Swisher C.C. III (2004). Early dispersals of Homo from Africa. *Annual Review of Anthropology* **33**, 271–296.

Aris-Brosou, S. and Excoffier, L. (1996). The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Molecular Biology and Evolution* **13**, 494–504.

Bamshad, M. (2005). Genetic influences on health: does race matter? *JAMA* **24**, 937–946.

Bamshad, M., Wooding, S., Salisbury, B.A. and Stephens, J.C. (2003). Deconstructing the relationship between genetics and race. *Nature Reviews Genetics* **5**, 598–609.

Barbujani, G. (2005). Human races: classifying people vs. Understanding diversity. *Current Genomics* **6**, 215–226.

Barbujani, G. and Belle, E.M.S. (2006). Genomic boundaries between human populations. *Human Heredity* **61**, 15–21.

Barbujani, G., Magagni, A., Minch, E. and Cavalli-Sforza, L.L. (1997). An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 4516–4519.

Barbujani, G., Oden, N.L. and Sokal, R.R. (1989). Detecting areas of abrupt change in maps of biological variables. *Systematic Zoology* **38**, 376–389.

Barbujani, G., Pilastro, A., De Domenico, S. and Renfrew, C. (1994). Genetic variation in North Africa and Eurasia: neolithic demic diffusion versus Paleolithic colonisation. *American Journal of Physical Anthropology* **95**, 137–154.

Bastos-Rodrigues, L., Pimenta, J.R. and Pena, S.D.J. (2006). The genetic structure of human populations studied through short insertion-deletion polymorphisms. *Annals of Human Genetics* **70**, 658–665.

Bateman, R., Goddard, I., O'Grady, R., Funk, V.A., Mooi, R., Kress, W.J. and Cannell, P. (1990). Speaking of forked tongues. *Current Anthropology* **31**, 1–24.

Beaumont, M.A. (1999). Detecting population expansion and decline using microsatellites. *Genetics* **153**, 2013–2029.

Beaumont, M.A. (2004). Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity* **92**, 365–379.

Beaumont, M.A. (2005). Adaptation and speciation: what can Fst tell us? *Trends in Ecology and Evolution* **20**, 435–440.

Beaumont, M.A. and Balding, D.J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969–980.

Beaumont, M.A. and Nichols, R. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London* **B 263**, 1619–1626.

Beaumont, M.A. and Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics* **5**, 251–261.

Beaumont, M.A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

Beerli, P. and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 4563–4568.

Beja-Pereira, A., Luikart, G., England, P.R., Bradley, D.G., Jann, O.C., Bertorelle, G., Chamberlain, A.T., Nunes, T.P., Metodiev, S., Ferrand, N. and Erhardt, G. (2003). Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nature Genetics* **35**, 311–313.

Bernasconi, R. (2001). *Concepts of Race in the Eighteenth Century*. Thoemmes Press, Bristol.

Binladen, J., Wiuf, C., Gilbert, M.T., Bunce, M., Barnett, R., Larson, G., Greenwood, A.D., Haile, J., Ho, S.Y., Hansen, A.J. and Willerslev, E. (2006). Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. *Genetics* **172**, 733–741.

Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R. and Cavalli-Sforza, L.L. (1994). High resolution of human evolutionary history trees with polymorphic microsatellites. *Nature* **368**, 455–457.

Bräuer, G. (1984). A craniological approach to the origin of anatomically modern human *Homo sapiens* and its implications for the appearance of modern Europeans. In *The Origins of Modern Humans*, F.H. Smith and F. Spencer, eds. Wenner-Gren Foundation for Anthropological Research New York, pp. 327–410.

Brown, R.A. and Armelagos, G.J. (2001). Apportionment of racial diversity: a review. *Evolutionary Anthropology* **10**, 24–20.

Burchard, E.G., Ziv, E., Coyle, N., Gomez, S.L., Tang, H., Karter, A.J., Mountain, J.L., Perez-Stable, E.J., Sheppard, D. and Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *The New England Journal of Medicine* **348**, 1170–1175.

Campbell, A.K., Waud, J.P. and Matthews, S.B. (2005). The molecular basis of lactose intolerance. *Science in Progress* **88**, 157–202.

Cann, R., Stoneking, M. and Wilson, A.J. (1987). Mitochondrial DNA and human evolution. *Nature* **325**, 31–36.

Caramelli, D., Lalueza-Fox, C., Condemi, S., Longo, L., Milani, L., Manfredini, A., de Saint Pierre, M., Adoni, F., Lari, M., Giunti, P., Ricci, S., Casoli, A., Calafell, F., Mallegni, F., Bertranpetit, J., Stanyon, R., Bertorelle, G. and Barbujani, G. (2006). A highly divergent mtDNA sequence in a neandertal individual from italy. *Current Biology* **16**, R630–R632.

Caramelli, D., Lalueza-Fox, C., Vernesi, C., Lari, M., Casoli, A., Mallegni, F., Chiarelli, B., Dupanloup, I., Bertranpetit, J., Barbujani, G. and Bertorelle, G. (2003). Evidence for a genetic

discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 6593–6597.

Cavalli-Sforza, L.L. (1966). Population structure and human evolution. *Proceedings of the Royal Society Section B* **164**, 362–379.

Cavalli-Sforza, L.L., Menozzi, P. and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.

Cavalli-Sforza, L.L., Menozzi, P., Piazza, A. and Mountain, J. (1988). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 6002–6006.

Chikhi, L. and Beaumont, M.A. (2005). Modelling human genetic history. *Encyclopaedia of Genetics, Genomics, Proteomics and Bioinformatics*. John Wiley & Sons, Chichester.

Chikhi, L. and Bruford, M.B. (2005). Mammalian population genetics and genomics. In *Mammalian Genomics*, A. Ruvinsky and J. Marshall Graves, eds. Chapter 21. CAB International Publishing, pp. 539–584.

Chikhi, L., Bruford, M.W. and Beaumont, M.A. (2001). Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**, 1347–1362.

Chikhi, L., Destro-Bisol, G., Bertorelle, G., Pascali, V. and Barbujani, G. (1998). Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 9053–9058.

Chikhi, L., Nichols, R.A., Barbujani, G. and Beaumont, M.A. (2002). Y genetic data support the Neolithic demic diffusion model. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 10008–10013.

Cho, M.K. and Sankar, P. (2004). Forensic genetics and legal, ethic and social implications beyond the clinic. *Nature Genetics Supplement* **36**, S8–S12.

Chung, C.S., Mi, M.P. and Morton, N.E. (1967). *Genetics of Interracial Crosses in Hawaii*. Karger, Basel.

Cohen, C. (1991). *Aux Origines d'Homo Sapiens*, J.-J. Hublin and A.M. Tillier, eds. Presses Universitaries de France, Paris, pp. 9–47.

Cole, S. (1965). *Races of Man*. British Museum (Natural History), London.

Collard, M. and Shennan, S. (2000). Processes of culture change in prehistory: a case study from the European Neolithic. In *Archaeogenetics: DNA and the Population Prehistory of Europe*, C. Renfrew and K. Boyle, eds. MacDonald Institute for Archaeological Research, Cambridge, pp. 89–97.

Collins, F. (2004). What we do and don't know about 'race', 'ethnicity', genetics and health at the dawn of the genome era. *Nature Genetics Supplement* **36**, S13–S15.

Conrad, D.F., Jakobsson, M., Coop, G., Wen, X., Eall, J.D., Rosenberg, N.A. and Pritchard, J.K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics* **38**, 1251–1260.

Coon, C.S. (1963). *The Origin of Races*. Alfred A. Knopf, New York.

Coon, C.S., Garn, S.M. and Birdsell, J.B. (1950). *A Study of the Problem of Race Formation in Man*. Charles Thomas, Springfield, Ill.

Cooper, R.S. (2005). Race and IQ: molecular genetics as deus ex machina. *The American Psychologist* **60**, 71–76.

Cooper, R.S., Kaufman, J.S. and Ward, R.W. (2003). Race and genomics. *The New England Journal of Medicine* **348**, 1166–1170.

Corander, J. and Marttinen, P. (2006). Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology* **15**, 2833–2843.

Corander, J., Marttinen, P. and Mäntyniemi, S. (2007). Bayesian identification of stock mixtures from molecular marker data. *Fishery Bulletin* (in press).

Corander, J., Waldmann, P., Marttinen, P. and Sillanpaa, M.J. (2004). BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**, 2363–2369.

Corander, J., Waldmann, P. and Sillanpaa, M.J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374.

Cordaux, R., Aunger, R., Bentley, G., Nasidze, I., Sirajuddin, S.M. and Stoneking, M. (2004). Independent origins of Indian caste and tribal paternal lineages. *Current Biology* **14**, 231–235.

Cornuet, J.-M. and Luikart, G. (1996). Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**, 2001–2014.

Crawford, M. (1998). *The Origin of Native Americans: Evidence from Archeological Genetics*. Cambridge University Press, Cambridge.

Currat, M. and Excoffier, L. (2004). Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biology* **2**, e421.

Currat, M., Excoffier, L., Maddison, W., Otto, S.P., Ray, N., Whitlock, M.C. and Yeaman, S. (2006). Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". *Science* **313**, 172.

Darwin, C.R. (1871). *The Descent of Man, and Selection in Relation to Sex*, Chapter VII. John Murray, London.

Dawson, K.J. and Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research* **78**, 59–77.

Dobzhansky, T. (1967). *Genetic Diversity and Human Behavior*, J.N. Spuhler, ed. Wenner-Gren Foundation for Anthropological Research, New York, pp. 1–19.

Dupanloup, I., Pereira, L., Bertorelle, G., Calafell, F., Prata, M.J., Amorim, A. and Barbujani, G. (2003). A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. *Journal of Molecular Evolution* **57**, 85–97.

Edmonds, C.A., Lillie, A.S. and Cavalli-Sforza, L.L. (2004). Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 975–979.

Edwards, A.W.F. (2003). Human genetic diversity: Lewontin's fallacy. *BioEssays* **25**, 798–801.

Edwards, A.W.F. and Cavalli-Sforza, L.L. (1965). A method for cluster analysis. *Biometrics* **21**, 362–375.

Eller, E. (2002). Population extinction and recolonisation in human demographic history. *Mathematical Biosciences* **177–178**, 1–10.

Eswaran, V. (2002). A diffusion wave out of Africa: the mechanism of the modern human revolution? *Current Anthropology* **43**, 749–774.

Eswaran, V., Harpending, H. and Rogers, A.R. (2005). Genomics refutes an exclusively African origin of humans. *Journal of Human Evolution* **49**, 1–18.

Evanno, G., Regnaut, S. and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611–2620.

Evans, P.D., Gilbert, S.L., Mekel-Bobrov, N., Vallender, E.J., Anderson, J.R., Vaez-Azizi, L.M., Tishkoff, S.A., Hudson, R.R. and Lahn, B.T. (2005). Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. *Science* **309**, 1717–1720.

Evans, P.D., Mekel-Bobrov, N., Vallender, E.J., Hudson, R.R. and Lahn, B.T. (2006). Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an archaic Homo lineage. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 18178–18183.

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Excoffier, L. (2002). Human demographic history: refining the recent African origin model. *Current Opinion in Genetics and Development* **12**, 675–682.

Excoffier, L., Estoup, A. and Cornuet, J.-M. (2005). Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**, 1727–1738.

Excoffier, L. and Hamilton, G. (2003). Comment on "Genetic structure of human populations". *Science* **300**, 1877.

Excoffier, L. and Schneider, S. (1999). Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 10597–10602.

Excoffier, L., Smouse, P.E. and Quattro, J.M. (1992). Analysis of molecular variance inferred from emtric distances among DNA haplotypes: application ti human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.

Falush, D., Stephens, M. and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.

Fay, J.C. and Wu, C.I. (1999). A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Molecular Biology and Evolution* **16**, 1003–1005.

Felsenstein, J. (1992). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research* **59**, 139–147.

Fischer, A., Pollack, J., Thalmann, O., Nickel, B. and Pääbo, S. (2006). Demographic history and genetic differentiation in apes. *Current Biology* **16**, 1133–1138.

Flint, J., Bond, J., Rees, D.C., Boyce, A.J., Roberts-Thompson, J.M., Excoffier, L., Clegg, J.B., Beaumont, M.A., Nichols, R.A. and Harding, R.M. (1999). Minisatellite mutational processes reduce F_{ST} estimates. *Human Genetics* **105**, 567–576.

Fu, Y.X. and Li, W.H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.

Gagneux, P., Wills, C., Gerloff, U., Tautz, D., Morin, P.A., Boesch, C., Fruth, B., Hohmann, G., Ryder, O.A. and Woodruff, D.S. (1999). Mitochondrial sequences show diverse evolutionary histories of African hominoids. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 5077–5082.

Garn, S.M. (1965). *Human Races*. Thomas, Springfield, Ill.

Garrigan, D. and Hammer, M.F. (2006). Reconstructing human origins in the genomic era. *Nature Reviews Genetics* **7**, 669–680.

Garrigan, D., Mobasher, Z., Kingan, S.B., Wilder, J.A. and Hammer, M.F. (2005). Deep haplotype divergence and long-range linkage disequilibrium at Xp21.1 provide evidence that humans descend from a structured ancestral population. *Genetics* **170**, 1849–1856.

Garza, J.C. and Williamson, E. (2001). Detection of reduction in population size using data from microsatellite DNA. *Molecular Ecology* **10**, 305–318.

Goldstein, D.B. and Chikhi, L. (2002). Human migrations and population structure: what we know and why it matters. *Annual Review of Genomics and Human Genetics* **3**, 129–152.

Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L. and Feldman, M.W. (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 6723–6727.

Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M. and Pääbo, S. (2006). Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330–336.

Griffiths, R.C. and Tavaré, S. (1994). Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131–159.

Grine, F.E., Bailey, R.M., Harvati, K., Nathan, R.P., Morris, A.G., Henderson, G.M., Ribot, I. and Pike, A.W. (2007). Late Pleistocene human skull from Hofmeyr, South Africa, and modern human origins. *Science* **315**, 226–229.

Hammer, M.F., Garrigan, D., Wood, E., Wilder, J.A., Mobasher, Z., Bigham, A., Krenz, J.G. and Nachman, M.W. (2004). Heterogeneous patterns of variation among multiple human x-linked Loci: the possible role of diversity-reducing selection in non-Africans. *Genetics* **167**, 1841–1853.

Harding, R.M., Fullerton, S.M., Griffiths, R.C., Bond, J., Cox, M.J., Schneider, J.A., Moulin, D.S. and Clegg, J.B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *American Journal of Human Genetics* **60**, 772–789.

Harding, R.M., Healy, E., Ray, A.J., Ellis, N.S., Flanagan, N., Todd, C., Dixon, C., Sajantila, A., Jackson, I.J., Birch-Machin, M.A. and Rees, J.L. (2000). Evidence for variable selective pressures at MC1R. *American Journal of Human Genetics* **66**, 1351–1361.

Harding, R.M. and McVean, G.A. (2004). A structured ancestral population for the evolution of modern humans. *Current Opinion in Genetics and Development* **14**, 667–674.

Harris, E.E. and Hey, J. (1999). Human demography in the Pleistocene: do mitochondrial and nuclear genes tell the same story? *Evolutionary Anthropology* **8**, 81–86.

Hey, J. (1997). Mitochondrial and nuclear genes present conflicting portraits of human origins. *Molecular Biology and Evolution* **14**, 166–172.

Hey, J. (2005). On the number of new world founders: a population genetic portrait of the peopling of the Americas. *PLoS Biology* **3**, 965–975.

Hey, J. and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**, 747–760.

Howells, W.W. (1976). Explaining modern man: evolutionists versus migrationists. *Journal of Human Evolution* **5**, 477–496.

Huxley, J. (1938). Clines: an auxiliary taxonomic principle. *Nature* **142**, 219–220.

Ingman, M., Kaessmann, H., Pääbo, S. and Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713.

Jablonski, N. (2006). *Skin. A Natural History*. University of California Press, Berkeley, CA.

Jorde, L.B., Watkins, W.S., Bamshad, M.J., Dixon, M.E., Ricker, C.E., Seielstad, M.T. and Batzer, M.A. (2000). The distribution of human genetic diversity: a comparison of mitochondrial, autosomal and Y-chromosome data. *American Journal of Human Genetics* **66**, 979–988.

Jorde, L.B. and Wooding, S.P. (2004). Genetic variation, classification, and 'race'. *Nature Genetics* **36**, S28–S33.

Kaessmann, H., Heissig, F., von Haeseler, A. and Pääbo, S. (1999). DNA sequence variation in a non-coding region of low recombination on the human X-chromosome. *Nature Genetics* **22**, 78–81.

Kayser, M., Brauer, S. and Stoneking, M. (2003). A genome scan to detect candidate regions influenced by local natural selection in human populations. *Molecular Biology and Evolution* **20**, 893–900.

Ke, Y., Su, B., Song, X., Lu, D., Chen, L., Li, H., Qi, C., Marzuki, S., Deka, R., Underhill, P., Xiao, C., Shriver, M., Lell, J., Wallace, D., Wells, R.S., Seielstad, M., Oefner, P., Zhu, D., Jin, J., Huang, W., Chakraborty, R., Chen, Z. and Jin, L. (2001). African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. *Science* **292**, 1151–1153.

Kimura, M. and Crow, J. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

Kitano, T., Schwarz, C., Nickel, B. and Pääbo, S. (2003). Gene diversity patterns at 10 X-chromosomal loci in humans and chimpanzees. *Molecular Biology and Evolution* **20**, 1281–1289.

King, M.C. and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116.

Klein, R.G. (1999). *The Human Career: Human Biological and Cultural Origins*. University of Chicago Press, Chicago, Ill.

Klein, R.G. (2005). Hominin dispersals in the old world. In *The Human Past*, C. Scarre, ed. Thames and Hudson Ltd, London, pp. 84–123.

Klopfstein, S., Currat, M. and Excoffier, L. (2006). The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution* **23**, 482–490.

Krings, M., Capelli, C., Tschentscher, F., Geisert, H., Meyer, S., von Haeseler, A., Grossschmidt, K., Possnert, G., Paunovic, M. and Paabo, S. (2000). A view of Neandertal genetic diversity. *Nature Genetics* **26**, 144–146.

Kuhner, M., Yamoto, J. and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**, 1421–1430.

Lahr, M.M. (1994). The multiregional model of modern human origins: a reassessment of its morphological basis. *Journal of Human Evolution* **26**, 23–56.

Lahr, M.M. and Foley, R.A. (1998). Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. *Yearbook of Physical Anthropology* **41**, 137–176.

Lander, E. and Budowle, B. (1994). DNA fingerprinting dispute laid to rest. *Nature* **371**, 735–737.

Latter, B.D.H. (1980). Genetic differences within and between populations of the major human subgroups. *American Naturalist* **116**, 220–237.

Lewin, R. and Foley, R.A. (2004). *Principles of Human Evolution*. Blackwell Science, Oxford.

Lewontin, R.C. (1972). The apportionment of human diversity. *Evolutionary Biology* **6**, 381–398.

Li, N. and Stephens, M. (2003). Modelling linkage disequilibrium and identifying recombination hotspots using single nucleotide polymorphism data. *Genetics* **165**, 2213–2233.

Liu, H., Prugnolle, F., Manica, A. and Balloux, F. (2006). A geographically explicit genetic model of worldwide human-settlement history. *American Journal of Human Genetics* **79**, 230–137.

Livingstone, F.B. (1962). On the nonexistence of human races. *Current Anthropology* **3**, 279–281.

Luis, J.R., Rowold, D.J., Regueiro, M., Caeiro, B., Cinnioglu, C., Roseman, C., Underhill, P.A., Cavalli-Sforza, L.L. and Herrera, R.J. (2004). The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. *American Journal of Human Genetics* **74**, 532–544.

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R., Cruciani, F., Taha, A., Shaari, N.K., Raja, J.M., Ismail, P., Zainuddin, Z., Goodwin, W., Bulbeck, D., Bandelt, H.J., Oppenheimer, S., Torroni, A. and Richards, M. (2005). Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* **308**, 1034–1036.

Madrigal, L. and Kelly, W. (2006). Human skin-color sexual dimorphism: a test of the sexual selection hypothesis. *American Journal of Physical Anthropology* **132**, 470–482.

Makova, K. and Norton, H. (2005). Worldwide polymorphism at the MC1R locus and normal pigmentation variation in humans. *Peptides* **26**, 1901–1908.

Manni, F., Guerard, E. and Heyer, E. (2004). Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. *Human Biology* **76**, 173–190.

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15324–15328.

Marth, G.T., Czabarka, E., Murvai, J. and Sherry, S.T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372.

Marth, G.T., Schuler, G., Yeh, R., Davenport, R., Agarwala, R., Church, D., Wheelan, S., Baker, J., Ward, M., Kholodov, M., Phan, L., Czabarka, E., Murvai, J., Cutler, C., Wooding, S., Rogers, A.R., Chakravarti, A., Harpending, H.C., Kwok, P.-Y. and Sherry, S.T. (2003). Sequence variations in the public human genome data reflect a bottlenecked population history. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 376–381.

Mayr, E. (1947). *Systematics and the Origin of Species*, 3rd edition. Columbia University Press, New York.

McEvoy, B., Beleza, S. and Shriver, M.D. (2006). The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. *Human Molecular Genetics* **15**, R176–R181.

McVean, G.A.T., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R. and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584.

Mekel-Bobrov, N., Gilbert, S.L., Evans, P.D., Vallender, E.J., Anderson, J.R., Hudson, R.R., Tishkoff, S.A. and Lahn, B.T. (2005). Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. *Science* **309**, 1720–1722.

Mekel-Bobrov, N., Posthuma, D., Gilbert, S.L., Lind, P., Gosso, M.F., Luciano, M., Harris, S.E., Bates, T.C., Polderman, T.J., Whalley, L.J., Fox, H., Starr, J.M., Evans, P.D., Montgomery, G.W., Fernandes, C., Heutink, P., Martin, N.G., Boomsma, D.I., Deary, I.J., Wright, M.J., de Geus, E.J. and Lahn, B.T. (2007). The ongoing adaptive evolution of ASPM and Microcephalin is not explained by increased intelligence. *Human Molecular Genetics* **16**, 600–608.

Mellars, P. (2006). Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* **313**, 796–800.

Menozzi, P., Piazza, A. and Cavalli-Sforza, L.L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786–792.

Meyer, U.A. (2004). Pharmacogenetics – five decades of therapeutic lessons from genetic diversity. *Nature Reviews Genetics* **5**, 669–676.

Molnar, S. (1975). *Human Variation. Races, Types and Ethnic Groups*. Prentice Hall, Upper Saddle River.

Montagu, M.F.A. (1941). The concept of race in the human species in the light of genetics. *The Journal of Heredity* **32**, 243–247.

Montpetit, A., Nelis, M., Laflamme, P., Magi, R., Ke, X., Rmm, M., Cardon, L., Hudson, T.J. and Metspalu, A. (2006). An evaluation of the performance of TAG SNPs derived from HapMap in a Caucasian population. *PLoS Genetics* **2**, e27.

Mountain, J.L. and Risch, N. (2004). Assessing genetic contributions to phenotypic differences among 'racial' and 'ethnic' groups. *Nature Genetics Supplement* **36**, S48–S53.

Mulligan, C.J., Hunley, K., Cole, S. and Long, J.C. (2004). Population genetics, history, and health patterns in native americans. *Annual Review of Genomics and Human Genetics* **5**, 295–315.

Neel, J.V. (1978). The population structure of an Amerindian tribe, the Yanomama. *Annual Review of Genetics* **12**, 365–418.

Nei, M., Chakravarti, A. and Tateno, Y. (1977). Mean and variance of F_{st} in a finite number of incompletely isolated populations. *Theoretical Population Biology* **11**, 291–306.

Nei, M., Maruyama, T. and Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution* **29**, 1–10.

Nei, M. and Roychoudhury, A.K. (1972). Gene differences between Caucasian, Negro and Japanese populations. *Science* **177**, 434–436.

Nielsen, R. and Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–896.

Nordborg, M. (1998). On the probability of Neanderthal ancestry. *American Journal of Human Genetics* **63**, 1237–1240.

Ordũnez, P., Bernal Munoz, J.L., Espinosa-Brito, A., Silva, L.C. and Cooper, R.S. (2005). Ethnicity, education, and blood pressure in Cuba. *American Journal of Epidemiology* **162**, 49–56.

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L. and Hofreiter, M. (2004). Genetic analyses from ancient DNA. *Annual Review of Genetics* **38**, 645–679.

Parra, F.C., Amado, R.C., Lambertucci, J.R., Rocha, J., Antunes, C.M. and Pena, S.D. (2003). Color and genomic ancestry in Brazilians. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 177–182.

Parra, E.J., Marcini, A., Akey, J., Martinson, J., Batzer, M.A., Cooper, R., Forrester, T., Allison, D.B., Deka, R., Ferrell, R.E. and Shriver, M.D. (1998). Estimating African American admixture proportions by use of population-specific alleles. *American Journal of Human Genetics* **63**, 1839–1851.

Plagnol, V. and Wall, J.D. (2006). Possible ancestral structure in human populations. *PLoS Biology* **2**, 972–979.

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A. and Feldman, M.W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**, 1791–1798.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Provine, W.B. (1986). Geneticists and race. *American Zoologist* **26**, 857–888.

Przeworski, M., Hudson, R.R. and Di Rienzo, A. (2000). Adjusting the focus on human variation. *Trends in Genetics* **16**, 296–302.

Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W. and Cavalli-Sforza, L.L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15942–15947.

Ramachandran, S., Rosenberg, N.A., Zhivotovsky, L.A. and Feldman, M.W. (2004). Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. *Human Genomics* **1**, 87–97.

Rannala, B. and Mountain, J.L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 9197–9201.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656.

Ray, N., Currat, M., Berthier, P. and Excoffier, L. (2005). Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Research* **15**, 1161–1167.

Ray, N., Currat, M. and Excoffier, L. (2004). Intra-deme molecular diversity in spatially expanding populations. *Molecular Biology and Evolution* **20**, 76–86.

Redd, A.J., Chamberlain, V.F., Kearney, V.F., Stover, D., Karafet, T., Calderon, K., Walsh, B. and Hammer, M.F. (2006). Genetic structure among 38 populations from the United States based on 11 U.S. core Y chromosome STRs. *Journal of Forensic Sciences* **51**, 580–585.

Reed, F.A. and Tishkoff, S.A. (2006). African human diversity, origins and migrations. *Current Opinion in Genetics and Development* **16**, 597–605.

Rees, J.L. (2003). Genetics of hair and skin color. *Annual Review of Genetics* **37**, 67–90.

Reich, D.E. and Goldstein, D.B. (1998). Genetic evidence for a Paleolithic human population expansion in Africa. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 8119–8123.

Relethford, J.H. (1999). Models, predictions, and the fossil record of modern human origins. *Evolutionary Anthropology* **8**, 7–10.

Relethford, J.H. (2001). Absence of regional affinities of Neandertal DNA with living humans does not reject multiregional evolution. *American Journal of Physical Anthropology* **115**, 95–98.

Relethford, J.H. (2002). Apportionment of global human genetic diversity based on craniometrics and skin color. *American Journal of Physical Anthropology* **118**, 393–398.

Relethford, J.H. (2003). Genetic history of the human species. In *Handbook of Statistical Genetics*, 2nd edition, D. Balding, M. Bishop and C. Cannings, eds. John Wiley & Sons, Chichester, pp. 793–829.

Relethford, J.H. (2004). Global patterns of isolation by distance based on genetic and morphological data. *Human Biology* **76**, 499–513.

Relethford, J.H. and Harding, R.M. (2001). Population genetics of modern human evolution. *Encyclopedia of Life Sciences*. John Wiley & Sons, Chichester. `http://www.els.net/`.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Risch, N. (2006). Dissecting racial and ethnic differences. *The New England Journal of Medicine* **354**, 408–411.

Risch, N., Burchard, E., Ziv, E. and Tang, H. (2002). Categorization of humans in biomedical research: genes, race and disease. *Genome Biology* **3**, 2007.1–2007.12.

Rogers, A.R. and Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* **9**, 552–569.

Romualdi, C., Balding, D., Nasidze, I.S., Risch, G., Robichaux, M., Sherry, S.T., Stoneking, M., Batzer, M.A. and Barbujani, G. (2002). Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Research* **12**, 602–612.

Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K. and Feldman, M.W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics* **1**, e70.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W. (2002). Genetic structure of human populations. *Science* **298**, 2381–2385.

Ryman, N., Chakraborty, R. and Nei, M. (1983). Differences in the relative distribution of human gene diversity between electrophoretic and red and white cell antigen loci. *Human Heredity* **33**, 93–102.

Sarich, V. (2000). The final taboo. Race differences in ability. *Skeptic* **8**, 38–43.

Sawyer, S.L., Mukherjee, N., Pakstis, A.J., Feuk, L., Kidd, J.R., Brookes, A.J. and Kidd, K.K. (2005). Linkage disequilibrium patterns vary substantially among populations. *European Journal of Human Genetics* **13**, 677–686.

Scarre, C. (2005). *The Human Past*. Thames and Hudson Ltd, London.

Seielstad, M.T., Minch, E. and Cavalli-Sforza, L.L. (1998). Genetic evidence for a higher female migration rate in humans. *Nature Genetics* **20**, 278–280.

Serre, D., Langaney, A., Chech, M., Teschler-Nicola, M., Paunovic, M., Mennecier, P., Hofreiter, M., Possnert, G. and Pääbo, S. (2004). No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biology* **2**, 313–317.

Serre, D. and Pääbo, S. (2004). Evidence for gradients of human genetic diversity within and among continents. *Genome Research* **14**, 1679–1685.

Sherry, S.T., Batzer, M.A. and Harpending, H. (1998). Modeling the genetic architecture of modern populations. *Annual Review of Anthropology* **27**, 153–169.

Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M. and Jones, K.W. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human Genomics* **1**, 274–286.

Shriver, M.D., Mei, R., Parra, E.J., Sonpar, V., Halder, I., Tishkoff, S.A., Schurr, T.G., Zhadanov, S.I., Osipova, L.P., Brutsaert, T.D., Friedlaender, J., Jorde, L.B., Watkins, W.S., Bamshad, M.J., Gutierrez, G., Loi, H., Matsuzaki, H., Kittles, R.A., Argyropoulos, G., Fernandez, J.R., Akey, J.M. and Jones, K.W. (2005). Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. *Human Genomics* **2**, 81–89.

Shriver, M.D., Parra, E.J., Dios, S., Bonilla, C., Norton, H., Jovel, C., Pfaff, C., Jones, C., Massac, A., Cameron, N., Baron, A., Jackson, T., Argyropoulos, G., Jin, L., Hoggart, C.J., McKeigue, P.M. and Kittles, R.A. (2003). Skin pigmentation, biogeographical ancestry and admixture mapping. *Human Genetics* **112**, 387–399.

Shriver, M.D., Smith, M.W., Jin, L., Marcini, A., Akey, J.M., Deka, R. and Ferrell, R.E. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics* **60**, 957–964.

Slatkin, M. and Hudson, R.R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–562.

Smith, F.H. (1985). Continuity and change in the origin of modern *Homo sapiens*. *Zeitschrift fur Morphologie und Anthropologie* **75**, 197–222.

Sokal, R.R., Jacquez, G.M. and Wooten, M.C. (1989). Spatial autocorrelation analysis of migration and selection. *Genetics* **121**, 845–855.

Spielman, R.S., Bastone, L.A., Burdick, J.T., Morley, M., Ewens, W.J. and Cheung, V.G. (2007). Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genetics* **39**, 226–231.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society* **B 62**, 605–635.

Storz, J.F. and Beaumont, M.A. (2002). Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* **56**, 154–166.

Storz, J.F., Payseur, B.A. and Nachman, M.W. (2004). Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Molecular Biology and Evolution* **21**, 1800–1811.

Stringer, C.B. (1978). Some problems in Middle Pleistocene hominid relationships. In *Recent Advances in Primatology*, D. Chivers and K. Joysey, eds. Academic Press, London, pp. 395–418.

Stringer, C.B. (1989). The origin of modern humans: a comparison of European and non-European evidence. In *The Human Revolution: Behavioural and Biological Perspectives on the Origins of Modern Humans*, P. Mellars and C. Stringer, eds. Edinburgh University Press, Edinburgh, pp. 123–154.

Stringer, C.B. and Andrews, P. (1988). Genetic and fossil evidence for the origin of modern humans. *Science* **239**, 1263–1268.

Tajima, F. (1989a). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Tajima, F. (1989b). The effect of change in population size on DNA polymorphism. *Genetics* **123**, 597–601.

Takahata, N., Lee, S.H. and Satta, Y. (2001). Testing multiregionality of human origins. *Molecular Biology and Evolution* **18**, 172–183.

Tang, H., Quertermous, T., Rodriguez, B., Kardia, S.L., Zhu, X., Brown, A., Pankow, J.S., Province, M.A., Hunt, S.C., Boerwinkle, E., Schork, N.J. and Risch, N.J. (2005). Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *American Journal of Human Genetics* **76**, 268–275.

Tattersall, I. (1995). *The Fossil Trail: How We Know What We Think We Know About Human Evolution*. Oxford University Press, Oxford.

Tattersall, I. (1997). Out of Africa again ... and again? *Scientific American* **276**, 60–67.

Tavaré, S., Balding, D.J., Griffiths, R.C. and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518.

Templeton, A.R. (1999). Human races: a genetic and evolutionary perspective. *American Anthropologist* **100**, 632–650.

Templeton, A.R. (2002). Out of Africa again and again. *Nature* **416**, 45–51.

Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A.J., Kidd, J.R., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A.S., Moral, P. and Krings, M. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387.

Tishkoff, S.A., Goldman, A., Calafell, F., Speed, W.C., Deinard, A.S., Bonne-Tamir, B., Kidd, J.R., Pakstis, A.J., Jenkins, T. and Kidd, K.K. (1998). A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *American Journal of Human Genetics* **62**, 1389–1402.

Tishkoff, S.A., Pakstis, A.J., Stoneking, M., Kidd, J.R., Destro-Bisol, G., Sanjantila, A., Lu, R.B., Deinard, A.S., Sirugo, G., Jenkins, T., Kidd, K.K. and Clark, A.G. (2000). Short tandem-repeat polymorphism/alu haplotype variation at the PLAT locus: implications for modern human origins. *American Journal of Human Genetics* **67**, 901–925.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., Ibrahim, M., Omar, S.A., Lema, G., Nyambo, T.B., Ghori, J., Bumpstead, S., Pritchard, J.K., Wray, G.A. and Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* **39**, 31–40.

Tishkoff, S.A. and Verrelli, B.C. (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annual Review of Genomics and Human Genetics* **4**, 293–340.

Toth, N. and Schick, K. (2005). African origins. In *The Human Past*, C. Scarre, ed. Thames and Hudson Ltd, London, pp. 46–83.

Trinkaus, E. (2005). Early modern humans. *Annual Review of Anthropology* **34**, 207–230.

Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., Yang, W.H., Kauffman, E., Bonne-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J.R., Mehdi, S.Q., Seielstad, M.T., Wells, R.S., Piazza, A., Davis, R.W., Feldman, M.W., Cavalli-Sforza, L.L. and Oefner, P.J. (2000). Y chromosome sequence variation and the history of human populations. *Nature Genetics* **26**, 358–361.

Vekua, A., Lordkipanidze, D., Rightmire, G.P., Agusti, J., Ferring, R., Maisuradze, G., Mouskhe-lishvili, A., Nioradze, M., Ponce de Leòn, M.S., Tappen, M., Tvalchrelidze, M. and Zollikofer, C.P.E. (2002). A new skull of early *Homo* from Dmanisi, Georgia. *Science* **297**, 85–89.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. and Wilson, A.C. (1991). African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507.

Voight, B.F., Adams, A.M., Frisse, L.A., Qian, Y., Hudson, R.R. and Di Rienzo, A. (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 18508–18513.

Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. *PLoS Biology* **4**, e72.

Waddle, D.M. (1994). Matrix correlation tests support a single origin for modern humans. *Nature* **368**, 452–454.

Wall, J.D. and Przeworski, M. (2000). When did the human population size start increasing? *Genetics* **155**, 1865–1874.

Wang, J.L. (2003). Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**, 747–765.

Watkins, W.S., Prasad, B.V., Naidu, J.M., Rao, B.B., Bhanu, B.A., Ramachandran, B., Das, P.K., Gai, P.B., Reddy, P.C., Reddy, P.G., Sethuraman, M., Bamshad, M.J. and Jorde, L.B. (2005). Diversity and divergence among the tribal populations of India. *Annals of Human Genetics* **69**, 680–692.

Watkins, W.S., Ricker, C.E., Bamshad, M.J., Carroll, M.L., Nguyen, S.V., Batzer, M.A., Harpending, H.C., Rogers, A.R. and Jorde, L.B. (2001). Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. *American Journal of Human Genetics* **68**, 738–752.

Watkins, W.S., Rogers, A.R., Ostler, C.T., Wooding, S., Bamshad, M.J., Brassington, A.M., Carroll, M.L., Nguyen, S.V., Walker, J.A., Prasad, B.V., Reddy, P.G., Das, P.K., Batzer, M.A. and Jorde, L.B. (2003). Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Research* **13**, 1607–1618.

Watterson, G.A. (1978). The homozygosity test of neutrality. *Genetics* **88**, 405–417.

Weatherall, D.J. (2001). Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. *Nature Reviews Genetics* **2**, 245–255.

Weber, W.W. (2001). Effect of pharmacogenetics on medicine. *Environmental and Molecular Mutagenesis* **37**, 179–184.

Weidenreich, F. (1943). The skull of *Sinanthropus pekinensis*: A comparative study of a primitive hominid skull. *Palaeontologica Sinica* **D10**, 1–485.

Weidenreich, F. (1947). Are human races in the taxonomic sense races or species? *American Journal of Physical Anthropology* **5**, 369–371.

Weiss, G. and von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics* **149**, 1539–1546.

Wilder, J.A., Kingan, S.B., Mobasher, Z., Metni Pilkington, M. and Hammer, M.F. (2004). Global patterns of human mtDNA and Y chromosome structure are not Influenced by higher rates of female migration. *Nature Genetics* **36**, 1122–1125.

Wilson, I.J. and Balding, D.J. (1998). Genealogical inference from microsatellite data. *Genetics* **150**, 499–510.

Wilson, I.J., Weale, M.E. and Balding, D.J. (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society A* **166**, 155–188.

Wilson, J.E., Weale, M.E., Smith, A.C., Gratrix, F., Fletcher, B., Thomas, M.G., Bradman, N. and Goldstein, D.B. (2001). Population genetic structure of variable drug response. *Nature Genetics* **29**, 265–269.

Wolpoff, M.H. (1999). The systematics of *Homo*. *Science* **284**, 1774–1775.

Wolpoff, M.H., Spuhler, J.N., Smith, F.H., Radovcic, J., Pope, G., Frayer, D.W., Eckhardt, R. and Clark, G. (1988). Modern human origins. *Science* **241**, 772–774.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1943). Isolation by distance. *Genetics* **28**, 139–156.

Yu, N., Chen, F.C., Ota, S., Jorde, L.B., Pamilo, P., Patthy, L., Ramsay, M., Jenkins, T., Shyue, S.K. and Li, W.H. (2003). Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* **161**, 269–274.

Zilhão, J. (2006). Neandertals and moderns mixed, and it matters. *Evolutionary Anthropology* **15**, 183–195.

Part 6

Genetic Epidemiology

Epidemiology and Genetic Epidemiology

P.R. Burton, J.M. Bowden and M.D. Tobin

Departments of Health Sciences and Genetics, University of Leicester, Leicester, UK

This chapter focuses on descriptive analysis in genetic epidemiology and some of its links with mainstream epidemiology. It starts with incidence and prevalence, briefly reviewing some of the relevant statistical methods, and then moves on to consider the analysis of correlated responses in mainstream epidemiology. The core of the chapter describes methods used in descriptive analysis in genetic epidemiology. These methods are primarily used to describe and draw inferences about the structure of phenotypic dependencies within families. The appropriate focus of traditional descriptive epidemiology on representative sampling frames and high response rates is paralleled by the equal importance of sampling schemes in descriptive genetic epidemiology and a full section is therefore devoted to non-random ascertainment.

32.1 INTRODUCTION

Epidemiology may be defined as 'the study of the distribution and determinants of health-related states and events in *populations*' (Last, 2001). The term *genetic epidemiology* is less clearly defined. Neel and Schull (1954) referred to *epidemiological genetics* almost 50 years ago and, since then, different authors have variously emphasised the focus of genetic epidemiology on familial aggregation (King *et al.*, 1984), inherited disease in populations (Morton and Chung, 1978), the genetic structure of populations (Roberts, 1985) and on gene–environment interactions (Cohen, 1980). Perhaps the most satisfactory single definition is that given by Morton when he defined genetic epidemiology as 'a science which deals with the aetiology, distribution and control of disease *in groups of relatives* and with *inherited causes of disease in populations*' (Morton, 1982). The italics in Last's definition of epidemiology and Morton's definition of genetic epidemiology are after Hopper (1992). Firstly, they draw attention to the fact that both disciplines have a focus on being able to draw inferences at the level of a population rather than at the level of an individual (or a single family). This is not to say that some of the analytical methods

used in genetic epidemiology cannot also be used to draw inferences in individual families, e.g. in genetic counselling, but that, like mainstream epidemiology, genetic epidemiology is primarily about the pooling of information across many individuals (or families) with the aim of drawing more powerful inferences about potentially weak effects at the level of a population. Secondly, they emphasise that a key difference between epidemiology and genetic epidemiology is that while the former usually (but not always) focuses on individuals, the latter often (though again not always) focuses on families and that interest often centres upon inherited causes of disease. In this regard, it is important to note that the term *inherited* does not necessarily imply 'genes', but that it also subsumes non-genetic causes of disease clustering within families and this includes cultural and environmental mechanisms of inheritance (Burton *et al.*, 2005).

In general terms, the studies conducted by mainstream epidemiological research groups may be divided into three broad categories: (1) those that aim to describe the distribution of a disease or a determinant at the level of a population of interest (often the general population); (2) those that attempt to investigate a potential aetiological link between one or more specific determinants and a disease of interest; and (3) those aimed at formally evaluating the effectiveness and/or cost of an intervention applied to individuals or groups of individuals in the general population. Historically, much of the work of genetic epidemiology may be viewed as having been of the first type and it is upon this category that this chapter will concentrate. However, many of the contemporary developments in genetic epidemiology address studies of the second type, and this category will be briefly discussed in Section 32.4. It is likely that, in the future, genetic epidemiologists will become increasingly involved in studies of the third type as they are required to evaluate interventions developed from the new genetics. However, the issues relevant to the undertaking of such studies go beyond the remit of this particular chapter.

We make no claim that this chapter represents a definitive review of all of the many issues it attempts to address. Its primary aim is to consider descriptive analysis in genetic epidemiology in the light of the analogous aims of mainstream epidemiology, and to point the reader to some of the highlights of a broad and complex literature. Some of the issues we raise are also addressed in other chapters in this handbook, and for reviews of much of the theory and practice of genetic epidemiology, we recommend the book *Biostatistical Genetics and Genetic Epidemiology* edited by Elston *et al.* (2002) and also the *Lancet Series in Genetic Epidemiology* (Burton *et al.*, 2005; Cordell and Clayton, 2005; Davey Smith *et al.*, 2005; Hattersley and McCarthy, 2005; Hopper *et al.*, 2005; Palmer and Cardon, 2005; Teare and Barrett, 2005).

32.2   DESCRIPTIVE EPIDEMIOLOGY

Descriptive studies in the traditional epidemiological setting typically focus upon the distribution of an outcome of interest (a disease, a health care event or possibly an aetiological exposure) in the general population, or in some defined subsection of the general population. A typical example of such a study might be a survey investigating the prevalence of symptoms related to asthma in children up to five years of age (Luyt *et al.*, 1993). Such a study might be undertaken for the purpose of determining appropriate health care provision, or alternatively as a lead into a potential aetiological study by asking

a question such as, 'has there been a change in prevalence over time?' (Kuehni *et al.*, 2000). In practice, such studies are difficult to conduct. This is because it is essential that the study population is representative of the target population to be described and careful study design and conduct are therefore crucial. It is essential that the sampling frame (the formal list of potential sampling units) is clearly defined and that it is appropriate, given the specific purpose of the study. It is also crucial that the response rate to the study (if it involves seeking some form of response from individuals) is high and that it is unbiased with respect to the outcome of interest. It is important that the outcome of interest is assessed in a valid and rigorous manner (Hennekens and Buring, 1987; Rothman and Greenland, 1998). Finally, analysis has to be undertaken with care, in particular, it is important to ensure that any potential correlation between observational units is recognised, investigated and taken into appropriate account. Such correlation is occasionally of interest in its own right and, even if it is not, failure to model it correctly will lead to incorrect standard errors and sometimes to biased parameter estimates.

All of these concepts apply equally to descriptive studies in genetic epidemiology (see Section 32.3), but for a variety of reasons, the issues are even more important in this latter setting. Firstly, there is a wider class of genetic epidemiological studies that may reasonably be viewed as being primarily descriptive in nature than in traditional epidemiology. Indeed, if one looks at the history of research in genetic epidemiology, much of the work that has been undertaken *has* been descriptive in nature. Historically, the key question that genetic epidemiologists often addressed was: '*is the distribution of disease D between and within families, consistent with there being an inherited component that might be consequent upon genes?*' Secondly, the option of rendering a study unbiased at the level of the general population in relation to the outcome of interest, while desirable, is sometimes not possible in genetic epidemiology. Many conditions of interest to genetic epidemiologists are simply too rare to be efficiently studied using random samples. Instead genetic epidemiologists often have to work with samples that are deliberately biased (by oversampling families with affected individuals) and must then deal with the implications of this non-random ascertainment in the analysis (Fisher, 1934; Elston and Sobel, 1979; Burton *et al.*, 2000). Thirdly, the correlation structure between family members, which is in part consequent upon powerful biological mechanisms, is complex. Furthermore, it is often of primary interest to genetic epidemiologists and must therefore be modelled carefully.

Figure 32.1 is a flow chart illustrating a 'stylised' overview of the sequence of tasks that might be taken on by a genetic epidemiologist in trying to determine whether one or more genes might exist that are likely to have a large enough effect upon the phenotype of interest to make it worthwhile to invest substantial resources to identify these genes. Several points should be made. Firstly, this is a rather simplified perspective of the way that research is (or has ever been) conducted in the real world. Secondly, Figure 32.1 emphasises the fact that this form of descriptive analysis should not be viewed as an end in itself; its primary objective is to help prioritise the expensive research that then needs to be undertaken in collaboration with bioscientists. Thirdly, the increasing focus on complex diseases and the plethora of biological knowledge and data that have followed in the wake of the human genome project have rendered such a stepped approach less common in recent times. The most expensive part of many contemporary studies is the collection of families, and the phenotyping (Thompson, 2002). Therefore once such data have been collected, there is a natural tendency to jump straight to tests of
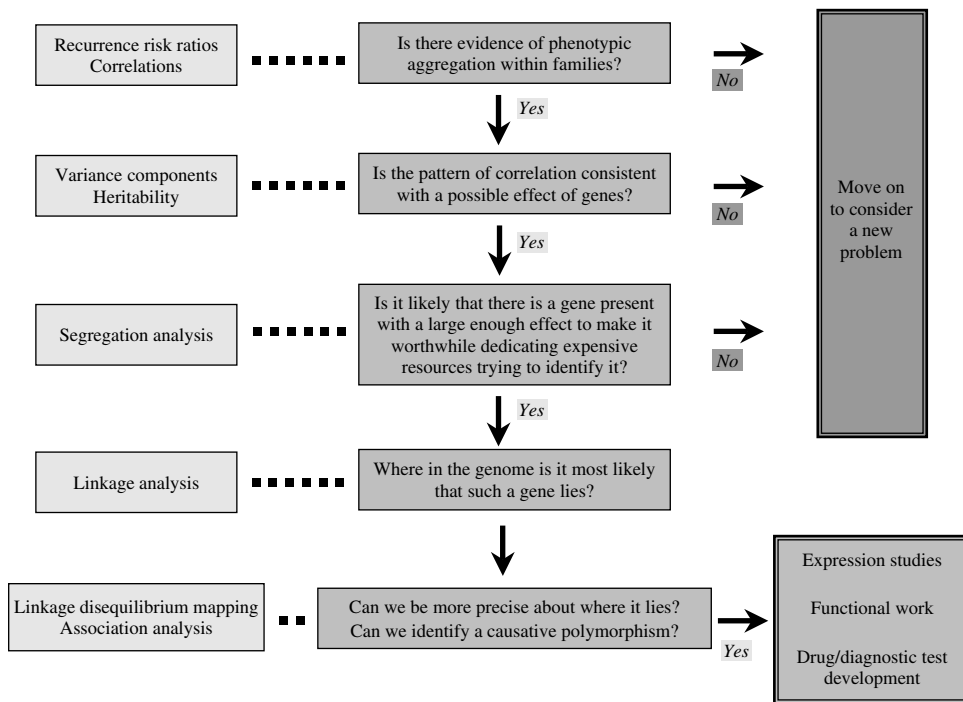
Figure 32.1 A stylised flow chart outlining a stepped approach in genetic epidemiology. (The failure to find evidence of genetic variant(s) with large enough effects will not necessarily impede contemporary aetiological studies shown in subsequent steps. See (Burton *et al.*, 2005) for a discussion of this issue.)

associations between phenotype and specific biomarkers (including genotypic markers) rather than investing extensive time and resources examining the familial distribution of the phenotype. Nevertheless, such analyses continue to be undertaken in an attempt to provide us with some initial guidance in our first attempts to understand the aetiology of the complex diseases (Duffy *et al.*, 1990; Palmer *et al.*, 2001). Fourthly, there are important overlaps with equivalent perspectives in agricultural genetics (see **Chapter 20**).

32.2.1 Incidence and Prevalence

Much of the descriptive analysis in epidemiology involves the investigation of prevalence or incidence.

Assuming that disease status is uncorrelated between individuals, the estimated prevalence of disease at time t in a population under study is defined as

$$\hat{\pi}_t = \frac{D_t}{N_t}, \tag{32.1}$$

where D_t is the number of individuals who are diseased at time t, and N_t is the total number of individuals who were evaluated at that same point in time. Because it relates to a specific point in time, prevalence is sometimes referred to as *point prevalence*. The

statistic $\hat{\pi}_t$ is a proportion and exact or asymptotic inferences may be based on the binomial log likelihood which (up to a constant) is given by

$$\ell = D_t \log(\pi_t) + (N_t - D_t) \log(1 - \pi_t). \tag{32.2}$$

Standard errors for $\hat{\pi}_t$ may be obtained in a variety of ways (Clayton and Hills, 1993). One of the most convenient methods (Clayton and Hills, 1993) is to use the Gaussian approximation to the binomial log likelihood and to treat $\text{logit}(\hat{\pi}_t)$ as being asymptotically normally distributed with standard error:

$$\sqrt{\frac{1}{D_t} + \frac{1}{N_t - D_t}}. \tag{32.3}$$

Incidence is a rate (λ) rather than a proportion. It is estimated as the number of new cases of disease (failures) occurring over a period of time (D) divided by the total number of person-years of at-risk exposure (Y) experienced by individuals in the study over that same period of time (Clayton and Hills, 1993):

$$\hat{\lambda} = \frac{D}{Y}. \tag{32.4}$$

The person-years of exposure is obtained by summing the subject-specific periods of at-risk exposure across all individuals. For most purposes, individuals are not viewed as being at risk of developing the disease once they have already developed the disease and/or following an event such as death. These events therefore abbreviate the total person-years of exposure. Analysis is typically based on the Poisson log likelihood which (up to a constant) can be expressed as

$$\ell \approx D \log(\lambda) - \lambda Y. \tag{32.5}$$

On the basis of a Gaussian approximation to the Poisson log likelihood, a standard error for $\log(\hat{\lambda})$ can conveniently be obtained (Clayton and Hills, 1993) as

$$\sqrt{\frac{1}{D}}. \tag{32.6}$$

Analyses of incidence and prevalence are easily extended to incorporate observed covariates by using generalised linear modelling (McCullagh and Nelder, 1989) in the form of logistic or Poisson regression (see also **Chapter 36**) (Breslow and Day, 1980; 1987; Clayton and Hills, 1993). Incidence data can also be modelled by treating the analysis as a failure time problem and using conventional survival time methods (Miller *et al.*, 1981) including Cox proportional hazards regression to incorporate covariates (Cox, 1972).

32.2.2 Modelling Correlated Responses

Correlation, covariance or association between the discrete outcomes in individual subjects in descriptive mainstream epidemiology usually reflects one of the following: (1) cluster sampling by design; (2) a multi-centre study design; (3) longitudinal repeated measurements over time; (4) a natural clustering structure in the study population, e.g. data related to patients registered with general practitioners (family physicians); or (5) latent

determinants of disease which have a tendency to cluster geographically or in time. Depending upon the purpose of analysis, and the structure of the data, a variety of different classes of models may be used to analyse such data. A comprehensive review of these models, and their implications for model interpretation would warrant a full chapter on its own (e.g. see: Neuhaus, 1992; Breslow and Clayton, 1993; Pendergast *et al.*, 2002) and we will restrict ourselves to a few brief comments.

The classes of model that may be used include (Neuhaus, 1992; Diggle *et al.*, 1994; Pendergast *et al.*, 2002): (1) conditional models in which the probability of response of an individual observational unit is conditioned on all other responses in a cluster; (2) transition models in which there is a logical ordering to the responses in a cluster and conditioning can be restricted to those responses that pre-date a response of interest; (3) marginal models in which one jointly estimates parameters underpinning the marginal mean and others reflecting the correlation or association between units in a cluster but, crucially, the model for the marginal mean does not include cluster-specific effects, and this results in parameters having a 'population-averaged' interpretation; (4) cluster-specific (random effects) models, which differ from marginal models in that they *do* have cluster-specific effects in the model for the mean, and parameters therefore have a 'subject-specific' interpretation. Many of the models that can be used to analyse correlated data may be viewed as being different types of generalised linear mixed model (GLMM) (Breslow and Clayton, 1993) and there are many different approaches to model fitting (Breslow and Clayton, 1993; Pendergast *et al.*, 2002). We will restrict our comments to some of the methods that are used most commonly in mainstream epidemiology.

If primary scientific interest focuses on the fixed effects and the correlation structure is essentially a nuisance, marginal models based on conventional (first order) generalised estimating equations (GEEs) provide a quick and convenient way to analyse the required data (Liang and Zeger, 1986; Zeger and Liang, 1986; 1992; Burton *et al.*, 1998; Pendergast *et al.*, 2002). In a conventional GEE the equations for the fixed effect parameters and the correlation parameters are assumed orthogonal and estimation is based solely on the mean and second order correlation components only. Provided the model for the mean is correctly specified and a robust sandwich-based estimator of the covariance matrix for the fixed parameter estimates is used (Huber, 1967), and provided inferences are then based on Wald tests and estimation intervals, the conventional GEE approach generates consistent (though not necessarily fully efficient) inferences for the fixed regression coefficients even if the covariance structure is incorrect or varies somewhat from cluster to cluster. Models based on GEEs can easily be fitted in many standard software packages including Stata (xtgee) (Stata Corporation, 2005) and SAS (Proc Genmod) (SAS Institute Inc., 2001).

Unfortunately, models fitted using first-order GEEs will often not provide good estimators of the parameters underlying the correlation/covariance structure itself (Pickles, 2002). In consequence, if the correlation is of primary interest in its own right, approaches based on multi-level modelling (Goldstein, 1986; 1991; Breslow and Clayton, 1993), structural equation modelling (Wright, 1921; Neale and Cardon, 1992) or higher order GEEs (Prentice and Zhao, 1991; Pendergast *et al.*, 2002; Pickles, 2002) are commonly used. In practice, models such as these are most often fitted using general purpose software such as MLWin (Rasbash *et al.*, 1999), Splus (lme, nlme) (MathSoft, 2000), Lisrel (Jöreskog and Sörbom, 1996) or Mx (Neale *et al.*, 2002) or specialist software such as GEE4 (Hanfelt, 1993). In recent times there has been an increasing use of Markov chain

Monte Carlo (MCMC) based approaches (Breslow and Clayton, 1993) particularly Gibbs sampling (Zeger and Karim, 1991; Breslow and Clayton, 1993; Spiegelhalter *et al.*, 2000).

Although most traditional epidemiological analysis is based upon the binomial and Poisson distributions, some quantitative outcomes (such as blood pressure) and exposures are more naturally modelled assuming normality either directly or following transformation. Standard statistical methods for Gaussian responses may then be used. These again include univariate methods (Armitage and Berry, 2002), generalised regression-based techniques (McCullagh and Nelder, 1989) and extensions to deal with correlated data (Goldstein, 1987; Breslow and Clayton, 1993; Burton *et al.*, 1998). Many mainstream epidemiologists use SAS (Proc Mixed) (SAS Institute Inc., 2001) or SPSS (GLM repeated measures) (SPSS Inc., 2002) to model correlated continuous responses.

32.3   DESCRIPTIVE GENETIC EPIDEMIOLOGY

Like any other epidemiologists, genetic epidemiologists sometimes need to undertake simple descriptive analyses based on prevalence, incidence or quantitative outcomes. Such analyses may be based on the methods outlined above. In addition, however, genetic epidemiologists need to be able to carry out the more specialised types of analysis that permit them to address the questions addressed in the stylised flow chart in Figure 32.1.

32.3.1   Is There Evidence of Phenotypic Aggregation within Families?

Simple familial aggregation can be studied in a number of different ways. For a binary phenotype (disease absent/present), one of the simplest forms of analysis is to estimate the recurrence risk ratio in relatives of type R of affected individuals (Risch, 1990). This risk ratio is defined as

$$\frac{\text{Pr(Disease in relative of type } R | \text{ affected case)}}{\hat{\pi}}, \tag{32.7}$$

where $\hat{\pi}$ is the estimated prevalence of the disease in the general population (Risch, 1990; Kopciuk and Bull, 2002). Most commonly it is the recurrence risk ratio for siblings ($\hat{\lambda}_S$) that is estimated. As well as evaluating the evidence in favour or against familial aggregation, Risch (1990) has shown that this statistic is an important determinant of the power of affected relative pair studies to detect linkage (see **Chapter 34**). Despite the simplicity of the λ_S statistic, some comments are warranted. Firstly, it reflects aggregation within a sibship regardless of its cause. It does *not* specifically estimate aggregation due to genes. Secondly, particularly in diseases with a late onset, any systematic differences between the age distribution in the siblings of the affected cases, and the age distribution in the population sample from which $\hat{\pi}$ was estimated could distort the statistic. Thirdly, in relation to diseases with a high population prevalence, such as asthma (Palmer *et al.*, 2001), $\hat{\lambda}_S$ has a theoretical upper bound at $\frac{1}{\pi}$ which means that one should be cautious in comparing values between populations that have different prevalences.

A variety of other risk ratios for relatives that may be used to assess familial aggregation have also been described (Kopciuk and Bull, 2002).

Familial aggregation can also be assessed by estimating appropriate intra-class and inter-class correlation coefficients. For quantitative phenotypes, this approach dates back more

than a century to Galton (1877; 1886) and Pearson (1896). From a modern perspective, the assessment of familial aggregation in this manner is no different to the more general problem of modelling clustered data with a simple correlation structure in mainstream epidemiology (see Section 32.2.2). All the same issues therefore apply and the same models and model fitting approaches may be used. It should however be noted that certain statistical packages, such as the fortran-based program FISHER (Lange *et al.*, 1988) which fits maximum-likelihood-based linear mixed models for normally distributed phenotypes, and Lisrel (Jöreskog and Sörbom, 1996) and Mx (Neale *et al.*, 2002) which may be used to fit structural equation models have been used much more commonly by genetic epidemiologists than by mainstream epidemiologists.

32.3.2   Is the Pattern of Correlation Consistent with a Possible Effect of Genes?

If and when it has been demonstrated that there is evidence of simple familial aggregation, the next logical step is to investigate the structure of the correlation underpinning that aggregation. The use of variance components models to explore complex correlation or covariance structures has a long and distinguished history in genetic epidemiology (Fisher, 1918; Wright, 1921; Jinks and Fulker, 1970; Neale and Cardon, 1992; Hopper, 1993; Khoury *et al.*, 1993a; Hopper, 2002b; Neale, 2002; Rao and Rice, 2002). More recently, extensive work has been carried out based upon mixed effects models fitted using MCMC methods (Guo and Thompson, 1992; Sobel and Lange, 1993; Gauderman and Thomas, 1994; Burton *et al.*, 1999; Scurrah *et al.*, 2000).

As in mainstream epidemiology, many of the relevant models may helpfully be viewed as being GLMMs (Breslow and Clayton, 1993). In order to facilitate a discussion of these models and the uses to which they may be put, we will now consider the structure of one such GLMM in more detail.

32.3.2.1   Model Structure

Combining the notation of several authors (Khoury *et al.*, 1993a; Burton, *et al.*, 1999; Hopper, 2002b), a general model with wide applicability may be written as

$$g(\mu_{ij}) = \eta_{ij} = \alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_{ij} + \xi_{ij}, \tag{32.8}$$

$$Y_{ij} \sim f(\mu_{ij}, \varpi),$$

$$\xi_{ij} \sim N(0, [\sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{D}} + \sigma^2_{\mathrm{C}}]),$$

$$\mathrm{Cov}(\xi_{ij}, \xi_{ik})[j \neq k] = 2\phi_{ij,ik}\sigma^2_A + \Delta_{ij,ik}\sigma^2_{\mathrm{D}} + \gamma_{ij,ik}\sigma^2_{\mathrm{C}},$$

where Y_{ij} is the observed phenotype in the jth member of the ith family, μ_{ij} is its expected value, $f(.)$ denotes an error distribution (from the exponential family) which may incorporate a nuisance parameter denoted ϖ. The expected value of the phenotype is predicted via a link function $g(.)$ applied (in inverse form) to a linear predictor (η_{ij}) comprising a baseline mean (α), a vector of observed covariates ($\mathbf{z}_{ij}$), a corresponding vector of unknown regression parameters ($\boldsymbol{\beta}$) and subject-specific random effects ξ_{ij} with an appropriate covariance structure. The components $\sigma^2_{\mathrm{A}}, \sigma^2_{\mathrm{D}}$ and σ^2_{C} represent, respectively, the variances arising from polygenic additive effects, polygenic dominance

effects and shared environmental effects (Fisher, 1918; Khoury *et al.*, 1993a; Hopper, 2002b). The term $\phi_{ij,ik}$ denotes the kinship coefficient between individuals ij and ik: the probability of randomly drawing a single allele in individual ij that is identical by descent (*ibd*) to a single allele at the same locus randomly drawn from individual ik. $\Delta_{ij,ik}$ is the probability that *both* alleles at a locus are shared *ibd* by individuals ij and ik. Table 32.1 details the $\phi_{ij,ik}$ and $\Delta_{ij,ik}$ values for selected relative pairs and the total genetic variances that these imply.

In many models, the elements $\gamma_{ij,ik}$ are simply binary indicators denoting whether two individuals live together ($\gamma_{ij,ik} = 1$) or apart ($\gamma_{ij,ik} = 0$). However, the effect of shared environment may be modelled in a more sophisticated manner. That is by allowing the impact of shared environment to increase over time living together and then to decline with time living apart (Hopper, 1993). Furthermore, one may choose to discriminate between the effects of living together *per se* from those of living together *as a child* by introducing an additional variance component to model shared sibling environment (σ^2_{Cs}) (Hopper, 1993; Burton *et al.*, 1999). However, σ^2_{Cs} is completely confounded with σ^2_{D} under most simple family designs (except those including monozygous (MZ) twins, double first cousins, adoptees or siblings reared apart). Given appropriate data, the genetic component of variance may be extended to incorporate gene:gene (epistatic) and/or gene:environment interaction effects.

Under model (32.8) (as it is specified) and regardless of the structure of the residual error, the distribution of the ξ_{ij} is multivariate normal within families. This parameterisation is often chosen for reasons of analytic and interpretational tractability (Burton *et al.*, 1999). However, depending how the model is fitted, alternative assumptions may be both possible and preferable. For example, γ frailties are often used in survival time models (Clayton, 1991; Hougaard and Thomas, 2002).

Table 32.2 lists combinations of error structures and link functions that are commonly used to model a number of important classes of trait (McCullagh and Nelder, 1989) (see also **Chapter 20**).

A variance components GLMM, such as model (32.8), invokes a number of critical assumptions (see Box 32.1). These include assumptions about the statistical model, assumptions about the underlying biology and assumptions about the epidemiological

Table 32.1 Genetic components of variance assuming random mating.

Relationship	ϕ	Δ	Genetic covariance
Same person	1/2	1	$\sigma^2_{\text{A}} + \sigma^2_{\text{D}}$
Parent–child	1/4	0	$1/2\sigma^2_{\text{A}}$
Full-siblings	1/4	1/4	$1/2\sigma^2_{\text{A}} + 1/4\sigma^2_{\text{D}}$
Half-siblings	1/8	0	$1/4\sigma^2_{\text{A}}$
Monozygous twins	1/2	1	$\sigma^2_{\text{A}} + \sigma^2_{\text{D}}$
Grandparent–grandchild	1/8	0	$1/4\sigma^2_{\text{A}}$
Uncle/aunt–nephew/niece	1/8	0	$1/4\sigma^2_{\text{A}}$
First cousins	1/16	0	$1/8\sigma^2_{\text{A}}$
Double first cousins	1/8	1/16	$1/4\sigma^2_{\text{A}} + 1/16\sigma^2_{\text{D}}$
Spouses	0	0	0

Table 32.2 Error-link combinations for common classes of phenotype.

Class of trait	$f(\mu_{ij}, \varpi)$	g(.)	Notes
Continuous normally distributed	$N(\mu_{ij}, \sigma^2_E)$	identity(.)	_
Binary	Bernoulli(μ_{ij})	logit(.)	probit(.) is a common alternative link
Count	Poisson(μ_{ij})	log(.)	_
Censored survival time	Poisson(μ_{ij})	log(.)	Response is the binary censoring indicator, log(survival time) is used as an offset

---

Mathematical assumptions

Conditional on fixed and random effects, all responses are independent
Random effects are mutually independent and are independent of the fixed effects
The structure of the model is appropriate (e.g. random effects *are* multivariate normal, and the assumed distribution of observed phenotype about its modelled predictions is correct).

Biological assumptions

The population is in Hardy–Weinberg equilibrium
The biological model is complete
The recorded family structures are correct (e.g. no unrecognised non-paternity).

Epidemiological assumptions

All recorded information (outcomes and covariates) is recorded faithfully
Ascertainment is either random (as assumed in model 1), or else the ascertainment model is appropriately reflected in the likelihood underpinning the model.

Box 32.1   **Modelling assumptions for variance components model.**

design. The risk of violation of each assumption and the relevance of a violation should it occur, will vary from trait to trait and will be influenced by the focus of scientific interest (Hopper, 1993).

32.3.2.2 *Model Fitting*

A GLMM equivalent to model (32.8), may be fitted using a range of different techniques. This includes many of the model-based approaches to analysing correlated data considered in Section 32.2.2. However, first-order GEE models are not ideal if the covariance structure is of primary interest. Model fitting using MCMC (for complex likelihoods in either a

Bayesian or non-Bayesian setting) also offers great flexibility and generalises relatively easily to a wide variety of non-normal traits and to complex extended pedigrees (Guo and Thompson, 1992; Hopper, 1993; Gauderman and Thomas, 1994; Burton *et al.*, 1999; Scurrah *et al.*, 2000). On the other hand, the use of MCMC can be demanding in terms of time commitment to setting up the data, undertaking the analysis and checking the MCMC process for convergence (Gilks *et al.*, 1996). Furthermore, it is often considered undesirable to use an MCMC approach when an exact likelihood-based solution is available.

32.3.2.3   *The Interpretation of Parameters*

Model (32.8) is a cluster-specific model that generates subject-specific inferences. This is because the random effects (ξ_{ij}) generating the covariance structure appear on the linear predictor. Under this model, $\hat{\beta}_x$ estimates the change in η_{ij}, or equivalently $g(\mu_{ij})$, associated with a one unit change in covariate $\mathbf{z}_x$ *in an individual*. The variance components also have a subject-specific interpretation (Burton *et al.*, 1999; Scurrah *et al.*, 2000). For example, in a model with a logit link, if $\hat{\sigma}_A^2 = 0.5$, then it suggests that a subject at the upper 95 % percentile for additive polygenic effects will have a predicted log(odds) of disease that will be $\Phi(0.95) \times \sqrt{0.5} = 1.645 \times \sqrt{0.5} \approx +1.16$ higher than it would have been if the same subject had been at the median for additive polygenic effects but nothing else had changed (i.e. observed covariates and other random effects were unaltered). This corresponds to an odds ratio of $\exp(1.16) \approx 3.19$.

In a marginal model, the variance components reflect the marginal covariance between the predicted means based solely on the estimated fixed effects (Breslow and Clayton, 1993; Pendergast *et al.*, 2002). Furthermore, $\tilde{\beta}_x$ (estimated under the marginal model) has a population-averaged interpretation reflecting the difference in the marginal expectation of η_{ij} associated with a one unit increase in $\mathbf{z}_x$. In effect, $\tilde{\beta}_x$ is averaged over the full range of frailties in the particular population under study (Diggle *et al.*, 1994). Under an identity link $\hat{\beta}_x = \tilde{\beta}_x$. Under logit and probit links $\hat{\beta}_x \geq \tilde{\beta}_x$. The difference increases as the variance of the frailties rises, and equality only holds if the frailty variance is zero (Neuhaus, 1992; Diggle *et al.*, 1994). The estimated variance components obtained from a marginal model are also shrunk relative to their equivalents in a cluster-specific model in GLMMs with logit and probit links. For a logarithmic link, there is an offset in the estimated marginal mean (Zeger *et al.*, 1988).

32.3.2.4   *Identifiability of Variance Components*

The joint estimation of the different variance components depends upon the availability of a data set containing a suitable range of different classes of familial relationship. For example, with no further information, a data set based on standard nuclear families allows the modelling of different covariances between two spouses, a parent and a child and between two siblings. This allows the simultaneous resolution of three variance components such as σ^2_A, σ^2_C and σ^2_{Cs} or σ^2_A, σ^2_C and σ^2_D (Burton *et al.*, 1999). Extension to three or more generations will not necessarily increase the number of variance components that may be identified. For example, despite the provision of an extra class of relation (grandparent:grandchild) addition of the grandparental generation

will still not permit simultaneous resolution of $\sigma^2_A, \sigma^2_C, \sigma^2_{Cs}$ and σ^2_D because σ^2_{Cs} and σ^2_D will still be linearly dependent (see Table 32.1). On the other hand, such an extension *will* increase the power to resolve σ^2_A, σ^2_C and σ^2_{Cs} and will markedly improve mixing in the MCMC setting (Scurrah *et al.*, 2000). In order to increase the number of components that may be identified, one must either incorporate individuals with different environmental exposures (e.g. some siblings living together and some living apart) or use designs including additional genetically informative relative types such as adoptees, half-siblings or MZ twins. In this context, a particularly attractive design is the twin-family study (Hansen *et al.*, 2000) which enrols MZ and dizygous (DZ) twin pairs with their parents and siblings. Even when all members of every family unit cohabit, such a design permits simultaneous identification of $\sigma^2_A, \sigma^2_C, \sigma^2_{Cs}$ and σ^2_D.

The classical twin design, MZ and DZ twins reared together (with no information on other family members) (Merriman, 1924; Siemans, 1924), provides a robust test of $H_0: \sigma^2_G = 0$. This is because under the null hypothesis and *assuming an equal environmental covariance*, the covariance between MZ twins and DZ twins should be the same and any genetic determinants, regardless of their nature, will tend to increase the covariance between MZ twins more than DZ twins. However, there are important interpretational limitations (Neale and Cardon, 1992; Neale, 2002). For example, in the absence of additional information (e.g. some twins are reared together and some apart) only two variance components may be resolved. This is because there are only two different covariances: that between a pairs of MZ twins and that between a pairs of DZ twins. Furthermore, the assumption that the shared environmental covariance between pairs of MZ and DZ twins is the same may be viewed as suspect. In contrast, if one is prepared to restrict the model to three covariance terms, the twin-family design (see above) enables this key assumption to be formally tested (Hansen *et al.*, 2000).

32.3.2.5 Twin Studies

There is no doubt that twin studies have been of great historical importance in the development of genetic epidemiology. This is principally because they provide a natural experiment that provides a straightforward means of separating the 'nature' and 'nurture' underlying phenotypic traits (Galton, 1875; Neale and Cardon, 1992; Spector *et al.*, 1999; Spector, 2000; Neale, 2002). That said, the classical twin design does have significant inferential limitations (see above), and it is important that the value of twin studies is not overstated. The potential role of a twin-based design in answering a scientific question of interest must be considered just as carefully as the corresponding role of any alternative design (Neale and Cardon, 1992; Neale, 2002).

From an epidemiological perspective, one of the great advantages of twin studies is that, as a population subgroup, twins are generally very proud of their status and are very motivated to join biomedical research studies. In consequence, there are many large twin registries around the world that provide an important opportunity for genomic and genetic epidemiological study (Strachan, 2002). A far from exhaustive list of major twin registries includes: the Australian Twin Register (`http://www.twins.org.au/`); GenomEUtwin, a biobank that pulls together twin registers across Europe (`http://www.genomeutwin.org/`); and the US Veterans Twins Study (`http://www.iom.edu/CMS/3795/4907.aspx`).

In the post-genomic era, the ability to work directly with characterised genotypes rather than having to infer genetic etiology indirectly from the covariance structure of a trait (Burton *et al.*, 2005) has meant that the role of twin registries is changing. They continue to provide an invaluable source of well-characterised epidemiological information on large numbers of subjects. But their role in providing a natural experiment to explore nature and nurture is becoming less crucial. In contrast increasing use is likely to be made of the special inferential opportunities that arise from twin-based studies. For example, MZ twins provide an ideal group in which to study gene–environment interactions. This is because they provide a 'natural experiment' in which the impact of different environments can be compared and contrasted between twins with exactly the same genetic background. Another opportunity arises from the fact that there is a clear need to increase understanding of the various sources of error in transcriptomics, proteomics and metabolomics. Twin studies, and in particular twin-family studies, provide a powerful design upon which to base such research and, in particular, to discriminate random measurement errors from variability arising from the systematic effect of modulating genes and environmental determinants.

32.3.2.6 Negative Variance Components

There is a debate as to whether a variance component such as σ^2_A should be allowed to take negative values. The biological motivation for variance components in genetic epidemiology usually invokes unobserved determinants (genetic or environmental) that are shared more commonly among close relatives. Under such a model, unless determinants are shared systematically *less* commonly among close relatives, truly negative correlations or covariances should not arise. However, unless one believes absolutely in the biological model motivating the variance components model, it can reasonably be argued that the only constraint that should be applied to the model is the fundamental requirement that a variance–covariance matrix should be positive definite; this requires, e.g. that $\sigma^2_A \geq -\sigma^2_{Cs}$ (Burton *et al.*, 1999) and is therefore less constraining that the requirement that it be non-negative. Further, it may be argued that it is only by permitting the variance components to take legal negative values that one can properly test the biological assumption that they *are* individually positive. The use of a less restrictive parameterisation (Burton *et al.*, 1999; Scurrah *et al.*, 2000) is particularly relevant for models fitted using an MCMC approach with inferences being based on the posterior mean, because one is then averaging the posterior distribution across iterations. If a variance component is in truth positive but close to zero, a conventional likelihood-based solution will be valid. On the other hand, unless the MCMC chain is allowed to traverse those legal parts of the parameter space associated with negative values, the MCMC generated estimate of a variance component will be positively biased (Zeger and Karim, 1991; Burton *et al.*, 1999).

32.3.2.7 Modelling Discrete Traits

Variance components models for discrete traits are often motivated by assuming an underlying normally distributed latent trait with one or more thresholds (Falconer, 1965; Hopper, 1993; Khoury *et al.*, 1993a; Todorov and Suarez, 2002). In the simple case

of a binary disease state, the unobserved value of the latent trait in the jth subject in the ith family may be denoted X_{ij} and the threshold in that same individual as τ_{ij}. The subject is modelled as 'affected' if $X_{ij} \geq \tau_{ij}$ and 'unaffected' otherwise. It has been argued that this model is fundamentally incoherent (Edwards, 1969; Todorov and Suarez, 2002) because very similar latent trait values on either side of a threshold imply a discordant affection status, while potentially very different values on the same side imply concordance. However, this concern has been overstated (Todorov and Suarez, 2002). In its full generality, the latent trait in the ijth individual could be viewed as being drawn from a $N(\theta_{ij}, \sigma^2_{ij})$ distribution. But, a model with subject-specific values for $\theta_{ij}, \sigma^2_{ij}$ and τ_{ij} would be overparameterised given that the X_{ij} are unobserved. Constraints must therefore be applied. If one arbitrarily specifies $\tau_{ij} = 0$ and $\sigma^2_{ij} = 1$ for all i and j, and sets $\theta_{ij} = \eta_{ij}$ (from model 32.8) then the latent trait model can be shown to be mathematically equivalent to a GLMM with a probit link. Under this model, $\mu_{ij} = \Phi(\eta_{ij})$, where $\Phi(.)$ is the cumulative distribution function for a standard normal random variable and μ_{ij} is the expected probability of disease in the ijth subject. Equivalently, under the latent trait model, the probability that X_{ij}, a random drawn from a $N(\eta_{ij}, 1)$ distribution, equals or exceeds $\tau_{ij} = 0$ is also $\Phi(\eta_{ij})$. This indicates that the latent threshold model provides a convenient motivation of what is fundamentally a coherent GLMM, with a probit link, under which the expected probability of affectation increases monotonically with the value of the linear predictor (η_{ij}).

32.3.2.8 Heritability

A starting point for many scientists investigating disease aetiology has often been to study the heritability of a health-related trait. Formally, the heritability of a continuous trait is defined as the proportion of its total variance (σ^2_{TOT}) that is attributable to genetic factors in a particular population. Narrow sense heritability is defined as $\sigma^2_A/\sigma^2_{TOT}$ and broad sense heritability as $\sigma^2_G/\sigma^2_{TOT}$, where σ^2_G includes all genetic components of variance.

Heritability is not about cause *per se*, but is about the causes of variation in a particular trait *in a particular population* (Hopper, 2002a). Heritability varies from study to study depending on the population being investigated (environmental exposures vary between populations), the structure of the analytic model and measurement error (Hopper, 1992; 2002a). Fisher (1951) pointed out that, in calculating heritability, whilst the numerator has a simple genetic meaning, the denominator does not. Scientists and the media sometimes treat heritability as meaning the proportion of a disease 'caused' by genetic factors. This is incorrect. If a disease process is entirely dependent upon the presence of a particular allele of a particular gene, but *everybody* in the population is homozygotic for that allele, variation at that locus will play no role in variation of the disease phenotype and it will therefore make no contribution to heritability. On the other hand, the gene is clearly implicated in the causal architecture of the disease. Equivalently, a near ubiquitous environmental exposure will make little or no contribution to the denominator, σ^2_{TOT}. Interpretational ambiguities are also introduced by decisions about the covariates to be included in a model and by interactions between genes and environmental determinants. For example, failure to include an important environmental covariate may well increase σ^2_E (the residual variance) and therefore σ^2_{TOT}, while leaving σ^2_A unchanged. This will apparently decrease the narrow sense heritability. For all of these reasons, it is often

preferable to quote the magnitude of the variance components (such as σ^2_A) individually, rather than relying solely on the overall value of the heritability itself (Fisher, 1951). The individual variance components each have a direct scientific interpretation, and their estimates can be compared meaningfully within and between populations (Hopper and Carlin, 1992; Burton *et al.*, 1999).

Because heritability is formally defined in terms of variation in a quantitative trait (Hopper, 2002a) there are specific interpretational problems that pertain to binary traits (Lichtenstein *et al.*, 2000; Spector, 2000). Although a binary trait can be modelled using a latent threshold model or equivalently a GLMM with a probit link (see above), this does not imply that the concept of heritability generalises directly from quantitative to binary traits. For example, in the setting of a binary disease state, the key assumption that the liability is normally distributed (Hopper, 2002a) cannot be tested. Furthermore, an important component of the variability in a binary problem is Bernoulli variation on the native scale of the trait. Mapping this to the scale of the linear predictor is a non-linear problem, and the required approximations are at their worst with binary responses when the binomial denominator is one (Breslow and Clayton, 1993). For a quantitative trait, a principal reason for calculating heritability, rather than quoting a raw value for the variance attributable to additive genetic effects, is that heritability 'standardises' σ^2_A for the 'intrinsic variability' of the trait under study. This is of obvious scientific utility. But for a binary phenotype, the intrinsic variability of the trait is determined principally by the Bernoulli error, which is fixed mathematically. Furthermore, for a typical complex trait, the variability will also include random unshared error terms that contribute directly to the assumed liability; e.g. those reflecting the impact of unmeasured aetiological determinants that are specific to an individual. In designs involving a single measurement of the trait, the additional variability arising from these error terms is generally non-identifiable for a binary phenotype (Neale and Cardon, 1992). In consequence, the 'variance of the liability' is fixed by assumption – it is not estimated from the data. There would appear to be little value in standardising σ^2_A for a quantity that is fixed by assumption, and this suggests that the calculation of 'heritability' for a binary trait in this setting is of rather limited utility.

Given the many pitfalls in the interpretation of heritability, why bother to calculate it at all? The power of most studies for discovering genes is positively associated with the heritability of the trait of interest; so, all else being equal and if the option exists, analytic efficiency may be enhanced by selecting a study population in which the heritability of the trait of interest is thought to be high. Furthermore, subject to all of the caveats above, knowledge that a trait of interest has a high heritability provides support for a study that proposes to investigate the genetic determinants of that trait. Equally, if heritability is low, the investigators and funders of the proposed study are forewarned that genetic effects may be difficult to find. In either case, interpretation demands expert understanding of the nature of the trait.

32.3.2.9 *Transition Models*

Most of the preceding discussion has focused on the use of marginal and cluster-specific models. However, in genetic epidemiology, transition models are also used to model within-family correlation structures. In particular, Bonney (1984) described a series of models that account for within-family dependencies by specifying a regression

relationship between an individual's phenotype and those of certain ancestors and older relatives. In special cases such a model reflects a simple first-order Markov process, but a range of alternative biologically sensible autoregressive dependency structures may also be specified. Classically, and although his nomenclature has changed somewhat over time, (Bonney, 1984; 2002; Hopper, 1995) described four major classes of 'regressive model': (Class A) regression on parental phenotypes and on spousal phenotype if spouse correlations are non-zero; (Class B) add regression on the oldest sibling; (Class C) add the immediately preceding sibling and (Class D) add all preceding siblings. The advantages of the regressive approach include the fact that the models can be fitted using standard regression software (simply by extending the design matrix), and that they may be applied quite naturally both to continuous normally distributed phenotypes (Bonney, 1984) and to binary traits (Bonney, 1986; 1987) by choosing an appropriate regression model. Furthermore, the properties of the models have been investigated extensively and specific analytical structures have been devised that reflect a wide range of different biological scenarios (Bonney, 2002). However, the models also have some disadvantages. These include a potential loss of efficiency arising from the conditioning on earlier responses (Diggle *et al.*, 1994), occasional ambiguities in how to order a pedigree and how to deal with missing data (Bonney, 1984), and the indirect manner in which inferences about the correlation structure are obtained. Regressive models for specialist applications in genetic epidemiology have been built into the software package SAGE (Elston, 2001).

32.3.3   Segregation Analysis

Having determined that the data are consistent with the presence of genetic effects, one may then move on to address the third question in Figure 32.1: 'Is it likely that there is a gene present with a large enough effect to make it worthwhile dedicating expensive resources trying to identify it?' This falls under the heading of segregation analysis and entails assessing whether there are important variants in one or a small number of major genes whose segregation (usually in a Mendelian manner) explains all or a substantial part of the observed variation in the trait of interest. Segregation analysis is already considered elsewhere in this handbook (**Chapter 19** and **Chapter 20**) and we will keep our comments brief.

Elston (1981) defines segregation analysis as: 'The statistical methodology used to determine from family data the mode of inheritance of a particular phenotype, especially with a view to elucidating single gene effects'. He goes on to emphasise that segregation analysis has four major components: (1) the genotypic distribution of mating individuals; (2) the relationship between phenotype and genotype; (3) the mode of inheritance and (4) the sampling scheme. He considers each of these components in detail (Elston, 1981).

Segregation analysis can conveniently be split into classical methods (Majumder, 2002) and more sophisticated model-based approaches which deal with complex segregation problems, particularly those in which more than one mating type is possible for a given sibship (Blangero, 2002). Classical segregation analysis often involves estimation of the *segregation ratio*: that is the probability that an offspring is affected given parental genotypes. That this is often not straightforward is generally a reflection of the complications introduced by non-random ascertainment. Classical estimators have been designed to deal with a number of simple ascertainment schemes (see Section 32.3.4). Weinberg (1912) proposed the *proband method* and derived a simple estimator for use under single ascertainment. The Li–Mantel estimator (Li and Mantel, 1968) may be

used under complete ascertainment and Davie (1979) proposed an efficient estimator for intermediate situations with incomplete ascertainment. In contrast, complex segregation methods reflect the fact that reality is more complicated and generally aim to build an analytical component reflecting one or more major genes into a more sophisticated model that simultaneously reflects other determinants of the phenotype of interest. Two approaches to complex segregation analysis that have been widely used are Morton and MacLean's mixed model (1974) and transition (regressive) models (see Section 32.3.2). Both of these approaches permit the modelling of a major gene against a background of polygenic and environmental effects.

Model fitting is often based upon regressive models or variance components models, sometimes incorporating a component of finite mixtures to reflect the major gene (Palmer *et al.*, 2001) (see also **Chapter 18**). Estimation may be based upon maximum likelihood, higher order GEEs or MCMC. Blangero provides an excellent contemporary review of complex segregation models (Blangero, 2002).

32.3.4   Ascertainment

Any descriptive study in mainstream epidemiology must invest considerable time and resource ensuring that its sampling frame is representative of the required target population, and that there is a high response rate. This emphasis highlights the importance of the recruitment process in descriptive epidemiology. Under many circumstances, the ideal sampling frame is one that permits the ascertainment of a random sample of the population to be studied. This is very convenient because it means that the ascertainment process can then be ignored in the analysis.

Unfortunately, genetic epidemiologists do not often have this luxury. If one is interested in the prevalence of disease Q amongst the offspring of parents who have mutually at-risk genotypes, then unless the at-risk genotype is relatively common, it is prohibitively expensive to recruit a random sample of the general population because most families will not be at risk and will therefore be uninformative. Consequently, it would be very useful to identify those families at risk and to restrict recruitment to this subset. If the at-risk genotype is expressed in the phenotype of the parents (e.g. if Q is caused by a dominant polymorphism with high penetrance at an early age) then it may be cheap and easy to identify relevant families through the parents. However, if the causative polymorphism (q) is rare and recessive, then almost all at-risk families will consist of doubly heterozygotic parents (Qq × Qq) who will both be unaffected. Consequently, in the absence of biological knowledge of the relevant genotypes, the only direct way to identify at-risk families is to sample families with affected offspring. From a traditional epidemiological perspective, this appears counter-intuitive: it would not normally be recommended that one should deliberately oversample affected individuals if one is trying to estimate the prevalence of a disease! It might be argued that if one obtained all or a representative sample of families with offspring affected by Q the problem would be circumvented. But this is not so. One group of at-risk families consists of parents who are both Qq heterozygotes, but (by chance) none of their offspring are qq homozygotes and, thus, none are affected by disease Q. By restricting recruitment to families with at least one affected child, this subset of families will be missed, and because amongst all those at risk these particular families have an unusually low prevalence (0 %), their exclusion will seriously bias a naïve estimate of the prevalence of Q in offspring. Ascertainment bias

is a recurring problem throughout genetic epidemiology and must always be considered both in study design and analysis.

Non-random ascertainment refers to a mode of sampling that depends systematically on the outcome being analysed as the dependent variable. Most commonly, families containing individuals affected by a disease of interest are preferentially oversampled. Non-random ascertainment has two important implications. Firstly, it distorts the distribution of the outcome in the sample being studied (see above). This may be referred to as *classical ascertainment bias* and its importance has been recognised for many years (Weinberg, 1928; Fisher, 1934; Morton, 1959). Secondly, in situations of aetiological heterogeneity, non-random ascertainment leads to the oversampling of families that truly have an unusually high risk of disease (Burton *et al.*, 2000; Epstein *et al.*, 2002; Glidden and Liang, 2002; Burton, 2003).

Classically, ascertainment has been addressed using the π model (Weinberg, 1928; Morton, 1959). This approach is based upon the concept of a proband who may be defined as: 'an affected person who at any time was detected independently of the other members of the family, and who would therefore be sufficient to assure selection of the family in the absence of other probands' (Morton, 1959). The ascertainment probability (π) is then defined as the probability that an affected individual becomes a proband and the ascertainment probability for a family as the probability that that family contains at least one proband. Thus, in an ascertainment scheme based upon offspring, a family containing s children where p_j denotes the probability of disease in the jth child under the current model, the probability that the family is ascertained (assuming that ascertainment events within sibships are conditionally independent) is given by

$$1 - \prod_{j=1}^{s} (1 - \pi p_j). \tag{32.9}$$

There are two important 'limiting' scenarios: (1) complete ascertainment ($\pi = 1$) under which *all* affected individuals are ascertained and the ascertained sample will include all families with at least one affected child; and (2) Single ascertainment ($\pi \to 0$) when the probability of ascertainment is very low, and the probability that any single family contains more than one proband becomes vanishingly small. Incomplete ascertainment refers to any intermediate situation ($0 < \pi < 1$).

A variety of classical estimators (see Section 32.3.3) have been devised to address the estimation of segregation ratios under these various ascertainment schemes (Hodge, 2002). However, the generality of the relevant theory goes beyond this. If one ignores the ascertainment process, the contribution of an ascertained family F to the likelihood under a statistical model (M) may be denoted $\Pr(\mathbf{Y_F}|M)$ where $\mathbf{Y_F}$ is the response vector for the family. Under non-random ascertainment, this should be replaced by the conditioned likelihood $\Pr(\mathbf{Y_F}|A_F^+, M)$ where A_F^+ denotes the event that family F has been ascertained. Standard probability theory implies

$$\Pr(\mathbf{Y_F}|A_F^+, M) = \frac{\Pr(A_F^+|\mathbf{Y_F}, M) \; \Pr(\mathbf{Y_F}|M)}{\Pr(A_F^+|M)}. \tag{32.10}$$

Furthermore, if the probability of ascertainment given the response vector is unaffected by the parameters in the current model, which it will be under many ascertainment schemes,

this simplifies to

$$\Pr(\mathbf{Y_F}|A_F^+, M) \propto \frac{\Pr(\mathbf{Y_F}|M)}{\Pr(A_F^+|M)}.$$

In words: *Up to a multiplicative constant, the likelihood component for family F conditioned for the fact that family F has been ascertained may be obtained as the likelihood assuming random ascertainment divided by the probability under the current model (for affection and for ascertainment given affection) that family F would have been ascertained.*

The ability to condition any likelihood for the probability of ascertainment has wide application and is an important feature of many analyses in genetic epidemiology. Of course, it is not always necessary. For example, non-random ascertainment is generally treated as being unimportant in linkage analysis provided ascertainment is based *either* on the trait of interest, *or* the marker, but not *both* (Hodge, 2002). On the other hand, there are other circumstances in which a failure to take proper account of non-random ascertainment can lead to an analysis that is so biased that conclusions are qualitatively incorrect. For example, Hodge (2002) describes an illustrative example consisting of a simple sibling-based design under complete ascertainment in which a true segregation ratio of 25 % (as expected for a rare recessive disease) is incorrectly estimated at 57.1 % (completely inconsistent with a recessive mode of inheritance) when the ascertainment scheme is ignored (Burton *et al.*, 2000; Kraft and Thomas, 2000). Failure to condition on the correct probability of ascertainment may also be due to reasons of computational infeasibility, rather than ignorance. A prime example is when classical bias and aetiological heterogeneity bias need to be accounted for in a complex variance component model of a binary response (Noh *et al.*, 2005). With this in mind, any method that can partially correct for ascertainment bias, that is also practically feasible should not be dismissed outright. Burton (Burton *et al.*, 2000; Burton, 2003) shows, e.g. that it is possible to remove bias due to classical ascertainment, without correcting for the effects of aetiological heterogeneity. However, after this partial correction, estimates for the prevalence of disease or the value of any variance component will pertain to the *ascertained sample* and not the *general population*. This method does not therefore provide a full correction, as noted by Epstein *et al.* (2002) and Glidden and Liang (2002), but it has been recently shown that parameter estimates for the ascertained population can be used to estimate general population parameters via a two-stage approach (Bowden *et al.*, 2006).

Under single ascertainment, adjustment for the sampling scheme can be based upon removal of the probands from all families (Khoury *et al.*, 1993b). However, this will generate inconsistencies under any ascertainment model in which the probability that an individual family contains a proband does *not* increase linearly with the number of affected members. Nevertheless, this is obviously related to the common practice of dividing the full likelihood without adjustment for ascertainment, by the likelihood (under the model) of the observed phenotype in each proband (Hopper, 2002b). Although the option to do this is built into a number of software applications, it is important that the true ascertainment mechanism is considered properly, and simple conditioning on the phenotype of the proband should only be undertaken if it makes sense in the given situation being considered.

In the real world ascertainment schemes are often poorly defined and may not involve unambiguous probands (Morton, 1959; Hodge, 2002). This means that although conditioning the likelihood for the probability that each family is ascertained remains

theoretically possible, it may, in practice, be impossible to express this in terms of the parameters in M and this may thwart adjustment. Instead, Ewens and Shute (Ewens and Shute, 1986) proposed an alternative 'resolution of the ascertainment sampling problem'. This involved 'conditioning the likelihood of the sample on that part of the data relevant to ascertainment'. For example, the number of affected children may represent one such component (Ewens and Shute, 1986). Provided one conditions for all such components of the data, it is '... clear that this method will lead to estimates of genetic parameters free of any specific mathematical ascertainment assumption' (Ewens and Shute, 1986). The problem is that conditioning in this manner can discard much of the information driving the analysis in the first place. It can therefore be inefficient (Ewens and Shute, 1986).

As in mainstream epidemiology, the importance of study sampling mechanisms to genetic epidemiology cannot be overstated (Fisher, 1934). Particularly for descriptive studies, it is essential that the ascertainment strategy is considered carefully during study design, and is then taken into proper account at the time of analysis. Where possible it is better for the design to include an explicitly stated sampling scheme (including any sequential rules for pedigree extension once a family has been ascertained (Hodge, 2002) as it is generally easier to correct for the impact of a precisely defined mechanism than for an informal process. It is also important to properly understand the scientific question being addressed and the mathematical model that is to be used, in order to ensure that the analysis is conditioned in such a way that inferences are valid and yet valuable information is not wasted.

We do not believe that there can ever be a single answer to the 'ascertainment problem' and it is likely that most genetic epidemiologists will continue to encounter circumstances in which ascertainment bias is severe (and hence cannot be ignored) and yet the true sampling model is obscure (and hence it is impossible to identify the ideal approach to adopt). As a practical solution in such circumstances, it is our belief that one has no option but to try a variety of tractable approaches to ascertainment adjustment that are consistent with what is known about the problem. This should potentially include adjusting for sampling schemes that are more extreme than single or complete (e.g. quadratic schemes under which the probability of ascertainment of a family increases with the square of the number of affected children (Ewens and Shute, 1986)). If these produce qualitatively different answers to one another, then there is true uncertainty and no firm conclusions can be drawn. On the other hand, if all approaches generate comparable results then one can at least feel more secure in taking action on the basis of those results. Nevertheless, one can never completely exclude the possibility that the true ascertainment mechanism was different, and that if one had been able to properly adjust for it, different conclusions would have resulted.

32.4   STUDIES INVESTIGATING SPECIFIC AETIOLOGICAL DETERMINANTS

To finish, we return briefly to the second class of studies that is fundamental to mainstream epidemiology: 'studies that attempt to investigate a potential aetiological link between one or more specific determinants and a disease of interest'. Most such studies are observational and are based upon cross-sectional, case–control or cohort designs, with analysis being based upon unconditional or conditional, binomial or Poisson likelihoods

(Breslow and Day, 1980; 1987; Clayton and Hills, 1993; Greenland and Rothman, 1998). That we have chosen not to focus our chapter on their equivalents in genetic epidemiology does not in any sense reflect a belief that such studies are unimportant. Indeed, in the wake of the human genome project, studies that investigate the relationships between a phenotype and genomic or proteomic biomarkers are fast becoming the dominant areas of research in genetic epidemiology. Our reason for choosing to focus elsewhere is because most of the key issues pertaining to such studies are already dealt with thoroughly in other chapters in this handbook. **Chapter 18** deals with association studies which investigate 'the detection and analysis of statistical association, at the population level, between a trait and a genotype'. Linkage analysis investigates 'the dependence in inheritance of genes at different genetic loci, on the basis of phenotype observations on individuals' (see **Chapter 33**), and model-free methods are addressed in **Chapter 34**. **Chapter 38** describes the classical transmission disequilibrium test. The TDT is a form of matched case pseudocontrol design, wherein the cases are alleles observed to have been transmitted to the offspring within familial trios consisting of two parents and an affected child, and the pseudocontrols are their unobserved counterparts that were not transmitted. One of the most useful characteristics of the TDT is that it permits the drawing of inferences pertaining to linkage disequilibrium from an, otherwise rather restrictive, family trio design in which an affected child has been genotyped and his/her parents have been genotyped but not necessarily phenotyped. Unfortunately, the TDT approach so rapidly gained widespread popularity that it is sometimes recommended and used unthinkingly and without proper regard to all of the cautionary notes originally described (Spielman *et al.*, 1993).

Contemporary methodological developments in genetic epidemiology are largely focusing upon ways to investigate specific aetiological determinants. Major areas of current methodological research and development include issues pertaining to: (1) the efficient and robust use of single nucleotide polymorphisms in both candidate regions and in genome-wide association studies (Balding, 2006); (2) haplotype analysis; (3) population admixture and stratification (Freedman *et al.*, 2004; Marchini *et al.*, 2004); (4) the use of data from genome-wide SNP genotyping arrays (Clayton *et al.*, 2005; de Bakker *et al.*, 2005; Barrett and Cardon, 2006), comparative genome hybridisation and precise genotyping technologies such as dynamic allele specific hybridisation (DASH) to infer copy number variation and ultimately to relate this to disease (Fredman *et al.*, 2004; Freeman *et al.*, 2006; Redon *et al.*, 2006) and (5) methods for genetic meta-analysis and cross-study synthesis. There is also increasing interest in the use of graphical models to provide a flexible and powerful analytic environment within which to develop the sophisticated models ultimately required to disentangle the complex aetiological architectures of current interest to international medical science. Much of the relevant theory relating to many of these developments is discussed in various places throughout this handbook.

32.5   THE FUTURE

There is a plethora of biological information pertaining to the genome and to the proteome that is following in the wake of the human genome project (Lander *et al.*, 2001; Venter *et al.*, 2001). There is also a growing recognition of the fundamental importance (and

serious resource implications) of ensuring that modern biomedical research is based upon studies that are not only large enough but also incorporate first-class phenotyping and a rigorous assessment of co-exposures and co-morbidities. As a result, there are a number of large national initiatives addressing epidemiology and genetics in several countries (Davey Smith, Ebrahim *et al.*, 2005) including the United Kingdom (BioBank (Vogel, 2002; Wellcome Trust and Medical Research Council, 2002)), Estonia (Frank, 2000) and Latvia (Abbott, 2001). Furthermore, many large pre-existing cohort studies such as the Framingham Study (Joost *et al.*, 2002) and the Nurses Health Study (Haiman *et al.*, 2002) in the United States, the Busselton Study in Australia (Palmer *et al.*, 2001), and european prospective investigation of cancer (EPIC) (Basham *et al.*, 2001), the avon longitudinal study of parents and children (ALSPAC) (Jones *et al.*, 2000; Golding *et al.*, 2001) and the 1958 National Birth Cohort in Britain (Jefferis *et al.*, 2002) all now have a significant focus on genetics. In addition, special family-based initiatives such as the deCODE project in Iceland (Gulcher *et al.*, 2001; Hakonarson *et al.*, 2002), and twin registries around the world (Hopper, 1995; Hansen, de Klerk *et al.* 2000; Strachan, 2002) will continue to produce important information. It is clear that those developing statistical methods in genetic epidemiology are unlikely to be underemployed. Nevertheless, it is crucial that genetic epidemiologists continue to work closely with mainstream epidemiologists and biostatisticians in order to ensure that we do not lose sight of the fact that we not only share methods but a fundamental scientific philosophy.

Acknowledgments

REFERENCES

Abbott, A. (2001). Hopes of biotech interest spur Latvian population genetics. *Nature* **412**, 468.

Armitage, P. and Berry, G. (2002). *Statistical Methods in Medical Research*. Blackwell Scientific, Oxford.

de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J. and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics* **37**(11), 1217–1223.

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Natutre Reviews Genetics* **7**(10), 781–791.

Barrett, J.C. and Cardon, L.R. (2006). Evaluating coverage of genome-wide association studies. *Nature Genetics* **38**(6), 659–662.

Basham, V.M., Pharoah, P.D., Healey, C.S., Luben, R.N., Day, N.E., Easton, D.F. Ponder, B.A.J. and Dunning, A.M. (2001). Polymorphisms in CYP1A1 and smoking: no association with breast cancer risk. *Carcinogenesis* **22**(11), 1797–1800.

Blangero, J. (2002). Segregation analysis, complex. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 696–708.

Bonney, G.E. (1984). On the statistical determination of major gene mechanisms in continuous human traits: regressive models. *American Journal of Medical Genetics* **18**, 731–749.

Bonney, G.E. (1986). Regressive logistic models for familial disease and other binary traits. *Biometrics* **42**(3), 611–625.

Bonney, G.E. (1987). Logistic regression for dependent binary observations. *Biometrics* **43**(4), 951–973.

Bonney, G.E. (2002). Regressive models. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 666–673.

Bowden, J.M., Thompson, J.R. and Burton, P.R. (2006). A two-step approach to ascertainment bias correction in complex genetic models with variance components. *Annals of Human Genetics* **71**, 220–229.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research. Volume 1 – The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

Breslow, N.E. and Day, N.E. (1987). *Statistical Methods in Cancer Research. Volume 2 – The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

Burton, P., Gurrin, L. and Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in Medicine* **17**(11), 1261–1291.

Burton, P.R. (2003). Correcting for non-random ascertainment in generalized linear mixed models (GLMMs) fitted using Gibbs sampling. *Genetic Epidemiology* **24**(1), 24–35.

Burton, P.R., Palmer, L.J., Jacobs, K., Keen, K.J., Olson, J.M. and Elston, R.C. (2000). Ascertainment adjustment: where does it take us? *American Journal of Human Genetics* **67**(6), 1505–1514. Erratum American Journal Human Genetics 69:672.

Burton, P.R., Tiller, K.J., Gurrin, L.C., Cookson, W.O., Musk, A.W. and Palmer, L.J. (1999). Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. *Genetic Epidemiology* **17**(2), 118–140.

Burton, P.R., Tobin, M.D. and Hopper, J.L. (2005). Key concepts in genetic epidemiology. *Lancet* **366**, 941–951.

Clayton, D. and Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.

Clayton, D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**(2), 467–485.

Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, L.M., Smink, L.J., Lam, A.C., Ovington, N.R., Stevens, H.E., Nutland, S., Howson, J.M., Faham, M., Moorhead, M., Jones, H.B., Falkowski, M., Hardenbol, P., Willis, T.D. and Todd, J.A. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics* **37**(11), 1243–1246.

Cohen, B.H. (1980). Chronic obstructive pulmonary disease: a challenge in genetic epidemiology. *American Journal of Epidemiology* **112**(2), 274–288.

Cordell, H.J. and Clayton, D.G. (2005). Genetic association studies. *Lancet* **366**(9491), 1121–1131.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B* **34**, 187–220.

Davey Smith, G., Ebrahim, S., Lewis, S., Hansell, A.L., Palmer, L.J. and Burton, P.R. (2005). Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* **366**(9495), 1484–1498.

Davie, A.M. (1979). The 'singles' method for segregation analysis under incomplete ascertainment. *Annals of Human Genetics* **42**(4), 507–512.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, New York.

Duffy, D.L., Martin, N.G., Battistutta, D., Hopper, J.L. and Mathews, J.D. (1990). Genetics of asthma and hay fever in Australian twins. *American Review of Respiratory Disease* **142**(6 Pt 1), 1351–1358.

Edwards, J.H. (1969). Familial predisposition in man. *British Medical Bulletin* **25**(1), 58–64.

Elston, R.C. (1981). Segregation analysis. *Advances in Human Genetics* **11**, 63–120, 372–373.

Elston, R.C. (2001). *S.A.G.E. Statistical Analysis for Genetic Epidemiology, S.A.G.E. 4.1.* Case Western Reserve University.

Elston, R.C., Olson, J.M. and Palmer, L.J. (2002). *Biostatistical Genetics and Genetic Epidemiology*. John Wiley & Sons, Chichester.

Elston, R.C. and Sobel, E. (1979). Sampling considerations in the gathering and analysis of pedigree data. *American Journal of Human Genetics* **31**(1), 62–69.

Epstein, M.P., Lin, X. and Boehnke, M. (2002). Ascertainment-adjusted parameter estimates revisited. *American Journal of Human Genetics* **70**(4), 886–895.

Ewens, W.J. and Shute, N.C. (1986). The limits of ascertainment. *Annals of Human Genetics* **50**(Pt 4), 399–402.

Falconer, D.S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics* **29**, 51–71.

Fisher, R. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399–433.

Fisher, R.A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics* **6**, 13–25.

Fisher, R.A. (1951). Limits to intensive production in animals. *British Agricultural Bulletin* **4**, 217–218.

Frank, L. (2000). Estonia prepares for national DNA database. *Science* **290**(5489), 31.

Fredman, D., White, S.J., Potter, S., Eichler, E.E., Den Dunnen, J.T. and Brookes, A.J. (2004). Complex SNP-related sequence variation in segmental genome duplications. *Nature Genetics* **36**(8), 861–866.

Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., Gabriel, S.B., Topol, E.J., Smoller, J.W., Pato, C.N., Pato, M.T., Petryshen, T.L., Kolonel, L.N., Lander, E.S., Sklar, P., Henderson, B., Hirschhorn, J.N. and Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics* **36**(4), 388–393.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., Carter, N.P., Scherer, S.W. and Lee, C. (2006). Copy number variation: new insights in genome diversity. *Genome Research* **16**(8), 949–961.

Galton, F. (1875). The history of twins, as criterion of the relative powers of nature and nurture. *Fraser's magazine* Nov, 566–576.

Galton, F. (1877). Typical laws of heredity. *Proceedings of the Royal Institution* **8**, 282–301.

Galton, F. (1886). Family likeness in stature. *Proceedings of the Royal Society* **40**, 42–73.

Gauderman, W.J. and Thomas, D.C. (1994). Censored survival models for genetic epidemiology: a Gibbs sampling approach. *Genet Epidemiol* **11**(2), 171–188.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Glidden, D.V. and Liang, K.Y. (2002). Ascertainment adjustment in complex diseases. *Genetic Epidemiology* **23**(3), 201–208.

Golding, J., Pembrey, M. and Jones, R. (2001). ALSPAC – the Avon longitudinal study of parents and children. I. Study methodology. *Paediatric and Perinatal Epidemiology* **15**(1), 74–87.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika* **73**(1), 43–56.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research*. Charles Griffin & Company, Ltd, London.

Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika* **78**(1), 45–51.

Greenland, S. and Rothman, K. (1998). Measures of effect and measures of association. *Modern epidemiology*, K. Rothman and G.S. Philadelphia, Lippincott-Raven, 47–64.

Gulcher, J., Kong, A. and Stefansson, K. (2001). The genealogic approach to human genetics of disease. *Cancer Journal* **7**(1), 61–68.

Guo, S.W. and Thompson, E.A. (1992). A Monte Carlo method for combined segregation and linkage analysis. *American Journal of Human Genetics* **51**(5), 1111–1126.

Haiman, C.A., Brown, M., Hankinson, S.E., Spiegelman, D., Colditz, G.A., Willett, W.C., Kantoff, P.W. and Hunter, D.J. (2002). The androgen receptor CAG repeat polymorphism and risk of breast cancer in the Nurses' Health Study. *Cancer Research* **62**(4), 1045–1049.

Hakonarson, H., Bjornsdottir, U.S., Halapi, E., Palsson, S., Adalsteinsdottir, E., Gislason, D., Finnbogason, G., Gislason, T., Kristjansson, K., Arnason, T., Birkisson, I., Frigge, M.L., Kong, A., Gulcher, J.R. and Stefansson, K. (2002). A major susceptibility gene for asthma maps to chromosome 14q24. *American Journal of Human Genetics* **71**(3), 483–491.

Hanfelt, J. (1993). *GEE4 Documentation, Department of Biostatistics*. Johns Hopkins University.

Hansen, J., de Klerk, N., Croft, M., Alessandri, P. and Burton, P. (2000). The Western Australian Twin Child Health (WATCH) study: work in progress. *Australian Epidemiologist* **7**(2), 16–20.

Hattersley, A.T. and McCarthy, M.I. (2005). What makes a good genetic association study? *Lancet* **366**(9493), 1315–1323.

Hennekens, C.H. and Buring, J.E. (1987). Descriptive studies. In *Epidemiology in medicine*, S.L. Mayrent, ed. Brown, Boston, Little, pp. 101–131.

Hodge, S.E. (2002). Ascertainment. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 20–28.

Hopper, J. (1993). Variance components for statistical genetics: applications in medical research to characteristics related to human diseases and health. *Statistical Methods in Medical Research* **2**, 199–223.

Hopper, J.L. (1992). The epidemiology of genetic epidemiology. *Acta Geneticae Medicae et Gemellologiae* **41**(4), 261–273.

Hopper, J.L. (1995). Australian NHMRC twin registry: a resource for pediatric research. *Pediatric Cardiology* **16**(2), 100.

Hopper, J.L. (2002a). Heritability. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 371–372.

Hopper, J.L. (2002b). Variance component analysis. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 778–788.

Hopper, J.L., Bishop, D.T. and Easton, D.F. (2005). Population-based family studies in genetic epidemiology. *Lancet* **366**(9494), 1397–1406.

Hopper, J.L. and Carlin, J.B. (1992). Familial aggregation of a disease consequent upon correlation between relatives in a risk factor measured on a continuous scale. *American Journal of Epidemiology* **136**, 1138–1147.

Hougaard, P. and Thomas, D. (2002). Frailty. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 279.

Huber, P.J. (1967). The behaviour of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 221–233.

Jefferis, B.J., Power, C. and Hertzman, C. (2002). Birth weight, childhood socioeconomic environment, and cognitive development in the 1958 British birth cohort study. *British Medical Journal* **325**(7359), 305.

Jinks, J.L. and Fulker, D.W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychological Bulletin* **73**(5), 311–349.

Jones, R.W., Ring, S., Tyfield, L., Hamvas, R., Simmons, H., Pembrey, M., and Golding, J. (2000). A new human genetic resource: a DNA bank established as part of the Avon longitudinal study of pregnancy and childhood (ALSPAC). *European Journal of Human Genetics* **8**(9), 653–660.

Joost, O., Wilk, J.B., Cupples, L.A., Harmon, M., Shearman, A.M., Baldwin, C.T., O'Connor, G.T., Myers, R.H. and Gottlieb, D.J. (2002). Genetic loci influencing lung function: a genome-wide scan in the Framingham study. *Americal Journal of Respiratory and Critical Care Medicine* **165**(6), 795–799.

Jöreskog, K.G. and Sörbom, D. (1996). *LISREL 8 User's Reference Guide*. Scientific Software International, Lincolnwood.

Khoury, M.J., Beaty, T.H. and Cohen, B.H. (1993a). Genetic approaches to familial aggregation: I. Analysis of heritability. *Fundamentals of Genetic Epidemiology* 200–232.

Khoury, M.J., Beaty, T.H. and Cohen, B.H. (1993b). Genetic approaches to familial aggregation: II. Segregation analysis. *Fundamentals of Genetic Epidemiology* 233–283.

King, M.C., Lee, G.M., Spinner, N.B., Thomson, G. and Wrensch, M.R. (1984). Genetic epidemiology. *Annual Review of Public Health* **5**, 1–52.

Kopciuk, K.A. and Bull, S.B. (2002). Risk Ratios. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 687–691.

Kraft, P. and Thomas, D.C. (2000). Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *American Journal of Human Genetics* **66**(3), 1119–1131.

Kuehni, C.E., Brooke, A.M. and Silverman, M. (2000). Prevalence of wheeze during childhood: retrospective and prospective assessment. *European Respiratory Journal* **16**(1), 81–85.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., N. Stange-Thomann, Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., L. Doucette-Stamm, Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., M. de la Bastide, Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T.,

Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., D. Thierry-Mieg, J. Thierry-Mieg, Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., Szustakowki, J., P. de Jong, Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S. and Chen, Y.J. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**(6822), 860–921.

Lange, K., Weeks, D. and Boehnke, M. (1988). Programs for pedigree analysis: MENDEL, FISHER, and dGENE. *Genetic Epidemiology* **5**(6), 471–472.

Last, J. (2001). *A Dictionary of Epidemiology*. Oxford University Press, New York.

Li, C.C. and Mantel, N. (1968). A simple method of estimating the segregation ratio under complete ascertainment. *American Journal of Human Genetics* **20**(1), 61–81.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**(1), 13–22.

Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E. Skytthe, A. and Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine* **343**(2), 78–85.

Luyt, D.K., Burton, P.R. and Simpson, H. (1993). Epidemiological study of wheeze, doctor diagnosed asthma, and cough in preschool children in Leicestershire. *British Medical Journal* **306**(6889), 1386–1390.

Majumder, P.P. (2002). Heritability. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 693–696.

Marchini, J., Cardon, L.R., Phillips, M.S. and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics* **36**(5), 512–517.

MathSoft, I. (2000). *Splus*. MathSoft, Inc., Cambridge, MA.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Merriman, C. (1924). The intellectual resemblance of twins. *Psychological Monographs* **33**, 1–58.

Miller, R.G., Gong, G. and Munoz, A. (1981). *Survival analysis*. John Wiley & Sons, New York.

Morton, N.E. (1959). Genetic tests under incomplete ascertainment. *American journal of Human Genetics* **11**, 1–16.

Morton, N.E. (1982). *Outline of Genetic Epidemiology*. Karger, London.

Morton, N.E. and Chung, C.S. (1978). *Genetic epidemiology*. Academic Press, New York.

Morton, N.E. and MacLean, C.J. (1974). Analysis of family resemblance. 3. Complex segregation of quantitative traits. *American Journal of Human Genetics* **26**(4), 489–503.

Neale, M.C. (2002). Twin analysis. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 743–756.

Neale, M.C., Boker, S.M., Xie, G. and Maes, H.H. (2002). *Mx: Statistical Modeling (6th Edition)*. Department of Human Genetics, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, U.S.A.

Neale, M.C. and Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*. Kluwer, Boston.

Neel, J.V. and Schull, W.J. (1954). *Human Heredity*. The University of Chicago Press, Chicago.

Neuhaus, J. (1992). Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research* **1**, 249–273.

Noh, M., Lee, Y. and Pawitan, Y. (2005). Robust ascertainment-adjusted parameter estimation. *Genetic Epidemiology* **29**(1), 68–75.

Palmer, L.J. and Cardon, L.R. (2005). Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* **366**(9492), 1223–1234.

Palmer, L.J., Cookson, W.O., James, A.L., Musk, A.W. and Burton, P.R. (2001). Gibbs sampling-based segregation analysis of asthma-associated quantitative traits in a population-based sample of nuclear families. *Genetic Epidemiology* **20**(3), 356–372.

Palmer, L.J., Knuiman, M.W., Divitini, M.L., Burton, P.R., James, A.L., Bartholomew, H.C., Ryan, G. and Musk, A.W. (2001). Familial aggregation and heritability of adult lung function: results from the Busselton health study. *European Respiratory Journal* **17**(4), 696–702.

Palmer, L.J., Rye, P.J., Gibson, N.A., Burton, P.R., Landau, L.I. and Lesouef, P.N. (2001). Airway responsiveness in early infancy predicts asthma, lung function, and respiratory symptoms by school age. *American Journal of Respiratory Cell and Molecular Biology* **163**(1), 37–42.

Pearson, K. (1896). Mathematical contributions to the theory of evolution – IIRegression, I., heredity, and panmixia. *Philosophical Transactions of the Royal Society, Series A* **187**, 253–318.

Pendergast, P.F., Gange, S.J. and Lindstrom, M.J. (2002). Correlated binary data. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 159–171.

Pickles, A. (2002). Generalized estimating equations. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 299–310.

Prentice, R.L. and Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**(3), 825–839.

Rao, D.C. and Rice, T. (2002). Path analysis. In *Biostatistical Genetics and Genetic Epidemiology.*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 606–619.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Draper, D., Healy, M. and Woodhouse, G. (1999). *A User's Guide to MLwiN*. Institute of Education, London.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., Gonzalez, J.R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J.R., Marshall, C.R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M.J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Armengol, L., Conrad, D.F., Estivill, X., C. Tyler-Smith, Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W. and Hurles, M.E. (2006). Global variation in copy number in the human genome. *Nature* **444**(7118), 444–454.

Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *American Journal of Human Genetics* **46**(2), 229–241.

Roberts, D.F. (1985). A definition of genetic epidemiology. In *Diseases of Complex Etiology in Small Populations: Ethnic Differences and Research Approaches*, R. Chakraborty and E.J.E. Szathmary, eds. Alan R Liss, New York, pp. 9–20.

Rothman, K. and Greenland, S. (1998). Accuracy considerations in study design. In *Modern epidemiology*, K. Rothman and S. Greenland, eds. Lippincott-Raven, Philadelphia, pp. 135–145.

SAS Institute Inc. (2001). *SAS Release 8.02.*, SAS Institute Inc, Cary, NC.

Scurrah, K.J., Palmer, L.J. and Burton, P.R. (2000). Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (GLMMs) and Gibbs sampling in BUGS. *Genetic Epidemiology* **19**(2), 127–148.

Siemans, H.W. (1924). *Zwillingspathologie: Ihre Bedeutung; Ihre Methodik, Ihre Bisherigen Ergebnisse.* Springer Verlag, Berlin.

Sobel, E. and Lange, K. (1993). Metropolis sampling in pedigree analysis. *Statistical Methods in Medical Research* **2**(3), 263–282.

Spector, N. (2000). Cancer, genes, and the environment. *New England Journal of Medicine* **343**(20), 1494. Discussion 1495-6.

Spector, T.D., Sneider, H. and MacGregor, A.J. (1999). *Advances in twin and sib pair analysis*. Greenwich Medical Media Ltd, London.

Spiegelhalter, D., Thomas, A. and Best, N. (2000). *WinBUGS Version 1.3 – User Manual*. MRC Biostatistics Unit, Cambridge.

Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**(3), 506–516.

SPSS Inc. (2002). *SPSS 11.0 Guide to Data Analysis*. SPSS Inc Headquarters, Chicago.

Stata Corporation. (2005). *Stata User's Guide. Version 9.0*. Stata Press, College Station, Texas.

Strachan, D.P. (2002). Twin registers. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 759–761.

Teare, M.D. and Barrett, J.H. (2005). Genetic linkage studies. *Lancet* **366**(9490), 1036–1044.

Thompson, E.A. (2002). Human Genetics, Overview. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J.M. Olson and L.J. Palmer, eds. Wiley, Chichester, pp. 386–390.

Todorov, A.A. and Suarez, B.K. (2002). Liability model. In *Biostatistical Genetics and Genetic Epidemiology*, R. Elston, J. Olson and L. Palmer, eds. John Wiley & Sons, Chichester, pp. 430–435.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., G. L. Gabor Miklos, Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., J. Abu-Threideh, Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., V. Di Francesco, Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., E. Winn-Deen, Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., J. Carnes-Stine, Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. (2001). The sequence of the human genome. *Science* **291**(5507), 1304–1351.

Vogel, G. (2002). Population studies. U.K.'s mass appeal for disease insights. *Science* **296**(5569), 824.

Weinberg, W. (1912). Weitere Beiträge zur Theorie der Vererbung. 5. Zur Vererbung der Anlage zur Blutenkranheit mit methodologischen Ergänzungen meiner Geschwistermethode. *Archiv für Rassen- und Gesellschafts-Biologie* **2**, 694–709.

Weinberg, W. (1928). Mathematische grundlagen der probandenmethode. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre* **48**, 179–228.

Wellcome Trust and Medical Research Council (2002). *Draft Protocol for BioBank UK: A study of genes, environment and health*. Medical Research Council, London.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20**, 557–585.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association* **86**(413), 79–86.

Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**(1), 121–130.

Zeger, S.L. and Liang, K.Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine* **11**(14–15), 1825–1839.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**(4), 1049–1060.

33

Linkage Analysis

E.A. Thompson

Department of Statistics, University of Washington, Seattle, WA, USA

Linkage analysis is the analysis of the dependence in inheritance of genes at different genetic loci, on the basis of phenotypic observations on individuals. The methods of linkage analysis that evolved over the years 1920 to 1970 closely followed more general developments in approaches to statistical inference. In recent years, the wealth of DNA markers and the near completion of the human genome project have led to changes in approaches to linkage analysis. New computational methods have been developed to make optimal use of available multilocus marker data. With the mapping of most simple Mendelian traits and the consequent emphasis on mapping the genes contributing to more complex traits, methods involving analysis of genome identity by descnt are increasingly used. In view of the limited numbers of meioses often available for inference, fine-scale mapping and estimation of accurate meiotic maps remain open challenges. As data on genetic variation becomes available at more detailed levels, sources of variation in the processes underlying meiosis can be investigated and more accurate genetic maps obtained. Genomic data at high resolution can allow models for meiosis outcomes to be tested and more accurate models developed. These may then be used in the mapping of trait genes.

33.1   INTRODUCTION

The last century has seen enormous change in methods of inference from genetic data, from the rediscovery of Mendel's laws in 1900, to near completion of Phase I of the Human Genome Project 100 years later. However, the scientific questions remain surprisingly constant: Where are the genes? What do they do? Genetics is the science of heritable variation, meiosis is the biological process whereby genetic information is transmitted from parent to offspring, and linkage analysis is inference concerning the outcomes of meioses from data on the genetic characteristics of individuals.

For the purposes of this chapter, linkage analysis is defined as the analysis of the dependence in inheritance of genes at different genetic loci, as evidenced in phenotypic observations on individuals in known pedigree relationships. A pedigree is a specification of the father and mother of each (nonfounder) individual. Individuals whose parents are

unspecified are *founders*: relationships are defined only relative to the specified pedigree. Dependence in inheritance of genes at different genetic loci reflects synteny of the loci, and the degree of dependence provides a measure of the genetic distance between them. When a genetic marker map is available, dependence between the inheritances of the trait and that at certain marker loci enables the locations of trait loci to be inferred.

Of the biological sciences, genetics is the one with the most clearly defined probability models, and hence the one in which classical parametric statistical inference has had the biggest role. Moreover, linkage analysis is the area of genetics which is most dependent on probability models and parametric inference. From the work of Fisher (1922b) onward, each major development in parametric statistical inference has been adopted by the developers of linkage analysis methods, and questions of genetic analysis have prompted new statistical developments. In many ways, statistical inference and genetic analysis have developed in parallel over the last 100 years.

33.2   THE EARLY YEARS

Modern genetics started with Mendel (1866), who postulated his two laws as a probability model. In modern terminology, they can be stated as follows:

- Every diploid organism has two factors [or *genes*] controlling a given trait, one from the mother, one from the father.
- When a (diploid) individual has an offspring, a copy of a randomly chosen one of his two genes is transmitted [or *segregates*] to the offspring.
- This transmission occurs independently of the other parent, independently for each child, and independently for each trait [or *locus*].

Mendel's work was rediscovered in 1900 and not long thereafter geneticists realized that independence of segregations for different traits is not always true. Instead, there are groups of traits, which are *linked*, the genes controlling them tending to be inherited by the offspring as a group, not independently. That is, the same combinations of alleles transmitted from grandparent to parent are transmitted to an offspring. The basic framework for linkage analysis was quickly established. The genetic material that underlies inherited characteristics consists of the chromosomes: linear structures of DNA in the form of a double helix contained within each cell nucleus. In a diploid organism, chromosomes come in pairs, one deriving from the genetic material of the mother and the other from the father. In *meiosis*, the process of formation of a gamete (sperm or egg cell), a parent provides one chromosome from each chromosome pair. In the formation of this chromosome, several *crossover* events may occur between the two parental chromosomes, such that the transmitted chromosome consists of alternating segments of the two parental chromosomes. Simple genetic characteristics are determined by the two segments of DNA sequence (*alleles*) at a specific location (*locus*) on a pair of chromosomes. Geneticists soon associated this dependence in inheritance (*linkage*) with the chromosomes. Sturtevant (1913) showed the patterns of coinheritance of genes for different sex-linked traits in *Drosophila* were best explained by a linear arrangement of genes and made the first *gene ordering* inference using methods analogous to those still used today.

Mathematically, we define

$$S_{i,j} = 0 \text{ or } 1 \tag{33.1}$$

to denote that in meiosis i at locus j the maternal or paternal gene (respectively) of the parent is transmitted to the offspring. For ease of notation we define the two sets of vectors

$$S_{\bullet,j} = (S_{i,j}, i = 1, ..., m) \text{ for locus } j, j = 1, \dots, l$$
$$S_{i,\bullet} = (S_{i,j}, j = 1, ..., l) \text{ at meiosis } i, i = 1, \dots, m. \tag{33.2}$$

These *meiosis indicators* or *switches* (Donnelly, 1983) specify the descent of genes in pedigrees. In experimental organisms, experiments can be designed such that the grandparental origin of each offspring allele is clear, and often large numbers of offspring can be observed. For two distinct locations on a chromosome, the *recombination frequency* θ is the probability that the alleles derive from different parental chromosomes; that is, have different grandparental origins. Between two loci, j and j', the recombination frequency is

$$\theta_{jl} = \Pr(S_{i,j} \neq S_{i,j'}).$$

The recombination frequency is often assumed not to depend on the specific meiosis i, or to depend only on the sex of the parent in which meiosis i occurs. In reality, many factors may affect the frequency of recombination.

For recombination to occur, there must be an odd number of crossovers between the two loci in the formation of the offspring gamete. For loci that are very close together on the chromosome, θ is close to 0, and alleles at the two loci will show strong dependence in their grandparental origins. Under most meiosis models, and apparently in nature, the value of θ increases with increasing length of chromosome intervening between the two loci, until for loci that are far apart on a chromosome, or are on different pairs of chromosomes, $\theta = \frac{1}{2}$ and the grandparental origins of alleles at the two loci are independent. Loci for which $\theta < \frac{1}{2}$ are said to be *linked*. Linkage analysis is the statistical analysis of genetic data with the goals of detecting whether $\theta < \frac{1}{2}$, of estimating θ, of ordering a set of genetic loci, and ultimately of placing the loci determining genetic traits of interest at correct locations in a genetic map.

An understanding of the population–genetic issues was also fundamental to the early development of statistical genetics. In 1908, G. Hardy and W. Weinberg independently established the idea of *Hardy–Weinberg equilibrium* showing that Mendelian segregation provided for the maintenance of stable genetic variation within a population. In the absence of selection, in an infinite population, allele frequencies remain constant. Moreover, for loci on the 22 pairs of autosomes, constant 'equilibrium' genotype frequencies are established by a single generation of random mating (see Hartl and Clark, 1997). For joint frequencies at two loci, this is not so. Denoting alleles at two diallelic loci A and B by a_i and b_i ($i = 1, 2$), there are four *haploypes*: a_1b_1, a_1b_2, a_2b_1, and a_2b_2. If the population frequencies of these are $p_k, k = 1, \dots, 4$ the standard measure of *linkage disequilibrium* or *allelic association* is

$$D = \Pr(a_1b_1) - \Pr(a_1)\,\Pr(b_1)$$

$$= p_1 - (p_1 + p_2)(p_1 + p_3)$$

$$= p_1 p_4 - p_2 p_3,$$

since $p_1 + p_2 + p_3 + p_4 = 1$. Robbins (1918) showed that, after one generation of random mating, in an infinite population, the allelic association is decreased to

$$D^* = D(1 - \theta). \tag{33.3}$$

Thus equilibrium haplotype frequencies are not obtained in a single generation even for unlinked loci ($\theta = \frac{1}{2}$). If linkage is tight ($\theta \approx 0$), allelic associations persist over many generations.

Haldane (1919) defined genetic map distance between two loci as the expected number of crossover events between them; this distance measure is additive regardless of the meiosis model. One Morgan is the length of chromosome in which one crossover event is expected; map distances are normally given in centiMorgans (1 cM = 0.01 M). He also related this additive *map distance* to recombination frequencies between loci on the assumption that crossovers occur as a homogeneous Poisson process. (This is the model of 'no genetic interference'; a given crossover event does not affect the probability distribution of the number and locations of other crossover events.) This sparked an exploration that still continues, on the relationship between physical distance (actual DNA length in base pairs (bp)) and map distance (the statistical measure based on the degree of dependence between genetic loci). There is no simple relationship; in terms of physical distance recombination rates vary widely over the genome. The early statistical geneticists made clear that only recombination frequencies, and hence genetic map distances, could be estimated from segregation data alone.

Fisher (1922b) used a different relationship between recombination frequencies and map distance, assuming one crossover precluded any other crossover in the region of chromosome he considered. Under this model of 'complete interference', recombination frequencies are themselves additive. With this simplifying assumption, he used the multinomial distribution of recombination counts in offspring as one of his first examples of maximum likelihood estimation, citing, and drawing on his theoretical work published in the same year Fisher (1922a). Thus, by 1925, questions of gene ordering, genetic maps, linkage disequilibrium, and the relationship between genetic and physical maps had all been raised.

33.3   THE DEVELOPMENT OF HUMAN GENETIC LINKAGE ANALYSIS

In the 1930s, three British mathematicians, statisticians, and geneticists laid the foundations of most of modern linkage analysis, and did so through the use of statistical models. Haldane (1934) and Fisher (1934) recognized that the same methods and ideas that were being increasingly applied to designed crosses of experimental organisms could be applied also to observed patterns of inheritance in human families. At the same time, Penrose (1935) pointed out that, with a sufficiently large sample, linkage could be inferred from associated sharing of traits in sib pairs (see **Chapter 34**). Haldane (1934) considered use of likelihood to detect linkage, while Fisher (1934) addressed the estimation

problem more generally, applying ideas of information and efficiency. Just as the work of the early years raised many still-topical map issues, the work of the 1930s raised some human trait issues. Trait heterogeneity in linkage detection was first addressed by Fisher (1936). While Fisher (1934) focused on the statistical issues of information and power to detect linkage, he also foresaw both the potential and the problems of using linked genetic markers in genetic counseling.

Statistically, the problem is one of missing data or latent variables. In human families, individuals may be unavailable for typing, the traits that are observed may have no direct correspondence with the underlying alleles, and even where they do the sharing of common alleles may make grandparental origins of alleles unclear. Since Fisher (1922b) and Haldane (1934), the usual approach to linkage analysis is through computation of likelihoods. Evaluation of the likelihood involves summing over all the latent events that could have led to the observed data. In the early approaches, the genotypes of individuals are considered the latent variables:

$$\Pr(\mathbf{Y}) = \sum_{\mathbf{G}} \Pr(\mathbf{Y}|\mathbf{G}) \, \Pr(\mathbf{G}) = \sum_{\mathbf{G}} \left(\prod_{\text{observed } i} \Pr(Y_{i,\bullet}|G_{i,\bullet}) \right) \Pr(\mathbf{G}), \qquad (33.4)$$

where

$$\Pr(\mathbf{G}) = \prod_{\text{founders } i} \Pr(G_{i,\bullet}) \prod_{\text{nonfounders } i} \Pr(G_{i,\bullet}|G_{M_i,\bullet}, G_{F_i,\bullet}). \qquad (33.5)$$

Here, and throughout, $\mathbf{Y}$ denotes phenotypic data at trait and marker loci: $Y_{i,\bullet}$ denotes all data on individual i. The bullet subscript denotes that multiple genetic loci may be involved. For ease of presentation, we assume that phenotypes $Y_{i,\bullet}$ are conditionally independent given the multilocus genotype $G_{i,\bullet}$ of individuals i. The population allele frequencies and allelic associations determine the probabilities $\Pr(G_{i,\bullet})$ of founder haplotypes, while the laws of Mendelian segregation and linkage relationships among the loci determine the transmission probabilities $\Pr(G_{i,\bullet}|G_{M_i,\bullet}, G_{F_i,\bullet})$ for the genotypes of nonfounder individuals i conditional on those of the parents M_i and F_i of i. Genotypes of individuals are conditionally independent given those of their parents. In the context of linkage analysis, recombination frequencies θ are the unknown parameters of the transmission probabilities, and $\Pr(\mathbf{Y})$ is the likelihood $L(\theta)$ for θ.

The next big advances came again with application to linkage analysis of new statistical approaches. Haldane and Smith (1947) developed the uses of the likelihood ratio and maximum likelihood estimation, including the computation (by hand) of likelihoods on extended pedigrees, while Smith (1953) investigated the properties of these likelihood methods and the information for linkage in various pedigree designs. Smith (1953) also introduced the *lod score* of Barnard (1949) to linkage analysis. The likelihood ratio for linkage at a given recombination frequency θ relative to that for unlinked loci $\theta = \frac{1}{2}$ is $L(\theta)/L\left(\frac{1}{2}\right)$. The *lod* score for linkage between two loci is

$$\max_{0 \leq \theta \leq \frac{1}{2}} \log \left(L(\theta)/L\left(\frac{1}{2}\right) \right).$$

Haldane and Smith (1947) also proposed use of a Bayesian prior probability for synteny of two randomly chosen loci, and for the locations of linked loci. The Bayesian approach was developed much further by Renwick (1969).

Statistical developments in hypothesis testing also influenced linkage analysis, continuing the distinction between testing and estimation first emphasized by Fisher (1934). Using logarithms to base 10, the *lod* score bounds for declaration of linkage (+3) and exclusion of linkage (−2) proposed by Morton (1955) derived from a consideration of type 1 and type II error following a sequential probability ratio test (SPRT) of Wald (1947). This base-10 *lod* score has become the standard criterion for assessing evidence of linkage. The use of base-10 logarithms also became standard, and is deeply embedded in current methodology in this area, although in some recent multipoint linkage approaches natural logarithms of likelihoods are also used (see *location score* below).

In the framework of testing, Smith (1963) formalized the ideas of Fisher (1936) to develop likelihood-based tests for linkage heterogeneity. Edwards (1971) also followed the likelihood and testing approach to linkage detection, referring to absence of linkage as 'the only true null hypothesis in biology'. He also considered questions of the amount of information in a set of observations, converting it to an 'equivalent number' of informative meioses on the basis of the curvature of the likelihood in the neighborhood of the maximum – in effect using the usual statistical measure of (Fisher) information.

33.4   THE PEDIGREE YEARS; SEGREGATION AND LINKAGE ANALYSIS

The issues facing linkage analysis in 1970 were primarily computational and statistical. Data had been collected on pedigrees segregating many Mendelian traits, some exhibiting more complex patterns of inheritance, and some quantitative traits. Markers remained few and mostly unmapped, but linkage analysis was increasingly attempted as digital computers began to provide the necessary computational power. Around 1970, Hilden (1970), Elston and Stewart (1971), Heuch and Li (1972) laid the basis for what would remain the primary approach to segregation and linkage analysis computations for the next 20 years. Using very similar ideas, they independently produced algorithms and programs for the computation of likelihoods on extended pedigrees. Very soon thereafter, Ott (1974) produced the linkage analysis program LIPED which is still in use today. This program used the Elston–Stewart algorithm to compute probabilities of data at two loci jointly, and hence likelihoods for linkage. Over the next several years, computers improved and computational algorithms were generalized to more complex pedigrees and more complex trait models (Cannings *et al.*, 1978; 1980).

Although more genetic markers that could be typed in human individuals were gradually accumulated, these methodological and computational advances would have meant little without the advent of DNA markers (Botstein *et al.*, 1980). Since 1980, new technology has made available a wealth of new types of genetic markers, on the basis of characteristics of the DNA sequence itself rather than on proteins and enzymes determined by the DNA. These markers rely primarily on length variations in DNA, either in terms of numbers of copies of a short repeat sequence, or lengths between occurrences of a given short motif. These genetic markers must first be mapped relative to each other. Then they can be used to localize the genes contributing to traits of interest. Now, there are thousands of such markers available, and the Human Genome Project goal of a marker map at 1 cM density has been achieved (Murray *et al.*, 1994). However, even a genetic length of

1 cM is approximately 1 million base pairs (bp) of DNA. More recent maps contain even more markers, including single-nucleotide-polymorphisms (SNPs). These may occur as frequently as 1 per 500 bp.

With the discovery of the first DNA markers, the restriction fragment-length polymorphisms (RFLPs), suddenly there was the prospect of markers throughout the genome. Instead of linkage groups, there would be genetic maps on which the loci contributing to traits could be placed. Questions of study design and map density arose: Who should be typed? For what markers? For many simple Mendelian traits, trait data on pedigrees either already existed or could be relatively easily collected. The marker typing was the expensive component. Again from the statistics arena, the *elod* entered the literature on genetic linkage (Thompson *et al.*, 1978). This is simply the expected log-likelihood difference or base-10 version of the Kullback–Leibler information (Kullback and Leibler, 1951)

$$elod(\theta) = \mathrm{E}_\theta \left(\log_{10} L(\theta) - \log_{10} L\left(\frac{1}{2}\right) \right),$$

where the expectation is over the probability distribution of trait and marker data under a recombination frequency θ between the trait and marker loci. The *elod* became the key measure of the information to be expected from data on a set of pedigrees, given a trait model. Since trait data were often already available, a more relevant measure of expected information for linkage is the *elod* conditional on these specific trait data. Ploughman and Boehnke (1989) solved the problem of providing these conditional expected *lod* scores in their program SIMLINK which became a mainstay for those involved in practical linkage analyses.

Another consequence of the existence of genetic marker maps, was the development of map-specific multipoint linkage analyses (Lathrop *et al.*, 1984). That is, the marker map is assumed known and a log-likelihood difference (or *location score*) is computed for each hypothesized location of the trait locus relative to the hypothesis that the trait locus is not linked to this segment of the marker map. Use of multiple marker loci increases the power of the analysis, combining information from markers that are informative in different meioses of the pedigree. Even *interval mapping*, in which a locus is mapped using data on two hypothesized flanking markers, provides much more information than mapping with each marker locus separately. These principles are embodied in programs such as VITESSE (O'Connell and Weeks, 1995) and FASTLINK (Cottingham *et al.*, 1993) which remain the standards for exact computation of lod scores on extended pedigrees, using a small number of markers.

With the use of data at larger numbers of marker loci, it becomes simpler to consider the *meiosis indicators* of (33.1) as the latent variables, rather than genotypes **G**. Then (33.4) becomes

$$\mathrm{Pr}(\mathbf{Y}) = \sum_{\mathbf{S}} \mathrm{Pr}(\mathbf{Y}|\mathbf{S})\mathrm{Pr}(\mathbf{S}),$$

where **S** is the total set of meiosis indicators $\{S_{i,j}; i = 1, \ldots, m, \ j = 1, \ldots, l\}$ for the m meioses of the pedigree and each of the l loci. For ease of presentation, we assume that the data are specific to known or putative genetic loci, and for convenience we suppose these to be ordered $1, 2, \ldots, l$ along a chromosome (or chromosomes). Then the data $Y_{\bullet,j}$, attributable to genotypes at locus j, depends only on the descent of genes at locus j. The

analogue of (33.4) and (33.5) is then

$$\Pr(\mathbf{Y}) = \sum_{\mathbf{S}} \left(\prod_{j=1}^{l} \Pr(Y_{\bullet,j}|S_{\bullet,j}) \right) \left(\prod_{i=1}^{m} \Pr(S_{i,\bullet}) \right), \tag{33.6}$$

since different meioses i are *a priori* independent. Although summation over the space of all $\mathbf{S}$ values is required, this is a smaller space than the space of all multilocus genotype configurations $\mathbf{G}$ on a pedigree, and has a simpler dependence structure. Moreover, probabilities $\Pr(\mathbf{S})$ depend only on the the pedigree structure and on the meiosis (chromosome) structure. Recombination parameters enter only into this term, while all other parameters relating to the genetic model for trait and marker loci affect only $\Pr(\mathbf{Y}|\mathbf{S})$.

In considering a location score curve for linkage of a particular trait, it is convenient to partition the data into the marker data $\mathbf{Y}_M$ and trait data Y_T, and likewise partition the meiosis indicators into those at the marker loci, $\mathbf{S}_M$, and those at the putative trait locus S_T. Often the genetic model for the marker phenotypes $\mathbf{Y}_M$ is assumed known. That is, the marker map Γ_M determining $\Pr(\mathbf{S}_M)$ and marker allele frequencies determining $\Pr(\mathbf{Y}_M|\mathbf{S}_M)$ are fixed. Also parameters that are often assumed to be known are the parameters β relating trait genes and trait phenotypes: $\Pr(Y_T|S_T)$ is a function of β. The only parameter which is varied is the location γ of the trait locus, with $\gamma = \infty$ corresponding to absence of linkage. Log likelihoods are plotted as a function of γ. The location score curve is thus twice the (natural) log of $L(\gamma)/L(\infty)$, where

$$L(\gamma) = \Pr(Y_T, \mathbf{Y}_M; \Gamma_M, \beta, \gamma) = \sum_{\mathbf{S}_M, S_T} \Pr(Y_T, \mathbf{Y}_M, \mathbf{S}_M, S_T; \Gamma_M, \beta, \gamma) \tag{33.7}$$

and

$$\Pr(Y_T, \mathbf{Y}_M, \mathbf{S}_M, S_T; \Gamma_M, \beta, \gamma) = \Pr(\mathbf{Y}_M|\mathbf{S}_M)\Pr(\mathbf{S}_M; \Gamma_M) \\ \times \Pr(Y_T|S_T; \beta)\Pr(S_T|\mathbf{S}_M; \gamma). \tag{33.8}$$

On the one hand, analysis is simplified by the fact that γ enters only into the final term $\Pr(S_T|\mathbf{S}_M; \gamma)$. On the other hand, analysis is complicated by the need to compute across the whole range of positions γ, and by the fact that there is often strong evidence that the trait locus is not coincident with any marker locus, leading to 0 or infinitesimal values of $L(\gamma)$ at some marker locations.

The development of genetic maps also initiated the era of genome scans, in which a test can be made for linkage of a trait with each available marker. The interpretation of *lod* scores under these multiple dependent tests is not straightforward. Even when the trait gene is in some map interval, data on a sufficient number of informative meioses will provide strong evidence for recombination between the trait locus and an adjacent marker. Thus, the log-likelihood curve for the location of the trait locus will normally have multiple peaks with sharp decreases at the marker locations.

The 1980s were the golden years of traditional linkage analysis. By 1990 simple Mendelian traits had been localized by linkage analysis, except for those so rare that too few data are available. Methods for fine-scale mapping were developed; many genes

for simple traits were identified. Ott (1999) covers many of the developments of the last two sections in his text. He also says:

> Linkage analysis is sometimes perceived as a matter of simply using the proper computer program, so that anyone with sufficient computer expertise could 'do the linkage analysis' after family data and marker typing have been obtained.... It is dangerous to have linkage analyses carried out by individuals without the necessary theoretical background.

A full understanding of linkage likelihoods, and particularly of location score curves, requires that many complex statistical questions be addressed.

33.5   LIKELIHOOD AND LOCATION SCORE COMPUTATION

The basis of the Elston–Stewart algorithm, and related algorithms, is to sum latent variables sequentially over a pedigree but jointly over loci. Equation (33.4) leads to computational algorithms that are linear in pedigree size, but exponential in pedigree complexity, as measured by the number of interconnected pedigree loops. Moreover, algorithms are linear in the number of potential multilocus genotype triples of a father–mother–child trio, and hence exponential in the number of genetic loci. Although the efficiency of computer algorithms for exact likelihood evaluation have been much improved (Cottingham *et al.*, 1993), and various approximate methods have been proposed (Curtis and Gurling, 1993), exact computation is intrinsically limited in the number of loci that can be considered. With multiple linked multiallelic loci, the approach rapidly becomes computationally infeasible.

A new approach to likelihood computation on pedigrees was provided by Lander and Green (1987). Their algorithm was based on using $S_{i,j}$ (33.1) as the latent variables (33.6). They called $S_{\bullet,j}$ the *inheritance vector* at locus j. Instead of proceeding sequentially over the pedigree and jointly over loci, the computation proceeds sequentially along a chromosome, jointly over all meioses. Under the assumption of no genetic interference, the random vectors $S_{\bullet,j}$ have a first-order Markov property, and the likelihood (33.6) may be rewritten as

$$\Pr(\mathbf{Y}) = \sum_{\mathbf{S}} \Pr(\mathbf{S}, \mathbf{Y}) = \sum_{\mathbf{S}} \Pr(\mathbf{Y}|\mathbf{S})\Pr(\mathbf{S})$$

$$= \sum_{\mathbf{S}} \left(\Pr(S_{\bullet,1}) \prod_{j=2}^{l} \Pr(S_{\bullet,j}|S_{\bullet,j-1}) \prod_{j=1}^{l} \Pr(Y_{\bullet,j}|S_{\bullet,j}) \right). \qquad (33.9)$$

Under this Markov assumption, the framework is that of a hidden Markov model (HMM) and the Baum algorithm provides a computational method which is linear in the number of loci (Baum, 1972), but exponential in pedigree size. Thus, the method allows computation of probabilities of marker phenotypes for any number of loci, but on small pedigrees.

The dependence structure of data on a pedigree is shown in Figure 33.1. This shows the conditional independencies which underlie both the Elston–Stewart and Lander–Green
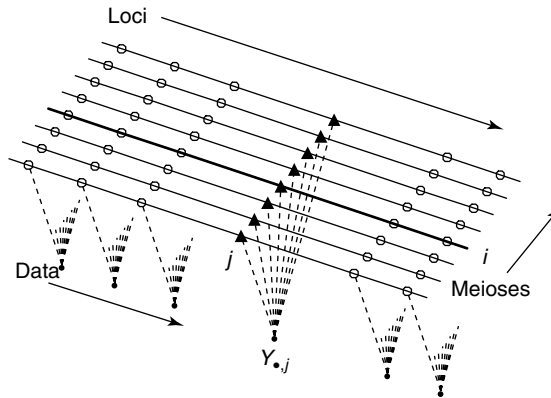
Figure 33.1   The dependence structure of data on a pedigree.

algorithms for likelihood computation on pedigrees. A pedigree is a collection of *a priori* independent meioses. In a given meiosis, transmissions are dependent along the chromosome. In the absence of interference this dependence is first-order Markov. Data are determined by the founder alleles and meioses at a given locus; observation of data at a locus creates *a posteriori* dependence of the meioses at that locus. The Elston–Stewart algorithm uses the *a priori* independence of meioses, which leads to conditional independence of genotypes (33.4). The Lander–Green uses the first-order Markov dependence of meioses along a chromosome (that is, of $S_{\bullet, j}$) in the absence of genetic interference.

Again the methods have been further developed, allowing for more efficient computation as implemented in the programs GENEHUNTER (Kruglyak *et al.*, 1996), ALLEGRO (Gudbjartsson *et al.*, 2000) and MERLIN (Abecasis *et al.*, 2002). A key recognition comes from placing the problem in the context of general graphical models (Lauritzen and Spiegelhalter, 1988), so that all dependencies in Figure 33.1 are treated equally. Due to the *a priori* independence of meioses, the structure is actually that of a *factored* HMM (Fischelson and Geiger, 2004). Thus performance is much improved by summing over the transitions of each meiosis in turn (from $S_{i, j-1}$ to $S_{i, j}$ for each $i = 1, ..., m$), rather than over the entire inheritance vector jointly (from $S_{\bullet, j-1}$ to $S_{\bullet, j}$). Regardless of the values of these transition probabilities (for example, using sex-specific maps), this factorization reduces the computation from order $2^m \times 2^m$ to order $m \times 2^m$. These and other computational advances are embodied in the program SUPERLINK (Fischelson and Geiger, 2004; Silberstein *et al.*, 2006), which has greatly increased the size of the pedigree on which exact computation of multimarker location score curve is feasible. However, Figure 33.1 also shows that computation is intrinsically exponential in the number of meioses m, so pedigree size is a limiting factor in this approach.

A disadvantage of the form (33.6) or (33.9) is the requirement to determine $\Pr(Y_{\bullet, j} | S_{\bullet, j})$, which is potentially more complex than the individual penetrance probabilities $\Pr(Y_{i, \bullet} | G_{i, \bullet})$ of (33.4). Genes are *identical-by-descent* (ibd) if they are copies of the same gene in some common ancestor. Disregarding mutation, such genes are necessarily of the same allelic type. The indicators $S_{\bullet, j} = (S_{i, j}; i = 1, ..., m)$ determine the descent of genes from the founders of the pedigree, and hence the pattern of gene *ibd*

among observed individuals at locus j. If k denotes a founder gene, $a(k)$ its allelic type, and $q_j(a)$ the population frequency of allele a at locus j, then

$$\Pr(Y_{\bullet,j}|S_{\bullet,j}) = \sum_{\mathcal{A}(j)} \left(\prod_k q_j(a(k)) \right). \tag{33.10}$$

(Thompson, 1974), where $\mathcal{A}(j)$ denotes an allocation of allelic types to the founder genes labels (FGL) k at locus j, which, given $S_{\bullet,j}$, is consistent with the data $Y_{\bullet,j}$. Computation requires an efficient algorithm for the summation over all feasible allocations $\mathcal{A}(j)$ at locus j. For marker genotype data observed without error, one such algorithm is given by Sobel and Lange (1996). Using this algorithm, computation of $\Pr(Y_{\bullet,j}|S_{\bullet,j})$ is readily accomplished, even on large pedigrees. For more general marker models or quantitative traits, the problem is again one of peeling, this time over FGL k on an *FGL graph*. The nodes of this graph are the FGL and there is a dependency link between each pair of FGL that are, given $S_{\bullet,j}$, both present in an observed individual. Details are given by Thompson (2005).

33.6   MONTE CARLO MULTIPOINT LINKAGE LIKELIHOODS

The structure of dependence shown in Figure 33.1 also lends itself to a variety of Monte Carlo approaches, in which the latent variables are realized from some importance-sampling distribution. Even when $\Pr(\mathbf{Y})$ cannot be computed exactly, due to the size or complexity of the pedigree and the number of linked genetic loci, Monte Carlo estimates of the likelihood or of posterior probabilities can be made.

Since, in linkage analysis, the parameter θ is normally reserved for recombination frequencies, we use ξ for the full set of parameters of the genetic model. For example, in the case of the location score curves of (33.7), $\xi = (\Gamma_M, \beta, \gamma)$. The likelihood of a genetic model indexed by parameters ξ, given data on a pedigree, can be written

$$L(\xi) = P_\xi(\mathbf{Y}) = \sum_{\mathbf{X}} P_\xi(\mathbf{X}, \mathbf{Y}) = \sum_{\mathbf{X}} P_\xi(\mathbf{Y}|\mathbf{X}) P_\xi(\mathbf{X}), \tag{33.11}$$

where $\mathbf{X}$ are latent variables, either the genotypes $\mathbf{G}$ of (33.4) or the meiosis indicators $\mathbf{S}$ of (33.6). Thus, for fixed observed data $\mathbf{Y}$,

$$L(\xi) = \mathrm{E}_\xi(P_\xi(\mathbf{Y}|\mathbf{X})), \tag{33.12}$$

where the expectation is over $\mathbf{X}$ having the distribution $P_\xi(\mathbf{X})$. Where the variable $\mathbf{X}$ is the set of latent genotypes $\mathbf{G}$, this is the form given by Ott (1979). In principle, $L(\xi)$ could be estimated by simulating $\mathbf{G}$ from the genotype distribution under model ξ and averaging the value of the penetrance probabilities $P_\xi(\mathbf{Y}|\mathbf{G})$ for the realized values of $\mathbf{G}$. This does not work well, except on very small pedigrees, since each realized $\mathbf{G}$ is almost certain to be inconsistent with data $\mathbf{Y}$, or at best to make an infinitesimal contribution to the likelihood.

Better ideas normally involve some form of importance sampling:

$$L(\xi) = \mathrm{E}_{P^*}\left(\frac{P_\xi(\mathbf{X}, \mathbf{Y})}{P^*(\mathbf{X})}\right),\qquad(33.13)$$

where now realizations are made from $P^*(\mathbf{X})$. An advantage of this approach is that a single set of realizations from $P^*(\mathbf{X})$ can be used to provide a Monte Carlo estimate of $L(\xi)$ over a range of models ξ. However, the importance sampling will be effective only if $P^*(\mathbf{X})$ is approximately proportional to $P_\xi(\mathbf{X}, \mathbf{Y})$. Since

$$P_\xi(\mathbf{X}|\mathbf{Y}) = \frac{P_\xi(\mathbf{X}, \mathbf{Y})}{P_\xi(\mathbf{Y})} \propto P_\xi(\mathbf{X}, \mathbf{Y}),\qquad(33.14)$$

the ideal choice of sampling distribution is $P_\xi(\mathbf{X}|\mathbf{Y})$. However the constant of proportionality is $L(\xi) = P_\xi(\mathbf{Y})$. If $P_\xi(\mathbf{X}|\mathbf{Y})$ can be computed, so also can $L(\xi) = P_\xi(\mathbf{Y})$. Thus the choice of sampling distribution is a balance between approximation to $P_\xi(\mathbf{X}|\mathbf{Y})$ and computational feasibility.

The following approach to choice of importance-sampling distribution $P^*(\mathbf{G})$ is due to Kong *et al.* (1994) and Irwin *et al.* (1994). Suppose, as before, there are data at l genetic loci (say a disease and $l-1$ markers) on a chromosome, and assume absence of genetic interference. Note that genotypes $G_{\bullet,j}$ satisfy the same first-order Markov dependence over loci as do the meiosis indicators $S_{\bullet,j}$ (Figure 33.1). Let $Y_{\bullet,j}$ again denote the data for locus j and $G_{\bullet,j}$ the underlying genotypes at that locus for all members of the pedigree. Let $Y^{(j)} = (Y_{\bullet,1}, \ldots, Y_{\bullet,j})$, and $G^{(j)}$ be analogously defined. For any specified ξ of interest, a realization $G^*_{\bullet,j}$ is obtained for each locus in turn from the sequential imputation distribution

$$P_\xi(G_{\bullet,j}|G^{(j-1)}, Y^{(j)}) = P_\xi(G_{\bullet,j}|G_{\bullet,1}, \ldots G_{\bullet,j-1}, Y_{\bullet,1}, \ldots, Y_{\bullet,j-1}, Y_{\bullet,j})$$
$$= w_j^{-1} P_\xi(Y_{\bullet,j}|G_{\bullet,j}) P_\xi(G_{\bullet,j}|G_{\bullet,j-1}),$$

where the predictive weight w_j is $P_\xi(Y_{\bullet,j}|G_{\bullet,j-1})$. Computation of w_j is a single-locus computation which may be done using the Elston–Stewart algorithm. Then it is readily shown that the joint simulation distribution for $\mathbf{G}^* = (G^*_{\bullet,1}, \ldots, G^*_{\bullet,l})$ is

$$P^*(\mathbf{G}^*) = \frac{P_\xi(\mathbf{Y}, \mathbf{G}^*)}{W_l(\mathbf{G}^*)}, \text{ where } W_l(\mathbf{G}^*) = \prod_{j=1}^{l} w_j.\qquad(33.15)$$

Thus

$$\mathrm{E}_{P^*}(W_l(\mathbf{G}^*)) = \sum_{\mathbf{G}^*} W_l(\mathbf{G}^*) P^*(\mathbf{G}^*) = P_\xi(\mathbf{Y}) = L(\xi).\qquad(33.16)$$

A Monte Carlo estimate of $L(\xi)$ is given by the mean value of $W_l(\mathbf{G}^*)$ over repeated independent repetitions of the sequential imputation process. Repeating the process for different trait-locus positions γ on the chromosome, keeping other components of ξ fixed, one can obtain an estimated location score curve (33.7). This procedure is implemented in the program SIMPLE (Skrivanek *et al.*, 2003), and works well for moderate numbers of loci.

Alternative methods use Markov chain Monte Carlo (MCMC) to sample directly from the conditional distribution of latent variables $\mathbf{X}$ given the data $\mathbf{Y}$. We shall not detail here the various MCMC methods that have been used to realize latent variables for location score estimation in linkage analysis. The early methods were primarily single-site methods using either genotypes $\mathbf{G}$ (Lange and Sobel, 1991) or meiosis indicators $\mathbf{S}$ (Thompson, 1994). More recently a variety of block-updating MCMC algorithms have been developed, including the locus-sampler (L-sampler) of Heath (1997) and the meiosis-sampler (M-sampler) of Thompson and Heath (1999), or their combination (Heath and Thompson, 1997). More recent methods allow the sampling of multiple meioses (Thomas *et al.*, 2000), and the use of sequential sampling distribution (33.15) as a Metropolis–Hastings proposal distribution (George and Thompson, 2003). These samplers require the combination of exact and Monte Carlo methods of analysis. Either the generalized Elston–Stewart algorithm (Cannings *et al.*, 1978) is used to perform the required single-locus computations conditional on neighboring loci, or the Lander–Green version of the Baum algorithm (Baum *et al.*, 1970) is used to resample some subset of the meioses.

The first MCMC-based linkage location score approach is due to Lange and Sobel (1991), who, using our current notation, write the likelihood (33.7) in the form

$$
\begin{aligned}
L(\beta, \gamma, \Gamma_M) = P_{\beta,\gamma,\Gamma_M}(Y_T, \mathbf{Y}_M) &\propto P_{\beta,\gamma,\Gamma_M}(Y_T|\mathbf{Y}_M) \\
&= \sum_{\mathbf{X}_M} P_{\beta,\gamma}(Y_T|\mathbf{X}_M) P_{\Gamma_M}(\mathbf{X}_M|\mathbf{Y}_M) \\
&= \mathrm{E}_{\Gamma_M}(P_{\beta,\gamma}(Y_T|\mathbf{X}_M)|\mathbf{Y}_M).
\end{aligned}
\tag{33.17}
$$

Now latent variables $\mathbf{X}_M$ are sampled from their conditional distribution given the marker data $\mathbf{Y}_M$. Provided exact computation of $P_{\beta,\gamma}(Y_T|\mathbf{X}_M)$ is possible for alternative trait models (β) and locations (γ), we have a Monte Carlo estimate of $L(\beta, \gamma, \Gamma_M)$, while comparison to the unlinked base point requires only $P_\beta(Y_T)$. Since Γ_M is fixed the Monte Carlo requires only a single set of N realizations $\mathbf{X}_M^{(\tau)}$, $\tau = 1, \ldots, N$. This approach is implemented in the program SIMWALK (Sobel and Lange, 1996).

The disadvantage of using (33.17) is that $P_{\beta,\gamma}(Y_T|\mathbf{X}_M^{(\tau)})$ must be computed for each such realization; this requires a single-locus peeling computation for the trait-locus data under the trait model. Further, this computation must be done not only for each realization $\mathbf{X}_M^{(\tau)}$ but also for each β and γ at which a likelihood estimate is required. In many cases, however, the gains outweigh the costs except when the simulation distribution $P_{\Gamma_M}(\mathbf{X}_M|\mathbf{Y}_M)$ is not close to proportional to the ideal importance-sampling target distribution $P_{\beta,\gamma,\Gamma_M}(\mathbf{X}_M|Y_T, \mathbf{Y}_M)$. This is particularly so for models (trait locations) γ which are not close to the truth, and for a trait which provides substantial information about the inheritance patterns of genes at the underlying trait locus, and hence also at linked marker loci (Thompson, 2000).

An alternative MCMC approach is based on the estimation of likelihood ratios between models that impose similar distributions on the latent variables conditional on the data, $\mathbf{Y}$. In this case MCMC is used to sample, not from $P_\xi(\mathbf{X}|\mathbf{Y})$ but from $P_{\xi_0}(\mathbf{X}|\mathbf{Y})$, where $\xi_0 \approx \xi$.

Then

$$P_\xi(\mathbf{Y}) = \sum_{\mathbf{X}} P_\xi(\mathbf{Y}, \mathbf{X}) = \sum_{\mathbf{X}} \frac{P_\xi(\mathbf{Y}, \mathbf{X})}{P_{\xi_0}(\mathbf{X}|\mathbf{Y})} P_{\xi_0}(\mathbf{X}|\mathbf{Y})$$

$$= \mathrm{E}_{\xi_0}\left(\frac{P_\xi(\mathbf{Y}, \mathbf{X})}{P_{\xi_0}(\mathbf{X}|\mathbf{Y})} \,\bigg|\, \mathbf{Y} \right) = P_{\xi_0}(\mathbf{Y}) \mathrm{E}_{\xi_0}\left(\frac{P_\xi(\mathbf{Y}, \mathbf{X})}{P_{\xi_0}(\mathbf{Y}, \mathbf{X})} \,\bigg|\, \mathbf{Y} \right).$$

Hence in genetic analysis, or in any latent-variable context, we have

$$\frac{L(\xi)}{L(\xi_0)} = \frac{P_\xi(\mathbf{Y})}{P_{\xi_0}(\mathbf{Y})} = \mathrm{E}_{\xi_0}\left(\frac{P_\xi(\mathbf{Y}, \mathbf{X})}{P_{\xi_0}(\mathbf{Y}, \mathbf{X})} \,\bigg|\, \mathbf{Y} \right). \tag{33.18}$$

(Thompson and Guo, 1991). In this expectation, $\mathbf{Y}$ is fixed, and the distribution of $\mathbf{X}$ is $P_{\xi_0}(\cdot|\mathbf{Y})$. If $\mathbf{X}^{(\tau)}$, $\tau = 1, ..., N$, are realized from this distribution then the likelihood ratio can be estimated by

$$\frac{1}{N} \sum_{\tau=1}^{N} \left(\frac{P_\xi(\mathbf{Y}, \mathbf{X}^{(\tau)})}{P_{\xi_0}(\mathbf{Y}, \mathbf{X}^{(\tau)})} \right).$$

For estimation a location score curve (33.7), the form that follows directly from (33.18) is

$$\frac{L(\beta, \gamma_1, \Gamma_M)}{L(\beta, \gamma_0, \Gamma_M)} = \mathrm{E}_{\xi_0}\left(\frac{P_{\xi_1}(Y_T, \mathbf{Y}_M|X_T, \mathbf{X}_M) P_{\xi_1}(X_T, \mathbf{X}_M)}{P_{\xi_0}(Y_T, \mathbf{Y}_M|X_T, \mathbf{X}_M) P_{\xi_0}(X_T, \mathbf{X}_M)} \,\bigg|\, Y_T, \mathbf{Y}_M \right).$$

for two trait-locus positions γ_1 and γ_0. Since only the position of the trait locus differs between numerator and denominator, the above equation reduces to

$$\frac{L(\beta, \gamma_1, \Gamma_M)}{L(\beta, \gamma_0, \Gamma_M)} = \mathrm{E}_{\xi_0}\left(\frac{P_{\gamma_1}(X_T|\mathbf{X}_M)}{P_{\gamma_0}(X_T|\mathbf{X}_M)} \,\bigg|\, Y_T, \mathbf{Y}_M \right). \tag{33.19}$$

Thus only the conditional probability of trait-locus latent variables given marker-loci latent variables appears explicitly in the estimator. Although realization of the latent variables is complex, and requires MCMC methods, computation of the estimate from the realizations is straightforward (Thompson and Guo, 1991).

For an accurate Monte Carlo estimator based on (33.18) the distribution of $\mathbf{X}$ given $\mathbf{Y}$ should be similar under ξ and ξ_0. In the context of the location score curve estimator (33.19), this means that trait locations γ_1 and γ_0 should lie in the same marker interval. While effective methods have been developed for combining local likelihood ratio estimates from (33.18) into a smooth likelihood surface, multipoint location score curves are far from smooth across marker positions. Attempts to use the estimator (33.19) directly for locations score curves have thus had only limited success (Thompson, 2000). However, see also Section 33.7, for another use of (33.18).

Another approach to estimating location score curves has been proposed by George and Thompson (2003). Since γ is a single scalar parameter, we may regain the likelihood from a posterior probability distribution:

$$L(\beta, \gamma, \Gamma_M) = \frac{P_{(\beta, \Gamma_M)}(\gamma \mid \mathbf{Y})}{\pi(\gamma)},$$

where (β, Γ_M) are held fixed, and $\pi(\gamma)$ is any chosen prior on trait location γ. Since $\pi(\gamma)$ is a pseudoprior in the sense of Geyer and Thompson (1995), it may be arbitrarily chosen to improve MCMC performance. In practice, it is chosen as a discrete prior on points at which the location score curve is to be estimated. To improve MCMC performance, a preliminary run is used to provide a prior approximately inversely proportional to the likelihood, so that in the main MCMC run the marginal posterior $P_{(\beta,\Gamma_M)}(\gamma \mid \mathbf{Y})$ is approximately uniform.

In Monte Carlo estimation, Rao-Blackwellization (Gelfand and Smith, 1990) can play an important role in improving estimators. This procedure replaces an estimator by its conditional expectation given some subset of the sampled variables. Thus the Rao-Blackwellized estimator requires more intensive exact computations, but benefits from reduced Monte Carlo sample size to achieve the same accuracy. For dependent realizations, Rao–Blackwellization does not guarantee reduced Monte Carlo variance (Liu *et al.*, 1994), but in practice it normally provides significant gains in computational efficiency. For example, suppose we have realizations $(\gamma^{(\tau)}, \mathbf{X}^{(\tau)})$, $\tau = 1, ..., N$ from $P_{(\beta,\Gamma_M)}(\gamma, \mathbf{X} \mid \mathbf{Y})$. Rather than estimating the marginal posterior $P_{(\beta,\Gamma_M)}(\gamma \mid \mathbf{Y})$ by

$$\frac{1}{N} \sum_{\tau=1}^{N} I(\gamma^{(\tau)} = \gamma),$$

we instead use the estimators

$$\frac{1}{N} \sum_{\tau=1}^{N} \mathrm{E}(I(\gamma^{(\tau)} = \gamma) \mid \mathbf{X}_M^{(\tau)}, \mathbf{Y}).$$

In fact, it is easily shown that

$$\mathrm{E}(I(\gamma^{(\tau)} = \gamma) \mid \mathbf{X}_M^{(\tau)}, \mathbf{Y}) = \frac{P(Y_T \mid \gamma, \mathbf{X}_M^{(\tau)})\pi(\gamma)}{\sum_{\gamma^*} P(Y_T \mid \gamma^*, \mathbf{X}_M^{(\tau)})\pi(\gamma^*)}, \qquad (33.20)$$

leading to an estimator similar in appearance to (33.17) in that both depend on the probabilities of trait data Y_T given sampled marker latent variables $\mathbf{X}_M$ computed for each desired trait location γ. However, note that here the trait data and trait model β enter into the sampling of variables $\mathbf{X}_M$.

The new MCMC methods can provide accurate location scores on large pedigrees with substantial missing data for two or three marker loci many times faster than the exact computational methods, and can estimate multipoint location score curves when exact computation is infeasible (George and Thompson, 2003). The MCMC lod score estimation methods of (33.17), (33.19), and (33.20) are all implemented in programs in the MORGAN package (Thompson, 2005).

33.7   LINKAGE ANALYSIS OF COMPLEX TRAITS

As the traits considered in linkage analyses become increasingly complex, there are two alternative approaches to analysis. One is to avoid explicit modeling of the

trait, using so-called 'nonparametric' genome-sharing computations for linkage detection (see **Chapter 34**). The other is to develop more complex models for more complex phenotypes. Several factors enhance the detection and resolution of genes contributing to complex genetic traits. These include joint analysis of data on members of extended pedigrees, joint analysis of data at multiple genetic markers, and selective sampling of pedigrees (Wijsman and Amos, 1997). Although data are increasingly available at multiple tightly linked genetic markers, often a high proportion of the individuals in an extended pedigree is not observed. Data may be available only on affected individuals, or even on only a few of their relatives. Moreover, selectively sampled extended pedigrees are often also complex. Often the traits of interest are not simple genetic traits. They may have delayed onset, or there may be incomplete penetrance; individuals carrying the disease allele may never show symptoms. Conversely, there may be individuals who apparently have the disease but do not carry the gene, there may be several different genetic causes of a disease, or even several genes interacting to produce the observed characteristics. In some cases, the traits of interest may be quantitative.

Because of difficulties of likelihood computation, trait model uncertainties, and the mass of marker data, often with different markers typed or informative in different families, a number of researchers moved away from likelihood analyses of linkage in the 1980s, and instead developed a variety of association tests for linkage detection. There are two main classes of such tests, those at the population level and those at the family or pedigree level. Genome-sharing methods are based on analysis of marker data, and rely on the fact that, regardless of the trait model provided it has some genetic component, related affected individuals or related individuals exhibiting extreme trait values will share genes *ibd* at trait loci with some increased probability. Hence also they will share genes *ibd* with increased probability at marker loci linked to those trait loci. While genome-sharing methods on individuals with known relationship derive first from Penrose (1935), and are best known in the form of sib pair analyses (Suarez *et al.*, 1978) and affected-relative methods (Weeks and Lange, 1988) they can also involve probability computations on a pedigree structure. For example, the method of *homozygosity mapping* for rare recessive traits developed by Lander and Botstein (1987) can be viewed either as computation of a linkage likelihood (Smith, 1953) or as an inference based on the genome sharing between the two haplotypes of inbred affected individuals.

Another area of linkage analysis which uses gene *ibd* is fine-scale localization of genes. The resolution of pedigree-based methods of linkage analysis is limited by the number of meioses in the pedigrees. However, when a new trait allele arises by mutation, it does so on some specific chromosome with a specific collection of alleles at nearby markers; that is, there is a specific marker *haplotype* that carries the new trait mutation. Where the loci are very tightly linked, associations of the disease allele with an allele at a linked marker locus may be maintained for many generations before decaying due to recombination (33.3). Thus, there will be population associations due to chance historical associations. The study of association can be very useful in narrowing the region in which a gene is located. For a general discussion of the use of association tests in linkage analysis, see **Chapter 34**. Here we focus on aspects relating to analysis of marker data on extended pedigrees.

There is much more information about genome-sharing patterns in the genome if all the available data on the pedigree are used to compute *ibd* probabilities, and if *ibd* is scored jointly among the several affected individuals of an extended pedigree, rather than

pair-wise. Exact computation of gene *ibd* patterns jointly among relatives in an extended pedigree, using data at multiple linked markers, is equivalent to likelihood computation for these marker data, as given by (33.4) or (33.9). Note that the gene *ibd* at any locus j is a function of the inheritance vector $S_{\bullet, j}$ (Section 33.5). For data at multiple marker loci, exact computation is thus feasible only on a small pedigree. Moreover, although likelihoods for relationship and the estimates of gene *ibd* patterns do incorporate the linkage information and data at all loci, the Baum algorithm (Baum *et al.*, 1970) gives only estimates of the probabilities $\Pr(S_{\bullet, j} \mid \mathbf{Y})$ marginally at each locus j. At best, on small pedigrees, it is possible to obtain the probabilities $\Pr(S_{\bullet, j-1}, S_{\bullet, j} \mid \mathbf{Y})$ for pairs of loci.

However, Monte Carlo sampling of the meiosis indicators $\mathbf{S}_M$ given marker data $\mathbf{Y}_M$ provides direct estimates of posterior probabilities of patterns of gene *ibd*. These realizations can be scored jointly over loci, and jointly over individuals, and hence provides for more general analyses of genome sharing in relatives (Thompson, 2000). Further, given each realization $\mathbf{S}_M^{(\tau)}$, realizations of the inheritance vector $S(\gamma)$ at location γ may be readily obtained. If γ lies between markers $j - 1$ and j

$$\Pr(S(\gamma) \mid \mathbf{S}_M, \mathbf{Y}_M) = \Pr(S(\gamma) \mid S_{\bullet, j-1}^{(\tau)}, S_{\bullet, j}^{(\tau)}).$$

That is, $S(\gamma)$ depends only on the realized inheritance vector at the two neighboring markers. On small pedigrees or for certain *ibd* patterns of interest exact probability computations given the realized marker inheritance vectors are often possible.

For linkage analysis, the statistical problem becomes, first, one of development of appropriate test statistics to detect linkage on the basis of estimated posterior *ibd* probabilities (for example, see McPeek, 1999). Second a way of assessing statistical significance of estimated *ibd* statistics is needed, taking into account both uncertainties in the *ibd* given marker data, that *ibd* patterns at linked locations with a region of the genome are highly dependent, and the fact that numerous regions of the genome may be tested for linkage to the trait of interest. Under the hypothesis of no linkage to the trait, resimulation of marker data is always possible, but analysis of multiple resimulated data sets is computationally intensive and this approach is sensitive to marker model misspecification. Permutation tests (Churchill and Doerge, 1994) are robust, but on an extended pedigree the set of available valid permutations may be small. A compromise is provided by the approach of Thompson and Geyer (2007), which, like a permutation test conditions on the observed marker data, and uses the probability distribution of the underlying latent variables $\mathbf{S}_M$ to provide a test of significance.

The alternative general approach for analysis of complex traits uses an explicit trait model. An early such model was the *mixed model* (Morton and MacLean, 1974), in which the phenotypic value is dependent both on the genotype at a 'major' Mendelian locus and on a heritable *polygenic* Gaussian random effect. While the polygenic component can be a convenient way to model additional heritable variation, likelihood computation for the mixed model is problematic, and linkage analysis more so. Increasingly, methods are being developed for *oligogenic models* in which several major genes contribute to a trait, and a Bayesian MCMC analysis is used to detect and localize these genes and estimate their effects (Heath, 1997). Although these methods have been directed primarily toward quantitative traits (see **Chapter 19**), they can also be used to analyze a qualitative trait controlled by a latent quantitative liability. Age-of-onset of a disease can also be analyzed, using these methods, by treating it as a censored quantitative trait (Daw *et al.*, 1999).

Robust genome-sharing methods for linkage detection and Bayesian methods of linkage estimation are gaining popularity in part because of the difficulties of both computation and interpretation of linkage likelihoods for complex traits. These difficulties of interpretation result from uncertainties in the trait model, heterogeneity in the genetic causes of the trait, and interpretation of location score curves. In terms of assessing sensitivity to trait, marker, or map model assumptions, (33.18) provides a useful MCMC approach. Recall that for any probability models, ξ and ξ_0

$$\frac{L(\xi)}{L(\xi_0)} = \frac{P_\xi(\mathbf{Y})}{P_{\xi_0}(\mathbf{Y})} = \mathrm{E}_{\xi_0}\left(\frac{P_\xi(\mathbf{Y}, \mathbf{X})}{P_{\xi_0}(\mathbf{Y}, \mathbf{X})} \,\bigg|\, \mathbf{Y}\right),$$

provided only that ξ_0 assigns positive probability to any latent values $\mathbf{X}$ that have positive probability under ξ. Thus a single set of Monte Carlo realizations at some assumed model ξ_0 can be used to obtain an estimate of the likelihood ratio $L(\xi)/L(\xi_0)$ for any model that is close to ξ. The models ξ may differ from ξ_0 in any, or several, parameters.

As data on more complex traits are analyzed, the problems of trait model misspecification and trait heterogeneity have assumed increasing importance. The excellent recent text by Ott (1999) gives a very thorough review of human genetic linkage, and methods of linkage analysis. The Bayesian approach, with a prior probability distribution for the location of a trait gene, avoids some of the problems of multiple tests for linkage. These methods can also be used to test for heterogeneity, providing a posterior probability that a particular pedigree is segregating a putative trait gene Ott (1999). Although Bayesian methods have a long history in linkage analysis (Haldane and Smith, 1947) and have recently attracted more attention, the problems of trait model misspecification and heterogeneity are no less with a Bayesian approach. Inference based on the *lod* score or location score curve is still the accepted practical approach. Further, new Monte Carlo–based approaches to *lod* score estimation can allow for more complex trait models, two such major loci, and a polygenic component all contributing jointly to a quantitative trait value (Sung *et al.*, 2006).

33.8   MAP ESTIMATION, MAP UNCERTAINTY, AND THE MEIOSIS MODEL

In the previous sections, we assumed the marker map Γ_M to be known. However, the usefulness of multipoint linkage analyses is dependent on the accuracy of both the marker map and of the trait model. Uncertainties about the marker map can have a strong impact on the location log-likelihood curve (Halpern and Whittemore, 1999; Daw *et al.*, 2000). Although there is now a wealth of DNA markers, their exact positions and the frequencies of their alleles in the study population are often uncertain. Thus, there is a need to compute location scores under alternative assumptions about the marker loci and to compute likelihoods relating to estimation of the marker map. Evaluation of multipoint lod scores is extremely computationally intensive in human pedigrees where there are often missing data and many alternative patterns of gene descent that are compatible with the observed data.

Estimation of an accurate meiotic map becomes harder as maps become denser. The same markers will not be typed in all studies; even if typed, the same will not be

informative for linkage. Moreover, rather than seeking to estimate recombination rates of perhaps 20 %, it is necessary to order markers between which recombination rates are less than 1 %. To analyze these small recombination frequencies, far more data are required. Typically, the amounts of data available in pedigree studies will not permit a finer scale of resolution than about 1 cM (Boehnke, 1994). Even 1 cM distance is approximately 1 million DNA base pairs – too great a length for current methods of physical mapping to be practical.

Many factors influence recombination frequencies. A major one is the sex of the parent; in humans the total female map length of the 22 chromosome pairs (not including the sex chromosomes) is 39 M, about 1.5 times the male map length (26.5 M). However, this ratio is not constant over the genome. In many linkage analyses, male and female recombination frequencies are estimated jointly, although sometimes they are constrained to be equal, or a fixed relationship between them may be assumed. Differences in male and female genetic map distances have an impact on multipoint location scores. However, even the best current meiotic maps (NIH-CEPH Collaborative Mapping Group, 1992) are based on very limited numbers of meioses particularly when sex-specific maps are desired. Often a large study may have as many meioses available for analysis as these standard maps. In fact, several recent large-scale genotyping studies, such as that of the Icelandic population by deCODE Genetics, provide large numbers of informative meioses, and will become a future resource for investigating the relationship between physical and genetic distance and sources of variation in genetic maps. For example a study of stroke by Gretarsdottir *et al.* (2002) yielded over 1000 informative meioses, over five times as many as in the standard CEPH panel. Map estimation is needed to refine maps in critical regions, where valid likelihoods for linkage detection or association analyses are desired.

The problem of map estimation is well suited to an EM algorithm approach (Dempster *et al.*, 1977). As earlier, suppose, we have l marker loci along a chromosome, now with recombination frequencies $\theta_{m,j-1}$ and $\theta_{f,j-1}$ in male and female meioses, respectively, between locus $j-1$ and locus j. With data $\mathbf{Y}$ and latent variables $\mathbf{S}$ consider the complete-data log-likelihood

$$\log \Pr(\mathbf{S}, \mathbf{Y}) = \log(\Pr(S_{\bullet,1})) + \sum_{j=2}^{l} \log(\Pr(S_{\bullet,j}|S_{\bullet,j-1})) + \sum_{j=1}^{l} \log(\Pr(Y_{\bullet,j}|S_{\bullet,j})) \quad (33.21)$$

(see equation (33.9)). Now, in the absence of interference, the recombination probabilities $\theta_{m,j-1}$ and $\theta_{f,j-1}$ enter only into the term $\log(\Pr(S_{\bullet,j} \mid S_{\bullet,j-1}))$ which takes the form

$$\log\left(\Pr\left(S_{\bullet,j}|S_{\bullet,j-1}\right)\right) = R_{m,j-1} \log\left(\theta_{m,j-1}\right) + \left(M_m - R_{m,j-1}\right) \log\left(1 - \theta_{m,j-1}\right)$$
$$+ R_{f,j-1} \log\left(\theta_{f,j-1}\right) + \left(M_f - R_{f,j-1}\right) \log\left(1 - \theta_{f,j-1}\right),$$

where $R_{m,j-1}$ is the number of recombinations in interval $(j-1, j)$ in male meioses, and M_m is the total number of male meioses scored in the pedigree. The recombination counts $R_{f,j-1}$, for $j = 2, \ldots, l$, and total meioses M_f are similarly defined for the female meioses. Thus computation of the expected complete-data log-likelihood requires only computation of

$$\tilde{R}_{m,j-1} = \mathrm{E}(R_{m,j-1}|\mathbf{Y}).$$

Since this is a simple binomial log likelihood, the M-step sets the new estimate of $\theta_{m,j-1}$ to $\tilde{R}_{m,j-1}/M_m$, and similarly for all intervals $j = 2, 3, \ldots, l$ and for both the male and female meioses. On a small pedigree, expectations can be computed exactly, and the EM algorithm is thus readily implemented to provide estimates of recombination frequencies for all intervals and for both sexes.

An alternative is Monte–Carlo EM (Guo and Thompson, 1994). Instead of computing the bivariate distributions of $(S_{\bullet,j-1}, S_{\bullet,j})$ (Baum *et al.*, 1970) in order to estimate the expected proportions of recombinations in interval $(j - 1)$, N realizations $\{S^{(\tau)}; \tau = 1, \ldots, N\}$ are obtained from the conditional distribution of $\Pr(S \mid Y)$ under the current parameter values. Recombinations and nonrecombinations are scored in the realized $S^{(\tau)}$, and Monte Carlo estimates of the expectations are obtained. This Monte Carlo EM is readily implemented, and, like many Monte Carlo EM procedures, performs as well as the deterministic version. Initially the Monte Carlo sample size N need not be large, although for the final EM steps it should be increased to gain precision in the final Monte Carlo estimates of the maximum likelihood estimates (MLEs) of the recombination frequencies. A generalization of this basic approach has been developed by Stewart and Thompson (2006), permitting map estimation on extended pedigrees with extensive missing data. These methods facilitate assessment of the accuracy and uncertainty in meiotic maps and provide an approach to combine data sets from several sources, including the large pedigrees used for disease–gene mapping.

MCMC realization of meiosis indicators S, or of multilocus genotypes G, on a pedigree under a genetic map given genetic marker data Y also leads directly to methods for detecting genotype or pedigree error. The meiosis indicators determine imputed multilocus haplotypes which can be used to detect probable genotyping errors that are apparent only when multilocus data are analyzed jointly. Methods for detection of such pedigree and typing errors are important as preliminaries to other multipoint analysis methods and have been a focus of research in recent years (McPeek and Sun, 2000; Epstein *et al.*, 2000; Sieberts *et al.*, 2002).

Another aspect of meiotic maps is the presence of genetic interference. In experimental organisms, genetic interference has been well known and well studied since Haldane (1919). For example, Weinstein (1936) investigated estimation of multilocus recombination pattern frequencies using data he had previously collected on seven linked loci in a sample of 28 239 meioses of *Drosophila* X chromosomes. Although Bailey (1961) developed the mathematical theory and models for the process of interference, only now do we have, in principle, the information to investigate interference on the basis of data on humans or other mammalian species. For example, King *et al.* (1991) have shown almost complete interference in mice over regions up to 10 cM. Genetic interference in meiosis is a more complex issue than that of sex-specific maps, since it destroys the first-order Markov conditional-independence-structure of the inheritance vectors along a chromosome. Although genetic interference exists, it is seldom incorporated into multilocus linkage computations. This can reduce the power to detect linkage (Goldstein *et al.*, 1995). In an analysis of data at multiple tightly linked markers from actual maternal meioses in loblolly pine, Thompson and Meagher (1998) have shown that interference can have a significant impact on patterns of cosegregation of genes and gene *ibd* at distances of 30–50 cM.

Almost all likelihood computations on pedigrees assume absence of interference, and hence the independence of recombination in different marker intervals. An earlier practice

was to transform estimated recombination fractions using some mapping function relating genetic distance d (in Morgans) to recombination fraction θ. For example, in the absence of interference

$$\theta(d) = \frac{1}{2}(1 - \exp(-2d)),$$

(Haldane, 1919), while for the Kosambi map function which exhibits a moderate amount of interference

$$\theta(d) = \frac{1}{2}\tanh(2d) = \frac{1}{2}\left(\frac{\exp(4d) - 1}{\exp(4d) + 1}\right).$$

Since for all natural map functions, $\theta \approx d$ when both are small, this transformation practice becomes increasingly futile as maps become denser. Interference must be accommodated within the multilocus computation itself. The question of the impact of interference on linkage analysis and linkage detection can only be resolved through likelihood computation and data analysis under both interference and noninterference models. Two methods of likelihood evaluation under interference have been proposed. Weeks *et al.* (1993) provides an approach for models of count interference (Liberman and Karlin, 1984), while Lin and Speed (1996) provides a method for the renewal process χ-square models of position interference (Zhao *et al.*, 1995). Both methods are computationally intensive, and limited to small numbers of loci and/or simple pedigrees.

MCMC and Monte Carlo likelihood methods greatly extend the feasibility of likelihood evaluation under interference models. If more complex meiosis models are to be considered, a meiosis-based inheritance-vector M-sampler has advantages over a locus-based genotypic L-sampler. Using the M-sampler, all the inheritance indicators for all the linked loci in an entire meiosis are resampled jointly and incorporation of an interference model is feasible. For a fixed marker map, it is feasible to precompute and store probabilities of all patterns of recombination and nonrecombination in a set of marker intervals for up to about 12 markers ($2^{11} = 2048$). However, as for any multilocus problem, exact likelihood computation on an extended or complex pedigree remains computationally infeasible. The MCMC M-sampler can, however, be extended. In that sampler, given marker data $\mathbf{Y}$ at loci $j = 0, \ldots, K$, meiosis indicators at meiosis i, $S_{i,\bullet} = (S_{i,0}, \ldots, S_{i,K})$ are realized from

$$P^{(H)}(S_{i,\bullet}|S_{k,\bullet}, k \neq i, \mathbf{Y}) \propto P(\mathbf{Y}|\mathbf{S})P^{(H)}(\mathbf{S}), \tag{33.22}$$

where $\mathbf{S}$ is the total set of meiosis indicators for all loci at all meioses of the pedigree, and the superscript (H) denotes the Haldane (no-interference) model. Using (33.22) as the proposal distribution, a Metropolis acceptance step (Metropolis *et al.*, 1953) can be added into the process to provide the correct conditional distribution of $S_{i,\bullet}$ under interference (denoted $P^{(I)}$). The required Hastings-ratio α (Hastings, 1970) for the acceptance probability $\min(1, \alpha)$ for current $\mathbf{S}$ and proposed $\mathbf{S}^{\dagger}$ is

$$\begin{aligned}
\alpha &= \frac{P^{(I)}(\mathbf{S}^{\dagger}, \mathbf{Y})}{P^{(I)}(\mathbf{S}, \mathbf{Y})} \frac{P^{(H)}(S_{i,\bullet}|S_{k,\bullet}, k \neq i, \mathbf{Y})}{P^{(H)}(S_{i,\bullet}^{\dagger}|S_{k,\bullet}^{\dagger}, k \neq i, \mathbf{Y})} \\
&= \frac{P^{(I)}(S_{i,\bullet}^{\dagger})}{P^{(I)}(S_{i,\bullet})} \frac{P^{(H)}(S_{i,\bullet})}{P^{(H)}(S_{i,\bullet}^{\dagger})}.
\end{aligned} \tag{33.23}$$

This considerable reduction, and consequent ease of computation of the acceptance probability $\min(1, \alpha)$ relies on the independence of segregation patterns at different meioses i (when not conditioned on data $\mathbf{Y}$), and of the fact that the probability of data $\mathbf{Y}$ given meiosis pattern $\mathbf{S}$ does not depend on the interference process giving rise to $\mathbf{S}$. Thompson (2000) gives an example of the use of this Metropolis–Hastings sampler to investigate the impact of interference on probabilities of patterns of gene identity by descent conditional on observed marker data on a pedigree.

33.9   THE FUTURE

Just as the questions of linkage analysis have remained unchanged, so also have the basic approaches. The ideas developed by Sturtevant (1913), Robbins (1918), Haldane (1919), and Fisher (1922b) are still the basis of analysis of the outcomes of meiosis, and hence of linkage analysis. With the completion of the Human Genome Project, the practical relevance of linkage analysis is sometimes questioned, particularly in view of the variable relationship between the physical map (DNA sequence lengths in base pairs) and the genetic map (recombination frequencies). However, this issue dates back to (Fisher, 1922b), and despite changing markers and changing traits of interest, analysis of the cotransmission of genes is still the only route to a fuller understanding of the meiotic map, and sources of variation in that map. Even when our markers become the DNA sequence, and traits are levels of gene expression in a given tissue or organ, patterns of similarity and difference among relatives still provide the basic information of genetic inheritance.

As data on multiple linked markers have become available, the computational paradigm has shifted from one based on genotypes of individuals in a pedigree structure (Elston and Stewart, 1971) to one based on the meiosis indicators along a chromosome (Lander and Green, 1987). As DNA data become available at even higher resolution, it may become impractical to study the huge number of marker loci available. Instead of considering recombination events between discrete marker loci, it becomes possible to analyze the precise crossover points in a segregation, or the segments of genome shared by relatives, or by individuals having a trait in common. Such models date back to Haldane (1919) but became more formalized with Fisher's theory of junctions (Fisher, 1949). In the context of data on pedigrees, Donnelly (1983) first developed continuous genome models in terms of the meiosis indicators of (33.1). As for classical linkage analysis, likelihoods must be computed, and the computational issues must become more complex. Monte Carlo estimation of likelihoods will have a role here also (Browning, 2000), in conjunction with other computational approaches.

From the situation in 1980, when many trait data were available but marker typing was hard and expensive, we have moved to an era in which the marker typing of available individuals is relatively cheap, fast, and easy. The major cost of a study of a complex trait is now in the family collection and trait phenotyping. The limits to resolution in linkage analysis now lies in the trait data rather than in the markers. Data on extended pedigrees are necessary to resolve questions of locus and allelic heterogeneity, but in an extended pedigree a high proportion of individuals may be unavailable. The robustness of linkage inferences not only to trait model deviations but also to genetic map and meiosis model

variations remains to be fully investigated. These questions raise further computational and statistical challenges to be met in the twenty-first century.

REFERENCES

Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2002). Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101.

Bailey, N.T.J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Clarendon Press, Oxford.

Barnard, G.A. (1949). Statistical inference. *Journal of the Royal Statistical Society Series B* **11**, 115–139.

Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions on Markov processes. In *Inequalities-III; Proceedings of the Third Symposium on Inequalities. University of California Los Angeles, 1969*, O. Shisha, ed. Academic Press, New York, pp. 1–8.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions on Markov chains. *Annals of Mathematical Statistics* **41**, 164–171.

Boehnke, M. (1994). Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. *American Journal of Human Genetics* **55**(2), 379–390.

Botstein, D., White, R.L., Skolnick, M.H. and Davis, R.W. (1980). Construction of a linkage map in man using restriction fragment polymorphism. *American Journal of Human Genetics* **32**, 314–331.

Browning, S. (2000). A Monte Carlo approach to calculating probabilities for continuous identity by descent data. *Journal of Applied Probability* **37**, 850–864.

Cannings, C., Thompson, E.A. and Skolnick, M.H. (1978). Probability functions on complex pedigrees. *Advances of Applied Probability* **10**, 26–61.

Cannings, C., Thompson, E.A. and Skolnick, M.H. (1980). Pedigree analysis of complex models. In *Current Developments in Anthropological Genetics*, J. Mielke and M. Crawford, eds. Plenum Press, New York, pp. 251–298.

Churchill, G.A. and Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**(3), 963–71.

Cottingham, R.W., Idury, R.M. and Schäffer, A.A. (1993). Faster sequential genetic linkage computations. *American Journal of Human Genetics* **53**, 252–263.

Curtis, D. and Gurling, H. (1993). A procedure for combining two-point lod scores into a summary multipoint map. *Human Heredity* **43**, 173–185.

Daw, E.W., Kumm, J., Snow, G.L., Thompson, E.A. and Wijsman, E.M. (1999). MCMC methods for genome screening. *Genetic Epidemiology* **17**(Suppl. 1), S133–S138.

Daw, E.W., Thompson, E.A. and Wijsman, E.M. (2000). Bias in multipoint linkage analysis arising from map misspecification. *Genetic Epidemiology* **19**, 366–380.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B* **39**, 1–37.

Donnelly, K.P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology* **23**(1), 34–63.

Edwards, J.H. (1971). The analysis of X-linkage. *Annals of Human Genetics* **34**, 229–259.

Elston, R.C. and Stewart, J. (1971). A general model for the analysis of pedigree data. *Human Heredity* **21**, 523–542.

Epstein, M., Duren, W. and Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. *American Journal of Human Genetics* **67**, 1219–1231.

Fischelson, M. and Geiger, D. (2004). Optimizing exact linkage computations. *Journal of Computational Biology* **11**, 263–275.

Fisher, R.A. (1922a). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London Series A* **222**, 309–368.

Fisher, R.A. (1922b). The systematic location of genes by means of crossover observations. *American Naturalist* **56**, 406–411.

Fisher, R.A. (1934). The amount of information supplied by records of families as a function of the linkage in the population sampled. *Annals of Eugenics* **6**, 66–70.

Fisher, R.A. (1936). Heterogeneity of linkage data for Friedreich's ataxia and the spontaneous antigens. *Annals of Eugenics* **7**, 17–21.

Fisher, R.A. (1949). *The Theory of Inbreeding*. Oliver and Boyd, Edinburgh.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

George, A.W. and Thompson, E.A. (2003). Multipoint linkage analyses for disease mapping in extended pedigrees: a Markov chain Monte Carlo approach. *Statistical Science* **18**, 515–531.

Geyer, C.J. and Thompson, E.A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90**, 909–920.

Goldstein, D.R., Zhao, H. and Speed, T.P. (1995). Relative efficiencies of chi-square models of recombination for exclusion mapping and gene ordering. *Genomics* **27**(2), 265–273.

Gretarsdottir, S., Sveinbjornsdottir, S., Jonsson, H.H., Jakobsson, F., Einarsdottir, E., Agnarsson, U., Shkolny, D., Einarsson, G., Gudjonsdottir, H.M., Valdimarsson, E.M., Einarsson, O.B., Thorgeirsson, G., Hadzic, R., Jonsdottir, S., Reynisdottir, S.T., Bjarnadottir, S.M., Gudmundsdottir, T., Gudlaugsdottir, G.J., Gill, R., Lindpaintner, K., Sainz, J., Hannesson, H.H., Sigurdsson, G.T., Frigge, M.L., Kong, A., Gudnason, V., Stefansson, K. and Gulcher, J.R. (2002). Localization of a susceptibility gene for common forms of stroke to 5q12. *American Journal of Human Genetics* **70**, 593–603.

Gudbjartsson, D., Jonasson, K., Frigge, M. and Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics* **25**, 12–13.

Guo, S.W. and Thompson, E.A. (1994). Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics* **50**(2), 417–432.

Haldane, J.B.S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**, 229–309.

Haldane, J.B.S. (1934). Methods for the detection of autosomal linkage in man. *Annals of Eugenics* **6**, 26–65.

Haldane, J.B.S. and Smith, C.A.B. (1947). A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Annals of Eugenics* **14**, 10–31.

Halpern, J. and Whittemore, A.S. (1999). Multipoint linkage analysis. A cautionary note. *Human Heredity* **49**, 194–196.

Hartl, D.L. and Clark, A.G. (1997). *Principles of Population Genetics*, 3rd edition. Sinauer, Sunderland, MA.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Heath, S.C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics* **61**(3), 748–760.

Heath, S. and Thompson, E.A. (1997). MCMC samplers for multilocus analyses on complex pedigrees. *American Journal of Human Genetics* **61**(Suppl.), A278.

Heuch, I. and Li, F.M.H. (1972). PEDIG – A computer program for calculation of genotype probabilities, using phenotypic information. *Clinical Genetics* **3**, 501–504.

Hilden, J. (1970). GENEX – An algebraic approach to pedigree probability calculus. *Clinical Genetics* **1**, 319–348.

Irwin, M., Cox, N. and Kong, A. (1994). Sequential imputation for multilocus linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 11684–11688.

King, T.R., Dove, W.F., Guénet, J., Hermann, B.G. and Shedlovsky, A. (1991). Meiotic mapping of murine chromosome 17: the string of loci around l(17)-2-Pas. *Mammalian Genome* **1**, 37–46.

Kong, A., Liu, J. and Wong, W.H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association* **89**, 278–288.

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* **58**(6), 1347–1363.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Annals of Statistics* **22**, 79–86.

Lander, E.S. and Botstein, D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567–1570.

Lander, E.S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans.*Proceedings of the National Academy of Sciences of the United States of America* **84**(8), 2363–2367.

Lange, K. and Sobel, E. (1991). A random walk method for computing genetic location scores. *American Journal of Human Genetics* **49**, 1320–1334.

Lathrop, G.M., Lalouel, J.M., Julier, C. and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 3443–3446.

Lauritzen, S.L. and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society Series B* **50**, 157–224.

Liberman, U. and Karlin, S. (1984). Theoretical models of genetic map functions. *Theoretical Population Biology* **25**(3), 331–346.

Lin, S. and Speed, T.P. (1996). Incorporating crossover interference into pedigree analysis using the χ^2 model. *Human Heredity* **46**, 315–322.

Liu, J., Wong, W.H. and Kong, A. (1994). A covariance struture of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

McPeek, M.S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology* **16**, 225–249.

McPeek, M. and Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *American Journal of Human Genetics* **66**, 1076–1094.

Mendel, G. (1866). Experiments in plant hybridisation. In *English Translation and Commentary by R. A. Fisher*, J.H. Bennett, ed. Oliver and Boyd, Edinburgh 1965.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Morton, N.E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7**, 277–318.

Morton, N.E. and MacLean, C.J. (1974). Analysis of family resemblance.III. Complex segregation of quantitative traits. *American Journal of Human Genetics* **26**, 489–503.

Murray, J.C., Buetow, K.H., Weber, J.L., Ludwigson, S., Scherpier-Heddema, T., Manion, F., Quillen, J., Sheffield, V., Sunden, S., Duyk, G.M., Weissenbach, J., Gyapay, G., Dib, C., Morrissette, J., Lathrop, G.M., Vignal, A., White, R., Matsunami, N., Gerken, S., Melis, R., Albertsen, H., Plaetke, R., Odelberg, S., Ward, D., Dausset, J., Cohen, D. and Cann, H. (1994). A comprehensive human linkage map with centimorgan density. *Science* **265**, 2049–2064.

NIH-CEPH Collaborative Mapping Group (1992). A comprehensive genetic linkage map of the human genome, *Science* **258**, 67–86.

O'Connell, J.R. and Weeks, D.E. (1995). The algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genetics* **11**(4), 402–408.

Ott, J. (1974). Estimation of the recombination frequency in human pedigrees: efficient computation of the likelihood for human linkage studies. *American Journal of Human Genetics* **26**, 588–597.

Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *American Journal of Human Genetics* **31**, 161–175.

Ott, J. (1999). *Analysis of Human Genetic Linkage*, 3rd edition. The Johns Hopkins University Press, Baltimore, MD.

Penrose, L.S. (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Annals of Eugenics* **6**, 133–138.

Ploughman, L.M. and Boehnke, M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics* **44**, 543–551.

Renwick, J.H. (1969). Progress in mapping the human autosomes. *British Medical Bulletin* **25**, 65–73.

Robbins, R.B. (1918). Some applications of mathematics to breeding problems.III. *Genetics* **3**, 375–389.

Sieberts, S.K., Wijsman, E.M. and Thompson, E.A. (2002). Relationship inference from trios of individuals in the presence of typing error. *American Journal of Human Genetics* **70**, 170–180.

Silberstein, M., Tzemach, A., Dovgolesky, N., Fishelson, M., Schuster, A. and Geiger, D. (2006). Online system for faster multipoint linkage analysis via parallel execution on thousands of personal computers. *American Journal of Human Genetics* **78**, 922–935.

Skrivanek, Z., Lin, S. and Irwin, M.E. (2003). Linkage analysis with sequential imputation. *Genetic Epidemiology* **25**, 25–35.

Smith, C.A.B. (1953). Detection of linkage in human genetics. *Journal of the Royal Statistical Society Series B* **15**, 153–192.

Smith, C.A.B. (1963). Testing for heterogeneity of recombination fractions in human genetics. *Annals of Human Genetics* **27**, 175–182.

Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics* **58**, 1323–1337.

Stewart, W.C.L. and Thompson, E.A. (2006). Improving estimates of genetic maps: a maximum likelihood approach. *Biometrics* **62**, 728–734.

Sturtevant, A.H. (1913). The linear association of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology* **14**, 43–59.

Suarez, B.K., Rice, J. and Reich, T. (1978). The generalized sib pair IBD distribution: its use in the detection of linkage. *Annals of Human Genetics* **42**, 87–94.

Sung, Y.J., Thompson, E.A. and Wijsman, E.M. (2006). MCMC-based multipoint linkage analysis for two loci plus a polygenic component and general pedigrees. *Genetic Epidemiology* **31**, 103–114.

Thomas, A., Gutin, A. and Abkevich, V. (2000). Multilocus linkage analysis by blocked Gibbs sampling. *Statistics and Computing* **10**, 259–269.

Thompson, E.A. (1974). Gene identities and multiple relationships. *Biometrics* **30**, 667–680.

Thompson, E.A. (1994). Monte Carlo likelihood in genetic mapping. *Statistical Science* **9**, 355–366.

Thompson, E.A. (2000). *Statistical Inferences from Genetic Data on Pedigrees*, *Vol. 6 of NSF-CBMS Regional Conference Series in Probability and Statistics*, Institute of Mathematical Statistics, Beachwood, OH.

Thompson, E.A. (2005). Chapter 4: MCMC in the analysis of genetic data on pedigrees. In *Markov Chain Monte Carlo: Innovations and Applications*, F. Liang, J.-S. Wang and W. Kendall, eds. World Scientific Co Pvt Ltd, Singapore, pp. 183–216.

Thompson, E.A. and Geyer, C.J. (2007). Fuzzy p-values in latent variable problems. *Biometrika* **94**, 49–60.

Thompson, E.A. and Guo, S.W. (1991). Evaluation of likelihood ratios for complex genetic models. *I.M.A. Journal of Mathematics Applied in Medicine and Biology* **8**(3), 149–169.

Thompson, E.A. and Heath, S.C. (1999). Estimation of conditional multilocus gene identity among relatives. In *Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology*, *IMS Lecture Note–Monograph Series*, Vol. 33, F. Seillier-Moiseiwitsch, ed. Institute of Mathematical Statistics, Hayward, CA, pp. 95–113.

Thompson, E.A., Kravitz, K., Hill, J. and Skolnick, M.H. (1978). Linkage and the power of a pedigree structure. In *Genetic Epidemiology*, N.E. Morton, ed. Academic Press, New York, pp. 247–253.

Thompson, E.A. and Meagher, T.R. (1998). Genetic linkage in the estimation of pairwise relationship. *Theoretical and Applied Genetics* **97**, 857–864.

Wald, A. (1947). *Sequential Analysis*. John Wiley & Sons, New York.

Weeks, D.E. and Lange, K. (1988). The affected pedigree member method of linkage analysis. *American Journal of Human Genetics* **42**, 315–326.

Weeks, D.E., Lathrop, G.M. and Ott, J. (1993). Multipoint mapping under genetic interference. *Human Heredity* **43**(2), 86–97.

Weinstein, A. (1936). The theory of multiple-strand crossing over. *Genetics* **21**, 155–199.

Wijsman, E.M. and Amos, C.I. (1997). Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. *Genetic Epidemiology* **14**, 719–735.

Zhao, H., Speed, T.P. and McPeek, M.S. (1995). Statistical analysis of crossover interference using the chi-square model. *Genetics* **139**(2), 1045–1056.

Non-parametric Linkage

P. Holmans

Department of Psychological Medicine, Cardiff University, Cardiff, UK

Model-free methods are commonly used to detect linkage to both dichotomous and quantitative traits. This review begins by discussing the principles underlying model-free methods for detecting linkage, and their merits relative to traditional lod-score methods (see also **Chapter 33**). Affected sib-pair methods and their extensions to include other typed unaffected relatives and extra affected siblings are discussed, together with model-free methods for the analysis of dichotomous traits on larger pedigrees. Extensions of the methods for the analysis of covariates, multiple marker loci and disease loci are reviewed, together with an investigation of methods for meta-analysis of genome scans. Finally, methods for the analysis of quantitative traits are briefly discussed.

34.1 INTRODUCTION

The basis of the initial (Penrose, 1935) and all subsequent model-free tests for linkage is that individuals concordant for a given genetic trait should show greater than expected concordance for another trait (or marker) if the two traits are linked. This is because concordant pairs are likely to have inherited the same alleles at the first trait locus, and are therefore likely to have inherited the same alleles at the second trait locus, if this is linked to the first locus. They are thus likely to be concordant for the second trait also. The expected degree of concordance if the traits are not linked can be calculated, and a statistical test performed.

The most commonly used measure of concordance of two individuals at a locus is the number of alleles they share *identical by descent* (*ibd*). This is illustrated for a pair of siblings in Figure 34.1. If the marker is not linked to a disease susceptibility locus, then, provided the Mendelian laws of inheritance hold, the probabilities of a sib pair sharing 0, 1 or 2 alleles ibd are 1/4, 1/2, 1/4 respectively. If the marker is linked to a disease locus, the probability of an affected sib pair sharing alleles ibd is increased. This is because an affected sib pair is likely to share (disease) alleles at the disease susceptibility locus, and therefore also at a linked marker locus. This is illustrated in Figure 34.2 for a recessive trait. The expected ibd sharing probabilities depend on the mode of inheritance. For a
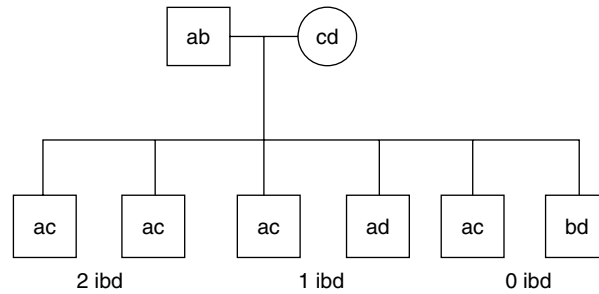
Figure 34.1 Two individuals are said to share an allele identical by descent if they both inherit a copy of that allele from a common ancestor.
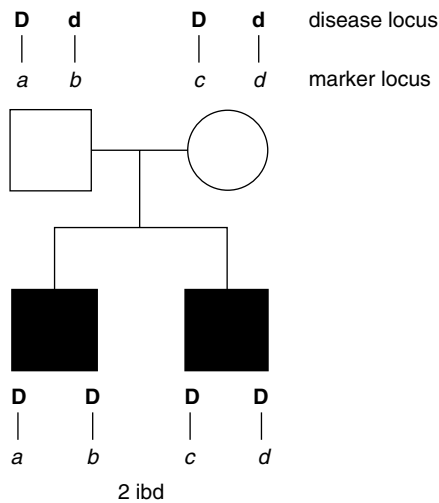
Figure 34.2 Two related affected individuals will exhibit increased ibd sharing at a marker locus linked to a disease susceptibility locus.

dominant or additive trait, the probability of 2 ibd increases above 1/4, but the probability of 1 ibd stays close to 1/2. For a recessive trait, the probability of 2 ibd increases, but the probability of 1 ibd decreases from 1/2 (Suarez, 1978). Concordant unaffected individuals should also exhibit increased ibd sharing. However, the increase is much smaller for such pairs, making them relatively uninformative (Suarez *et al.*, 1978), so in practice, such pairs are not included in linkage analyses of dichotomous traits. For a review of sib-pair methods for dichotomous traits, see also Holmans (1998). A more general overview of model-free methods is provided by Elston and Cordell (2001).

34.2 PROS AND CONS OF MODEL-FREE METHODS

A major advantage of model-free methods over the traditional lod-score approach (Ott, 1974) is that they do not require specification of the disease model. Lod-score methods

are more powerful if the model is known accurately, but their power is reduced if an incorrect model is used, and the resulting estimate of the recombination fraction may be biased (Clerget-Darpoux *et al.*, 1986). A common procedure is to perform the analysis under a number of different disease models, which involves multiple testing and necessitates adjustment of the significance level of the lod score (MacLean *et al.*, 1993; Risch, 1991). An alternative is to maximise the lod score over (single-locus) models. For example, Curtis and Sham (1995) vary the penetrance of the heterozygote in their program MFLINK. Greenberg *et al.* (1998) recommend performing analyses with a dominant and a recessive model, each with 50 % penetrance, and taking the larger lod score. However, the performance of such methods when the true disease model involves several interacting loci is unclear. Some model-free methods also make implicit assumptions regarding the disease model (Whittemore, 1996). For example, the mean test has been shown to be equivalent to a lod-score analysis assuming a recessive model (Knapp *et al.*, 1994). However, they require only one test to be performed, avoiding multiple testing and making it easier to interpret the significance of the test statistics.

Another advantage is that the large multiply affected pedigrees for which traditional lod-score approaches are most informative are relatively rare, particularly for complex traits with late onset. However, large samples of relatively small families may still be available. Indeed, it has been suggested (Risch, 1994) that, for multilocus traits, sib pairs may be even more informative for linkage than multiply affected pedigrees. This is because the presence of a large number of affecteds in a pedigree increases the chance that one or more of the founders was homozygous for the disease allele at the locus being analysed, which will reduce the amount of ibd sharing. In practice, the decision regarding the correct method of analysis will be influenced by the structures of the available families. For complex multifactorial traits, where model-free methods may be advantageous, the majority of the sample will be affected sib pairs or other small pedigrees. For single-locus traits, however, large pedigrees containing multiple affecteds will be more likely, suggesting the use of parametric analyses. A major advantage of model-free methods in the analysis of complex traits is that it is relatively easy to include covariates, since their effect on disease penetrance need not be explicitly specified.

Model-free linkage methods also have some drawbacks – they do not give estimates of the recombination fraction, and are thus of limited use for mapping genes, although this defect may be overcome to some extent by the use of multipoint analyses. Neither do they explicitly allow for heterogeneity in the analysis, resulting in greatly reduced power when heterogeneity is present. Heterogeneity in linkage between *samples* (e.g. samples collected in different countries) can be overcome to some extent. For example, the ibd sharing probability can be modelled as a logistic function including a covariate with different levels for each sample (Rice, 1997; Levinson *et al.*, 2000). However, it is not possible to model heterogeneity in linkage between *individual pedigrees*, unlike model-based methods.

The conclusion seems to be that model-free analyses such as affected sib-pair methods are most advantageous in analysing complex multilocus traits where the mode of inheritance is unclear, particularly if the genetic model involves environmental or clinical covariates. For single-locus, high-penetrance traits, a lod-score analysis on large pedigrees will be more powerful. Issues raised in the analysis of complex traits are more fully reviewed by Lander and Schork (1994) and Weeks and Lathrop (1995), and the relative merits of model-free and model-based linkage analyses are discussed by Elston (1998).

It is important to note that the power of all linkage methods applied to dichotomous traits depends on the frequency of the trait in the population, decreasing as this frequency increases (Blangero *et al.*, 1998). For a common trait, it might be advantageous to determine a related quantitative measure and analyse that instead. In particular, one should not attempt to dichotomise a quantitative trait by applying a threshold, since this may greatly reduce power (Blangero *et al.*, 1998; Duggirala *et al.*, 1997).

34.3 MODEL-FREE METHODS FOR DICHOTOMOUS TRAITS

34.3.1 Affected Sib-pair Methods

The simplest approach is to count up the number of pairs sharing 0, 1 and 2 ibd and compare these to the expected frequencies under the hypothesis of no linkage using a χ^2 test (Cudworth and Woodrow, 1975). Alternatively, one could compare the average number of alleles shared ibd by the affected pairs (the 'mean test') or the number of pairs sharing 2 alleles ibd with their expected values. In each case, an excess of ibd sharing is taken as evidence for linkage. The powers of these 'counting' methods were investigated by Blackwelder and Elston (1985). Although the relative efficiencies of the methods depend on the disease model, the mean test was found to perform well in many situations.

All the above methods assume that the ibd status of the sib pairs can be determined unequivocally. This is often impossible, except for highly polymorphic markers such as the HLA system, even when the parents are typed. For an example, see Figure 34.3. Simply omitting pairs for which the ibd status is not known with certainty is likely to bias the results against detecting linkage, since it is easier to unequivocally determine pairs sharing 0 alleles ibd than pairs sharing 1 or 2 ibd (for sib pairs with no typed relatives, it is *only* possible to determine 0 sharers unequivocally). The mean test has been adapted to deal with this problem by estimating the ibd score as the average of the ibd scores under the various possible parental genotype combinations, weighted by their probabilities, and
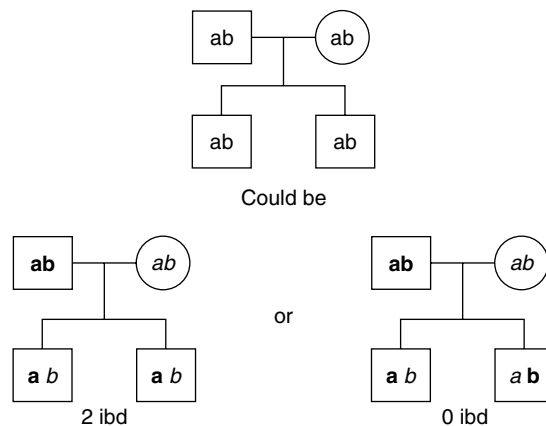
Figure 34.3 Two alleles which look the same are not necessarily shared ibd.

is implemented in packages such as SAGE (2006). Whilst the mean test is optimal under an additive genetic model, it can lose power in certain situations (e.g. under a recessive model). Whittemore and Tu (1998) considered the range of 1 degree of freedom (df) tests defined by varying the weight assigned to pairs sharing one allele ibd, and proposed the 'minimax test', which counts the 1 sharers as 0.55, rather than 1 as in the means test.

Another commonly used technique is the likelihood-ratio method proposed by Risch (1990b; 1990c). Here, the likelihood of the observed marker data is expressed as a linear function of the (unknown) ibd probabilities. The likelihood is then maximised with respect to these probabilities, and a likelihood-ratio test performed. The power of this method can be increased by restricting the maximisation to the set of ibd probabilities consistent with possible genetic models (Holmans, 1993), i.e. $\Pr(0\,\text{ibd}) \leq 1/4$, $2\Pr(0\,\text{ibd}) \leq \Pr(1\,\text{ibd})$. For a more detailed description, see Risch (1990b; 1990c) and Holmans (1993). If the dominance variance component of the genetic model is small, holding $\Pr(1\,\text{ibd}) = 0.5$ may improve the power still further (Lunetta and Rogus, 1998). The method is implemented in the packages SPLINK (Holmans and Clayton, 1995), MAPMAKER/SIBS (Kruglyak and Lander, 1995) and ASPEX (Hinds and Risch, 1996). Risch has also extended the method to other pairs of relatives (Risch, 1990b).

Another approach to the problems caused by incomplete polymorphic marker loci and/or untyped parents is to consider the number of alleles shared *identical by state* (*ibs*). Two individuals are said to share an allele ibs if they both have a copy of that allele, regardless of from whom it was inherited. It follows that the ibs status of a pair of typed individuals can always be calculated with certainty, no matter how uninformative the marker. A number of ibs methods have been suggested, of which the most notable is the affected-pedigree-member (APM) method of Weeks and Lange (1988). However, they are generally less powerful than ibd methods, especially when parental data is available (Davis and Weeks, 1997), and are thus not in common usage nowadays.

Note that the methods described above are designed for use on *autosomal* chromosomes. Variations of the methods are available for X-linked data (e.g. Cordell *et al.*, 1995a) and are implemented in most analysis software. Particular care must be taken for *pseudoautosomal* data (Dupuis and Van Eerdewegh, 2000).

34.3.2   Parameter Estimation and Power Calculation Using Affected Sib Pairs

The ibd sharing probabilities may be expressed in terms of the parameters of the genetic model in a number of different ways. The observed ibd sharing probabilities can then be used to obtain estimates of the parameters, or to infer the mode of inheritance. A major drawback of using affected sib pairs for parameter estimation is that at most two parameters may be estimated uniquely, since there are only 2 df in the data (3 ibd probabilities which are constrained to sum to 1). It is therefore necessary to make assumptions about the other parameters. For example, Thomson and Bodmer (1977) assumed that the penetrance of the normal trait genotype was zero, in order to test whether the trait exhibited recessive inheritance. This assumption was also made by Thomson (1986) to test recessive vs additive inheritance, and to estimate disease gene frequency. Day and Simons (1976) used sib pairs to estimate both disease gene frequency and relative risk, but had to specify the mode of inheritance in advance. With data from larger groups of affected siblings, it is possible to estimate more parameters. Payami *et al.* (1985) used affected trios to estimate disease gene frequency in addition to the relative

risks associated with both heterozygous and homozygous disease genotypes, and to test various modes of inheritance.

The ability of sib-pair data to yield reliable estimates of genetic parameters may also be influenced by other factors, such as the method of ascertainment (Risch, 1983; Olson and Cordell, 2000), selection (Payami *et al.*, 1984) and the presence of other susceptibility loci (Risch, 1983). For these reasons, it is advisable to be cautious when using affected sib pairs for estimating genetic model parameters, although it is possible to correct for ascertainment bias (Cordell and Olson, 2000).

Although it may not be possible to estimate model parameters from the sib-pair sample itself, they may be available from previous studies. In such situations, the expected values of the ibd sharing may be calculated, and the power of the study estimated. Formulae which express the ibd probabilities in terms of quantities which may be estimated from population data are particularly useful, since power calculations may then be made without the need for prior segregation analyses. Such formulae were provided by Suarez *et al.* (1978), who expressed the ibd probabilities in terms of the prevalence and the additive and dominance variances, and Risch (1990a), who used the recurrence risks in various types of relatives of a proband. Power calculations using Risch's parameterisation are given by Weeks and Lathrop (1995). Caution should be exercised when using estimates of ibd sharing or sibling recurrence risks estimated directly from genome-wide scans since these are prone to upward bias (Goring *et al.*, 2001). Correcting fully for this bias is difficult, although partial solutions are available (Sun and Bull, 2005).

34.3.3 Typing Unaffected Relatives in Sib-pair Analyses

The power of any sib-pair method is increased by increasing the amount of information available regarding the ibd status of the affected pairs. For this reason, typing the parents increases power, especially for relatively unpolymorphic markers (Risch, 1990c; 1992; Holmans and Clayton, 1995). If parents are unavailable, a degree of information about their genotypes may be obtained by typing unaffected siblings. Again, the increase in power is greatest when the information content of the marker is low. For the likelihood-ratio method, typing both parents was found to reduce the sample size needed for a given power by at most one-third, while doubling the amount of genotyping. Typing just one parent or one unaffected sib reduced the sample size by at most one-sixth while increasing the amount of genotyping by one-half (Holmans and Clayton, 1995). It can therefore be seen that typing unaffected relatives actually *increases* the number of individuals who must be genotyped to obtain a given power. It should be noted that genotyping unaffected siblings will be necessary if one wishes to analyse discordant sibling pairs (e.g. Xing *et al.*, 2006).

Typing unaffected relatives has some other major advantages. If there are a limited number of affected pairs, it may be important to get the maximum amount of information out of each pair. Most importantly, if both parents are typed, the probability of obtaining false-positive linkage evidence will not be increased by population stratification, Hardy−Weinberg disequilibrium at the marker locus, consanguinity in the population or incorrect specification of marker allele frequencies. Typing unaffected relatives increases the chance of detecting genotyping errors in the family (Douglas *et al.*, 2002), thereby increasing power (since genotyping errors tend to reduce power). Inappropriate correction for genotyping errors may also inflate Type I error rate (Seaman and Holmans, 2005). Typing unaffected siblings enables the assumption of Mendelian inheritance to be tested,

by investigating the distribution of alleles shared ibd in discordant pairs (Khoury *et al.*, 1991). This depends on the disease penetrance; for low penetrance one would expect ibd sharing close to the expected values, while for high penetrance one might observe an excess of pairs sharing 0 ibd. In any case, the proportion of alleles shared ibd by discordant pairs should be less than for affected pairs. If discordant pairs are found to show an excess of ibd sharing, then the Mendelian assumption may be false (e.g due to transmission-ratio distortion), and the usual sib-pair tests invalid. In this situation, one possibility is to compare haplotype sharing in concordant and discordant pairs (Flanders and Khoury, 1991; Wang and Elston, 2005). Typing unaffected siblings may also increase power substantially if there is assortative mating (Sribney and Swift, 1992).

For these reasons, it is desirable to genotype unaffected relatives, particularly both parents. However, affected pairs without relatives available for typing should not be discarded (Holmans and Clayton, 1995). If some of the parents in the sample are missing, the analysis should be repeated on the subset of pairs with both parents typed, to check that any positive result is not a result of incorrect assumptions about the marker locus.

If parental data is missing, all ibd methods require the specification of marker allele frequencies. It is possible to use frequencies obtained from a previous study, or a database such as CEPH; however, if these are incorrect for the population being studied, the false-positive rate may be increased (Ott, 1992; Freimer *et al.*, 1993). A safer procedure is to estimate the marker allele frequencies from the sample itself. This eliminates the risk of using frequencies estimated from a different population, and does not lose much power compared to the situation where the true frequencies are known (Holmans, 1993).

34.3.4   Application of Sib-pair Methods to Multiplex Sibships

The methods described above were originally designed for use on affected sibling *pairs*. In practice, it is likely that the sample collected for an affected sib-pair study will contain a number of sibships containing three or more affected siblings. There are a number of ways to analyse such sibships. For some methods, sibships of any size can be analysed without modification. An example is the binomial maximum-likelihood approach of Abel *et al.* (1998), which models the number of affected siblings inheriting a given parental allele as a binomial distribution, enabling a likelihood-ratio test to be performed.

However, many commonly used methods, such as the likelihood-ratio methods of Risch (1990b; 1990c), restrict the analysis to sib pairs only. Data from multiply affected sibships must therefore be broken up. The most common ways to do this, as implemented in GENEHUNTER (Kruglyak *et al.*, 1996) are: (1) to choose just one affected pair from each sibship (2) to choose one affected sibling and include all pairs containing this sibling in the analysis (3) to include all possible sib pairs from the sibship in the analysis. For example, an affected trio of sibs A, B and C would contribute the affected pair $A - B$ under (1), the two pairs $A - B$ and $A - C$ under (2), and the three pairs $A - B$, $A - C$ and $B - C$ under (3). The main problem with choosing just one sib pair from each sibship is that a large proportion of the data is discarded, reducing power. Furthermore, the result may depend on which pair is chosen, giving rise to the possibility, in extreme cases, of two researchers applying the same analysis to the same dataset, but reaching different conclusions. The same drawbacks apply to the method of analysing all pairs containing a given individual.

It is therefore usual to include all possible affected pairs from each sibship in the analysis. However, pairs obtained in this way from the same sibship are no longer

independent. A number of weighting schemes for such pairs have been suggested (e.g. Hodge, 1984) to deal with this problem. The most commonly used weighting scheme is that of Suarez and Hodge (1979), in which pairs formed from a sibship containing n affected individuals are weighted by a factor of $2/n$. This scheme has the intuitively appealing result that the total weight assigned to pairs formed from a n-affected sibship is $n - 1$ ($= n(n - 1)/2$ pairs $\times 2/n$), which is the number of *independent* pairs which could be formed from the sibship.

In fact, the lack of independence does not inflate the false-positive rate for 'counting' methods such as the mean test. This is because the number of alleles shared ibd by two pairs of affected sibs are independent, even if the sib pairs share a common member (Blackwelder and Elston, 1985). The variance of the test statistics is therefore unaltered by the dependence among pairs from the same sibship. For the likelihood-ratio method, however, there is evidence that analysing pairs from the same sibship as if they were independent may alter the false-positive rate. If the genotypes of the affected sibs in the sibship which are not part of the pair being analysed are used to infer missing parental genotypes (as is the case for programs such as MAPMAKER/SIBS), the test is conservative even without weighting (Holmans, 2001). Otherwise, the absence of parental genotypes increased the false-positive rate if no weighting was applied. However, the $2/n$ weighting scheme of Suarez and Hodge is too conservative. These effects become more pronounced as marker informativeness decreases. If parents are genotyped, there is a slight inflation of false-positive rate without weighting, most pronounced for small nominal p values (≤ 0.001). The $2/n$ weighting scheme of Suarez and Hodge was found to be conservative in all situations.

The behaviour of likelihood-ratio methods when applied to multiplex sibships depends on both family structure and marker informativeness, so no single weighting scheme can be correct in all circumstances. Thus, it is advisable to estimate the significance of observed results by simulation, rather than relying on pre-defined test criteria based on asymptotic theory, when the sample contains a high proportion of multiply affected sibships.

Even when the method of analysis does not require the use of weighting to ensure the correct false-positive rate, it may still be possible to increase power by assigning different weights to pairs formed from sibships containing different numbers of affected siblings (Suarez and Van Eerdewegh, 1984; Sham *et al.*, 1997).

34.3.5 Methods for Analysing Larger Pedigrees

The APM method was proposed by Weeks and Lange (1988) for the model-free linkage analysis of general pedigrees. It compares the sum of the observed number of alleles shared ibs by each pair of affected relatives, weighted according to the frequency of the shared allele(s), to its expected value in the absence of linkage. The method was extended to include unaffected relatives by Ward (1993).

One drawback of the method was that the results were dependent on the (arbitrary) choice of the weighting function. Another was that, since the APM method used only ibs information, it was less powerful than methods which inferred ibd, especially when other relatives were typed (Goldin and Weeks, 1993). Therefore, Davis *et al.* (1996) developed SimIBD, which counts the ibd between affected pairs and computes an empirical p value using conditional simulation. This method was found to perform poorly when analysing

sibships without typed parents (Davis and Weeks, 1997), although it may be useful for larger pedigrees.

An alternative approach, proposed by Kruglyak *et al.* (1996), performs a non-parametric test (NPL) based on the observed inheritance patterns of the affected individuals, and is implemented in the program GENEHUNTER. The significance levels of the NPL statistic calculated by GENEHUNTER assume complete ibd information, and the test is conservative when this is not the case (Kruglyak *et al.*, 1996), resulting in a loss of power. This may make the NPL statistics less efficient than other methods for analysing sib-pair data (Davis and Weeks, 1997), although it appears to be a powerful method of non-parametric analysis on large pedigrees. The problem of the conservativeness of the NPL statistics has been addressed by Kong and Cox (1997) in their program GENEHUNTER-PLUS, and is also implemented in ALLEGRO (Gudbjartsson *et al.*, 2005). Two scoring functions are commonly available for NPL-type statistics: S_{pairs}, which is based on numbers of alleles shared ibd by *pairs* of affected relatives, and S_{all}, which is based on ibd patterns among all affecteds in a pedigree. S_{pairs} is equivalent to the mean test when applied to sibship data. These and other scoring functions were reviewed by McPeek (1999). The optimal scoring function depends on the (unknown) disease model, but S_{pairs} was found to perform well over all models tested.

The weighted pairwise correlation (WPC) method (Commenges, 1994) is also designed to perform non-parametric analysis on arbitrarily large pedigrees. Its basic premise is that, under linkage, the correlation of the residuals (i.e. the observed trait value minus its expected value) of a pair of related individuals will increase with the number of alleles shared ibd. The WPC method can be applied to either quantitative or dichotomous traits provided the form of the residuals is suitably chosen.

34.3.6 Extensions to Multiple Marker Loci

The power of all methods for detecting linkage increases with the informativeness of the marker, since this enables the ibd status of the affected individuals to be determined more accurately. Extra information regarding ibd sharing may be obtained by extending the methods to analyse two or more linked marker loci simultaneously, thereby increasing power in both dichotomous (Holmans and Clayton, 1995; Olson, 1995) and quantitative (Fulker and Cardon, 1994) traits. A further advantage of analysing multiple linked markers is that a greater area of the chromosome can be studied, and an estimate of the location of the disease locus provided, although this may not always be very precise, even for large samples, when the genetic effect is small for complex traits (Roberts *et al.*, 1999, Cordell, 2001). The likelihood-ratio method of Risch and Holmans has been extended to multiple linked markers in the program GENEHUNTER (Kruglyak *et al.*, 1996), which also allows multipoint NPL and parametric analyses of larger pedigrees. The maximum size of pedigree that GENEHUNTER can analyse is determined by the quantity $2n - f$, where n is the number of individuals in the pedigree and f the number of founders, and is limited by the memory of the computer. If the pedigree is too large, either individuals must be removed or the pedigree must be split, and in some cases the way this is done may lose power (Romero-Hidalgo *et al.*, 2005). The ALLEGRO package (Gudbjartsson *et al.*, 2005) performs GENEHUNTER and GENEHUNTER-PLUS analyses, but is faster and capable of handling larger pedigrees, as is MERLIN (Abecasis *et al.*, 2002). However, very large pedigrees must still be split, or model-based analyses performed using programs such as FASTLINK (Cottingham *et al.*, 1993), although these are restricted in the number

of marker loci they can analyse. Alternatively, Monte-Carlo techniques, such as those implemented in SIMWALK2 (Sobel and Lange, 1996; Sobel *et al.*, 2002) may be used.

34.3.7   Multipoint Analysis with Tightly Linked Markers

It should be noted that all programs for multipoint analysis assume the marker loci to be in linkage equilibrium when calculating haplotype frequencies. The presence of linkage disequilibrium might therefore lead to spurious results in the absence of typed parents. This is only likely to be a problem if the marker loci are close together (< 1 cM), which was rarely the case for grids of microsatellite markers. Recent advances in genotyping techniques have made it feasible to perform genome-wide linkage studies using dense maps of SNPs. These give more linkage information than the microsatellite grids, and thus can increase power (Schaid *et al.*, 2004). However, linkage disequilibrium (LD) between the SNPs can also result in false-positive linkages (Schaid *et al.*, 2004), and the extent to which Type I error rate is inflated by inter-marker LD has been investigated by several authors (Huang *et al.*, 2004; 2005; Boyles *et al.*, 2005; Levinson and Holmans, 2005).

One approach to dealing with inter-marker LD is to measure the LD between all pairs of SNPs, and then remove SNPs until the LD between all remaining pairs of SNPs is below a pre-set criterion. The choice of criterion differs from study to study: Schaid *et al.* used $|D'| < 0.7$, whereas Levinson and Holmans used $r^2 < 0.06$. The optimal criterion for controlling Type I error rate while losing as little linkage information as possible is still unclear (and is likely to depend on the structure of the pedigrees – in particular, the number of typed parents).

An alternative approach has been implemented in recent versions of MERLIN (Abecasis and Wigginton, 2005). Here, consecutive SNPs are grouped into clusters such that all pairs of SNPs within a cluster satisfy a user-selected criterion based either on LD (e.g. $r^2 < 0.1$) or genetic distance (e.g. less than 0.1 cM). Haplotype frequencies are calculated within clusters without assuming linkage equilibrium, while different clusters are assumed to be in linkage equilibrium. Recombination within clusters is assumed to be zero. If a recombination is observed, data from the entire cluster in that pedigree is dropped from the analysis, which could reduce power. The relative performances of these approached under varying LD structures have not as yet been assessed.

If unaffected siblings are typed, an analysis comparing ibd sharing in affected sib pairs to that in discordant sib pairs can be performed, and this is robust to inter-marker LD (Xing *et al.*, 2006). However, such an approach is less powerful than a study of affected sib pairs alone (Blackwelder and Elston, 1985).

34.3.8   Inclusion of Covariates

Since there is often epidemiological evidences showing that measurable environmental or clinical factors influence disease risk, it is desirable to allow for the effects of these factors as covariates in a model-free linkage analysis. One approach is to use logistic regression. Greenwood and Bull (1997; 1999) propose modelling the probability that a sib pair shares 0, 1 or 2 ibd as a multinomial logistic regression, into which parameters modelling covariate effects can be incorporated. A similar approach has been taken by Olson (1999). How best to apply constraints to the ibd probabilities when maximising the likelihood, as the constraints may not necessarily hold for all covariate values is one issue with this approach. Olson *et al.* (2001; 2002) recommend a linear constraint based on the

results of Whittemore and Tu (1998). Alternatively, one can model the probability that an affected sib pair share a given parental allele ibd as a logistic regression, and to assume that the ibd probabilities of maternal and paternal alleles are independent (Rice, 1997; Rice *et al.*, 1999). This avoids the constraint problems, and also requires fewer degrees of freedom, but may lose power if the true genetic model violates the independence assumption.

Another approach is to consider the sample as a mixture of linked and unlinked sib pairs, in a similar way to heterogeneity analysis in parametric linkage (Devlin *et al.*, 2002). The probability that a sib pair is linked can be modelled in terms of the covariates (e.g. by logistic regression). Whilst this approach is efficient for modelling linkage heterogeneity, it does not allow pairs to share fewer alleles ibd than expected under the null (as might happen under certain gene – environment interaction models where the pairs are discordant for the environmental factor).

All of the methods described above operate on one sib pair at a time, which may cause problems when applying them to multiplex sibships, as noted earlier. The binomial likelihood method of Abel *et al.* (1998) has been extended (Alcais and Abel, 2001) to incorporate covariates by performing a logistic regression of disease status on the covariates, then analysing the residuals as a quantitative trait (Alcais and Abel, 1999). This method deals easily with multiplex sibships, but would not be able to analyse covariates defined only on *pairs* of sibs (e.g. ibd sharing at a second locus).

A final approach, applicable to general pedigrees, is ordered subsets analysis (Hauser *et al.*, 2004). Here, pedigrees are ranked in order of some covariate (e.g. earliest age at onset). Pedigrees are added to the sample in rank order, and the linkage analysis repeated after each. The largest of the resulting linkage statistics is taken as the overall test statistics, with significance estimated by randomly permuting the order in which pedigrees are added. Again, this method would not be able to analyse covariates defined on pairs of sibs. Also, it is limited to the analysis of one covariate at a time, unlike the other approaches. However, it makes no assumptions (other than monotonicity) on the relationship of linkage evidence with covariate, so may be advantageous if this is non-(log)linear.

Power comparisons of these methods were carried out by Tsai and Weeks (2006) under a variety of models incorporating gene – environment interactions. Pairwise/familywise covariates were defined as the average of the covariates of the individuals involved. This raises the question of whether other definitions of pairwise covariate (e.g. differences) might be superior. This has not yet been studied.

In theory, one can analyse several covariates simultaneously in the logistic-regression and mixture model approaches. In practice, power may be reduced due to the large number of degrees of freedom corresponding to the regression parameters. An interesting way of dealing with this problem is the propensity score method (Doan *et al.*, 2006). Here, logistic regression of disease status is performed on all covariates simultaneously. Residuals for the affecteds are used as covariates in a logistic-regression linkage analysis. A drawback of this approach is that sufficient numbers of unaffected siblings must be genotyped and measured for the covariates.

Issues regarding the inclusion of covariates in linkage analysis are further discussed by Schaid *et al.* (2003).

34.3.9 Multiple Disease Loci

For many complex traits, there is evidence that more than one locus is involved in the aetiology, e.g.: schizophrenia (Risch, 1990a), bipolar disorder (Craddock *et al.*, 1995) and type 1 diabetes (Davies *et al.*, 1994). Schork *et al.* (1993) extended the lod-score approach to analyse two trait loci simultaneously, including two marker loci, one linked to each trait locus, in the analysis. In certain situations, they found a substantial increase in power over standard analysis involving just one trait locus. However, if just one marker locus is used, including a second trait locus in the analysis does not greatly increase power (Goldin, 1992; Vieland *et al.*, 1992). Knapp *et al.* (1994) found that sib-pair analysis applied to two marker loci, each linked to a different trait locus, could be much more powerful than analysing the loci separately. The maximum-likelihood method used by Knapp *et al.* was refined by Cordell *et al.* (1995b) to include the restrictions on ibd sharing proposed by Holmans (1993), and used not only to test for linkage with type 1 diabetes, but also to distinguish between models of interaction between the disease loci. This approach was extended to multiple *linked* trait loci by Farrall (1997). Cordell *et al.* (2000) extended the method further to allow likelihood models based on variance components to be fitted to ibd data from pairs of affected relatives from extended pedigrees at an arbitrary number of loci, linked or unlinked. This enables the evidence for a disease locus to be tested while conditioning on the presence of other loci, and for hypotheses regarding epistatic interactions to be tested, but it is unclear how other covariates could be included. The logistic-regression method of Olson (1999) is capable of performing two-locus analyses (Olson *et al.*, 2002). Alternatively, the proportion of alleles shared ibd by each pair at a second locus may be included as a covariate in the Rice model (Holmans, 2002).

Another approach to testing for epistasis was proposed by Cox *et al.* (1999). The principle underlying this approach is that linkage statistics (e.g. NPL score, ibd proportion) for each sampling unit (pedigree, affected sib pair) at two loci A and B should be positively correlated if A and B interact epistatically, but negatively correlated if the loci conform to a heterogeneity model. Then, one can select the families whose test statistics at locus A is greater (if testing for epistasis) or less (if testing for heterogeneity) than a pre-defined criterion, and perform analysis at locus B using these families only. Significance is estimated by randomly sampling replicate sets containing the same number of families used in the actual analysis without reference to their linkage score at locus A. This method was used by Cox *et al.* to detect interactions between susceptibility loci for diabetes. This is an intuitively appealing procedure, but it is often by no means clear *a priori* what criterion should be used in the analysis. Whilst the issue of multiple testing caused by the use of several criteria giving non-independent tests was addressed by Buhler *et al.* (1997), allowing for this may reduce power. Furthermore, it is not clear how to apply this method when loci A and B are linked, since one would then expect the linkage scores to be positively correlated whatever the underlying disease model.

A problem with using a two-trait-locus approach is that the number of possible pairs of marker loci rapidly becomes large, leading to multiple testing problems. For this reason, such approaches probably ought not to be used in an initial genome scan, but rather to test sets of candidate loci, or to follow-up loci to which (possibly weak) evidence of linkage has been found using single trait-locus analysis (Schork *et al.*, 1993). Furthermore, when only linkage data (rather than a candidate gene) is available, it is unlikely that a two-trait-locus analysis could give a significant lod without some evidence of linkage being visible from single-locus analyses (Holmans, 2002).

34.3.10   Significance Levels for Genome Scans

Advances in genotyping technology mean that genome scans for linkage are now very common. However, they still involve a large number of linkage tests, and it is thus necessary to apply a correction for multiple testing to reduce the number of false positives. Simple corrections such as Bonferroni are inappropriate due to the non-independence of these tests. Lander and Kruglyak (1995) devised theoretical criteria for linkage statistics to be deemed genome-wide 'significant' (i.e. obtained in fewer than 5 % of genome scans in the absence of disease loci) or 'suggestive' (obtained less than once per genome scan). These criteria depend on the analysis method. Whilst these criteria are useful rules of thumb, they are based on the assumption of complete linkage information, and are conservative in more realistic situations (Sawcer *et al.*, 1997). In addition, as noted earlier, the distribution of linkage statistics may also depend on the number of affected sibs, the number of typed parents, and the inclusion of covariates. Thus, significance levels for observed results should be obtained by simulation.

34.3.11   Meta-analysis of Genome Scans

Genetic effects underlying complex traits are likely to be small. Thus, large-sample sizes will be required to achieve power, often more than can be collected by any single group. Ideally, several groups would pool their raw genotype and phenotype data and perform a single analysis on the combined sample (e.g. Levinson *et al.*, 2000). Often, however, raw data is unavailable, but summary measures of individual studies (e.g. linkage statistics) can be obtained. In such situations, studies can be combined via a meta-analysis.

One method for performing meta-analysis (Badner and Goldin, 1999; Badner and Gershon, 2002) is based on Fisher's method for combining p values (i.e. $-2\sum_{i=1}^{n} \ln p_i \sim \chi^2_{2n}$ if none of the studies is a true positive). Zaykin *et al.* (2002) extended the Fisher method to use only p values below a certain threshold the truncated product method (TPM), which may be useful if studies only report significant regions. Some linkage analysis methods (e.g. likelihood-ratio sib-pair analysis) truncate the test statistics at zero, which can bias the Fisher method. This bias was overcome by Province (2001).

Another commonly used meta-analysis method is the genome search meta-analysis (GSMA) proposed by Wise *et al.* (1999). Here, the genome is split into a number of 'bins' of approximately equal width. Within each study, bins are ranked according to the most significant test statistics occurring in that bin. Ranks for each bin are summed across studies and compared to the expected distribution in the absence of susceptibility genes (i.e. ranks assigned randomly). This gives a bin-wise p value p_{bin}. Correction for testing multiple bins may be performed by randomly permuting ranks within each study. This yields a p value p_{ord} depending on the position of the bin in the ranked p_{bin} values (e.g. most significant, second-most significant etc.) Combining p_{bin} and p_{ord} was shown to increase power to detect true linkages, as was weighting by sample size (Levinson *et al.*, 2003). The GSMA has been applied to numerous complex traits, e.g. schizophrenia (Lewis *et al.*, 2003), and software is now available (Pardi *et al.*, 2005).

The final class of methods involves combining Z scores (i.e. linkage statistics with a standard normal distribution, such as NPL scores) across studies (Loesgen *et al.*, 2001). Statistics which do not have a standard normal distribution (such as lod scores) can be transformed into Z scores (Etzel *et al.*, 2005). Weighting schemes can be applied, such as by study size or linkage information content (Dempfle and Loesgen, 2004).

All of these methods are generally applicable, since they do not require that the studies use the same linkage analysis method, or the same set of markers (although clearly the different maps must be combined such that their locations agree). The performance of the Z-score, Fisher, TPM and GSMA methods was compared on simulated data by Dempfle and Loesgen (2004). The Z-score method was found to be most powerful when all studies were simulated under the same linkage model. However, if there is linkage heterogeneity, the TPM method may be preferable (Zaykin *et al.*, 2002), while the power of the GSMA is reduced (Levinson *et al.*, 2003). Methodolgy has been developed to test for linkage heterogeneity in the GSMA (Zintzaras and Ioannidis, 2005), although its power is questionable (Lewis and Levinson, 2006).

34.4 MODEL-FREE METHODS FOR ANALYSING QUANTITATIVE TRAITS

This topic is also covered in **Chapter 19**, and thus only a brief overview of the most commonly used methods for linkage analysis of quantitative traits in humans will be given.

A commonly used method for analysing quantitative traits is the Haseman–Elston method (Haseman and Elston, 1972). In this method, the squared trait differences between pairs of siblings is regressed on the proportion of alleles that the pair is estimated to share their ibd with. This method retains large-sample validity if all distinct sib pairs from sibships with three or more members are included in the analysis as independent pairs (Amos *et al.*, 1989; Collins and Morton, 1995), does not require the trait to be normally distributed, and allows the testing of multiple trait loci, including interactions, by multiple linear regression. It was pointed out (Wright, 1997) that using the sib-pair difference ignores information available in the sib-pair trait sum. The method was therefore 'revisited' to use the mean-corrected cross-product of trait values instead (Elston *et al.*, 2000), thereby increasing power in some situations. This method has been implemented in the program SIBPAL2, part of the SAGE package. Power may be increased further by regressing the sum of the squared trait sum and squared trait difference, each weighted inversely proportional to their variance, of each pair on ibd. This is equivalent to variance-components analysis on sib-pair data (Sham and Purcell, 2001).

The other major approach to analysing quantitative traits is to use variance-components analysis (Goldgar, 1990; Amos, 1994; Amos *et al.*, 1996; Blangero and Almasy, 1996). Here, the covariance matrix for the trait is partitioned into components attributable to additive genetic variance at each of the trait loci, residual genetic effects and random environmental effects. The size of each effect is estimated, and its significance tested by means of a likelihood-ratio test. The main advantage of this approach is its flexibility: it is applicable to general pedigrees, it enables the effects of covariates, epistasis between multiple trait loci and gene – environment interactions to be modelled (Blangero and Almasy, 1997). Multiple traits may be analysed simultaneously, both quantitative and qualitative (Almasy *et al.*, 1997; Williams *et al.*, 1999). The method appears to be more powerful than the original (Williams and Blangero, 1997) and 'revisited' (Xu *et al.*, 2000). Haseman–Elston methods, but requires that the traits being analysed have a normal distribution, with non-normality inflating Type I error rates (Allison *et al.*, 1999).

Applying transformations (e.g. Box – Cox) to the data in an attempt to approximate normality appears not to be efficient at correcting the Type I error rate, so estimating significance by simulation is recommended (Deutsch *et al.*, 2005). Variance-components analysis is implemented in the SOLAR package (Almasy and Blangero, 1998), and in MERLIN.

An interesting alternative approach (Sham *et al.*, 2002) is to regress the estimated ibd sharing of affected relative pairs on the squared sums and differences of their trait values (i.e. reversing the direction of regression used by Haseman–Elston and variance-components methods). This has the potential to combine the power of the variance-components approach with robustness to non-normality (e.g. it can be applied to samples selected on their trait values). However, the mean and variance of the trait in the population must be specified. The method is implemented in MERLIN-REGRESS.

34.5   CONCLUSIONS

When the mode of inheritance is known, such as for high-penetrance Mendelian traits, model-based maximum-likelihood methods applied to multigenerational pedigrees are the preferred form of analysis, due to their extra power. However, model-free methods do have some advantages over traditional model-based parametric analyses for the detection of linkage to complex traits. Firstly, they do not require specification of a disease model, thereby evading the problem of multiple testing caused by analysing a number of different (incorrect) models. Secondly, large, multigenerational pedigrees containing multiple affected members are usually rare for complex traits, especially those with late onset. Small pedigrees, such as nuclear families or affected sib pairs are relatively common, enabling large samples to be collected. Model-free methods may be advantageous in the analyses of such pedigrees. Whilst model-free methods do not allow direct estimation of the recombination fraction, estimates of the location of the disease locus are available by using multipoint analysis.

Given the small locus-specific genetic effects observed in complex traits, together with likely genetic heterogeneity, it is important that future linkage analyses incorporate clinical subphenotypes and/or environmental risk factors as covariates. Model-free methods are very suitable for this, as they do not require the relationship between the covariate and the disease model to be explicitly specified. Numerous methods have been proposed, but optimal procedures have yet to be determined.

Small genetic effect sizes also require large-sample sizes to achieve power. This, in turn, increases the importance of efficient meta-analysis methods, and there is scope for future work on these.

To conclude: carefully designed linkage studies making appropriate use of clinical and/or epidemiological data, and information from other studies, still have an important role to play in genetic epidemiology. Their results can be useful in informing the design of follow-up association studies, and samples of multiply affected pedigrees can be a powerful resource for such studies.

Related Chapters

Chapter 19; **Chapter 33**.

REFERENCES

Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101.

Abecasis, G.R. and Wigginton, J.E. (2005). Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *American Journal of Human Genetics* **77**, 754–767.

Abel, L., Alcais, A. and Mallet, A. (1998). Comparison of four sib-pair linkage methods for analysing sibships with more than two affecteds: interest of the binomial maximum likelihood approach. *Genetic Epidemiology* **15**, 371–390.

Alcais, A. and Abel, L. (1999). Maximum-Likelihood-Binomial method for genetic model-free linkage analysis of quantitative traits in sibships. *Genetic Epidemiology* **17**, 102–117.

Alcais, A. and Abel, L. (2001). Incorporation of covariates in multipoint model-free linkage analysis of binary traits: how important are unaffecteds? *European Journal of Human Genetics* **9**, 613–620.

Allison, D.B., Neale, M.C., Zannolli, R., Schork, N.J., Amos, C.I. and Blangero, J. (1999). Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *American Journal of Human Genetics* **65**, 531–544.

Almasy, L. and Blangero, J. (1998). Multipoint quantitative trait linkage analysis in general pedigrees. *American Journal of Human Genetics* **62**, 1198–1211.

Almasy, L., Dyer, T.D. and Blangero, J. (1997). Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. *Genetic Epidemiology* **14**, 953–958.

Amos, C.I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics* **54**, 535–543.

Amos, C.I., Elston, R.C., Wilson, A.F. and Bailey-Wilson, J.E. (1989). A more powerful robust sib-pair test of linkage for quantitative traits. *Genetic Epidemiology* **6**, 435–449.

Amos, C.I., Zhu, D.K. and Boerwinkle, E. (1996). Assessing genetic linkage and association with robust components of variance approaches. *Annals of Human Genetics* **60**, 143–160.

Badner, J.A. and Gershon, E.S. (2002). Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Molecular Psychiatry* **7**, 405–411.

Badner, J.A. and Goldin, L.R. (1999). Meta-analysis of linkage studies. *Genetic Epidemiology* **17**(Suppl. 1), S485–S490.

Blackwelder, W.C. and Elston, R.C. (1985). A comparison of sib-pair linkage tests for disease susceptibility loci. *Genetic Epidemiology* **2**, 85–97.

Blangero, J. and Almasy, L. (1997). Multipoint oligogenic linkage analysis of quantitative traits. *Genetic Epidemiology* **14**, 959–964.

Blangero, J., Almasy, L. and Williams, J.T. (1998). Tragedy of the commons: common misconceptions about mapping genes for common diseases. *American Journal of Human Genetics* **63**, A45.

Boyles, A.L., Scott, W.K., Martin, E.R., Schmidt, S., Li, Y.J., Ashley-Koch, A., Bass, M.P., Schmidt, M., Pericak-Vance, M.A., Speer, M.C. and Hauser, E.R. (2005). Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. *Human Heredity* **59**, 220–227.

Buhler, J., Owerbach, D., Schaffer, A.A., Kimmel, M. and Gabbay, K.H. (1997). Linkage analyses in type I diabetes using CASPAR, a software and statistical program for conditional analysis of polygenic diseases. *Human Heredity* **47**, 211–222.

Clerget-Darpoux, F., Bonaiti-Pellie, C. and Hochez, J. (1986). Effects of misspecifying genetic parameters in lod score analysis. *Biometrics* **42**, 393–399.

Collins, A. and Morton, N.E. (1995). Nonparametric tests for linkage with dependent sib pairs. *Human Heredity* **47**, 333–353.

Commenges, D. (1994). Robust genetic linkage analysis based on a score test of homogeneity: the weighted pairwise correlation statistic. *Genetic Epidemiology* **11**, 189–200.

Cordell, H.J. (2001). Sample size requirements to control for stochastic variation in magnitude and location of allele-sharing linkage statistics in affected sibling pairs. *Annals of Human Genetics* **65**, 491–502.

Cordell, H.J., Kawaguchi, Y., Todd, J.A. and Farrall, M. (1995a). An extension of the maximum lod score method to X-linked loci. *Annals of Human Genetics* **59**, 435–449.

Cordell, H.J. and Olson, J.M. (2000). Correcting for ascertainment bias of relative-risk estimates obtained using affected-sib-pair linkage data. *Genetic Epidemiology* **18**, 307–321.

Cordell, H.J., Todd, J.A., Bennett, S.T., Kawaguchi, Y. and Farrall, M. (1995b). Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes. *American Journal of Human Genetics* **57**, 920–934.

Cordell, H.J., Wedig, G.C., Jacobs, K.B. and Elston, R.C. (2000). Multilocus linkage tests based on affected relative pairs. *American Journal of Human Genetics* **66**, 1273–1286.

Cottingham, R.W. Jr., Idury, R.M. and Schaffer, A.A. (1993). Faster sequential genetic linkage computations. *American Journal of Human Genetics* **53**, 252–263.

Cox, N.J., Frigge, M., Nicolae, D.L., Concannon, P., Hanis, C.L., Bell, G.I. and Kong, A. (1999). Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genetics* **21**, 213–215.

Craddock, N., Khodel, V., Van Eerdewegh, P. and Reich, T. (1995). Mathematical limits of multilocus models: the genetic transmission of bipolar disorder. *American Journal of Human Genetics* **57**, 690–702.

Cudworth, A.G. and Woodrow, J.C. (1975). Evidence for HLA-linked genes in "juvenile" diabetes mellitus. *British Medical Journal* **3**, 133–135.

Curtis, D. and Sham, P.C. (1995). Model-free linkage analysis using likelihoods. *American Journal of Human Genetics* **57**, 703–716.

Davies, J.L., Kawaguchi, Y., Bennett, S.T., Copeman, J.B., Cordell, H.J., Pritchard, L.E., Reed, P.W., Gough, S.C.L., Jenkins S.C., Palmer, S.M., Balfour, K.M., Rowe, B.R., Farrall, M., Barnett, A.H., Bain, S.C., Todd, J.A. (1994). A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371**, 130–136.

Davis, S., Schroeder, M., Goldin, L.R. and Weeks, D.E. (1996). Nonparametric simulation-based statistics for detecting linkage in general pedigrees. *American Journal of Human Genetics* **58**, 867–880.

Davis, S. and Weeks, D.E. (1997). Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation. *American Journal of Human Genetics* **61**, 1431–1444.

Day, N.E. and Simons, M.J. (1976). Disease susceptibility genes-their identification by multiple case family studies. *Tissue Antigens* **8**, 109–117.

Dempfle, A. and Loesgen, S. (2004). Meta-analysis of linkage studies for complex diseases: an overview of methods and a simulation study. *Annals of Human Genetics* **68**, 69–83.

Deutsch, S., Lyle, R., Dermitzakis, E.T., Attar, H., Subrahmanyan, L., Gehrig, C., Parand, L., Gagnebin, M., Rougemont, J., Jongeneel, C.V. and Antonarakis, S.E. (2005). Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. *Human Molecular Genetics* **14**, 3741–3749.

Devlin, B., Jones, B.L., Bacanu, S.A. and Roeder, K. (2002). Mixture models for linkage analysis of affected sibling pairs and covariates. *Genetic Epidemiology* **22**, 52–65.

Doan, B.Q., Sorant, A.J., Frangakis, C.E., Bailey-Wilson, J.E. and Shugart, Y.Y. (2006). Covariate-based linkage analysis: application of a propensity score as the single covariate consistently improves power to detect linkage. *European Journal of Human Genetics* **14**, 1018–1026.

Douglas, J.A., Skol, A.D. and Boehnke, M. (2002). Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *American Journal of Human Genetics* **70**, 487–495.

Dupuis, J. and Van Eerdewegh, P. (2000). Multipoint linkage analysis of the pseudoautosomal regions, using affected sibling pairs. *American Journal of Human Genetics* **67**, 462–475.

Duggirala, R., Williams, J.T., Williams-Blangero, S. and Blangero, J. (1997). A variance component approach to dichotomous trait linkage analysis using a threshold model. *Genetic Epidemiology* **14**, 987–992.

Elston, R.C. (1998). Linkage and association. *Genetic Epidemiology* **15**, 565–576.

Elston, R.C., Buxbaum, S., Jacobs, K.B. and Olson, J.M. (2000). Haseman and Elston revisited. *Genetic Epidemiology* **19**, 1–17.

Elston, R.C. and Cordell, H.J. (2001). Overview of model-free methods for linkage analysis. *Advances in Genetics* **42**, 135–150.

Etzel, C.J., Liu, M. and Costello, T.J. (2005). An updated meta-analysis approach for genetic linkage. *BMC Genetics* **6**(Suppl. 1), S43.

Farrall, M. (1997). Affected sibpair linkage tests for multiple linked susceptibility genes. *Genetic Epidemiology* **14**, 103–115.

Flanders, W. and Khoury, M. (1991). Extensions to methods of sib-pair linkage analysis. *Genetic Epidemiology* **8**, 399–408.

Freimer, N., Sandkuijl, L. and Blower, S. (1993). Incorrect specification of marker allele frequencies: effects on linkage analysis. *American Journal of Human Genetics* **52**, 1102–1110.

Fulker, D.W. and Cardon, L.R. (1994). A sib-pair approach to interval mapping of quantitative trait loci. *American Journal of Human Genetics* **54**, 1092–1103.

Goldgar, D.E. (1990). Multipoint analysis of human quantitative genetic variation. *American Journal of Human Genetics* **47**, 957–967.

Goldin, L.R. (1992). Detection of linkage under heterogeneity: comparison of two-locus vs. admixture models. *Genetic Epidemiology* **9**, 61–66.

Goldin, L.R. and Weeks, D.E. (1993). Two-locus models of disease: comparison of likelihood and nonparametric linkage methods. *American Journal of Human Genetics* **53**, 908–915.

Goring, H.H., Terwilliger, J.D. and Blangero, J. (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics* **69**, 1357–1369.

Greenberg, D.A., Abreu, P. and Hodge, S.E. (1998). The power to detect linkage in complex disease by means of simple LOD-score analyses. *American Journal of Human Genetics* **63**, 870–879.

Greenwood, C.M.T. and Bull, S.B. (1997). Incorporation of covariates into genome scanning using sib pair analysis in bipolar affective disorder. *Genetic Epidemiology* **14**, 635–640.

Greenwood, C.M.T. and Bull, S.B. (1999). Analysis of affected sib pairs, with covariates – with and without constraints. *American Journal of Human Genetics* **64**, 871–885.

Gudbjartsson, D.F., Thorvaldsson, T., Kong, A., Gunnarsson, G. and Ingolfsdottir, A. (2005). Allegro version 2. *Nature Genetics* **37**, 1015–1016.

Haseman, J.K. and Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* **2**, 3–19.

Hauser, E.R., Watanabe, R.M., Duren, W.L., Bass, M.P., Langefeld, C.D. and Boehnke, M. (2004). Ordered subset analysis in genetic linkage mapping of complex traits. *Genetic Epidemiology* **27**, 53–63.

Hinds, D.A., Risch, N. (1996). *The ASPEX package: affected sib-pair exclusion mapping.* `http://aspex.sourceforge.net/`.

Hodge, S.E. (1984). The information contained in multiple sibling pairs. *Genetic Epidemiology* **1**, 109–122.

Holmans, P. (1993). Asymptotic properties of affected sib-pair linkage analysis. *American Journal of Human Genetics* **52**, 362–374.

Holmans, P. and Clayton, D. (1995). Efficiency of typing unaffected relatives in an affected sib-pair linkage study with single locus and multiple tightly-linked markers. *American Journal of Human Genetics* **57**, 1221–1232.

Holmans, P. (1998). Affected sib-pair methods for detecting linkage to dichotomous traits: a review of the methodology. *Human Biology* **70**, 1025–1040.

Holmans, P. (2001). Likelihood-ratio affected sib-pair tests applied to multiply affected sibships: issues of power and type I error rate. *Genetic Epidemiology* **20**, 44–56.

Holmans, P. (2002). Detecting gene-gene interactions using affected sib pair analysis with covariates. *Human Heredity* **53**, 92–102.

Huang, Q., Shete, S. and Amos, C.I. (2004). Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *American Journal of Human Genetics* **75**, 1106–1112.

Huang, Q., Shete, S., Swartz, M. and Amos, C.I. (2005). Examining the effect of linkage disequilibrium on multipoint linkage analysis. *BMC Genetics* **6**(Suppl. 1), S83.

Khoury, M., Flanders, W., Lipton, R. and Dorman, J. (1991). The affected sib-pair method in the context of an epidemiologic study design. *Genetic Epidemiology* **8**, 277–282.

Knapp, M., Seuchter, S.A. and Baur, M.P. (1994). Linkage analysis in nuclear families. II. Relationship between affected sib-pair tests and lod score analysis. *Human Heredity* **44**, 44–51.

Kong, A. and Cox, N.J. (1997). Allele sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics* **61**, 1179–1188.

Kruglyak, L., Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics* **57**, 439–454.

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* **58**, 1347–1363.

Lander, E.S. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for reporting linkage results. *Nature Genetics* **11**, 241–247.

Lander, E.S. and Schork, N.J. (1994). Genetic dissection of complex traits. *Science* **265**, 2037–2048.

Levinson, D.F. and Holmans, P. (2005). The effect of linkage disequilibrium on linkage analysis of incomplete pedigrees. *BMC Genetics* **6**(Suppl. 1), S6.

Levinson, D.F., Holmans, P., Straub, R.E., Owen, M.J., Wildenauer, D.B., Gejman, P.V., Pulver, A.E., Laurent, C., Kendler, K.S., Walsh, D., Norton, N., Williams, N.M., Schwab, S.G., Lerer, B., Mowry, B.J., Sanders, A.R., Antonarakis, S.E., Blouin, J.L., DeLeuze, J.F. and Mallet, J. (2000). Multicenter linkage study of schizophrenia candidate regions on chromosomes 5q, 6q, 10p and 13q: schizophrenia linkage collaborative group III. *American Journal of Human Genetics* **67**, 652–663.

Levinson, D.F., Levinson, M.D., Segurado, R. and Lewis, C.M. (2003). Genome scan meta-analysis of schizophrenia and bipolar disorder, part I: methods and power analysis. *American Journal of Human Genetics* **73**, 17–33.

Lewis, C.M. and Levinson, D.E. (2006). Testing for genetic heterogeneity in the genome search meta-analysis method. *Genetic Epidemiology* **30**, 348–355.

Lewis, C.M., Levinson, D.F., Wise, L.H., DeLisi, L.E., Straub, R.E., Hovatta, I., Williams, N.M., Schwab, S.G., Pulver, A.E., Faraone, S.V., Brzustowicz, L.M., Kaufmann, C.A., Garver, D.L., Gurling, H.M., Lindholm, E., Coon, H., Moises, H.W., Byerley, W., Shaw, S.H., Mesen, A., Sherrington, R., O'Neill, F.A., Walsh, D., Kendler, K.S., Ekelund, J., Paunio, T., Lonnqvist, J., Peltonen, L., O'Donovan, M.C., Owen, M.J., Wildenauer, D.B., Maier, W., Nestadt, G., Blouin, J.L., Antonarakis, S.E., Mowry, B.J., Silverman, J.M., Crowe, R.R., Cloninger, C.R., Tsuang, M.T., Malaspina, D., Harkavy-Friedman, J.M., Svrakic, D.M., Bassett, A.S., Holcomb, J., Kalsi, G., McQuillin, A., Brynjolfson, J., Sigmundsson, T., Petursson, H., Jazin, E., Zoega, T. and Helgason, T. (2003). Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *American Journal of Human Genetics* **73**, 34–48.

Loesgen, S., Dempfle, A., Golla, A. and Bickeboller, H. (2001). Weighting schemes in pooled linkage analysis. *Genetic Epidemiology* **21**(Suppl. 1), S142–S147.

Lunetta, K.L. and Rogus, J.J. (1998). Strategy for mapping minor histocompatibility genes involved in graft-versus-host disease: a novel application of discordant sib pair methodology. *Genetic Epidemiology* **15**, 595–607.

MacLean, C., Bishop, D.T., Sherman, S. and Diehl, S. (1993). Distribution of lod scores under uncertain mode of inheritance. *American Journal of Human Genetics* **52**, 354–361.

McPeek, M.S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology* **16**, 225–249.

Olson, J.M. (1995). Multipoint linkage analysis using sib pairs: an interval mapping approach for dichotomous outcomes. *American Journal of Human Genetics* **56**, 788–798.

Olson, J.M. (1999). A general conditional-logistic model for affected relative pair linkage studies. *American Journal of Human Genetics* **65**, 1760–1769.

Olson, J.M. and Cordell, H.J. (2000). Ascertainment bias in the estimation of sibling genetic risk parameters. *Genetic Epidemiology* **18**, 217–235.

Olson, J.M., Goddard, K.A. and Dudek, D.M. (2001). The amyloid precursor protein locus and very-late-onset Alzheimer disease. *American Journal of Human Genetics* **69**, 895–899.

Olson, J.M., Goddard, K.A. and Dudek, D.M. (2002). A second locus for very-late-onset Alzheimer disease: a genome scan reveals linkage to 20p and epistasis between 20p and the amyloid precursor protein region. *American Journal of Human Genetics* **71**, 154–161.

Ott, J. (1974). Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *American Journal of Human Genetics* **26**, 588–597.

Ott, J. (1992). Strategies for characterizing highly polymorphic markers in human gene mapping. *American Journal of Human Genetics* **51**, 283–290.

Pardi, F., Levinson, D.F. and Lewis, C.M. (2005). GSMA: software implementation of the genome search meta-analysis method. *Bioinformatics* **21**, 4430–4431.

Payami, H., Thomson, G. and Louis, E. (1984). The affected sib method. III. Selection and recombination. *American Journal of Human Genetics* **36**, 352–362.

Payami, H., Thomson, G., Motro, U., Louis, E. and Hudes, E. (1985). The affected sib method. IV. Sib trios. *Annals of Human Genetics* **49**, 303–314.

Penrose, L. (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Annals of Eugenics* **6**, 133–138.

Province, M.A. (2001). The significance of not finding a gene. *American Journal of Human Genetics* **69**, 660–663.

Rice, J.P. (1997). The role of meta-analysis in linkage studies of complex traits. *American Journal of Medical Genetics* **74**, 112–114.

Rice, J.P., Rochberg, N., Neuman, R.J., Saccone, N.L., Liu, K.Y., Zhang, X. and Culverhouse, R. (1999). Covariates in linkage analysis. *Genetic Epidemiology* **17**(Suppl. 1), S703–S708.

Risch, N. (1983). The effects of reduced fertility, method of ascertainment, and a second unlinked locus on affected sib-pair marker allele sharing. *American Journal of Medical Genetics* **16**, 243–259.

Risch, N. (1990a). Linkage strategies for genetically complex traits: I. Multilocus models. *American Journal of Human Genetics* **46**, 222–228.

Risch, N. (1990b). Linkage strategies for genetically complex traits: II. The power of affected relative pairs. *American Journal of Human Genetics* **46**, 229–241.

Risch, N. (1990c). Linkage strategies for genetically complex traits: III. The effect of marker polymorphism on analysis of affected relative pairs. *American Journal of Human Genetics* **46**, 242–253.

Risch, N. (1991). A note on multiple testing procedures in linkage analysis. *American Journal of Human Genetics* **48**, 1058–1064.

Risch, N. (1992). Corrections to "Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs". *American Journal of Human Genetics* **51**, 673–675.

Risch, N. (1994). Mapping genes for psychiatric disorders. In *Genetic Approaches to Mental Disorders*, E.S. Gershon, C.R. Cloninger and J.E. Barrett, eds. American Psychiatric Press, Washington, DC, pp. 47–61.

Roberts, S.B., MacLean, C.J., Neale, M.C., Eaves, L.J. and Kendler, K.S. (1999). Replication of linkage studies of complex traits: an examination of variation in location estimates. *American Journal of Human Genetics* **65**, 876–884.

Romero-Hidalgo, S., Rodrigues, E.R., Gutierrez-Pena, E., Riba, L. and Tusie-Luna, M.T. (2005). GENEHUNTER versus SimWalk2 in the context of an extended kindred and a qualitative trait locus. *Genetica* **123**, 235–244.

SAGE (2006). *Statistical Analysis for Genetic Epidemiology* 5.2.0 Computer package available from the Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, MetroHealth campus. Case Western Reserve University, Cleveland, OH. `http://darwin.cwru.edu`.

Sawcer, S., Jones, H.B., Judge, D., Visser, F., Compston, D.A.S., Goodfellow, P.N. and Clayton, D. (1997). Empirical genomewide significance levels established by whole genome simulations. *Genetic Epidemiology* **14**, 223–229.

Schaid, D.J., Guenther, J.C., Christensen, G.B., Hebbring, S., Rosenow, C., Hilker, C.A., McDonnell, S.K., Cunningham, J.M., Slager, S.L., Blute, M.L., Thibodeau, S.N. (2004). Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility Loci. *American Journal of Human Genetics* **75**, 948–965.

Schaid, D.J., Olson, J.M., Gauderman, W.J. and Elston, R.C. (2003). Regression models for linkage: issues of traits, covariates, heterogeneity, and interaction. *Human Heredity* **55**, 86–96.

Schork, N.J., Boehnke, M., Terwilliger, J.D. and Ott, J. (1993). Two -trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *American Journal of Human Genetics* **53**, 1127–1136.

Seaman, S.R. and Holmans, P. (2005). Effect of genotyping error on type-I error rate of affected sib pair studies with genotyped parents. *Human Heredity* **59**, 157–164.

Sham, P.C. and Purcell, S. (2001). Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *American Journal of Human Genetics* **68**, 1527–1532.

Sham, P.C., Purcell, S., Cherny, S.S. and Abecasis, G.R. (2002). Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American Journal of Human Genetics* **71**, 238–253.

Sham, P.C., Zhao, J.H. and Curtis, D. (1997). Optimal weighting scheme for affected sib-pair analysis of sibship data. *Annals of Human Genetics* **61**, 61–69.

Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *American Journal of Human Genetics* **58**, 1323–1337.

Sobel, E., Papp, J.C. and Lange, K. (2002). Detection and integration of genotyping errors in statistical genetics. *American Journal of Human Genetics* **70**, 496–508.

Sribney, W.M. and Swift, M. (1992). Power of sib-pair and sib-trio linkage analysis with assortative mating and multiple disease Loci. *American Journal of Human Genetics* **51**, 773–784.

Suarez, B.K. (1978). The affected sib pair IBD distribution for HLA-linked disease susceptibility genes. *Tissue Antigens* **12**, 87–93.

Suarez, B.K. and Hodge, S.E. (1979). A simple method to detect linkage for rare recessive diseases: an application to juvenile diabetes. *Clinical Genetics* **15**, 126–136.

Suarez, B.K., Rice, J. and Reich, T. (1978). The generalized sib pair IBD distribution: its use in the detection of linkage. *Annals of Human Genetics* **42**, 87–94.

Suarez, B.K. and Van Eerdewegh, P. (1984). A comparison of three affected sib-pair scoring methods to detect HLA-linked disease susceptibility genes. *American Journal of Medical Genetics* **18**, 135–146.

Sun, L. and Bull, S.B. (2005). Reduction of selection bias in genomewide studies by resampling. *Genetic Epidemiology* **28**, 352–367.

Thomson, G. (1986). Determining the mode of inheritance of RFLP-associated diseases using the affected sib-pair method. *American Journal of Human Genetics* **39**, 207–221.

Thomson, G. and Bodmer, W. (1977). The genetic analysis of HLA and disease associations. In *HLA and Disease*, J. Dausset and A. Svejgaard, eds. Munksgaard, Copenhagen, pp. 84–93.

Tsai, H.J. and Weeks, D.E. (2006). Comparison of methods incorporating quantitative covariates into affected sib pair linkage analysis. *Genetic Epidemiology* **30**, 77–93.

Vieland, V.J., Hodge, S.E. and Greenberg, D.A. (1992). Adequacy of single-locus approximations for linkage analysis of oligogenetic traits. *Genetic Epidemiology* **9**, 45–59.

Wang, T. and Elston, R.C. (2005). The bias introduced by population stratification in IBD based linkage analysis. *Human Heredity* **60**, 134–142.

Ward, P. (1993). Some developments on the affected pedigree-member method of linkage analysis. *American Journal of Human Genetics* **52**, 1200–1215.

Weeks, D.E. and Lange, K. (1988). The affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics* **42**, 315–326.

Weeks, D.E. and Lathrop, G.M. (1995). Polygenic disease: methods for mapping complex disease traits. *Trends in Genetics* **11**, 513–519.

Whittemore, A.S. (1996). Genome scanning for linkage: an overview. *American Journal of Human Genetics* **59**, 704–716.

Whittemore, A.S. and Tu, I.-P. (1998). Simple, robust linkage tests for affected sibs. *American Journal of Human Genetics* **62**, 1228–1242.

Williams, J.T. and Blangero, J. (1997). Comparison of variance-components and sibpair methods for quantitative trait linkage analysis in sibships and nuclear families. *Genetic Epidemiology* **14**, 543.

Williams, J.T., Van Eerdewegh, P., Almasy, L. and Blangero, J. (1999). Joint multipoint linkage analysis of multivariate quantitative traits. I. Likelihood formulation and simulation results. *American Journal of Human Genetics* **65**, 1134–1147.

Wise, L.H., Lanchbury, J.S. and Lewis, C.M. (1999). Meta-analysis of genome searches. *Annals of Human Genetics* **63**, 263–272.

Wright, F.A. (1997). The phenotypic difference discards sib-pair QTL linkage information. *American Journal of Human Genetics* **60**, 740–742.

Xing, C., Sinha, R., Xing, G., Lu, Q. and Elston, R.C. (2006). The affected-/discordant-sib-pair design can guarantee validity of multipoint model-free linkage analysis of incomplete pedigrees when there is marker-marker disequilibrium. *American Journal of Human Genetics* **79**, 396–340.

Xu, X., Weiss, S., Xu, X. and Wei, L.J. (2000). A unified Haseman-Elston method for testing linkage with quantitative traits. *American Journal of Human Genetics* **67**, 1025–1028.

Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H. and Weir, B.S. (2002). Truncated product method for combining P-values. *Genetic Epidemiology* **22**, 170–185.

Zintzaras, E. and Ioannidis, J.P. (2005). Heterogeneity testing in meta-analysis of genome searches. *Genetic Epidemiology* **28**, 123–137.

Population Admixture and Stratification in Genetic Epidemiology

P.M. McKeigue

Conway Institute, University College Dublin, Dublin, Ireland

Population admixture and stratification generally occur together. Admixture between subpopulations generates allelic associations that decay with map distance, whereas stratification generates allelic associations that are independent of map distance. The autocorrelation of ancestry on gametes inherited from parents of mixed descent can be exploited to localize genes in which the pool of disease risk alleles is differentially distributed between subpopulations. Tests for linkage are based on testing for association of the disease or outcome of interest with locus ancestry, conditioning on parental admixture proportions to eliminate confounding by genetic background. This approach, known as *admixture mapping*, is an extension of the principles underlying linkage analysis of an experimental cross between inbred strains. Most current approaches to modelling admixture are based on a standard model in which the stochastic variation of ancestry on gametes inherited from admixed parents is generated by independent Poisson arrival processes. For such models, the posterior distribution of locus ancestry and parental admixture proportions can be generated by Markov chain Monte Carlo simulation, given a sample of individuals typed at ancestry-informative marker loci. Tests for linkage are constructed by averaging over this posterior distribution. The same statistical model can be used with unselected marker loci to detect and control for population stratification in ordinary genetic association studies. For studies using arrays of closely spaced tag SNPs, this approach has limitations as it requires that the marker loci be spaced far enough apart for all allelic association to be attributable to admixture and stratification. An alternative approach is to use principal components analysis to infer a few underlying axes of variation that summarize the allelic associations in the dataset. This approach is computationally efficient, and can be extended to datasets with closely spaced markers using a simple adjustment for short-range allelic associations. Tests of the number of underlying latent variables can be constructed; for a given F_{ST} distance between subpopulations, there is a threshold size of dataset at which stratification can be detected. With either modelling approach, confounding by stratification can be controlled by adjusting for

genetic background when testing for allelic association; alternative 'genomic control' approaches based on correcting the variance of the test statistic has serious limitations.

35.1 BACKGROUND

Population admixture and stratification are often considered together, because they usually occur together and because the two phenomena can be modelled with similar statistical methods. Admixture between subpopulations with different allele frequencies generates gametes that consist of a mosaic of segments inherited from each of the ancestral subpopulations. On a gamete inherited from a parent of mixed descent, there will be autocorrelation of ancestry states: the shorter the distance between two loci, the higher the probability that the subpopulation of ancestry will be the same at these two loci. Where allele frequencies vary between ancestral subpopulations, this autocorrelation of ancestry gives rise to allelic association that decays with distance. This makes it possible to infer recent admixture from data on a sample of individuals typed at linked marker loci. Where admixture is recent, the distances over which this association is detectable are far longer than the few hundred kilobases (kb) over which haplotype block structure is detectable.

Genetic stratification, on the other hand, generates allelic associations that are independent of map distance. In principle, it is possible for genetic stratification to occur without admixture: one possible scenario is a total population consisting of discrete endogamous subpopulations, with strong social barriers to mating between subpopulations. Another possible scenario for stratification without admixture is a geographic cline of allele frequencies, with no migration between regions. Such extreme situations would usually be obvious to a researcher, without the need to infer stratification from genetic data. In a stratified population with no admixture, the strength of allelic associations is independent of map distance beyond the very short distances (typically a few hundred kb in human populations) over which haplotype structure is detectable. At the other extreme, it is possible in principle for an admixed population to have no stratification, so that admixture proportions – the proportions of the individual's genome that have ancestry from each ancestral subpopulation – are the same in all individuals. A possible scenario for this would be an island population where an initial pulse of admixture has been followed by isolation, with no new gene flow, and mating that has not been assortative with respect to individual admixture proportions. Usually, stratification and recent admixture occur together. Stratification in admixed populations can be maintained by continuing gene flow from one or more of the unadmixed ancestral subpopulations, or by assortative mating resulting from socioeconomic stratification on the basis of genetic background. See Excoffier (**Chapter 29**) and Beaumont (**Chapter 30**) for further discussion of population stratification and admixture.

Admixture and stratification present both opportunities and challenges to genetic epidemiology. There are opportunities to exploit the genetic structure of admixed and stratified populations for more efficient approaches to mapping and identification of disease susceptibility loci. The challenges are to control for hidden population stratification as a confounder in genetic association studies, and to control for the long-range associations generated by admixture when undertaking fine mapping of a disease locus. This chapter begins by discussing the theory of admixture mapping, which exploits the genetic

structure of admixed populations to localize genes that underlie variation in disease risk
between ethnic groups.

35.2 ADMIXTURE MAPPING

35.2.1 Basic Principles

Because the allelic associations generated by admixture decay with map distance,
admixture generates information about linkage. Early suggestions for exploiting this
information were based on testing the allelic associations generated by admixture in a
manner similar to classical linkage disequilibrium mapping of a Mendelian disease locus
in an isolated population (Chakraborty and Weiss, 1988; Stephens *et al.*, 1994). With this
approach, the information available is limited by the magnitude of allelic associations,
which depend upon the size of the allele frequency differentials between the ancestral
subpopulations. Admixture mapping, in contrast, uses the observed allelic associations
to infer the underlying states of locus ancestry. This can be viewed as an extension,
to outbred admixed populations, of the principles underlying linkage analysis of an
experimental cross between inbred strains (see **Chapter 18**). In family linkage studies, it
is the segregation indicators at each locus, not the marker alleles, that convey information
about linkage (see **Chapter 33**). In admixture mapping, it is the locus ancestry states, not
the marker alleles, that convey information about linkage. When exploiting admixture to
localize genetic effects, the marker genotypes are irrelevant once all available information
about locus ancestry has been extracted.

To exploit the information about linkage that is generated by admixture, it is necessary
to model the underlying stochastic variation of ancestry on chromosomes inherited from
admixed parents. To extract information about ancestry, it is desirable to select markers
that are informative for ancestry in that they have large allele frequency differentials
between the ancestral populations.

Admixture mapping is complementary to association studies and family linkage studies:
each has the objective of identifying genes that underlie between-population variation in
disease risk. Where admixture mapping can be applied, it has several advantages in
comparison with linkage or association study designs.

1. To detect loci that generate between-population variation in disease risk, statistical
    power is higher for admixture mapping than for family linkage studies. For example,
    to detect a locus that underlies a twofold risk ratio between populations by admixture
    mapping, the required sample size is much lower than that required to detect a locus
    that underlies a twofold sibling recurrence risk ratio in an affected sib-pair study.
    The low statistical power of family linkage studies to detect loci of modest effect is
    a fundamental property of this study design (Lander and Schork, 1994). To explain
    this, consider a study in which the disease locus itself has been typed. The most direct
    test for an effect at this locus is to test for association of outcome with genotype.
    Family linkage analysis of a single pedigree can be viewed as an association study
    in which each gene copy in each founder has been coded as a separate allele, and
    the association of outcome with the genotypes defined by these 'alleles' is examined
    within that pedigree. The direction of association (defined by the coding of founder

copies as 'alleles') is fixed within pedigrees, but random between pedigrees. Unless the pedigrees are very large and the number of founders is small, tests that combine information across pedigrees have low statistical power because the direction of association is integrated out of the likelihood function. In admixture mapping, by contrast, the direction of association between outcome and locus ancestry is fixed across all individuals; analysis of an experimental cross between inbred strains can be viewed as linkage analysis of a single large sibship.

2. The required marker density is much lower for admixture mapping than for genome-wide association studies. Genome-wide admixture mapping studies typically require about 2000 ancestry-informative markers, compared with at least 300 K for genome-wide association studies. Now that high-density genotyping arrays are relatively inexpensive, this is no longer a key advantage.

3. Admixture mapping, like family linkage studies, is not dependent upon the assumption of low allelic heterogeneity, as Terwilliger and Weiss (1998) have pointed out. Association studies using haplotype–tagging SNPs depend critically upon the common disease–common variant hypothesis (Reich and Lander, 2001): if there are many disease alleles distributed over all the modal haplotypes in each gene, SNP-based association studies are unlikely to detect them. In contrast, admixture mapping studies can detect a locus as long as the total pool of risk alleles is differentially distributed between the ancestral populations that have contributed to the admixed population under study, whether there are many rare risk alleles or only a few common ones in the disease susceptibility gene.

In a classic linkage analysis of an experimental cross, inbred strains produced by brother–sister mating over many generations are crossed over two generations to generate a sample of $F2$ intercrosses or $F1$ backcrosses. These animals are typed at a set of marker loci at which different alleles have become fixed in each strain, and tested for association between the trait and the marker genotypes at each locus. In extending this approach to admixed human populations, three main problems arise.

1. The history of admixture is not under experimental control or even known. In an experimental cross, the design can be specified so that, for instance, we can study a sample of individuals that consists entirely of $F2$ intercrosses. In human admixed populations, admixture proportions – the proportions of the genome that have ancestry from each continental group – vary between individuals. This confounds any associations between a trait and locus ancestry. This confounding can be controlled by conditioning on parental admixture proportions when testing for association of the trait with locus ancestry. In an affected-only design, we can compare the observed ancestry state frequencies at the locus under study with the expected frequencies given parental admixture proportions of each individual. In a cross-sectional study, we can fit a regression model for the dependence of the trait upon parental admixture proportions, and test for association of adjusted trait values with ancestry state frequencies.

2. Human subpopulations are not inbred strains. In a cross between two inbred strains, the strains of ancestry of the two gene copies at each locus can be inferred directly from the observed genotypes, as the marker alleles are differentially fixed in the two strains. In humans, the F_{ST} distance between continental populations is typically of

the order of 0.1–0.15 (Cavalli-Sforza *et al.*, 1994); another way of expressing this is that human continental groups have lost only 10–15 % of their shared ancestral allelic diversity. Loci at which alleles have become differentially fixed in each continental group are rare: one of the few examples is the *FY* (Duffy antigen) locus, at which a null allele is fixed in populations originating from sub-Saharan Africa and rare in other populations. Even when hundreds of thousands of markers are screened to identify a subset with extreme allele frequency differentials between continental groups, the average information content for ancestry of these ancestry-informative markers is typically no more than 40 %. Thus, we cannot directly infer locus ancestry from the genotype at that locus. As in classical linkage analysis, this problem can be overcome by a multipoint analysis that combines all genotype information on each chromosome to infer the posterior distribution of ancestry at each marker locus.

3. The ancestral populations are not available for study. In an experimental cross, the marker allele frequencies in the ancestral strains are known, or can be estimated directly in the case of a cross between outbred populations. In human admixture, this is generally not the case: for instance, we cannot sample the exact mix of West African subpopulations that contributed genes to the modern African-American population. This problem can be overcome by re-estimating ancestry-specific allele frequencies within the admixed population under study, as described below.

35.2.2 Statistical Power and Sample Size

To examine the construction of hypothesis tests for admixture mapping, and the statistical power of studies based on these tests, initially we assume that parental admixture proportions, and ancestry at each locus can be inferred without error. For statistical power calculations, this assumption is reasonable, as any region of putative linkage detected in an initial analysis can be saturated with additional markers to reduce any uncertainty in the inference of locus ancestry.

In a study of a binary trait, the parameter under test at each locus is the ancestry risk ratio r, defined as the risk ratio between individuals with 2 and 0 gene copies with ancestry from the high-risk population at the locus under study. For simplicity, we assume a multiplicative model in which there is a linear relationship between the log risk and the number a of gene copies at the disease susceptibility locus that have ancestry from the high-risk population. For an individual whose parents' genomes have proportions μ_1 and μ_2 of ancestry from the high-risk population, the probabilities p_0, p_1, p_2 of 0, 1, 2 gene copies from the high-risk population at any given locus are $(1 - \mu_1)(1 - \mu_2), \mu_1(1 - \mu_2) + (1 - \mu_1)\mu_2, \mu_1\mu_2$, respectively. For an affected individual, given that we observe x gene copies at the locus under study that have ancestry from the high-risk population, the likelihood L as a function of r, conditional on parental admixture proportions is

$$L(r) = \frac{p_x r^{x/2}}{p_0 + p_1\sqrt{r} + p_2 r},$$

where p_x is the probability of x gene copies with ancestry from the high-risk population at the locus under study, defined as above in terms of the parental admixture proportions μ_1, μ_2. For calculations of statistical power, and construction of score tests, it is convenient to evaluate the likelihood as a function of $\theta = \log r$, which makes the asymptotic approximation of the log likelihood by a quadratic function more accurate.

In a cross-sectional study design, the parameter under test is the regression slope θ for the dependence of the trait value upon locus ancestry, adjusted for parental admixture proportions μ. Given data on a single individual with observed trait value y and x gene copies at the locus under study that have ancestry from the high-risk population, the log likelihood as a function of θ is

$$L\left(\theta\right) = -\frac{1}{2}\nu\left(y - \alpha - x\theta\right)^{2},$$

where α is the expected value of y under the null hypothesis $\theta = 0$, and ν is the residual precision (inverse variance) in the regression model, after including parental admixture and any other covariates that are adjusted for.

We can approximate the statistical power of any study design that tests the null hypothesis that a scalar parameter θ has value θ_0 from the Fisher information V, defined as the expectation over repeated experiments of $-\mathrm{d}\log L\left(\theta\right)/\mathrm{d}\theta$ (Clayton and Hills, 1993). This calculation assumes that the second derivative of the log likelihood is approximately constant over the plausible range of the parameter under test (equivalent to approximating the log likelihood by a quadratic function of θ). For one-sided Type 1 error probability α, Type 2 error probability β, and effect size θ, the required sample size n is given by

$$n = \left(\frac{Z_{1-\alpha} + Z_{\beta}}{\theta}\right)^{2}/V,$$

where Z_p is the pth quantile of the standard normal distribution.

For 90 % power ($Z_{\beta} = 1.28$) to detect an effect of size $\theta = 1$ at a one-sided threshold p-value of 10^{-5} ($Z_{1-\alpha} = 4.27$), we require the Fisher information V to be 30.8. The required sample size for a given study design and effect size $\theta = 1$ can be calculated by dividing 30.8 by the average Fisher information contributed by a single individual. The statistical power for any other effect size can be calculated by dividing by the square of the effect size.

For an affected-only design testing the null hypothesis that $\theta = 0$ (where θ is the log ancestry risk ratio), the Fisher information contributed by a single individual whose parental gametes have proportionate admixture μ is $\frac{1}{2}\mu(1-\mu)$. Thus, with $\mu = 0.2$ (a typical value for mean European admixture in African-American populations), a single affected individual contributes information of 0.08 and 385 individuals are required to detect a locus that accounts for an effect size $\theta = 1$. The number of individuals required to detect an ancestry risk ratio $r = 2$ (equivalent to $\theta = 0.69$) is about 800. A practical lower limit for the ancestry risk ratio generated by a locus to be detectable with realistic sample sizes (no more than a few thousand affected individuals) by admixture mapping is about $r = 1.5$.

Where cases and controls for a rare disease have been sampled, almost all information about linkage with locus ancestry is contributed by the cases. If we can assume that ancestry state frequencies do not vary systematically across the genome within the admixed population under study, we can compare observed with expected ancestry state frequencies (calculated from parental admixture proportions) in affected individuals only, and ignore the controls. If the number of founders of the admixed population is large, and the effects of selection and drift since admixture can be ignored, the assumption of no systematic variation of ancestry state frequencies within the admixed population under

study should be valid. Alternatively, we can relax this assumption, and compare ancestry state frequencies in cases and controls, adjusting for expected ancestry state frequencies. This case–control test, however, has lower statistical power (for a given number of cases) because it is testing for a difference between the observed minus expected ancestry state frequency in the cases and the observed minus expected frequency in the controls, rather than comparing the observed ancestry state frequency in the cases to the expected frequency (which can usually be inferred quite accurately). For a case–control comparison of locus ancestry to have the same statistical power as an affected-only design, four times as many individuals must be genotyped as in an affected-only study. It is, however, useful to sample at least some controls to allow a check on the assumption of no systematic variation of ancestry state frequencies, and more generally as a sanity check. Controls may also be required to obtain additional information about ancestry-specific allele frequencies.

The statistical power of a test for association with a quantitative trait can be calculated by similar arguments (Hoggart *et al.*, 2004). The null hypothesis is that $\theta = 0$ (where θ is the slope of regression of the trait upon locus ancestry (coded as the proportion of copies $(0, 1/2, 1)$ from the high-risk population). The Fisher information contributed by a single individual whose parental gametes have proportionate admixture μ, for a trait with residual precision λ, is $\frac{1}{2}\lambda\mu(1 - \mu)$. Thus for a cross-sectional study of a quantitative trait, the sample size required to detect a locus that accounts for an effect size equal to the residual standard deviation $(1/\sqrt{\lambda})$ is the same as that required to detect an effect size $\theta = 1$ equivalent to ancestry risk ratio $r = 2.7$ in an affected-only study. In practice, effects of locus ancestry on a quantitative trait are unlikely to be as large as this; with the exception of skin melanin content, there are few quantitative traits for which the mean difference between continental populations exceeds one standard deviation. Thus admixture mapping studies of quantitative traits, using unselected cross-sectional samples from the admixed population under study, typically require larger sample sizes than studies of extreme phenotypes for which affected-only and case–control designs can be used. For instance, to detect genes underlying ethnic differences in blood pressure, a study based on ascertainment of individuals with the extreme phenotype (hypertension) via clinical records is likely to be more efficient than a cross-sectional study of blood pressure.

In practice, individual admixture proportions and locus ancestry cannot be inferred without error from marker genotypes. This is a 'missing-data' problem, similar to the problem of inferring segregation indicators from marker genotypes in classical linkage studies (Thompson **Chapter 33**). A general approach to such problems is to generate the posterior distribution of the missing data – admixture proportions and locus ancestry – given the observed data – marker genotypes and trait values – under the null hypothesis of no effect of any locus on the outcome. We can then construct tests based on the likelihood functions given above by averaging over the posterior distribution of the missing data under the null hypothesis, as described later.

35.2.3 Distinguishing between Genetic and Environmental Explanations for Ethnic Variation in Disease Risk

Admixture mapping is most likely to work where there are large ethnic differences in disease risk and these differences are at least partly attributable to genetic factors. The classical epidemiological approach to distinguishing between genetic and environmental

explanations for between-population differences in disease rates is to study migrants (Reid, 1971). When migrant groups living in the same environment have different disease rates that are not accounted for by adjusting for known environmental determinants of disease risk, it is reasonable to consider genetic explanations. Genetic explanations are most likely where differences in disease rates persist even in migrants who have been settled outside the home country for several generations, and where such differences are consistently found in all countries where the migrant group has settled. On these criteria, genetic factors are likely to underlie the high rates of coronary heart disease and non-insulin-dependent diabetes, which have been reported in people of South Asian (Indian, Pakistani, Bangladeshi and Sri Lankan) descent settled overseas (McKeigue *et al.*, 1989). In contrast, migrant studies suggest that genetic factors are unlikely to account for the low rates of coronary heart disease, colorectal cancer and breast cancer in native Japanese compared with European-ancestry populations in the United States; although initial studies showed that the rates in first-generation Japanese migrants to the United States were similar to Japanese in Japan, in second- and third-generation migrants the rates of these diseases are close to those in European-ancestry individuals in the United States (Dunn, 1975).

A more definitive approach to distinguishing between environmental and genetic explanations for ethnic differences in disease risk is to study populations in which the proportionate admixture of genes from high-risk and low-risk populations varies between individuals. If disease risk is dependent upon the proportion of the individual's genome that has ancestry from the high-risk population, independently of environmental factors, this suggests genetic influences. However, as with any epidemiological study design, inference of an independent relationship depends on the ability to control for confounding factors such as socioeconomic status. The shape of the relationship between disease risk and individual admixture proportions may also be informative. Even if the difference in disease risk is accounted for by a single locus with an autosomal recessive or dominant mode of action, the relationship of disease risk to individual admixture proportions will be approximately linear. It is possible to construct genetic models in which the relationship of disease risk to admixture proportions is non-linear: for instance, we could postulate a model in which the risk of disease is highest in heterozygous individuals. This might be biologically plausible for an autoimmune disease, as for instance in experimental models of lupus (Rudofsky and Lawrence, 1999). Under such a model, the risk of disease would show an inverse U-shaped relationship to admixture, with highest risk in those who have one parent from each of the two ancestral subpopulations. If a model that allows admixture proportions to differ between the two parents is fitted, such a relationship can be modelled and tests for linkage can be easily constructed.

Some examples of ethnic differences in diseases or traits for which epidemiological evidence points to a genetic explanation are listed in Table 35.1. It is remarkable that non-insulin-dependent diabetes, obesity and hypertension should be so prominent among the conditions for which genetic differentiation has apparently led to ethnic differences in disease risk. This may be because genes influencing risk of these diseases also influence traits on which there has been differential selection pressure, such as the ability to survive famine (Neel *et al.*, 1998). Whatever the explanation, these diseases are likely to be the ones most amenable to admixture mapping.

Table 35.1 Ethnic differences in disease rates that are likely to have a genetic component.

Disease or trait	High-risk groups	Low-risk groups	Typical risk ratios between high-risk and low-risk groups
Non-insulin-dependent diabetes	Pacific islanders, Native Americans, Native Australians, South Asians (Zimmet, 1992)	Europeans	5–12
Hypertension	West Africans (Prineas and Gillum, 1985)	Europeans	2–5
Hypertensive renal disease	West Africans (Qualheim *et al.*, 1991)	Europeans	6–20
Generalised obesity	Native Americans, Pacific islanders, West African women (Hodge and Zimmet, 1994)	Europeans	1–2 SD (continuous scale
Central adiposity	South Asians (McKeigue *et al.*, 1991), Native Australians (O'Dea *et al.*, 1993)	Europeans	1 SD (continuous scale)
Coronary heart disease	South Asians (McKeigue *et al.*, 1989)	West Africans (Miller *et al.*, 1989)	2–3
Systemic lupus erythematosus	West Africans (Hopkinson *et al.*, 1994)	Europeans	10
Prostate cancer	West Africans (Merrill and Brawley, 1997)	Europeans	2–3
Angle-closure glaucoma	Alaskan Natives and other Inuit (Congdon *et al.*, 1992)	Europeans	5–10

35.3 STATISTICAL MODELS

35.3.1 Modelling Admixture

To combine information from linked marker loci to infer admixture proportions and locus ancestry, we have to model the stochastic variation of ancestry on gametes generated by parents of mixed descent. Most current programs for modelling admixture, including
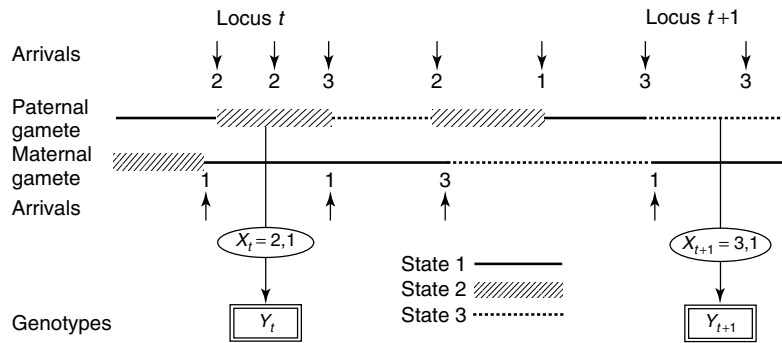
Figure 35.1 Structure of the standard statistical model of admixture with three ancestral populations. Genotypes at each locus (two are shown) depend on the ancestral states, which change along gametes according to a Poisson arrival process.

ADMIXMAP, STRUCTURE and ANCESTRYMAP, specify a model of admixture between K subpopulations in which the stochastic variation of ancestry on each gamete is generated by K independent Poisson arrival processes. In this chapter, this is referred to as the 'standard statistical model' of admixture.

In this model (Figure 35.1), the K arrival processes have intensities $(\mu_1\rho, \ldots, \mu_K\rho)$, where $(\mu_1, \ldots, \mu_K)$ is the vector of proportions of the parent's genome that have ancestry from each of the K subpopulations, and ρ is the sum of intensities of the K arrival processes. The probability of a transition to ancestry state j at locus $t + 1$, given probability p_j of ancestry state j at locus t, is then $\left[e^{-\rho d_t} p_j + \left(1 - e^{-\rho d_t}\right) \mu_j \right]$, where d_t is the distance in morgans from locus t to locus $t + 1$. If admixture has occurred in a single pulse, the arrival rate parameter ρ can be interpreted as the number of generations back to unadmixed ancestors (Falush *et al.*, 2003). More generally, we can interpret ρ as the effective number of generations since admixture. The assumption of independent Poisson arrival processes is convenient for modelling purposes, but has no basis in any biological model of recombination. Even if we assume a Haldane mapping function – in which meiotic crossovers occur as a Poisson process with intensity 1 per morgan – the stochastic variation of ancestry on admixed gametes does not in general follow a Markov process, except on gametes produced by an $F1$ individual who has one parent from each of two unadmixed ancestral populations (McKeigue, 1998). However, this simple model has been found to perform well in inferring admixture proportions, locus ancestry, and number of generations since admixture. A key assumption of the model is that there is no allelic association between markers loci within any of the ancestral subpopulations, and thus that all allelic association between linked loci can be attributed to admixture. This assumption is generally valid as long as the spacing between marker loci is at least 1 cM.

35.3.2 Modelling Stratification

To model stratification as well as admixture, we allow the admixture proportion vector $\boldsymbol{\mu}$ to vary between individuals. If we assume that mating in the parental generation is completely assortative with respect to admixture, we can specify that the admixture proportion vector $\boldsymbol{\mu}$ has the same value on both parental gametes. Alternatively, if we assume random mating within the parental generation, we can specify a model in which
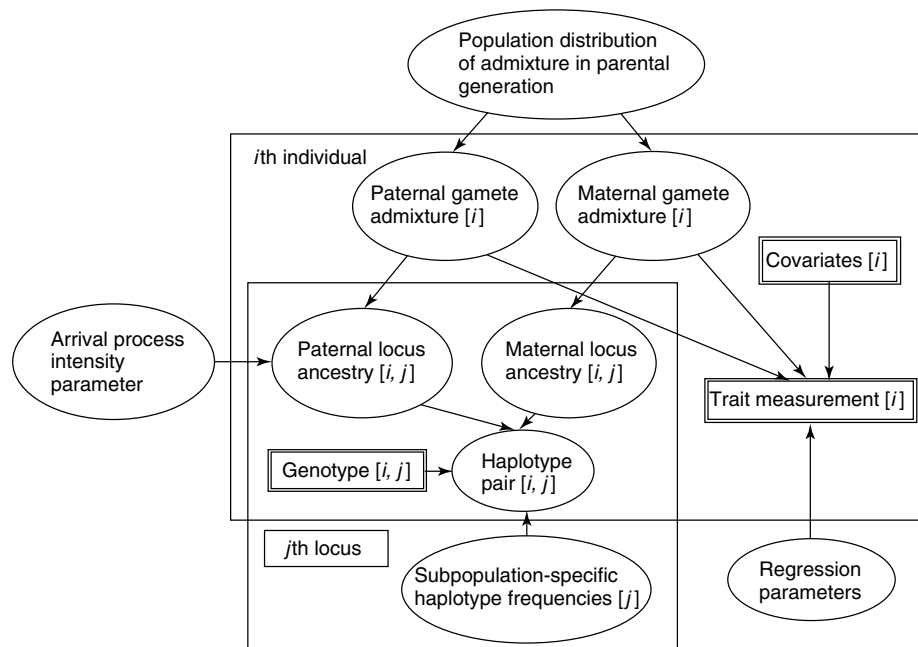
Figure 35.2 Graph representing the statistical model for individual admixture and locus ancestry. This figure does not show the dependence of locus ancestry between linked loci.

the admixture proportions of the two parents are drawn independently from the same population distribution. As the conjugate prior for a probability vector is a Dirichlet distribution, it is natural to use this distribution to model the distribution of admixture proportions over individuals (or gametes) in the population. We place a prior on the parameters of this Dirichlet distribution, and infer the posterior distribution of these parameters from the data. If all elements of the Dirichlet parameter vector α are greater than 1, the distribution of individual admixture proportions in the population is unimodal. If all elements are less than 1, the distribution of individual admixture proportions is U-shaped: that is, most individuals have admixture proportions in which one element of the proportion vector is close to 1 and the others are 0. This is a hierarchical model, as shown in Figure 35.2; inference about each individual's admixture proportions uses not only the information from that individual's genotype data, but also the inferred distribution of admixture proportions based on other individuals sampled from the same population. Where a trait or outcome variable has been measured, the standard model of admixture can be extended to model the possible dependence of the trait value upon individual admixture proportions; this extension is implemented in the ADMIXMAP program. The dependence of trait value on individual admixture can be modelled with logistic regression (for a binary trait) or linear regression (for a quantitative trait). This is useful not only when we want to examine the relationship of an outcome variable to individual admixture proportions (for instance, when attempting to distinguish between genetic and environmental contributions to ethnic variation in disease risk) but also because the likelihood from the regression model feeds back into the inference of individual admixture proportions, enhancing the ability to detect subtle degrees of population stratification.

In principle, it is possible to specify more complex models for the distribution of individual admixture proportions in the population – for instance, we could specify a mixture of two Dirichlet distribution – but this is not implemented in programs currently available. Where enough information is available from marker data to infer the arrival rate parameter ρ on each gamete, we can similarly extend the model to allow the arrival rate to vary across gametes, using a γ distribution to model the distribution of arrival rates over gametes in the population.

35.3.3 Modelling Allele Frequencies

For admixture mapping, the ancestry-specific allele frequencies are usually estimated from samples of modern unadmixed descendants of the populations that contributed genes to the admixed population under study. To allow for sampling error in these estimates, we can specify Dirichlet priors based on the posterior that would be obtained by combining the observed counts with a 'reference' prior on allele frequencies. For a diallelic locus, the 'reference' prior is $\beta(0.5, 0.5)$.

More generally, we may be uncertain as to whether the unadmixed modern descendants are representative of the subpopulations that contributed to the admixed population: for instance, it is not possible to sample the exact mix of Native American populations that contributed to the modern Mexican population. To allow for this uncertainty, we can specify a 'dispersion' model in which both the ancestry-specific allele frequencies in the admixed population and the allele frequencies in the corresponding modern unadmixed population are drawn from a Dirichlet distribution with parameter vector $(\alpha_1 \eta, \ldots, \alpha_K \eta)$, where the mean parameter $\boldsymbol{\alpha}$ is a proportion vector, and η is the precision parameter. The mean parameter vector is allowed to vary across loci, but the precision parameter is assumed to be constant across loci. The biological basis for this is the Wright–Fisher model for drift of allele frequencies in subpopulations that have split from an ancestral total population; the precision parameter η is related to the F_{ST} distance between the two subpopulations by $F_{ST} = 1/(1 + \eta)$ (Lockwood *et al.*, 2001). Thus, for instance, if we assume that the dispersion of allele frequencies between modern West African populations and the African-ancestry gene pool is of similar magnitude to the F_{ST} distance between West African subpopulations (estimated to be about 0.02 (Cavalli-Sforza *et al.*, 1994)) we might specify a prior mean of about 50 $(0.98/0.02 = 49)$ for the African-specific allele frequency precision parameter in African-Americans. For the European-specific allele frequency precision parameter, we might specify a prior mean of about 500 based on estimates that the F_{ST} distance between European subpopulations is typically about 0.002 (Cavalli-Sforza *et al.*, 1994).

With large samples from both the admixed population under study and the modern unadmixed descendants, and a highly informative marker panel, the allele frequency precision parameter η can be estimated for each of the continental groups under study. In the few studies that have estimated this parameter for admixed populations such as African-Americans, the posterior mode for the African-specific precision parameter is large, implying that the dispersion of allele frequencies between modern unadmixed West Africans and the African gene pool in the African-American population is fairly small (Patterson *et al.*, 2004). One reason for this may be that markers that show large frequency variation between West African subpopulations are typically excluded from panels of ancestry-informative markers. The practical implication of these results is that it may only rarely be necessary to specify a dispersion model for allele frequencies, as

the additional uncertainty generated by dispersion of allele frequencies is usually small compared with the uncertainty in the Dirichlet prior based on the observed allele counts.

Once we have studied one sample from an admixed population, we can re-use the posterior distribution of allele frequencies obtained from that study to specify priors on allele frequencies for any study of new samples from the same admixed population. For simplicity, we can approximate the posterior distribution of allele frequencies by a Dirichlet distribution with the same mean and variance, and use the parameters of this Dirichlet distribution to specify the prior. From this stage onwards, it is unnecessary to specify a dispersion model, because the posterior distribution from the initial study has already taken into account our uncertainty about the extent to which we can rely on modern unadmixed descendants to estimate ancestry-specific allele frequencies in the admixed population under study. In effect, the ancestry-specific allele frequencies have been re-estimated within the admixed population under study. To re-use samples from the posterior distributions of allele frequencies, we can approximate these posteriors by Dirichlet distributions, equating the means and variances of the Dirichlet distribution with the mean and variance of the sample.

35.3.4 Fitting the Statistical Model

Two general approaches to model fitting have been used: a fully Bayesian approach in which the posterior distribution of model parameters is generated by Markov chain Monte Carlo (MCMC) simulation, and a maximum likelihood approach in which the model parameters are held at their maximum likelihood values. The Bayesian approach is more computationally intensive, but has the advantage that samples from the posterior distribution (under the null) can be used to construct hypothesis tests as outlined below. As the transitions of ancestry are generated by Poisson arrival processes, the variation of ancestry along the chromosome is a Markov process. If we assume no allelic association between loci other than that attributable to the autocorrelation of locus ancestry, this is a hidden Markov model (HMM). For such models, standard 'message-passing' algorithms are available to calculate the likelihood of the model parameters $\boldsymbol{\mu}$ and ρ given the data, to sample the joint distribution of hidden states (ancestry) at all loci, and to calculate the marginal distribution of hidden states at each locus, conditional on the model parameters (MacDonald and Zucchini, 1997). In a Bayesian MCMC approach, we can use the likelihood calculated from the HMM forward recursion (combined with the prior specified by the population-level parameters) in a Metropolis algorithm to sample the model parameters for each individual. If we also calculate the HMM backward recursions, we can calculate the marginal distribution of ancestry states at each locus, conditional on the realized model parameters. Alternatively, in the maximum likelihood approach, we can use the HMM forward–backward algorithm to calculate the marginal expectations of the numbers of 'arrivals' of each state of ancestry, and maximize the likelihood by an expectation-maximization (EM) algorithm.

The HMM recursions account for most of the computational burden of the MCMC simulations. When standard HMM algorithms are used to compute the forward and backward recursions by matrix multiplication, these recursions have complexity $O\left(K^4\right)$ for diploid individuals, as the transition matrix for each interval between adjacent loci is of order K^2 and each element of this matrix has to be multiplied by the corresponding element of the vector of 'emission' probabilities. As in other situations where the transition probabilities in the HMM arise from a simpler stochastic process, the algorithms for the

forward and backward recursions can be rewritten in a more efficient form that does not require matrix multiplication, and has complexity $O\left(K^2\right)$ for diploid individuals. Updating the population-level parameters at each iteration requires less computation as the number of these parameters does not depend upon the number of individuals. The parameters of the Dirichlet distribution of individual admixture proportions can be updated with a simple Metropolis random walk. After the ancestry states have been sampled on each gamete, and the ordered genotypes or haplotype pairs have been sampled conditional on the realized ancestry states, the allele frequency parameters can be updated with a conjugate Dirichlet update.

35.3.5   Model Comparison

For any given problem, we have a wide choice of the model to be specified. For instance, we can specify more subpopulations, we can allow admixture proportions to differ on the two parental gametes, or we can specify a 'dispersion model' for the ancestry-specific allele frequencies. Two general approaches are available to model choice: construction of diagnostic tests for specific inadequacies of the model, or evaluating the marginal likelihood of the model given the data.

35.3.5.1   Model Diagnostics

One approach to model choice is to construct diagnostic tests for specific inadequacies of the model. For instance, we can construct a test for residual allelic association between unlinked loci, implying residual population stratification not explained by the model. To account for this residual stratification, we might have to specify a model with more subpopulations. Another example would be to test for lack of fit of the observed allele frequencies to the prior, implying that a dispersion model should be fitted. Where the diagnostic test can be set up in the form of a test of a null hypothesis for a continuous parameter, it is possible to evaluate a score test of this null hypothesis by averaging over the posterior distribution. For example, to test for departure from Hardy–Weinberg equilibrium that is not explained by the fitted model, we can set up a test of the null hypothesis of zero inbreeding coefficient. For any realization of the complete data, the score is then the observed frequency of heterozygotes at a locus minus the expected frequency given the realized locus ancestry states and ancestry-specific allele frequencies. Where the diagnostic test cannot be set up as a null hypothesis for a scalar parameter, an alternative approach is to construct a test based on the posterior predictive distribution of some scalar variable T that captures the type of deviation from the model that we are looking for. For instance, to test for dispersion of allele frequencies when the model specifies no dispersion between ancestry-specific allele frequencies and the corresponding allele frequencies in unadmixed modern descendants, we can calculate at each realization of the model parameters a scalar variable T_{obs} as the log likelihood of the prior parameters (of the Dirichlet distribution of allele frequencies) given the realized allele frequencies. Although no obvious sampling distribution (over repeated experiments) is available for the posterior mean of T_{obs}, we can compare the posterior distribution of T_{obs} with the posterior distribution of a quantity T_{rep} calculated by the same method from a replicate dataset drawn conditional on the model parameters at each iteration (Rubin, 1984). We can calculate a posterior predictive check probability ("Bayesian p value") as the frequency

with which T_{rep} is more extreme than T_{obs}. Under the null hypothesis, T_{obs} and T_{rep} are exchangeable, and the posterior predictive check probability has expectation 0.5 over repeated experiments. This provides a useful diagnostic check for bad loci: at loci where the observed genotype data are incompatible with the prior on ancestry-specific allele frequencies, the posterior predictive check probability is usually close to 0.

35.3.5.2 *Evaluation of the Marginal Likelihood*

In a Bayesian framework, formal model comparison is based on the marginal likelihood $P(Y \mid M)$ of each model M, given the observed data Y. The ratio of marginal likelihoods between models M_1 and M_0 is the Bayes' factor, which by Bayes theorem is the ratio of posterior to prior odds in favour of M_1 versus M_0. The marginal likelihood can be evaluated by subtracting a measure of 'complexity' (the information conveyed by the data, defined as the decrease in entropy when passing from the prior to the posterior) from a measure of 'fit' (posterior mean of the log likelihood of the model parameters, given the data). The Bayesian framework for hypothesis testing thus incorporates a penalty for complexity (Mackay, 2003). It is straightforward to evaluate the posterior mean of the log likelihood of the model parameters, given the data, from MCMC simulation. Evaluation of the information is more computationally demanding. In the program STRUCTURE, the information is approximated by the posterior variance of the log likelihood of the model parameters (Falush *et al.*, 2003). Although this has been widely used to evaluate how many subpopulations are required to model admixed and stratified subpopulations, it may give answers that are rather different from direct calculation of the marginal likelihood. It is possible, although computationally intensive, to calculate the information from a series of MCMC runs at different temperatures by the method of thermodynamic integration (Neal, 1993).

35.3.6 Assembling and Evaluating Panels of Ancestry-informative Marker Loci

Admixture mapping depends critically upon the availability of markers that are informative for ancestry. Simulations and empirical studies show that to extract about 70 % of information about locus ancestry across the genome in a population where the effective number of generations back to unadmixed ancestors is 5, markers with average information content for ancestry of 40 % are required at average spacing of 1–2 cM (Hoggart *et al.*, 2004). For admixture between populations originating in different continents, where the F_{ST} distance between these populations is at least 0.1, it is feasible to assemble such marker panels by screening large numbers of SNPs and selecting those with the most extreme allele frequency differentials. Before such markers can be used in an admixture mapping study, it is necessary to re-estimate the allele frequencies in independent samples of unadmixed individuals. Marker panels informative for West African/European admixture are already available (Smith *et al.*, 2004), and marker panels for admixture between other continental populations are being assembled.

In principle, it is possible to model admixture with unselected marker panels at higher density. Thus, for African/European admixture, using unselected SNP markers (for which the average marker information content for ancestry is about 0.08) would require only about five times higher density than using markers informative for ancestry (average information content for ancestry about 0.4). With arrays that can score tens of thousands

of markers in parallel, the additional genotyping costs of typing more markers may not be a barrier. In practice, however, with such a dense panel of markers, it is difficult to ensure that there is no allelic association between adjacent markers in any of the ancestral subpopulations, which is a key modelling assumption. Another disadvantage, of course, is the greater computational burden.

35.4 TESTING FOR LINKAGE WITH LOCUS ANCESTRY

The model of admixture and locus ancestry is based on the null hypothesis that the effect size parameter θ is 0 at any locus. Such a null hypothesis can be tested by averaging over the posterior distribution generated by MCMC simulation. Two general approaches to constructing tests are available: score tests based on the gradient of the log likelihood at the null value of the parameter, and direct calculation of the likelihood ratio. To calculate a score test, we calculate for each realization of the missing data X the gradient $\mathrm{d}\log P(Y,X)/\mathrm{d}\theta$ and second derivative $\mathrm{d}^2\log P(Y,X)/\mathrm{d}\theta^2$ of the log-likelihood function. We evaluate the score U as the posterior mean of the realized score, the missing information as the posterior variance of the realized score, and the complete information as the posterior mean of the realized information. The observed information V is calculated by subtracting the missing information from the complete information (Louis, 1982). A chi-square test statistic can then be evaluated as $UV^{-1}U$. An attractive feature of this algorithm for calculating score tests by averaging over the posterior distribution is that the ratio of observed information to complete information – the 'proportion of information extracted' – can be interpreted as a measure of the efficiency of the study design, relative to one in which individual admixture proportions and locus ancestry are inferred without uncertainty.

Alternatively, we can calculate, for each locus and for specified values of the effect size parameter θ, the likelihood ratio $P(Y\mid\theta)/P(Y\mid\theta_0)$, where Y is the observed data. We make use of a standard result that $P(Y\mid\theta)/P(Y\mid\theta_0)$ is the expectation of the complete-data likelihood ratio $P(Y,X\mid\theta)/P(Y,X\mid\theta_0)$ over the posterior distribution of the missing data X under the null hypothesis $\theta=\theta_0$ (Thompson and Guo, 1991). In linkage analysis of a pedigree, the missing data X are the segregation indicators at the locus under test; in an admixture mapping study, the missing data for a single individual are the ancestry states at the locus under test and the parental admixture proportions. In a multipoint linkage analysis with known marker map, there is no need to include the recombination fraction as an unknown parameter in the disease model as we can test each position on the genome and specify a disease model in which the effect is mediated only through the locus under test. Under this assumption, $P(Y\mid X,\theta)$ is independent of θ. The likelihood ratio then simplifies to

$$\frac{P(Y\mid\theta)}{P(Y\mid\theta_0)}=\left\langle\frac{P(Y,X\mid\theta)}{P(Y,X\mid\theta_0)}\right\rangle_{P(X\mid Y,\theta_0)}=\left\langle\frac{P(Y\mid X,\theta)\,P(X\mid\theta)}{P(Y\mid X,\theta_0)\,P(X\mid\theta_0)}\right\rangle_{P(X\mid Y,\theta_0)}$$

$$=\left\langle\frac{P(X\mid\theta)}{P(X\mid\theta_0)}\right\rangle_{P(X\mid Y,\theta_0)},$$

where angled brackets denote that expectation is evaluated over the distribution specified in the subscript.

In an admixture mapping study, we can treat the individual's admixture proportions $\boldsymbol{\mu}$ as fixed by the design, and evaluate the likelihood ratio as

$$\left\langle \frac{P\left(x, \mid \boldsymbol{\mu}, \theta\right)}{P\left(x \mid \boldsymbol{\mu}, \theta_0\right)} \right\rangle_{P(x, \boldsymbol{\mu}|Y, \theta_0)},$$

where x is the individual's ancestry at the locus under study (Patterson *et al.*, 2004). Averaging this ratio over the prior on θ yields a Bayes factor (ratio of integrated likelihoods) comparing the hypothesis of an effect at the locus under test with the null hypothesis of no disease locus. The log to base 10 of the Bayes factor can be interpreted as equivalent to a log score in a classical linkage analysis in which the genetic model parameters are fixed. An extension of this argument is to average the Bayes factor over all positions on the genome: equivalent to averaging the likelihood over a diffuse prior on location. This yields a Bayes factor comparing the hypothesis of a single disease locus at an unknown location with the null hypothesis of no disease locus (Patterson *et al.*, 2004). With this Bayesian approach to hypothesis testing, it is unnecessary to correct for multiple testing: averaging over a diffuse prior on location imposes the correct penalty for not specifying the position of the disease locus.

In an affected-only study, the parameter under test is the ancestry risk ratio r and the complete-data likelihood can be calculated at each iteration from the realized values of x and $\boldsymbol{\mu}$ using the expression given earlier. For a score test, the asymptotic properties of the test statistic are improved by evaluating the score and information as functions of $\theta = \log r$. For a single individual, the realized score at $\theta = 0$ is $x - \left(\mu_1 + \mu_2\right)$, where x is the number of gene copies at the locus under test that have ancestry from the high-risk population, and μ_1, μ_2 are the parental admixture proportions. This is simply the observed minus expected proportion of gene copies that have ancestry from the high-risk population. The complete information is $\frac{1}{4}\left(\mu_1\left[1 - \mu_1\right] + \mu_2\left(1 - \mu_2\right)\right)$.

In a case–control or cross-sectional study, we can calculate a conventional test for association with locus ancestry, conditional on parental admixture proportions. This can be implemented as a test for the effect size parameter in a regression model, or simply as a test for dependence of the trait on the observed minus expected proportion of gene copies that have ancestry from the high-risk population. In comparison with the affected-only test, the case–control test has lower statistical power but does not depend upon the assumption of no heterogeneity of ancestry state across the genome within the admixed population under study.

With either test, the computational burden can be reduced by using the conditional distribution of locus ancestry given the model parameters λ and averaging over the posterior distribution of model parameters, rather than using the realized states of locus ancestry. The conditional distribution of hidden states (locus ancestry) is calculated from the forward and backward recursions of the hidden Markov model, as described elsewhere (MacDonald and Zucchini, 1997). To calculate the posterior variance of the score from posterior samples of the conditional mean $\langle U \mid \lambda \rangle$ and variance $V\left(U \mid \lambda\right)$ of the realized score, we use the identity

$$\mathrm{V}\left(U\right) = \left\langle \mathrm{V}\left(U \mid \lambda\right)\right\rangle_{P(\lambda)} + \mathrm{V}\left(\langle U \mid \lambda \rangle\right)_{P(\lambda)},$$

where angled brackets denote taking expectations. This calculation provides a useful breakdown of the missing information into two components: the expectation $\langle V(U \mid \lambda) \rangle_{P(\lambda)}$ of the score variance conditional on model parameters λ is a measure of how much missing information is attributable to uncertainty about locus ancestry. To reduce this component of missing information, we can increase the density of markers in the genomic region containing the locus under study. The variance $V(\langle U \mid \lambda \rangle)_{P(\lambda)}$ of the conditional expectation of the score is a measure of how much missing information is attributable to uncertainty about model parameters such as parental admixture proportions. To reduce this component of missing information, we would have to increase the density or information content of the marker panel in other regions unlinked to the locus under study.

35.4.1   Modelling Population Stratification

Population stratification is characterized by the presence of allelic associations between unlinked loci; to model it we have to specify the underlying latent variables that generate these associations. Where no prior information is available about demographic background or allele frequencies in ancestral subpopulations, this is an unsupervised learning problem. Given a dataset in which a sample of individuals from a possibly heterogeneous subpopulation have been typed at marker loci, one possible modelling approach is to use principal components analysis to infer a few uncorrelated latent variables that explain most of the observed allelic associations and to calculate the coordinates of each individual with respect to these latent variables (Patterson *et al.*, 2006). An alternative approach is to fit the standard statistical model of admixture and stratification described above and infer the admixture proportions of each individual (Falush *et al.*, 2003). These two approaches are described below.

35.4.1.1   Modelling Stratification with Principal Components Analysis

Principal components analysis is a transformation of the data that does not depend upon any specific statistical model. The theory of this approach is briefly outlined below. Given a data matrix in which M variables have been measured on N units of observation, principal components analysis performs an orthogonal rotation of the axes defined by the M observed variables, so as to define M new variables that are linear combinations of the original variables and that maximize the proportion of residual variance explained by each successive latent variable. Algebraically, this reduces to finding the eigenvalues and eigenvectors of the $M \times M$ covariance matrix, and ranking them in descending order of magnitude of the eigenvalues. The first eigenvalue λ_1 measures the variance explained by the first principal component, and the corresponding eigenvector represents the loadings (weights) of each variable on that component. Where the M observed variables have been measured on different scales, it is usual to standardize them to zero mean and unit variance. If we specify a statistical model in which the latent variables and measurement errors have Gaussian distributions, it is not possible to infer anything about how the axes should be rotated, because orthogonal rotations do not change the multivariate Gaussian likelihood. It is, however, possible to test hypotheses about the number of independent latent variables that underlie the observed covariance between observed variables. The null hypothesis of no latent variable specifies that the M observed variables correspond

to M independent Gaussian variables. This can be compared with alternative hypotheses such as, for instance, a model in which a single latent variable gives rise to covariance between the observed variables. If, for instance, a single latent variable underlies the observed covariances, we expect that the first eigenvalue λ_1 will be large compared with the other $M - 1$ eigenvalues; we can thus use the ratio of the first eigenvalue to the sum of the M eigenvalues $\lambda_1 / \sum_m \lambda_m$ as a test statistic.

In the context of an MCMC simulation, we can calculate at each realization a test statistic T_{obs} as the ratio of the first eigenvalue λ_1 to the sum of the M eigenvalues, and compare it with a test statistic calculated from a replicate dataset drawn from the distribution defined by the realized model parameters. This yields a posterior predictive check probability (Rubin, 1984). More recently it has been shown that if M and N are large, the distribution of the the test statistic $\lambda_1 / \sum_m \lambda_m$ (appropriately normalized for the size of the dataset) under the null hypothesis of no latent variable has has a density discovered by Tracy and Widom from which tests of significance can be obtained (Johnstone, 2001; Patterson $et\ al.$, 2006).

Where the observed variables are allele frequencies or allele counts, and all loci are unlinked, inferring that the number of latent variables is greater than 0 is equivalent to detecting population stratification. Where a model for stratification has already been fitted, the expected values can be subtracted from the observed values before calculating the covariance matrix for the residuals. We can thus test for residual stratification not explained by the fitted model. In the first applications of principal components analysis to infer genetic structure, the data matrices consisted of allele frequency estimates in geographically defined subpopulations. The object was to study geographic clines in genetic background (Cavalli-Sforza $et\ al.$, 1994). Such analyses demonstrated that a high proportion of geographic variation of allele frequencies within the European continent could be explained by the first principal component, which varies from South-East to North-West Europe. More recently, principal components analysis has been applied to individual genotype data, such that each row of the data matrix contains allele counts (scored as 0, 1, 2), standardized so that each column has zero mean and unit variance (Price $et\ al.$, 2006). To allow tightly linked loci to be included as if they were unlinked, the observed allele counts at locus t can be regressed on the observed counts at loci $t - 1, \ldots, t - T$, where T is chosen to be the smallest number that makes the residual allelic associations independent of map distance.

Patterson $et\ al.$ (2006) has demonstrated how tests for stratification can be constructed using the asymptotic distribution of the magnitude of the first eigenvalue under the null. If the first eigenvalue is declared to be significant, the test procedure can be applied to the remaining eigenvalues $\lambda_2, \ldots, \lambda_M$. This procedure is repeated until $K - 1$ significant eigenvalues have been retained and the Kth eigenvalue is not significant. The number K inferred by this test for stratification corresponds to the number of subpopulations in a mixture model of population stratification (Patterson $et\ al.$, 2006). Patterson $et\ al.$ (2006) exploit recent work on the distribution of the largest eigenvalue of a sample covariance matrix to examine the number of individuals and marker loci required to detect significance. They show that where the total population consists of two equally sized subpopulations, the ability to detect stratification with a sample of N individuals typed at M marker loci depends critically upon whether the fixation index F_{ST} between the two subpopulations is greater than $1/\sqrt{MN}$. If F_{ST} is less than the threshold defined by the data size, the test statistic will have the distribution expected under the null,

and stratification will be undetectable. Above this threshold, evidence for stratification will accumulate very rapidly as the data size is increased. They denote this as a phase-change phenomenon. Thus for a sample of 10^3 individuals typed at 10^5 marker loci, stratification into two equally sized subpopulations will be detectable if the F_{ST} distance between subpopulations is greater than 10^{-4}. Although this result was derived for detecting stratification by principal components analysis, it appears to apply to any method for detecting stratification from marker genotypes alone (Patterson *et al.*, 2006).

A key advantage of using principal components analysis to model stratification is that the approach is computationally feasible even for very large datasets. Another advantage is that it is possible to adjust for allelic association between tightly linked loci simply by regressing the genotype on the last few loci as outlined above. This makes it possible to analyse data from whole-genome tag SNP arrays. A limitation of principal components analysis is that it ignores the additional information that may be available from marker positions. In admixed populations, the decay of allelic associations with map distance yields additional information about the latent variables that underlie these associations. In admixed populations, the presence of long-range allelic associations that are not attributable to stratification will invalidate tests for stratification unless the analysis is restricted to unlinked loci, although principal components analysis may still be effective in defining a few axes that summarize most variation in genetic background. In principle, it is possible to augment the genotype data matrix with phenotypic traits or demographic variables relevant to genetic stratification, so that inference of genetic structure is based on 'supervised' learning.

35.4.1.2 *Modelling Stratification with a Mixture Model*

The standard model of admixture and stratification described above can also serve as a model for population stratification, and programs such as STRUCTURE and ADMIXMAP can thus be used to infer stratification. In contrast to a principal components analysis, where the number of 'dimensions' to be retained is not fixed at the start of the analysis, for the admixture/stratification model it is necessary to specify the number K of discrete subpopulations in advance. Where the demographic background of the population is not known, or we do not want to make assumptions about the ancestral subpopulations, the subpopulation-specific allele frequencies can be specified with uninformative priors. Stratification is inferred from allelic associations between unlinked loci, and admixture from the decay of allelic associations with map distance. Thus, in the extreme situation where a population consists of discrete endogamous strata, the strength of allelic association will be independent of map distance (except over very short distances where allelic associations are generated by haplotype structure). In the admixture model, this implies that all individuals have admixture proportions such that one element of the proportion vector is close to 1 and the others are close to 0. This distribution can be represented as a Dirichlet distribution with low precision parameter; the mean vector represents the proportion of the population assigned to each of the K subpopulations. The problem of inferring the number K of ancestral subpopulations, and the posterior assignment probabilities of each individual to one of these subpopulations, then resembles the classical 'K-means' clustering problem. To infer K, we can evaluate the marginal likelihood of models with different values of K as described above, or we

can construct diagnostic tests for residual stratification not explained by the model, using principal component analyses of the residuals as described above.

Other plausible models can also be subsumed within the standard model of admixture and stratification. Thus, a gradient of allele frequencies across a geographic cline can be represented by an admixture model in which the distribution of individual admixture proportions in the population is specified by a Dirichlet distribution with moderately high precision, and the arrival rate parameter is large (so that allelic association does not decay with distance). Thus, for instance, the north–south gradient of allele frequencies in the native British population could be modelled as a gradient of admixture between two ancestral subpopulations corresponding to 'southern' and 'northern' .

Where the population under study is admixed, and its demographic history is known, as for African-American and Hispanic-American populations, this information can be taken into account when specifying the model, so that for instance we can specify two or three ancestral continental populations and can set informative priors on the ancestry-specific allele frequencies. If the demographic history of the population is unknown or we do not wish to make any assumptions about demographic history, uninformative 'reference' priors on the allele frequencies can be specified. In this situation, the K subpopulations are not identifiable, and inference that relies on assuming that the subpopulation labels do not switch during a sampling run may be unreliable. When using unselected markers to infer stratification and admixture between less genetically distant subpopulations, with no prior information about allele frequencies, it is necessary to specify in the model our prior expectation that the allele frequencies at each locus will be correlated between subpopulations. This improves the ability to detect subtle degrees of stratification (Falush *et al.*, 2003). Specifying correlated allele frequencies is equivalent to specifying at each locus a Dirichlet-multinomial model with Dirichlet parameter vector $\boldsymbol{\alpha} = \eta\boldsymbol{\mu}$ for the distribution of allele frequencies across subpopulations. The proportion parameter vector $\boldsymbol{\mu}$ is sampled from a flat Dirichlet prior independently for each locus, with a dispersion parameter η that is the same across all loci. The Dirichlet parameters are given by $\alpha_i = \eta\mu_i$. Where the allelic associations between unlinked loci are weak, the ability to detect stratification may be improved further by specifying an informative prior on the allele frequency precision parameter. With randomly chosen markers, the average correlation of allele frequencies between subpopulations is given by the fixation index F_{ST} between those subpopulations. Values of F_{ST} in the range 0.1–0.2 are typical for subpopulations originating in different continents, and values less than 0.05 are typical for subpopulations originating in the same continent. This implies that plausible prior values for η, in studies of stratification within populations originating from the same continent, are at least 20.

Where a trait that is strongly dependent upon genetic background has been modelled, it is possible to allow for this dependence in the model, for instance by specifying a linear or logistic regression model as implemented in ADMIXMAP. Any demographic variable, such as self-reported ethnicity or geographic origin, can be included in the model as a phenotypic trait. This improves the ability to detect subtle degrees of stratification, because information from the observed trait value is fed back into the admixture model via the contribution of the regression model to the likelihood.

For modelling population stratification, the standard model of admixture/stratification has two main limitations: it is computationally intensive, and the assumption of no allelic association other than that generated by admixture or stratification requires a minimum

spacing between markers: typically about 1 cM within European-ancestry populations. Advantages of this approach to modelling are that both marker–trait and marker–marker associations can be used simultaneously to infer stratification, and that it allows tests of association based on averaging over the posterior distribution to be constructed as described below.

35.4.1.3 Controlling for Population Stratification as a Confounder in Genetic Association Studies

With either of the two modelling approaches described above, we can obtain an estimate of the genetic background of each individual: the coordinates with respect to the retained components of a principal components analysis, or the posterior means of individual admixture proportions in a model of admixture with K subpopulations. To control for possible confounding by genetic background in genetic association studies, we can simply adjust for estimated genetic background in regression models of the effect of genotype on outcome. This has been denoted the 'structured association' approach, and is implemented in the program STRUCTURE/STRAT (Pritchard and Donnelly, 2001). A variant of this approach, implemented in the program ADMIXMAP (Hoggart *et al.*, 2003), is to construct score tests for association based on averaging over the joint posterior distribution of individual admixture proportions and parameters of the model for regression of outcome on admixture proportions. In theory, this one-step procedure has the advantage that it allows correctly for uncertainty in the inference of individual admixture proportions, where two-step procedures may fail to allow for residual confounding where the information about genetic background available from marker loci is limited and the confounding effect of genetic background is strong. Uncertainty in the inference of genetic background is quantified as missing information in the score test, and problems of lack of identifiability of the K subpopulations are eliminated. Where many loci are to be tested for association, stored samples from the posterior distribution of individual admixture proportions and regression parameters (inferred from genotypes at a subset of ancestry-informative marker loci) can be used without having to include all the tested loci in the model of admixture. In association studies that use DNA pools from multiple individuals, it is possible to control for stratification if an initial panel of markers has been typed on individuals to obtain estimates of genetic background, simply by forming pools that are stratified by genetic background.

A somewhat different method for using marker genotype data is the 'genomic control' approach proposed by (Devlin *et al.*, 2003). In this approach, confounding by population stratification is treated as a random effect that inflates the variance of the χ-square test statistic. The magnitude of this random effect is inferred from trait–marker associations: the information available from marker–marker associations is ignored. In effect, genomic control rescales the crude p values so that the observed Type 1 error rate (over all loci) approximates the target Type 1 error rate. The re-scaling of the crude χ-square statistic is equivalent to multiplying the log p values by a constant factor, because the right tail probability of a chi-square distribution approximates an exponential curve. A serious limitation of the genomic control approach is that it assumes that the locus under study and the marker loci are exchangeable with respect to F_{ST}. If, for instance, we are studying a candidate gene that influences a trait such as drug metabolism or immune response, we might expect allele frequency differentials to be larger for functional polymorphisms in this candidate gene than for randomly chosen neutral SNPs. In this situation, the

genomic control method may fail to control for confounding. A recent example is the demonstration of an association between height and a lactase polymorphism in European-Americans (Campbell *et al.*, 2005). Adjustment for marker–trait associations at 178 SNPs by the genomic control method did not reduce the size of the crude effect. However, the association appeared to be at least partly explained by confounding by demographic origin (defined on an axis from north-western to south-eastern Europe).

Another limitation of the genomic control approach is that adjustment necessarily increases the Type 2 error rate. Where the confounding effect of population stratification and the true effect of genotype at the locus under study on the outcome variable are opposite in direction, adjustment of the test statistic by the genomic control approach will mask the true effect, whereas adjustment by the structured association approach will unmask the true effect. These limitations of the genomic control approach are consequences of modelling confounding as a random effect, rather than measuring the confounder (genetic background) and modelling its effect on the outcome variable.

35.5   CONCLUSIONS

The statistical problems of admixture mapping with samples of unrelated individuals have now been largely solved. Panels of ancestry-informative markers are available for West African/European admixture, and will soon be available for other continental groups also. Several programs are available to model individual admixture and locus ancestry, using either Bayesian approaches to generate the posterior distribution of model parameters (Falush *et al.*, 2003; Patterson *et al.*, 2004; Hoggart *et al.*, 2004) or fitting of model parameters by maximizing the likelihood. An unsolved problem is how to combine modelling admixture with other types of genetic model: for instance, modelling pedigree data, or modelling extended haplotypes using genotypes at tag SNPs scored on arrays. Modelling admixture and inheritance in pedigrees simultaneously is possible in theory, by extending the standard HMM algorithms. Thus, we could specify a model in which the hidden state space is defined by the ancestry of the gametes in each founder in the pedigree, and by the segregation indicators in each meiosis. Even for a sib pair, the four founder gametes and two meioses would define a hidden state vector of length $2K^4$ at each locus, for which the HMM recursions would be computationally demanding. Modelling admixture and extended haplotypes simultaneously is even less computationally feasible with current methods. With more than a few thousand marker loci, it becomes increasingly difficult to maintain the assumption of no allelic association between marker loci other than that generated by admixture and stratification. To exploit the information about locus ancestry and individual admixture that is available from ancestry-informative markers, one possible approach would be a two-stage analysis in which estimates of individual admixture or locus ancestry are 'plugged in' to a second step in which extended haplotypes are modelled.

For inferring population stratification with tag SNP arrays that include tens of thousands or hundreds of thousands of markers, ordination methods such as principal components analysis are the most feasible approach. The recent development of formal methods to test for residual stratification unexplained by the retained components strengthens this approach. Statistical modelling of admixture and stratification may, however, yield results that are more directly interpretable in terms of genetic structure and population history:

for instance, the model parameters can be interpreted in terms of the length of time over which admixture has occurred. One possible approach would be to use ordination methods to preselect for each continental group, a subset of markers that are highly informative for stratification, and then to use these markers to model stratification and admixture.

REFERENCES

Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G. and Hirschhorn, J.N. (2005). Demonstrating stratification in a European American population. *Nature Genetics* **37**, 868–872.

Cavalli-Sforza, L.L., Menoozz, P. and Piazzi, A. (1994). *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.

Chakraborty, R. and Weiss, K.M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 9119–9123.

Clayton, D. and Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press.

Congdon, N., Wang, F. and Tielsch, J.M. (1992). Issues in the epidemiology and population-based screening of primary angle-closure glaucoma. *Survey of Ophthalmology* **36**, 411–423.

Devlin, B., Roeder, K. and Wasserman, L. (2003). Analysis of multilocus models of association. *Genetic Epidemiology* **25**, 36–47.

Dunn, J.E. (1975). Cancer epidemiology in populations of the United States–with emphasis on Hawaii and California–and Japan. *Cancer Research* **35**, 3240–3245.

Falush, D., Stephens, M. and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.

Hodge, A.M. and Zimmet, P.Z. (1994). The epidemiology of obesity. *Bailliere's Clinical Endocrinology and Metabolism* **8**, 577–599.

Hoggart, C.J., Parra, E.J., Shriver, M.D., Bonilla, C., Kittles, R.A., Clayton, D.G. and McKeigue, P.M. (2003). Control of confounding of genetic associations in stratified populations. *American Journal of Human Genetics* **72**, 1492–1504.

Hoggart, C.J., Shriver, M.D., Kittles, R.A., Clayton, D.G. and McKeigue, P.M. (2004). Design and analysis of admixture mapping studies. *American Journal of Human Genetics* **74**, 965–978.

Hopkinson, N.D., Doherty, M. and Powell, R.J. (1994). Clinical features and race-specific incidence/prevalence rates of systemic lupus erythematosus in a geographically complete cohort of patients. *Annals of the Rheumatic Diseases* **53**, 675–680.

Johnstone, I. (2001). On the distribution of the largest principal component. *Annals of Statistics* **29**, 295–327.

Lander, E.S. and Schork, N.J. (1994). Genetic dissection of complex traits. *Science* **265**, 2037–2048.

Lockwood, J.R., Roeder, K. and Devlin, B. (2001). A Bayesian hierarchical model for allele frequencies. *Genetic Epidemiology* **20**, 17–33.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series A* **44**, 226–232.

MacDonald, I.L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series – Monographs on Statistics & Applied Probability*. Chapman and Hall/CRC.

Mackay, D.J. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge.

McKeigue, P.M. (1998). Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *American Journal of Human Genetics* **63**, 241–251.

McKeigue, P.M., Miller, G.J. and Marmot, M.G. (1989). Coronary heart disease in south Asians overseas: a review. *Journal of Clinical Epidemiology* **42**, 597–609.

McKeigue, P.M., Shah, B. and Marmot, M.G. (1991). Relation of central obesity and insulin resistance with high diabetes prevalence and cardiovascular risk in South Asians. *Lancet* **337**, 382–386.

Merrill, R.M. and Brawley, O.W. (1997). Prostate cancer incidence and mortality rates among white and black men. *Epidemiology* **8**, 126–131.

Miller, G.J., Beckles, G.L., Maude, G.H., Carson, D.C., Alexis, S.D., Price, S.G. and Byam, N.T. (1989). Ethnicity and other characteristics predictive of coronary heart disease in a developing community: principal results of the St James Survey, Trinidad. *International Journal of Epidemiology* **18**, 808–817.

Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Neel, J.V., Weder, A.B. and Julius, S. (1998). Type II diabetes, essential hypertension, and obesity as "syndromes of impaired genetic homeostasis": the "thrifty genotype" hypothesis enters the 21st century. *Perspectives in Biology and Medicine* **42**, 44–74.

O'Dea, K., Patel, M., Kubisch, D., Hopper, J. and Traianedes, K. (1993). Obesity, diabetes, and hyperlipidemia in a central Australian aboriginal community with a long history of acculturation. *Diabetes Care* **16**, 1004–1010.

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O'Brien, S.J., Altshuler, D., Daly, M.J. and Reich, D. (2004). Methods for high-density admixture mapping of disease genes. *American Journal of Human Genetics* **74**, 979–1000.

Patterson, N., Price, A.L. and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics* **2**, e190.

Price, A.L., Patterson, N.J. Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909.

Prineas, R.J. and Gillum, R. (1985). *US Epidemiology of Hypertension in Blacks*, chapter 2. Year Book, Chicago, IL, pp. 17–36.

Pritchard, J.K. and Donnelly, P. (2001). Case-control studies of association in structured or admixed populations. *Theoretical Population Biology* **60**, 227–237.

Qualheim, R.E., Rostand, S.G., Kirk, K.A., Rutsky, E.A. and Luke, R.G. (1991). Changing patterns of end-stage renal disease due to hypertension. *American Journal of Kidney Diseases* **18**, 336–343.

Reich, D.E. and Lander, E.S. (2001). On the allelic spectrum of human disease. *Trends in Genetics* **17**, 502–510.

Reid, D.D. (1971). The future of migrant studies. *Israel Journal of Medical Sciences* **7**, 1592–1596.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.

Rudofsky, U.H. Lawrence, D.A. (1999). New Zealand mixed mice: a genetic systemic lupus erythematosus model for assessing environmental effects. *Environmental Health Perspectives* **107**, Suppl. 5, 713–721.

Smith, M.W., Patterson, N., Lautenberger, J.A., Truelove, A.L., McDonald, G.J., Waliszewska, A., Kessing, B.D., Malasky, M.J., Scafe, C., Le, E., de Jager, P.L., Mignault, A.A., Yi, Z., de The, G., Essex, M., Sankale, J.L., Moore, J.H., Poku, K., Phair, J.P., Goedert, J.J., Vlahov, D., Williams, S.M., Tishkoff, S.A., Winkler, C.A., De La Vega, F.M., Woodage, T., Sninsky, J.J., Hafler, D.A., Altshuler, D., Gilbert, D.A., O'Brien, S.J. and Reich, D. (2004). A high-density admixture map for disease gene discovery in African-Americans. *American Journal of Human Genetics* **74**, 1001–1013.

Stephens, J.C., Briscoe, D. and O'Brien, S.J. (1994). Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *American Journal of Human Genetics* **55**, 809–824.

Terwilliger, J.D. and Weiss, K.M. (1998). Linkage disequilibrium mapping of complex disease: fantasy or reality? *Current Opinion in Biotechnology* **9**, 578–594.

Thompson, E.A. and Guo, S.W. (1991). Evaluation of likelihood ratios for complex genetic models. *IMA Journal of Mathematics Applied in Medicine and Biology* **8**, 149–169.

Zimmet, P.Z. (1992). Kelly west lecture 1991. Challenges in diabetes epidemiology–from West to the rest. *Diabetes Care* **15**, 232–252.

Population Association

D. Clayton

Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK

This chapter relates the analysis of population-based studies of associations between genetic polymorphism and disease to the established literature on the analysis of epidemiological case-control studies. Dealing with confounding by use of stratified analyses and regression methods is discussed. The possible influence of hidden population structure is compared with the more generic problem of unmeasured confounding in epidemiology.

36.1 INTRODUCTION

This chapter concerns the detection and analysis of statistical association, *at the population level*, between a binary trait and genotype. Although the main interest is in disease traits, many of the methods are relevant to any binary trait. The only methods that are not more generally applicable are those that depend on the 'rare disease' assumption (that the relative frequency of disease in the population is small).

  Population association between genotype at a particular locus and disease can arise in three ways:

1. the locus may be causally related to the disease, different alleles carrying different risks (*direct* association);

2. the locus may not itself be causal, but may be sufficiently close to a causal locus so as to be in linkage disequilibrium with it (*indirect* association); and

3. the association may be due to *confounding* by population stratification or admixture.

If the association has arisen for the last mentioned reason, it is of little scientific interest. It is therefore important to attempt to exclude such spurious association by appropriate design and/or analysis of studies.

  The problem of confounding is discussed in every textbook of epidemiology. In the present context, confounding would arise if the population contained several ethnic groups,

if allele frequencies at the locus of interest differed between groups, and if disease frequency also differed between groups for reasons quite unrelated to the locus of interest. Such reasons could include genetic differences at a locus in another part of the genome or differences in exposure to environmental causes due to differing customs. It should not be forgotten that such confounding may act to create population association in the absence of a causal link (or linkage disequilibrium with a causal locus), but it may also act in the reverse manner to obscure a causal relationship.

A brief plan of the chapter is as follows. Section 36.2 introduces some notation and discusses statistical measures of association between disease risk and genotype at a diallelic locus. In Section 36.3 the epidemiological case-control study is briefly described and its application to genetic association studies discussed. Section 36.4 discusses simple statistical tests for association in such studies, and Section 36.5 shows how likelihood-based analyses can be carried out using standard software for logistic regression and the log-linear model. When stratification of the population is manifest, control for confounding in case-control studies may be achieved either by post-stratification during analysis or, at the design stage, by matching; these options are discussed in Section 36.6. But neither option is available if population stratification is hidden; Section 36.7 analyses the seriousness of this problem in practice and reviews approaches to dealing with it. The next two sections discuss problems arising as a result of increased polymorphism; Section 36.8 extends the discussion to the case of loci with more than two alleles, and Section 36.9 considers association with haplotypes formed by several closely linked loci. Finally, Section 36.10 briefly discusses some outstanding problems, notably the extension of these ideas to quantitative traits.

Sections 36.2–36.6 draw heavily on the epidemiological literature. The present discussion is of necessity brief, concentrating on matters particularly relevant to studies in genetic epidemiology; for more detailed treatments of this material, see Clayton and Hills (1993) or Breslow and Day (1980).

36.2 MEASURES OF ASSOCIATION

Table 36.1 introduces some notation for the case of a diallelic locus with alleles A and a. On the left, Table 36.1(a) shows the probability distributions of disease conditional upon genotype. Thus, π_{AA}, π_{Aa} and π_{aa} are the *penetrances* of the genotypes AA, Aa and aa respectively. On the right, Table 36.1(b) shows the distributions of genotype given presence or absence of disease in the population. These are denoted by $\gamma^{(1)}, \gamma^{(0)}$ and $\gamma^{(\cdot)}$, respectively.

Association is defined by differences between the penetrances π_{AA}, π_{Aa} and π_{aa} or, equivalently, between the distributions $\gamma^{(1)}$ and $\gamma^{(0)}$. However, the *strength* of association is most naturally expressed in terms of contrasts in the prevalences. A popular measure in epidemiology is the *relative risk* in which each prevalence is expressed relative to the risk in some 'reference' category. If allele a is the more common form, it would be natural to take genotype aa as reference and express the strength of association by the two relative risks:

$$\theta_{AA} = \frac{\pi_{AA}}{\pi_{aa}}, \quad \theta_{Aa} = \frac{\pi_{Aa}}{\pi_{aa}}.$$

Table 36.1 Population distributions of (a) disease given genotype, and (b) genotype given disease.

	Disease				Disease		
Genotype	Yes	No		Genotype	Yes	No	All
AA	π_{AA}	$1 - \pi_{AA}$		AA	$\gamma_{AA}^{(1)}$	$\gamma_{AA}^{(0)}$	$\gamma_{AA}^{(\cdot)}$
Aa	π_{Aa}	$1 - \pi_{Aa}$		Aa	$\gamma_{Aa}^{(1)}$	$\gamma_{Aa}^{(0)}$	$\gamma_{Aa}^{(\cdot)}$
aa	π_{aa}	$1 - \pi_{aa}$		aa	$\gamma_{aa}^{(1)}$	$\gamma_{aa}^{(0)}$	$\gamma_{aa}^{(\cdot)}$
	(a)				(b)		

However, we shall see that it is more convenient to measure strength of association in terms of *odds ratios*. The odds of disease contrasts the probability that disease is present with the probability that it is absent. Thus, for penetrance π, the odds of disease is $\pi/(1 - \pi)$ and the two odds ratios that describe association between disease and genotype are

$$\theta_{AA}^* = \frac{\pi_{AA}}{1 - \pi_{AA}} \bigg/ \frac{\pi_{aa}}{1 - \pi_{aa}}, \quad \theta_{Aa}^* = \frac{\pi_{Aa}}{1 - \pi_{Aa}} \bigg/ \frac{\pi_{aa}}{1 - \pi_{aa}}.$$

The convenience of the odds ratios stems from the relationship between the two parameterisations set out in Table 36.1(a) and (b). Denoting genotype by i, the connection between these two is given by

$$\gamma_i^{(1)} = \frac{\pi_i \gamma_i^{(\cdot)}}{\sum_i \pi_i \gamma_i^{(\cdot)}}, \quad \gamma_i^{(0)} = \frac{(1 - \pi_i) \gamma_i^{(\cdot)}}{\sum_i (1 - \pi_i) \gamma_i^{(\cdot)}},$$

so that

$$\frac{\gamma_i^{(1)}}{\gamma_i^{(0)}} = K \frac{\pi_i}{1 - \pi_i},$$

where K is constant across the three genotypes. Thus, ratios of genotype relative frequencies between people with and without disease are proportional to the corresponding odds of disease, and the odds ratio measures of association can also be written as

$$\theta_{AA}^* = \frac{\gamma_{AA}^{(1)}}{\gamma_{AA}^{(0)}} \bigg/ \frac{\gamma_{aa}^{(1)}}{\gamma_{aa}^{(0)}} = \frac{\gamma_{AA}^{(1)}}{\gamma_{aa}^{(1)}} \bigg/ \frac{\gamma_{AA}^{(0)}}{\gamma_{aa}^{(0)}}, \quad \theta_{Aa}^* = \frac{\gamma_{Aa}^{(1)}}{\gamma_{Aa}^{(0)}} \bigg/ \frac{\gamma_{aa}^{(1)}}{\gamma_{aa}^{(0)}} = \frac{\gamma_{Aa}^{(1)}}{\gamma_{aa}^{(1)}} \bigg/ \frac{\gamma_{Aa}^{(0)}}{\gamma_{aa}^{(0)}}.$$

The practical advantage of odds ratios is that the distributions $\gamma^{(1)}$ and $\gamma^{(0)}$ can be estimated from samples of persons with disease (cases) and of persons free of disease (controls). Particularly when the disease is rare, such studies are much more efficient than total population surveys.

When all penetrances are small, there is very little difference between relative risks and odds ratios. Further, the distribution of genotypes in disease-free subjects (controls) differs little from the population distribution (i.e. $\gamma^{(0)} \approx \gamma^{(\cdot)}$). This will normally be the case, although it may not be so for traits other than diseases.

Before moving on to discuss the design and analysis of population-based case-control studies, we should consider the case where we can assume that the population distribution

of genotypes is in Hardy–Weinberg equilibrium (HWE). Then, denoting the relative frequencies of alleles A and a by $\alpha_A^{(\cdot)}$ and $\alpha_a^{(\cdot)}(= 1 - \alpha_A^{(\cdot)})$, respectively,

$$\gamma_{AA}^{(\cdot)} = (\alpha_A^{(\cdot)})^2, \quad \gamma_{Aa}^{(\cdot)} = 2\alpha_A^{(\cdot)}\alpha_a^{(\cdot)}, \quad \gamma_{aa}^{(\cdot)} = (\alpha_a^{(\cdot)})^2.$$

In general, if there is association between disease and genotype, this law will not hold within cases of disease or in controls (although in the latter case the discrepancy may be imperceptible for the reasons discussed above). For this reason, deviation from HWE in cases of disease is often taken as preliminary evidence for association. However, in one special case, association does not lead to deviation from HWE in cases. This occurs under the *multiplicative penetrance* model in which each copy of allele A multiplies risk by a factor ψ. Then $\theta_{Aa} = \psi$ and $\theta_{AA} = (\psi)^2$ and the distribution of genotypes in cases can be shown to follow the Hardy–Weinberg law, with modified allele frequencies:

$$\alpha_A^{(1)} = \frac{\psi\alpha_A^{(\cdot)}}{\alpha_a^{(\cdot)} + \psi\alpha_A^{(\cdot)}}, \quad \alpha_a^{(1)} = \frac{\alpha_a^{(\cdot)}}{\alpha_a^{(\cdot)} + \psi\alpha_A^{(\cdot)}}.$$

When disease is rare, $\alpha^{(\cdot)} \approx \alpha^{(0)}$ and ψ is very closely approximated by the odds ratio

$$\psi^* = \frac{\alpha_A^{(1)}}{\alpha_A^{(0)}}\Big/\frac{\alpha_a^{(1)}}{\alpha_a^{(0)}} = \frac{\alpha_A^{(1)}}{\alpha_a^{(1)}}\Big/\frac{\alpha_A^{(0)}}{\alpha_a^{(0)}}.$$

It seems most logical to refer to the parameters θ_{AA} and θ_{Aa} as *genotype relative risks* and the parameter, ψ, of the multiplicative model as a *haplotype relative risk*, although these terms have been used somewhat confusingly in the literature.

36.3 CASE-CONTROL STUDIES

In case-control studies of factors that might be associated with disease, the distributions of such factors are compared in a series of cases of the disease and in a control series. The fundamental assumption of such studies is that these two series of subjects may be used to provide unbiased estimates of the corresponding distributions amongst affected and unaffected members of some underlying population. This underlying population of interest is often called the *study base*. When this assumption is met, odds ratio measures of the strength of associations may be estimated, making use of the fact, demonstrated in Section 36.2, that odds ratios are stable whether defined in terms of the distribution of disease conditional upon risk factor or in terms of the distribution of risk factor conditional upon disease.

However, for many reasons, this fundamental assumption of the case-control study may not be met in practice, leading to biased findings. In the late 1960s and early 1970s there emerged an extensive literature cataloguing reasons for such bias. These fall into two broad classes:

1. *selection bias* caused by inappropriate sampling of cases and controls, and

2. *information bias* caused by differential measurement errors in cases and controls arising because measurements of risk factors are usually made when disease status is known to both subject and interviewer.

Studies of genetic association are not immune to either of these problems. For example, cases may be obtained by advertising nationally among clinicians treating the disease of interest, while controls may be recruited in the locality of the investigating centre. Such a design may easily introduce selection bias. A better procedure is to attempt to draw cases from a disease register that seeks to capture all cases of disease in a defined geographical area. The control group should then be a representative sample of the (disease-free) population of the same area. In most countries, however, there are no easy ways to draw such a sample. An alternative approach is to use matched designs; these will be discussed in Section 36.6.

Since genotype is invariant throughout life, genetic association studies are intrinsically less prone to information bias than are studies of environmental influences on disease. The only real danger is posed by poor laboratory procedures; if disease status of subjects is known to the technician who is scoring alleles, or if all cases and all controls are genotyped at different times, bias may result. Such bias is easily excluded by genotyping in random order, remaining blind to disease status.

Another problem in the interpretation of case-control studies has been termed *inci-dence/prevalence bias*. This is not really a bias at all but a failure to carefully define the nature of the disease variable. Thus, a table such as Table 36.1 could describe two rather different types of studies:

1. a cross-sectional study in which a population is surveyed at a fixed point in time and *current disease status* recorded, and

2. a longitudinal study in which a population of subjects, initially free of disease, are followed for a defined period and *new disease occurrence* recorded.

In the former case, π_{AA}, π_{Aa} and π_{aa} are disease *prevalences*, while in the latter case they are *incidences*. Incidence and prevalence are different quantities and may be influenced in different ways. Most importantly, a factor that is related to the duration for which a subject with disease remains in the population before death or migration will be reflected in prevalence even if it has no influence on incidence. Case-control studies based upon currently prevalent cases in a population measure effects of risk factors upon disease prevalence, while studies based around newly occurring cases in a defined period measure effects on disease incidence.

An example of the data arising from a case-control (incidence) study is shown in Table 36.2. The study was reported by Dunning *et al.* (1997) and concerns possible asso-ciation of *common* polymorphisms in the *BRCA1* gene with breast cancer. Table 36.2(a) shows the frequency distribution of the Pro871Leu genotype in 800 cases and 572 controls. These frequencies can be used to estimate the corresponding population distributions and hence the odds ratios. Taking the Pro/Pro genotype as reference, the odds ratio estimates are

$$\hat{\theta}^*_{LL} = \frac{89}{342} \Big/ \frac{56}{266} = 1.236, \quad \hat{\theta}^*_{LP} = \frac{369}{342} \Big/ \frac{250}{266} = 1.148.$$

Since occurrence of breast cancer is a rare event, the odds ratios θ^*_{LL} and θ^*_{LP} closely approximate the corresponding genotype relative risks θ_{LL} and θ_{LP} in the population.

Table 36.2(b) shows allele frequencies in the 1600 chromosomes of cases and the 1142 chromosomes of controls. These can be used to estimate the allele distributions $\alpha^{(1)}$ and $\alpha^{(0)}$. Again, since breast cancer is rare, the allele frequencies in healthy controls are very

Table 36.2 Distributions of Pro871Leu polymorphism in the *BRCA1* gene yin breast cancer cases and in population controls.

| Pro871Leu | Subjects | | Pro871Leu | Chromosomes | |
genotype	Case	Control	allele	Case	Control
Leu/Leu	89	56	Leu	547	362
Leu/Pro	369	250	Pro	1053	782
Pro/Pro	342	266			
Total	800	572	Total	1600	1142
	(a)			(b)	

nearly the same as in the population at large so that $\alpha^{(0)} \approx \alpha^{(\cdot)}$ and the haplotype relative risk may be estimated by the odds ratio

$$\hat{\psi}^* = \frac{547}{1053} \Big/ \frac{362}{782} = 1.122.$$

The corresponding genotype relative risks, θ_{LL} and θ_{LP}, predicted by the multiplicative model are 1.254 and 1.122, respectively. The next section deals more formally with statistical inference concerning these parameters.

36.4 TESTS FOR ASSOCIATION

The statistical model underlying data such as those of Table 36.2(a) is that of two samples drawn from multinomial distributions. Denoting observed frequencies by f and genotype by the subscript i, the corresponding log likelihood is

$$\sum_i \left(f_i^{(1)} \log \gamma_i^{(1)} \right) + \sum_i \left(f_i^{(0)} \log \gamma_i^{(0)} \right).$$

By re-expressing the case genotype probabilities, $\gamma^{(1)}$, in terms of $\gamma^{(0)}$ and θ^*, this may be written as a function $\ell(\theta^*, \gamma^{(0)})$ of the control probabilities and the odds ratios. Since, by definition, $\theta_{aa}^* = 1$ and since $\sum_i \alpha_i^{(0)} = 1$, there are effectively only two free scalar parameters in each of the vectors θ^* and $\gamma^{(0)}$, the former being the parameter of interest and the latter being a 'nuisance parameter'.

There are two standard asymptotic tests of the hypothesis of no association, $H_0 : \theta^* = \mathbf{1}$:

1. *The log-likelihood ratio (LLR) test*. Denoting the maximum- likelihood (ML) estimates of $\gamma^{(0)}$ and θ^* by $\hat{\gamma}^{(0)}$ and $\hat{\theta}^*$, and the ML estimate of $\gamma^{(0)}$ under H_0 by $\hat{\hat{\gamma}}^{(0)}$, this test statistic is defined as twice the difference between the corresponding log likelihoods:

$$2 \left[\ell(\hat{\theta}^*, \hat{\gamma}^{(0)}) - \ell(\mathbf{1}, \hat{\hat{\gamma}}^{(0)}) \right].$$

2. *The score test*. If $\mathbf{u}$ represents the value of the vector of first derivatives of the log likelihood with respect to $\boldsymbol{\theta}^*$ evaluated at $\boldsymbol{\theta}^* = \mathbf{1}$ and $\boldsymbol{\gamma}^{(0)} = \hat{\boldsymbol{\gamma}}^{(0)}$, and $\mathbf{V}$ represents an estimate of its variance under H_0 (obtained by standard arguments from the matrix of second derivatives of the log-likelihood function), this test is defined by the quadratic form:

$$\mathbf{u}^t \mathbf{V}^\ominus \mathbf{u},$$

where $\ominus$ denotes a generalised inverse matrix.

Both of these statistics are asymptotically distributed as χ^2 on 2 degrees of freedom (df), and both can be expressed as simple functions of the observed frequencies, $f_i^{(1)}$ and $f_i^{(0)}$, and the corresponding 'expected' frequencies, $e_i^{(1)}$ and $e_i^{(0)}$, fitted under H_0:

$$\text{LLR test} = 2 \sum_i \left(f^{(1)} \log \frac{f^{(1)}}{e^{(1)}} + f^{(0)} \log \frac{f^{(0)}}{e^{(0)}} \right),$$

$$\text{Score test} = \sum_i \left[\frac{(f^{(1)} - e^{(1)})^2}{e^{(1)}} + \frac{(f^{(0)} - e^{(0)})^2}{e^{(0)}} \right].$$

The expected frequencies are calculated in the usual manner for tests of independence in contingency tables:

$$e_i^{(1)} = \frac{f_{\cdot}^{(1)} f_i^{(\cdot)}}{f_{\cdot}^{(\cdot)}}, \quad e_i^{(0)} = \frac{f_{\cdot}^{(0)} f_i^{(\cdot)}}{f_{\cdot}^{(\cdot)}}$$

(where $\cdot$ in subscript or superscript denotes summation). In our example of Table 36.2, the LLR and score tests are 2.056 and 2.055 respectively, corresponding to a P-value of approximately 0.36, so these tests do not suggest statistically significant association between breast cancer incidence and this polymorphism.

Asymptotic confidence intervals for estimates of odds ratios are provided by standard likelihood theory. It is normal to assume that the logarithm of an odds ratio estimate is asymptotically normally distributed with standard error estimated by the square root of the sum of reciprocals of the four frequencies used in the estimate. For example, for our estimate of θ_{LL} obtained in Section 36.3, the standard error of the log odds ratio is

$$\sqrt{\frac{1}{89} + \frac{1}{342} + \frac{1}{56} + \frac{1}{266}} = 0.189.$$

The approximate 95 % confidence limits for the log odds ratio are $\log 1.236 \pm 1.96 \times 0.189$. These correspond to limits on the odds ratio of 0.85 and 1.79. Thus there is no suggestion that the Leu/Leu genotype is associated with increased risk of breast cancer.

Analysis of the chromosome count table, Table 36.2(b), follows very similar lines. However, for us to be able to legitimately ignore subject and treat chromosomes as independent observations, the assumption of HWE is essential. The test of the null hypothesis only requires that we assume HWE in the population since, under H_0, this ensures that both case and control distributions of genotype obey the Hardy–Weinberg law. However, for validity of the standard method of calculating a confidence interval for the odds ratio, we require the additional assumptions of (1) a rare disease (so that controls

will be in HWE), and (2) the multiplicative model for penetrances (to ensure that cases are also in HWE).

For the data of Table 36.2(b), the LLR and score test statistics are 1.954 and 1.949 respectively. However, since these are tests of only a single parameter, they should be compared with the χ^2 distribution on 1 df, yielding a P value of approximately 0.16. The standard error of the log odds ratio estimate is

$$\sqrt{\frac{1}{547} + \frac{1}{1053} + \frac{1}{362} + \frac{1}{782}} = 0.0826.$$

Similar calculations as before lead to 95 % confidence limits for the haplotype relative risk of 0.95 and 1.32. As before, there is little evidence for association between breast cancer and this polymorphism.

Since the tests based on chromosomes has only 1 df, it must be expected to be more powerful that the 2 df tests – at least against alternative hypotheses, which are close to the multiplicative model. However, the need to assume HWE might be regarded as undesirable. If so, this assumption can be avoided by analysing the full genotype data of Table 36.2(a) using the multiplicative model

$$\theta_{\mathrm{LP}}^* = \psi^*, \quad \theta_{\mathrm{LL}}^* = (\psi^*)^2.$$

ML estimation of the parameter of this model can be carried out using logistic regression (see Section 36.5) and, by comparing maximised log likelihoods under null and alternative hypotheses, LLR tests are readily computed using this method. The score test of $H_0 : \psi^* = 1$ is also relatively well known, being equivalent to the Cochran–Armitage χ^2 test (1 df) for trend in the proportion of cases across rows of Table 36.2(a) (Armitage, 1955). In our example, the ML estimate of ψ^* is 1.125 and the standard error of its logarithm is 0.834, leading to confidence limits of 0.95 and 1.32 for ψ^*. The LLR test statistic is 1.991, while the score test takes the value 1.984. These results agree quite closely with those obtained by analysis of the chromosome counts (Table 36.2b).

Although the resistance of this approach to deviations from HWE is to be welcomed, such deviations cannot be entirely ignored. Deviation from HWE constitutes evidence of population stratification and, when this is present, there is a danger that such stratification could confound the association between disease and genotype. We will return to this problem in Section 36.7.

Just as there is a case for relaxing the HWE assumption while carrying out haplotype relative risk analyses based on the multiplicative model, there can be a case for making an HWE assumption when carrying out an analysis of genotype relative risks. The first three columns of Table 36.3 shows some data (also drawn from Dunning *et al.*, 1997) concerning another polymorphism in BRCA1.

The 2 df LLR test statistic is 11.718 and the (perhaps implausible) suggestion is that the homozygous Arg/Arg genotype is protective against breast cancer. However, because the Arg allele is uncommon, the frequency of the Arg/Arg genotype is very low in controls and, therefore, poorly estimated. Lathrop (1983) pointed out that the power of the 2 df test could be improved by assuming HWE in the population. With this assumption, and assuming a rare disease, both cases and controls will be in HWE at the null hypothesis, while under the alternative hypothesis only controls will be in HWE. The remaining columns show the expected frequencies under these assumptions. The LLR

Table 36.3 Distributions of Gln356Arg polymorphism in the *BRCA1* gene in breast cancer cases and in population controls, with expected frequencies assuming HWE.

Gln356Arg genotype	Observed		Expected (H_0)		Expected (H_1)	
	Case	Control	Case	Control	Case	Control
Arg/Arg	0	7	2.80	2.31	0	3.07
Arg/Gln	81	74	87.00	71.77	81	81.86
Gln/Gln	684	550	675.19	556.92	684	546.07

test for association compares the log likelihoods for these two fitted models and has 2 df. In this example this test statistic is 7.94 – a rather less extreme value than obtained with the simple 2 df test, which is inflated by the (probably chance) excess of Arg/Arg genotypes in controls.

The tests described so far in this section rely on asymptotic approximations which may be rather poor when there are small cell frequencies such as in Table 36.3. However, exact tests may be easily computed. The natural probability model is that the distribution of genotypes (or alleles) for case and control subjects (chromosomes) follow two multinomial distributions, which are identical under H_0. However this model is unsatisfactory for computing the null distribution of test statistics, owing to unknown 'nuisance' parameters – the common genotype (allele) population relative frequencies. This problem can be avoided by arguing conditionally upon the marginal genotype (allele) frequencies. The tables then follow hypergeometric distributions under H_0. This is the argument used in the construction of Fisher's exact test for 2×2 tables such as Table 36.2(b), but the same argument may be used to calculate exact *P* values for any of the statistics discussed above. In practice, complete evaluation of hypergeometric distributions is often too demanding and a simulation approach must be used. This proceeds by assigning the observed genotypes to cases and controls at random, recalculating the test statistic each time and counting the proportion of the time the statistic exceeds its value for the observed data. Random assignment of genotypes amongst subjects can be carried out in two ways. For a total of *N* subjects, we might

1. assign the *N* observed genotypes to cases and controls at random, or

2. assign the 2*N* observed alleles to cases and controls with no consideration of the pairing of alleles in the original genotypes.

The second of these approaches yields the *P* value under the HWE assumption, while the first yields a *P* value that makes no such assumption; this is more generally appropriate, and has the additional advantage of being rather easier to compute.

One further test should be discussed. This is a 1 df test in which the alternative model assumes dominance of one allele. This is a conventional test for association in the 2×2 contingency table in which cases and controls are classified as carriers of the dominant allele or non-carriers (i.e. homozygous for the recessive allele). In the absence of prior knowledge, this strategy will require two tests to be carried out since either allele could be dominant.

The question is, which test should be used? The 1 df tests based on the dominance model is frequently used in the study of single gene, or 'Mendelian' disorders while the

1 df multiplicative Cochran–Armitage test is more often preferred in studies of 'complex' disorders. This largely follows empirical experience and, perhaps, reflects the different mechanisms that apply in these settings. An additional consideration is that, when the polymorphism studied is not itself causal but related to disease via linkage disequilibrium with a nearby causal variant, dominance tends to be masked and the observed risk relationship becomes closer to the multiplicative penetrance prediction. Many authors have compared the power of these various tests under different model assumptions. Except in the case where the minor allele frequency is low, so that there are few subjects homozygous for the minor allele, the 2 df test against unrestricted alternatives behaves reasonably well against all alternatives.

36.5 LOGISTIC REGRESSION AND LOG-LINEAR MODELS

In Section 36.4 it was mentioned that maximisation of likelihoods for some models could be achieved by use of a computer program for logistic regression. However, since the 'response' variable measured in cases and controls is a genotype, which has *three* possible values, the relevance of logistic regression is not immediately apparent. This section explains this and also shows how another standard statistical technique, log-linear modelling, may be used to reproduce all the analyses described in Section 36.4. This will be useful for the extensions of this methodology to be discussed in later sections.

Although the natural way to model case-control data is in terms of probability distributions of genotype conditional upon disease status, reflecting the manner in which the data are generated, it has been shown that identical results are obtained from a likelihood-based analysis in which case-control status is regarded as the random outcome (Prentice and Pyke, 1979). The response variable is taken as the proportions of subjects with each genotype who are cases, and this is related via a logistic regression model to a design matrix expressing the precise model to be fitted. If π_i^* represents the probability that a subject, drawn at random from those subjects in the case-control study with genotype i, is a case. The logistic regression model is

$$\log \frac{\pi_i^*}{1 - \pi_i^*} = \mathbf{x}_i^{\mathrm{T}} \beta,$$

where $\mathbf{x}_i^{\mathrm{T}}$ is the ith row of the design matrix. With suitable choice of design matrix, the regression coefficients, β, are the logarithms of the odds ratio parameters discussed in Section 36.2.

Table 36.4 illustrates how two analyses of the data of Table 36.2 could be carried out using logistic regression. The first design matrix fits a fully saturated model and corresponds to the conventional 2 df test for association in the 3×2 contingency table. The coefficients of g_1 and g_2 in the model are the logarithms of the odds ratios θ_{LP}^* and θ_{LL}^* respectively, and the 2 df LLR test can be carried out by dropping g_1 and g_2 from the model. The second model is the multiplicative model in which $\theta_{\mathrm{LL}}^* = (\theta_{\mathrm{LP}}^*)^2$. The coefficient of g in this model corresponds to $\log \psi^*$ in our earlier notation, and the 1 df LLR test for trend can be carried out by dropping g from the model. The dominance

Table 36.4 Logistic regression of the Pro871Leu data.

	Design matrices				
Observed proportion	2 df model			1 df model	
	Constant	g_1	g_2	Constant	g
89/145	1	0	1	1	2
369/619	1	1	0	1	1
342/606	1	0	0	1	0

model can also be fitted in logistic regression, simply by coding the indicator g as 1 for carriers of the dominant allele and 0 otherwise (not shown).

This approach argues conditionally upon genotype and hence it is impossible to incorporate the HWE assumption, which is a model for the distribution of genotypes. A more flexible, but rather less convenient, alternative – log-linear modelling – allows a wider class of models to be fitted. In this approach, frequencies in a contingency table are assumed to be distributed as Poisson variates, the logarithms of their expected values obeying a linear model. Again it can be shown that this approach leads to likelihood inferences for odds ratios that are identical to those obtained under the assumption of multinomial distributions of genotypes conditional upon disease status.

Table 36.5 demonstrates how log-linear modelling can be used to reproduce two analyses discussed in Section 36.4. The cell frequencies of Table 36.2 are the response variable, and the first design matrix represents a saturated model in which disease status, genotype and their interaction are included. The parameters representing interaction in this model are the coefficients of $d.g_1$ and $d.g_2$ and these are the logarithms of the odds ratios θ^*_{LP} and θ^*_{LL} respectively. Dropping both these variables from the model provides the 2 df test for association between disease and genotype.

The second design matrix listed in Table 36.5 is a more restrictive model which assumes (1) the multiplicative mode for odds ratios, and (2) HWE equilibrium in both cases and controls. Note that this model must include an 'offset' (equivalent to the inclusion of a variable whose coefficient is constrained to take the value 1). Use of this design matrix reproduces the simple analysis of the 2×2 table of chromosome counts; the coefficient of the variable $d.g$ (the disease–genotype interaction term) is the odds ratio in this 2×2

Table 36.5 Log-linear modelling of the Pro871Leu data.

	Design matrices										
Observed frequency	2 df model						1 df + HWE model				
	Constant	d	g_1	g_2	$d.g_1$	$d.g_2$	Offset	Constant	d	g	$d.g$
89	1	1	0	1	0	1	0	1	1	2	2
369	1	1	1	0	1	0	log 2	1	1	1	1
342	1	1	0	0	0	0	0	1	1	0	0
56	1	0	0	1	0	0	0	1	0	2	0
250	1	0	1	0	0	0	log 2	1	0	1	0
266	1	0	0	0	0	0	0	1	0	0	0

table. An advantage of fitting the model to the full table of genotype counts rather than to the collapsed table of chromosome counts is that this provides tests of fit of modelling assumptions that are implicit in the latter analysis. An additional advantage is that the alternative analyses discussed in Section 36.4 can also be carried out quite easily. Thus, if the columns $d.g_1$ and $d.g_2$ in the first matrix are replaced by the single column $d.g$, the model assumes the multiplicative (1 df) model but does not assume HWE. Conversely, if the single column $d.g$ in the second matrix is replaced by the two columns $d.g_1$ and $d.g_2$, we have Lathrop's model which assumes HWE in controls but allows 2 df for disease–genotype association. As before, dominance models may also be considered with appropriate coding of the indicator g.

36.6   STRATIFICATION AND MATCHING

In the introduction to this chapter, the problem of confounding was briefly discussed; association between disease and genotype might be attributable to a third variable, not on a causal path between genotype and disease, which is independently related to both variables. In genetic epidemiology, most concern is usually related to the possibility of confounding by admixture or ethnic stratification of populations, but there are other possibilities. For example, a gene that is related to the tendency to smoke cigarettes (and such genes have been found) may be associated with lung cancer. In one sense such a gene is indeed a cause of lung cancer but, in the context of the question as to whether the gene is directly involved in the biology of cancer, cigarette smoking would be more likely to be regarded as a confounder.[1]

When the confounding variable has been measured in the study, it is relatively straightforward to deal with the problem during analysis. The classical method in epidemiology is by *stratification* of the analysis by the potentially confounding variable and testing for association between factor of interest (here genotype) and disease *within strata*. Figure 36.1 illustrates this idea.

Each stratum (ethnic group, for example) contributes a contingency table of the form discussed in Section 36.4 – either a 3×2 table of counts of subject by genotype or a 2×2 table of counts of chromosomes by allele.

Since each stratum, taken alone, may contain insufficient data to test for association, it is necessary to pool the evidence for association over strata. An obvious way of doing this
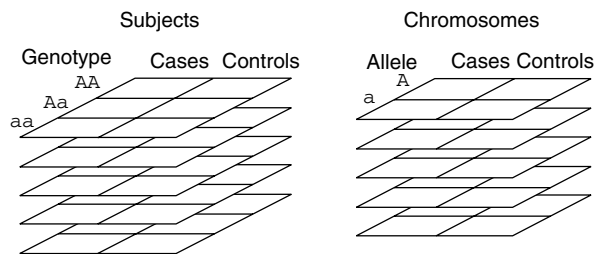
Figure 36.1   Stratified analysis of case–control studies.

---

[1] This example also illustrates why the concept of confounding is rather difficult to define precisely, since this depends crucially on the nature of causality.

would be to simply add the χ^2 values over strata, remembering to also add their degrees of freedom. However, the proliferation of degrees of freedom with this approach indicates that it is based on a very flexible model for association, with either two or one different parameter for each stratum. Since all these parameters will be imprecisely estimated, this approach must be expected to lack power.

An alternative approach is to adopt the model in which the parameters measuring association between genotype and disease are assumed to be constant across strata, although the distributions of both disease and genotype are allowed to vary among strata. With this model as the alternative hypothesis, tests of association have either 1 or 2 df. In the case of the 1 df tests, the score tests are well known and can be carried out using hand calculator. The test for association in a stack of 2×2 tables of allele counts is the Mantel–Haenszel test and an easily calculated estimate of their common odds ratio, ψ^*, is also available (Mantel and Haenszel, 1959). If we wish to avoid the HWE assumption, the 1 df score test is the stratified test for trend in proportions in the stack of 3×2 tables of genotype counts (Mantel, 1963); this is often known as the *Mantel extension test*. The 2 df score test for the stack of 3×2 tables is rather more difficult to calculate and is not explicitly described in the literature. The more general case of the $k - 1$ df test in a stack of $k \times 2$ tables has been described, but calculation of the variance of the score vector requires matrix inversion. An approximate procedure is to (1) calculate expected frequencies under the null hypothesis in each stratum separately, (2) add observed and expected frequencies over strata and (3) calculate a chi-squared test in this margin using the usual formula $\sum (O - E)^2 / E$. This approximation can be shown to be conservative – the value of χ^2 obtained this way is always slightly smaller than that obtained if the variance–covariance matrix of the score vector is calculated correctly.

LLR tests are more difficult to calculate than score tests, since the ML estimates of common odds ratios require iterative computation. However, the logistic regression and log-linear modelling approaches of Section 36.5 are easily extended to encompass stratified analyses.

Since logistic regression argues conditionally on genotype, this method explicitly allows association between stratum and genotype and to control for confounding it is necessary only to allow disease status to depend on stratum by including stratum in the regression model as a categorical variable (or 'factor'). If there are S strata, this introduces a further $S - 1$ parameters in the regression model. Thus the model of no association between disease and genotype within strata has S parameters, and the alternative model introduces a further 1 or 2 parameters.

Log-linear modelling is slightly more complicated since the relationship between genotype and stratum must be explicitly modelled; we require $S - 1$ parameters for the distribution of subjects between strata and, if HWE is assumed within strata, we require a further $S - 1$ parameters to model variation of allele relative frequencies amongst strata. As in the logistic regression approach, $S - 1$ parameters are needed to model association between disease status and stratum. Thus the model of no association between disease and genotype within stratum has a total of $3S$ parameters. Again, the alternative model introduces a further 1 or 2 parameters for association between disease and genotype. The design matrix for a two-stratum analysis is illustrated in Table 36.6.

Dropping the last column from the model produces a 1 df LLR test for association between genotype and disease within strata.

Table 36.6 Stratified analysis by log-linear modelling.

Stratum	Disease	Genotype	Offset	Design matrix						
				Constant	s	d	g	$s.d$	$s.g$	$(d.g)$
1	Case	AA	0	1	1	1	2	1	2	2
1	Case	Aa	$\log 2$	1	1	1	1	1	1	1
1	Case	aa	0	1	1	1	0	1	0	0
1	Control	AA	0	1	1	0	2	0	2	0
1	Control	Aa	$\log 2$	1	1	0	1	0	1	0
1	Control	aa	0	1	1	0	0	0	0	0
2	Case	AA	0	1	0	1	2	0	0	2
2	Case	Aa	$\log 2$	1	0	1	1	0	0	1
2	Case	aa	0	1	0	1	0	0	0	0
2	Control	AA	0	1	0	0	2	0	0	0
2	Control	Aa	$\log 2$	1	0	0	1	0	0	0
2	Control	aa	0	1	0	0	0	0	0	0

Computation of exact P values based on any of these test statistics may be obtained by simulation, randomly assigning genotypes to subjects within strata while ignoring disease status.

The use of regression models to control for confounding has much to recommend it, since such approaches extend without difficulty to deal with several potential confounders. For this reason, their use has become widespread in epidemiology. But whether analysis is by regression or by the more classical Mantel–Haenszel methods, it should be noted that the need to control confounding may result in some loss in precision of the study if the stratifying variable is strongly related to disease. In this case, some strata may have many fewer controls than cases, while others may have many fewer cases than controls. An optimal design would maintain the ratio of cases to controls across strata, and a study that is carried out so that this is so is called a *group matched* study.

Within a group matched study, there is no relationship between stratum and disease status and for some time it was believed that the matched design eliminates confounding so that the analyses discussed in this section are no longer necessary. Unfortunately this is not the case, owing to the non-collapsibility of the odds ratio as a measure of association; even when the distribution of disease status is constant over strata, and when the odds ratios are constant across strata, the odds ratios in the marginal table are closer to one than are the stratum-specific odds ratios. Likewise, the test for association in the marginal table is incorrect. Thus it is still necessary to include stratum effects when modelling data from a matched study. The only exception to this rule occurs in the degenerate case in which the distribution of genotype also does not vary among strata.

At the limit, matching can be so fine that each single case has its own control(s). This is an *individually matched* case-control study. Although, in principle, the analysis of such studies is the same as above, we encounter a new technical difficulty; since we must include stratum effects in the model and since we introduce a new stratum with each new case, the number of parameters in the model increases just as fast as the total study size. Unfortunately, the asymptotic properties of likelihood inference in logistic regression break down in these circumstances. The solution is to use *conditional* logistic regression – a variant of logistic regression analysis in which the parameters expressing

stratum effects are eliminated from the likelihood by use of a conditional argument rather than by attempting to estimate them. Before concluding the topic of individually matched studies, it should be noted that an important special case occurs when the matching is by nuclear family so that cases are compared with unaffected sibling controls. This type of study is the 'sib TDT' study described by Spielman and Ewens (1998); such studies and their analysis are discussed in **Chapter 38** of this handbook.

We have seen that matching controls to cases at the design stage does not simplify analysis. It has also been shown elsewhere that the gains in efficiency achieved by matching are modest except when there is very strong confounding. These two facts might lead us to question its usefulness. However, there is another important motivation for matched designs. In Section 36.3 attention was drawn to the problem of selection bias in case-control studies and in particular to the lack of appropriate sampling frames for controls in most countries. However, matching can simplify the sampling problem. If controls are matched to cases, e.g. by general practitioner or by neighbourhood, it can be much easier to ensure that (1) all eligible cases within a given stratum are indeed captured in the study, and (2) that controls representative of the stratum are selected. It is these considerations rather than the desire for efficient control of confounding that accounts for the popularity of matching in case-control studies.

36.7   UNMEASURED CONFOUNDING

A criticism frequently levelled at case-control studies in genetic epidemiology is the possibility of confounding by *unmeasured* population stratification or admixture. But the possibility of unmeasured confounding is not unique to genetic epidemiology. Indeed, one of the most celebrated controversies of epidemiology surrounded Fisher's 'constitutional hypothesis', which suggested that unmeasured genetic factors could confound the observed association between smoking and lung cancer. Ultimately, such explanations can always be offered for observed associations. In general epidemiology, the only counter arguments are in terms of plausibility. An explanation in terms of unmeasured confounding must be plausible, both *biologically* and *quantitatively*. The quantitative plausibility of explanations based on unmeasured confounding has, not surprisingly, been the subject of much attention in the epidemiological literature, but has been largely ignored in the debate concerning population-based vs family-based association studies in genetic epidemiology. An early finding of such work was that putative confounders must be very strongly related to both disease and risk factor for their confounding effect to be appreciable (Bross, 1967).

To carry these arguments over into genetic epidemiology, it is most convenient to assume HWE and the multiplicative model for association, so that the theory can be expressed in terms of properties of 2×2 tables of chromosome counts. Let there be two hidden strata and assume that the odds ratios between disease status and allele is ϕ^* in each stratum. Then it can easily be shown that the *marginal* odds ratio is given by

$$\phi^* \left[\frac{\zeta p_A + (1 - p_A)}{\zeta p_a + (1 - p_a)} \right],$$

where p_A represents the proportion of control A alleles drawn from the high-risk stratum, p_a is the proportion of a alleles drawn from this stratum and ζ is the odds ratio between

stratum and disease. The term in square brackets represents the factor by which the true odds ratio is inflated or deflated by confounding and has been termed the *confounding risk ratio* (Miettinen, 1972). Breslow and Day (1980) give examples of such calculations.

The problem of unmeasured confounding is no more serious when the factor of interest is a gene than when it is an environment exposure. Indeed it could be argued that it is considerably less serious, since it is possible to collect *evidence* for or against the existence of substantial admixture within a population. This can be checked by typing a number of genetic markers that are sufficiently distant from the locus of interest to be assumed to be in linkage disequilibrium with it. The presence of admixture is then indicated by deviation from HWE at each locus, by population associations amongst pairs of loci and, if disease risk differs among strata, by widespread association between genotype and disease – attributable to confounding. Although this possibility has been recognised for some time, formal methods of analysis have not been proposed until recently. Pritchard and Rosenberg (1999) propose testing for confounding by unobserved stratification by summation of χ^2 tests for association over markers. They then seek to adjust the P value for the candidate locus. A full description of approaches to modelling stratification and admixture is to be found in **Chapter 35** of this handbook.

A further approach, 'genomic control', is due to Devlin and Roeder (1999), who suggested that the effect of admixture and stratification is to inflate the 1 df Cochran–Armitage marker/disease association test statistics by a constant factor, which they designated λ. They showed that λ is expected to increase with the study size, essentially reflecting the increased sensitivity of large studies to hidden confounding, but argued that, given a large number of tests for markers widely spread on the genome, λ can be estimated empirically and the distribution of test statistic corrected. A simple estimate of λ is the ratio of the mean of the tests falling in the lower part of the distribution with its expectation under the χ^2 assumption. Devlin and Roeder suggested use of the smallest 50 % of test values, but others have suggested that 90 % can be used.

There are a number of difficulties with the idea of genomic control. First, although the type 1 error is corrected, this correction is achieved at the price of a loss of power to detect true associations. Secondly, there is an assumption of 'exchangeability' of markers; the estimate of λ will reflect inflation of tests due to small differences in allele frequencies at a large number of loci, due to 'genetic drift' at the point of separation of ancestral populations, However, some loci could have been under selection pressures, resulting in substantially larger differences. A final difficulty, although less serious given recent advances in high-throughput genotyping methods, is that accurate estimation of λ requires very large numbers of loci to be studied.

Devlin and Roeder (1999) discussed the case of 'cryptic Relatedness' between study subjects, leading to violation of the assumption of independent data points. They showed that this is serious when the relatedness is stronger between pairs of cases (and possibly pairs of controls) than it is between a case-control pair. Again, they suggested that the effect of this is to inflate the test statistics by a factor λ that increases with the size of the study. In principle, cryptic relatedness is no different from the stratification and admixture effects discussed by other authors. Cryptic relatedness is concerned with the effect of more recent coalescence of the genealogy of study subjects, while stratification represents ancient coalescence into a few ancestral populations. The difference is that, whereas the latter can be approached by modelling structure by using ancestry informative markers,

cryptic relatedness will not be corrected by such an approach and genomic control remains the only remedy.

36.8 MULTIPLE ALLELES

The discussion to this point has been limited to the case of a diallelic genetic locus. This section discusses the case of a locus with $K > 2$ alleles. The most serious consequence that must be considered is the proliferation of possible genotypes that follows. A test based on genotype relative risks has $K(K + 1)/2 - 1$ df and lacks power even for quite modest values of K. However, the multiplicative model generalises naturally and allows tests based on haplotype relative risks. Denoting the genotype with alleles j and k by (j, k), the multiplicative model for genotype relative risks is

$$\theta_{(j,k)} = \psi_j \psi_k,$$

where $\psi_j, \psi_k (j, k = 1 \ldots K)$ are haplotype relative risks. One allele, usually the most common, is taken as reference so that the corresponding ψ_j takes the value 1.0 by definition. The global test of no association, $H_0 : \psi_j = 1, j = 1 \ldots K$ has $K - 1$ df. In the context of case-control studies, genotype and haplotype relative risks are closely approximated by the corresponding odds ratio parameters. In particular, taking allele 1 as reference, ψ_j is closely approximated by

$$\psi^* = \frac{\alpha_j^{(1)}}{\alpha_j^{(0)}} \bigg/ \frac{\alpha_1^{(1)}}{\alpha_1^{(0)}} = \frac{\alpha_j^{(1)}}{\alpha_1^{(1)}} \bigg/ \frac{\alpha_j^{(0)}}{\alpha_1^{(0)}}.$$

All of the methods discussed above in the case of a diallelic locus extend naturally. For example, the model that combines HWE and multiplicative assumptions can be fitted using the $K \times 2$ table of chromosome counts. Relaxation of the HWE assumption is most conveniently achieved by logistic regression. Denoting the probability that a subject, drawn at random from those subjects in a case-control study with genotype (j, k), is a case by $\pi^*_{(j,k)}$, the multiplicative model corresponds with the logistic regression model

$$\log \frac{\pi^*_{(j,k)}}{1 - \pi^*_{(j,k)}} = \beta_0 + \beta_j + \beta_k.$$

The design matrix has K columns, the first column being the unit vector, and the elements of remaining columns $j = 2 \ldots K$ taking values 0, 1 or 2 reflecting the number of times allele j occurs in each genotype. The dominance model is also readily extended to the multiple allele case, although there are more versions of the model to consider (since various combinations of alleles could be dominant). For tests for indirect association, the model of multiplicative effects of alleles is the one most frequently used.

Tests for association can be calculated by using logistic regression to calculate likelihood ratio tests. The score test, which generalises the Cochran–Armitage test, is essentially Hotelling's T^2 test, comparing the vector of means for the allele indicator variables between cases and controls (Xiong *et al.*, 2002; Chapman *et al.*, 2003; Fan and Knapp, 2003). To deal with observed potential confounders, the regression model can be extended

by entering them into the regression model. The equivalent generalisation of the score test is a stratified version of the Hotelling's T^2 statistic.

The multiplicative model may also be defined by the property that the genotype relative risk for a heterozygous genotype is the geometric mean of the two homozygous genotypes defined by its alleles. In terms of odds ratios,

$$\theta^*_{(j,k)} = \sqrt{\theta^*_{(j,j)}\theta^*_{(k,k)}}.$$

A less restrictive approach, which nevertheless avoids the large number of degrees of freedom of the test between genotype frequencies, is to assume, in the case of a diallelic locus, only that the heterozygous genotype relative risk falls in the interval bounded by the two homozygous relative risks. The model could be fitted under these order constraints, but the asymptotic distribution of the LLR test statistic would be complex. An alternative approach is suggested by noting that, under the multiplicative model, the effects of the alleles on the two chromosomes are additive on the log odds scale. This can be extended by assuming additivity of effects *on some other scale*. The problem of the scale on which two effects are additive has received some attention in the epidemiological literature, since it is central to the discussion of *synergism* of risk factors.

A natural approach is suggested by generalised linear models (Nelder and Wedderburn, 1972); we may replace the logit transformation of the probabilities π^* by a more general 'link function', $g\left(\dfrac{\pi^*}{1-\pi^*}\right)$:

$$g\left(\frac{\pi^*_{(j,k)}}{1-\pi^*_{(j,k)}}\right) = \beta_0 + \beta_j + \beta_k,$$

a convenient choice of link function being the Box–Cox transformation of the odds (Box and Cox, 1964)

$$g(x) = \frac{x^\rho - 1}{\rho}, \quad \rho \neq 0,$$
$$= \log x, \quad \rho = 0.$$

This extension to the logistic regression model was proposed by Guerrero and Johnson (1982). In the present context, as $\rho \to \infty$, $\pi^*_{(j,k)}$ tends to the larger of $\pi^*_{(j,j)}$ and $\pi^*_{(k,k)}$, while as $\rho \to -\infty$ it tends to the smaller of these values. These extreme cases generalise the ideas of dominant and recessive inheritance respectively. When $\rho = 0$ the model reverts to the simple multiplicative model. For given ρ, this link function is either available or easily implemented in most generalised linear modelling programs. An important property of the Box–Cox transformation is that it is continuous in ρ through $\rho = 0$ so that, in principle, ρ can be treated as an extra model parameter and estimated by ML. In practice, it is more usual to repeat the analysis on a grid of values for ρ and to plot the resultant profile log likelihood. Note, however, that the usual asymptotic theory does not hold for LLR tests for association based upon maximisation with respect to ρ, since the likelihood is flat with respect to ρ under H_0.

A disadvantage of the above approach is that the effect parameters β_j, β_k are not invariant under case-control sampling and, following Breslow and Storer (1985), other authors have concentrated on generalised *relative* risk models, which, in the present

context, take the form

$$\log \frac{\pi^*_{(j,k)}}{1 - \pi^*_{(j,k)}} = \beta_0 + \log \theta^*_{(j,k)},$$

$$h(\theta^*_{(j,k)}) = \beta_j + \beta_k.$$

With this model, a heterozygous genotype relative risk is a generalised mean of the two homozygous genotype relative risks:

$$\theta^*_{(j,k)} = h^{-1} \left[\frac{h\left(\theta^*_{(j,j)}\right) + h\left(\theta^*_{(k,k)}\right)}{2} \right].$$

Moolgavkar and Venzon (1987) have reviewed such approaches, criticising some earlier proposals on the grounds that the functional form of the model is not maintained under change of reference category. Their preferred choice of relative risk function is equivalent to the approach of Guerrero and Johnson (1982).

A widely used alternative to the $K - 1$ df test is to carry out K separate 1 df tests, each one focusing on a specific allele and combining all other alleles. This is an efficient approach when it is reasonable to assume that association is limited to a single allele, but it is necessary to correct the P value for multiple testing. However, this is not a straightforward matter since the tests are not independent. In the context of TDT tests this problem was discussed by Morris *et al.* (1997), who pointed out that the P value for the largest of the K correlated 1 df test statistics may be calculated by simulation from the randomisation distribution. The same approach can be applied in population case-control studies.

36.9   MULTIPLE LOCI

When the polymorphism of interest is not by itself functional but may be in linkage disequilibrium with a causal locus, it is sometimes easier to demonstrate association between disease and this region of the genome by using a more polymorphic marker. For example, a polymorphic marker provides a more detailed discrimination between haplotypes present in the population, with a correspondingly better chance of identifying the haplotype(s) carrying a causal variant. But it may happen that there are no highly polymorphic markers in the region of interest. However, single nucleotide polymorphisms (SNPs), which are diallelic, occur very frequently throughout the human genome, and an alternative to the use of a single marker is to relate disease to a *haplotype* formed by several closely linked SNPs. Even when a polymorphic marker is available, recent advances in genotyping technology have been such that, usually, typing several SNPs will be preferred. In this section, for simplicity, diallelic loci are assumed.

In general, the analysis of association involving extended haplotypes brings new possibilities and difficulties. However, in one case, the analytical problem becomes identical to that explored in the previous section. This is the case where K loci to be tested are in 'complete' linkage disequilibrium, defined by all pair-wise values of Lewontin's D' measure of linkage disequilibrium equal to 1. In this case, the markers represent K mutations, with no mutation occurring more than once and no recombination occurring

in the region in the population history. In this case, only $K + 1$ different haplotypes are observed and the collection of markers behave in exactly the same way as a single polymorphic marker with $K + 1$ alleles. In this case, there is no phase uncertainty when resolving haplotypes (*i.e.* when assigning alleles along each chromosome). As an illustration, the logistic regression analysis of the model of multiplicative effects is achieved by regressing the disease status binary variable on indicator variables for each diallelic marker genotype, coded 0, 1 or 2. Similarly, the Hotelling's T^2 test with the same alternative model compares the vectors of means of these genotype scores for cases and controls.

Chapman *et al.* (2003) proposed that, when the pair-wise D' values are high but fall somewhat short of 1, this testing strategy can still be preferred to more detailed reconstruction of haplotypes. This is because there is little power to detect association with rare recombinant haplotypes, so that the expenditure in the test of the extra degrees of freedom that they require is not justified. With the availability of very large numbers of SNP markers and falling typing costs, it is likely that this situation will become commonplace. In the presence of more common recombinant haplotypes, however, full haplotype-based approach may be justified.

The main difficulty with haplotype-based analyses is that haplotypes are not directly observed owing to phase uncertainty. Family-based studies of association are not immune to this problem (Clayton, 1999), but studies based upon single individuals may be more seriously affected. If, in such a study, a subject is heterozygous at H of the loci considered, the observed genotype data are consistent with 2^H possible haplotype assignments. Methods based upon likelihood can be extended to deal with this, most conveniently by use of the EM algorithm. Each observed ambiguous genotype is expanded into all possible phases and, at the E step, its total observed frequency is divided between the possible phases according to their posterior probability, assuming HWE and current estimates of haplotype probabilities. A high-dimensional contingency table of imputed haplotypes by disease status can then be calculated, and the log-linear model fitted to this table (M step). However, some words of caution are necessary. Firstly, it is well known that the likelihood does not always have a unique maximum. When there are multiple maxima, the EM algorithm will converge to a local maximum and this may not be the global maximum. It is usually wise to repeat the model fit starting from different initial estimates. A second consequence of irregular likelihood surfaces is that LLR tests may not conform well to standard asymptotic theory. Nevertheless, exact P-values may be obtained by simulation by randomly permuting case-control status between subjects in the study (although this may be rather demanding in terms of computer time).

Score tests are generalised rather more easily to the case of unknown phase. The value of the 'score' vector, which would have been tested if the phase had been observed, is replaced by its posterior distribution of phase assignments calculated under the null hypothesis (Schaid *et al.*, 2002).

There is an extensive literature on further haplotype-based approaches to analysis that attempt to use estimated ancestral relationships between haplotypes to increase the power to detect and localise causal variants. These may be informal, and based on 'cladistic' classifications of haplotypes (see, e.g. Seltman *et al.*, 2001) or on more formal approximate models for ancestral recombination graphs. For detecting rare variants, several authors have proposed tests based upon extended haplotype sharing in cases (e.g. Tzeng *et al.*, 2003).

36.10 DISCUSSION

The discussion of previous sections has demonstrated that the analysis of population-based association studies can involve many of the statistical methods that have been developed for the analysis of multivariate categorical data. The main factors that give such analyses a distinctive flavour are

1. the fact that chromosomes are paired in subjects, and the implications of the HWE assumption (or its avoidance), and

2. the problem of unknown haplotype phase.

The latter problem, in particular, presents some challenging technical difficulties.

Most of this chapter has been concerned with inference from case-control studies under the 'rare disease' assumption. Although many of the methods described are applicable for general binary traits, those methods requiring the HWE assumption (in order, e.g. to resolve unknown haplotype phase) are more problematic. The log-linear models discussed here assume HWE conditional upon disease status, and this can only be expected to hold under the rare disease assumption or under the hypothesis of no association. In other cases, it would only be legitimate to assume HWE *marginally*. Since HWE holds under the null hypothesis, the size of tests constructed under conditional HWE models would be expected to be correct, but LLR and score tests may not have optimal properties and parameter estimation will be incorrect. For data from representative population samples, models for the joint distribution of genotype and disease phenotype must be specified in terms of the factorisation

$$\text{Prob (disease|genotype)} \times \text{Prob (genotype)},$$

where the HWE assumption is introduced in the second term. Because case-control sampling distorts the marginal distribution of genotypes, these studies require the rare disease assumption, which predicts the *conditional* assumption of HWE.

Lack of space prevents discussion of population association between quantitative traits and genotype. However, such a discussion would largely parallel that for discrete traits. Regression methods based on distribution of trait conditional upon genotype would usually be approached by assuming (or transforming to) normality of conditional distributions and using the classical (Gaussian) linear model. Proliferation of degrees of freedom may be avoided by assuming additivity of haplotype effects (i.e. zero dominance variance), and there is scope for extending such models by the incorporation of flexible 'link' functions. For methods that require us to assume HWE, the approach outlined in the previous paragraph will be satisfactory for the analysis of data drawn from representative population samples. However, it will be more difficult to incorporate such assumptions if sampling has been 'response based', e.g. if extreme trait values have been deliberately oversampled. There is scope for further work in this area.

Acknowledgments

REFERENCES

Armitage, P. (1955). Test for linear trend in proportions and frequencies. *Biometrics* **11**, 375–386.

Box, G. and Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B* **26**, 211–252.

Breslow, N. and Day, N. (1980). *Statistical methods in cancer research. volume I – the analysis of case-control studies*. IARC scientific publications. IARC, Lyon.

Breslow, N. and Storer, B. (1985). General relative risk functions for case-control studies. *American Journal of Epidemiology* **122**, 149–162.

Bross, I. (1967). Pertinency of an extraneous variable. *Journal of Chronic Diseases* **20**, 487–497.

Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. *Human Heredity* **56**, 18–31.

Clayton, D. (1999). A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. *American Journal of Human Genetics* **65**, 1170–1177.

Clayton, D. and Hills, M. (1993). *Statistical models in epidemiology*. Oxford University Press, Oxford.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997–1004.

Dunning, A., Chiano, M., Neil, R., Dearden, J., Gore, M., Oakes, S., Wilson, C., Stratton, M., Peto, J., Easton, D., Clayton, D. and Ponder, B. (1997). Common *BRCA1* variants and susceptibility to breast and ovarian cancer in the general population. *Human Molecular Genetics* **6**, 285–289.

Fan, R. and Knapp, M. (2003). Genome association studies of complex diseases by case-control designs. *American Journal of Human Genetics* **72**, 850–868.

Guerrero, V. and Johnson, R. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika* **69**, 309–14.

Lathrop, G. (1983). Estimating genotype relative risks. *Tissue Antigens* **22**, 160–166.

Mantel, N. (1963). Chi-square tests with one degree of freedom: extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association* **58**, 690–700.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**, 719–48.

Miettinen, O. (1972). Components of the crude risk ratio. *American Journal of Epidemiology* **96**, 168–172.

Moolgavkar, S. and Venzon, D. (1987). General relative risk regression models for epidemiological studies. *American Journal of Epidemiology* **126**, 949–961.

Morris, A., Curnow, R. and Whittaker, J. (1997). Randomization tests of disease-marker associations. *Annals of Human Genetics* **61**, 49–60.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A* **135**, 370–384.

Prentice, R. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

Pritchard, J. and Rosenberg, N. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* **65**, 220–228.

Schaid, D., Rowland, C., Tines, D., Jacobson, R. and Poland, G. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70**, 425–434.

Seltman, H., Roeder, K. and Devlin, B. (2001). Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *American Journal of Human Genetics* **68**, 1250–1263.

Spielman, R. and Ewens, W. (1998). A sibship test for linkage in the presence of association: the sib transmission disequilibrium test. *American Journal of Human Genetics* **62**, 450–458.

Tzeng, J.-Y., Devlin, B., Roeder, K. and Wasserman, L. (2003). On the identification of disease mutations by the analysis of haplotype matching and goodness-of-fit. *American Journal of Human Genetics* **72**, 891–902.

Xiong, M., Zhao, J. and Boerwinkle, E. (2002). Generalized t^2 test for genome association studies. *American Journal of Human Genetics* **70**, 1257–1268.

37

Whole Genome Association

A.P. Morris and L.R. Cardon

Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

Whole genome association (WGA) studies have been widely recognised as having great potential to identify genetic polymorphisms contributing to complex human diseases. With recent advances in single nucleotide polymorphism (SNP) genotyping technology, WGA studies using $> 10^5$ markers are being undertaken by many research groups worldwide with samples large enough to detect the modest genetic effects we expect for complex diseases. In this chapter, we review the key issues for the analysis of data from population-based WGA studies, building on the concepts introduced by Clayton (**Chapter 36**). We briefly discuss design issues and describe how to assess genotype quality from WGA genotyping technology. We discuss techniques for single-locus analysis and appropriate corrections that can be made to allow for multiple testing of the thousands of SNPs used in WGA studies. We describe how these methods could be extended to allow for environmental and other non-genetic risk factors, multiple SNPs, haplotypes, and epistasis. Finally, we emphasise the importance of replication of the results from WGA studies and discuss the prospects of this approach for complex disease gene mapping.

37.1 INTRODUCTION

The traditional approach to mapping disease genes has been linkage analysis, which studies the co-segregation of marker alleles with disease within large pedigrees or smaller family units such as affected sib pairs (see **Chapters 33** and **34**). This approach has proved to be successful for locating genes contributing to simple Mendelian disorders such as cystic fibrosis and Huntington's disease, where there is a clear relationship between phenotype and genotypes at the underlying functional polymorphism(s). However, linkage studies have proved less reliable for complex diseases, e.g. type 2 diabetes, where disease status may be difficult to define from multiple intermediate phenotypes, there may be multiple interacting genes underlying these phenotypes, and the effects of these genes may differ according to exposure to environmental and other non-genetic risk factors such as diet and smoking. As a result, individuals affected by complex diseases are less concentrated within families and affected family members are less likely to

share the same variants at the underlying functional polymorphisms than for Mendelian disorders.

Population-based association studies are more powerful than linkage studies for identifying genetic polymorphisms contributing to complex diseases, provided that the underlying causative variants are not very rare (Risch and Merikangas, 1996; Zondervan and Cardon, 2004). Association studies focus on identifying genetic markers that occur with different frequencies in samples of unrelated affected cases and unaffected controls, exploiting the fact that it is easier to ascertain large groups of affected individuals sharing a genetic risk factor for a complex disease across the whole population than within individual families, which would be required for linkage.

The success of association studies for disease gene mapping relies, in part, on genotyping the functional polymorphism(s) themselves (so-called *direct* association), or flanking genetic markers that are highly correlated with the functional polymorphism(s) (*indirect* association). Direct association studies focus on genotyping candidate loci with a relatively high prior probability of functional relevance, including non-synonymous polymorphisms, splice-site variants, and copy number polymorphisms. Conversely, indirect association studies incorporate panels of single nucleotide polymorphisms (SNPs), each of which is unlikely to be directly of functional relevance, but at sufficiently high density one or more is likely to be correlated with the underlying causative variants. This correlation is referred to as *linkage disequilibrium* (LD), generated as a result of the shared ancestry of a population of chromosomes at proximal loci. As a result, alleles at flanking loci tend to occur together, in cis, on the same chromosome, with each specific combination of alleles known as a *haplotype*.

37.1.1 Linkage Disequilibrium and Tagging

There are numerous measures of LD between a pair of SNPs, most based around the statistic $D = h - p_1 p_2$. In this expression, h denotes the population frequency of the *MM* haplotype, and p_1 and p_2 denote the population frequencies of the *M* allele at each SNP, where at each locus *M* and *m* denote the major and minor alleles, respectively. Under gametic phase equilibrium, $h = p_1 p_2$ so that $D = 0$. More generally, D takes values in the range $[-1, 1]$, but is highly dependent on population allele frequencies. To reduce this dependence, two commonly used measures of LD have been proposed:

$$r^2 = \frac{D}{p_1(1 - p_1)p_2(1 - p_2)},$$

and

$$D' = \begin{cases} \dfrac{D}{\max[-p_1 p_2 - (1 - p_1)(1 - p_2)]} & \text{if } D < 0 \\ \dfrac{D}{\min[p_1(1 - p_2), (1 - p_1)p_2]} & \text{if } D \geq 0 \end{cases}.$$

The statistic D' can take values in the range $[0, 1]$, where $D' = 1$ is consistent with no recombination between a pair of SNPs in the time since the mutations generating the polymorphisms occurred. This is referred to as *complete* LD, and implies that at least one of the four possible haplotypes has frequency 0. The squared correlation coefficient, r^2, between a pair of SNPs also takes values in $[0, 1]$. However, $r^2 = 1$ corresponds to *perfect* LD, where genotypes at one SNP can be used as proxies for genotypes at a

second SNP, referred to as *genetically identical* by Lawrence *et al.* (2005). The value of r^2 is directly related to the power to detect association of a disease with a genetic marker in LD with a flanking functional polymorphism. For a more detailed discussion of LD measures, see **Chapter 27**.

With the increasing availability of high-density SNP genotyping technology, many empirical studies have been undertaken to characterise the extent and distribution of LD throughout the genome in different populations. Initial studies focused on specific genes and small genomic regions (Clark *et al.*, 1998; Johnson *et al.*, 2001; Reich *et al.*, 2001; Gabriel *et al.*, 2002), but later large genomic regions (Taillon-Miller *et al.*, 2000; Dunning *et al.*, 2000; Abecasis *et al.*, 2001) and whole chromosomes were screened (Patil *et al.*, 2001; Dawson *et al.*, 2002; Phillips *et al.*, 2003). The clear conclusion from these studies is that the extent of LD is extremely variable throughout the genome, and across different populations. Further, much of common human genetic variation can be arranged in blocks of SNPs in strong LD with each other, maintained by low levels of recombination, bounded by hotspots of meiotic crossover activity.

Knowledge of the patterns of LD throughout the genome aids study design, since markers can be selected so as to guarantee coverage of all common SNPs with some predetermined threshold of r^2. The advantage of this approach is that we need not genotype all SNPs in a candidate gene or region, or even the whole genome, but can focus on a smaller number of so-called tag SNPs from which we can recover much of the information about common human genetic variation. Tag SNPs can be selected within blocks of strong LD by selecting combinations of SNPs that jointly define all common haplotype variation. Alternatively, SNPs can be allocated to *bins* of strong LD, based on pair-wise r^2 measurements, and a single tag selected from each bin (Carlson *et al.*, 2004).

One of the key advances in the design of population-based association studies has been the publication of data from the International HapMap project (The International HapMap Consortium, 2005). The initial phase of the project genotyped more than 1 million evenly spaced SNPs genome-wide, in samples from four populations: (1) 30 Yoruba trios (two parents plus offspring) from Ibadan, Nigeria; (2) 30 trios from the CEPH collection from Utah, USA, all with north and west European ancestry; (3) 45 unrelated individuals from Beijing, China; and (4) 44 unrelated individuals from Tokyo, Japan. The second phase of the project genotyped a further 4.6 million SNPs in the same samples, so that the average inter-SNP spacing was less than 1 Kb. Genotype data from the project are publicly available, and can be downloaded for detailed analysis of LD to aid marker selection in association study design, and interpretation of the results of analysis.

37.1.2 Current WGA Studies

A key determinant of the success of population-based association studies to map complex disease genes is sample size. We expect the alleles underlying complex disease phenotypes to each have small individual marginal effects, requiring samples of the order of thousands of cases and controls. Improvements in high-throughput SNP genotyping technology have revolutionised the field, making such sample sizes feasible for large candidate regions, hundreds or thousands of candidate genes, or most recently, whole genome association (WGA) studies. The latest generation of mapping arrays consist of 10^5–10^6 genome-wide SNPs. The Affymetrix 500K GeneChip is based on randomly selected SNPs, and tags approximately 65 % of HapMap phase II polymorphisms with $r^2 > 0.8$ in the CEPH samples, although coverage is lower for the Yoruba samples, reflecting less extensive

LD in African populations (Barrett and Cardon, 2006). The Illumina HumanHap300K BeadArray includes more than 300K tag SNPs, selected to capture common variation among north and west European populations, covering about 75 % of HapMap phase II SNPs in the CEPH samples, but only 28 % among the Yoruba samples.

Despite the potential of association studies to identify polymorphisms contributing susceptibility to complex diseases, the success of initial screens of candidate genes or larger candidate regions was limited to a handful of major gene effects including *APOE* for Alzheimer's disease (Rubinsztein *et al.*, 1999) and *NOD2* for Crohn's disease (Hugot *et al.*, 2001). With more recent appreciation of the importance of sample size, study design and genotype calling the list of reported associations is rapidly expanding. Several candidate genes have been associated with type 2 diabetes in multiple studies, including *PPARG*, *KCNJ11*, and *TCF7L2* (Parikh and Groop, 2004; Grant *et al.*, 2006). However, several reported associations have been difficult to replicate, including the $G972R$ polymorphism in the *IRS1* gene (Almind *et al.*, 1993), which has more recently been demonstrated to have no effect on the risk of type 2 diabetes in population-based studies (Florez *et al.*, 2004; van Dam *et al.*, 2004; Zeggini *et al.*, 2004). This demonstrates the need to develop statistical methods for WGA studies, with increased power to detect disease association, while minimising the occurrence of false positives.

One success story for WGA studies was the identification of a causal variant for age-related macular degeneration (AMD) among Europeans in the complement factor H (*CFH*) gene (Klein *et al.*, 2005). An initial genome scan of more than 100 K SNPs on the Affymetrix 111K GeneChip in just 96 cases and 50 controls revealed a strong association of a common intronic polymorphism in *CFH* with AMD (nominal p value $< 10^{-7}$). Investigation of the patterns of LD flanking this signal using HapMap phase I data revealed this SNP to be part of a block contained wholly within *CFH*. Resequencing of the block identified a polymorphism in strong LD with the associated SNP that represents a tyrosine–histidine change at amino acid 402. This effect size is much larger than we expect for complex diseases (relative risk of 7.4 for individuals homozygous for the mutant allele compared to those homozygous for the wild-type allele), and mapping for complex diseases, in general, is unlikely to be so straightforward. Nevertheless, with such a small sample size, and the poor coverage of the Affymetrix 111K GeneChip (31 % in HapMap phase II CEPH samples), this result is encouraging for WGA studies.

A recent WGA study for type 2 diabetes has identified four novel susceptibility loci in a Fench case–control cohort of more than 1300 individuals (Sladek *et al.*, 2007). The samples were initially genotyped using the Illumina Infinium Human1 BeadArray, which assays 110K gene-centric SNPs, in addition to the Illumina HumanHap300K BeadArray. Markers demonstrating significance of the order of $p < 10^{-4}$ for the gene-centric array and $p < 5 \times 10^{-5}$ for the 300K array were carried forward for genotyping in a second cohort of more than 2500 cases and almost 2900 controls. These SNPs included known associations in the *TCF7L2* gene (Grant *et al.*, 2006), but did not include polymorphisms in other previously identified genes such as *PPARG*. This demonstrates the difficulties in replicating results from WGA studies, where effect sizes may be small, studies may be underpowered, and there may be variable coverage of the genome by the genotyping technologies used.

The Wellcome Trust Case Control Consortium (WTCCC) represents one of the largest WGA studies to date. The main arm of the study consists of an indirect WGA screen of 2000 unrelated cases each of seven diseases (type 1 diabetes, type 2 diabetes, coronary heart disease, hypertension, bipolar disorder, rheumatoid arthritis, and inflammatory bowel

disease) from across the United Kingdom, together with 1500 unrelated controls each from the 1958 British birth cohort and recruits from the UK national blood service. All samples have been genotyped using the Affymetrix 500K GeneChip Mapping Array. An additional direct association study of more than 15K non-synonymous coding SNPs for 1000 cases each of four additional diseases (breast cancer, autoimmune thyroid disease, multiple sclerosis, and ankylosing spondylitus) 1500 controls from the 1958 British birth cohort has also been undertaken as part of the WTCCC using a custom Infinium chip from Illumina. Analysis of the data generated by the WTCCC will point to the design of further studies for each disease, focusing on the most interesting positive signals for subsequent investigation.

In this chapter, we describe statistical methods for the analysis and interpretation of results from WGA studies. We begin by describing procedures to assess the quality of genotypes obtained from WGA genotyping technology. Then, we discuss techniques for single-locus analyses to detect association with disease, including appropriate multiple-testing corrections. Next, we describe extensions to allow for environmental effects and to include tightly linked multiple SNPs or haplotypes, and epistasic interactions across more widely spaced SNPs. Finally, we emphasise the importance of replication of the results from WGA studies, and discuss their prospects for complex disease gene mapping. Balding (2006) provides an additional review of the design of, and methods for the analysis of, WGA studies.

37.2   GENOTYPE QUALITY CONTROL

Data filtering to identify genotype errors is a critical aspect of WGA analyses, which can determine whether real discoveries are made or false positives plague interpretation. No experimental system involving biological material is without error, and with large numbers of both SNPs and study subjects in a single study, even the modest error rates of $<0.25\%$ expected with current technology can be important. If such errors were distributed randomly across both genotypes and phenotypes, their effect would be limited to a small loss of statistical power. However, because of the nature of the experimental technologies available and factors such as DNA quality and preparation, specific experimental conditions, skill of experimenter, incorrect automated assignment (or 'calling') of experimental intensity values into discrete genotype classes, and stochastic variation, errors are not always distributed randomly. Non-random distribution of errors can inflate type I error rates and/or reduce power. The difficulty often lies with the designation of a heterozygous genotype, which is incorrectly classified as homozygous or the genotype is labelled as missing because of assignment ambiguity.

Animal models of continuous characters, human data in complex traits and long-standing theory in biometrical genetics (Wright and Hastie, 2001; Valdar *et al.*, 2006) all suggest that genetic influences on most multifactorial phenotypes follow an L-shaped distribution of effects, with a small number of alleles with large effects and a larger number of small effect sizes. For discrete human diseases, an odds ratio (OR) above about 1.25 might now be considered large and many more effects are expected to be smaller than this. These effect sizes mean that most WGA studies aim to identify very small differences in allele frequencies, often only a few percent, among phenotype groups. Accordingly, even small amounts of experimental error can have profound effects on the outcomes (Clayton *et al.*, 2005; Barrett and Cardon, 2006), particularly in the presence

of rare alleles. It may seem that an experimental ideal would be to identify individual genotype errors and correct them one at a time, but this is often difficult or impossible in the WGA setting. Therefore, many filtering procedures aim to identify specific SNPs yielding errors in multiple individuals (a problem with the marker) or individuals in the sample with errors across multiple SNPs (a problem with the DNA sample), and simply exclude them from the study.

Neutral genetic variants in a large random-mating population are expected to display Hardy–Weinberg equilibrium (HWE), under which expected genotype frequencies satisfy $\mathbf{E}(f_{MM}) = p^2, \mathbf{E}(f_{Mm}) = 2p(1 - p)$, and $\mathbf{E}(f_{mm}) = (1 - p)^2$, where p is the population frequency of allele M. Genotyping error can alter the observed frequencies from the expected proportions, and thus tests of deviations from HWE comprise a traditional approach to detecting genotyping errors and excluding markers with significant deviations (Weir, 1996). Such a test can be constructed using a Pearson goodness-of-fit statistic,

$$\mathrm{HWE}X^2 = \sum_{g=mm,Mm,MM} \frac{(f_g - \mathbf{E}(f_g))^2}{\mathbf{E}(f_g)},$$

having an approximate χ^2 distribution under the null hypothesis of HWE, with 1 degree of freedom (df) because p is derived from the observed data.

There are three serious problems with the use of HWE tests in this manner: (1) natural selection and copy-number variants can also lead to significant deviations, and these may underlay true causal associations, which will be missed if the SNP is excluded because it 'fails' a HWE test; (2) the test is insensitive to the modest deviations that are most often observed in WGA studies; and (3) in the WGA setting of $> 10^5$ markers, an appropriate threshold of 'significance' is difficult to determine. As a result of these considerations, the most prudent use of HWE tests for genotyping error may be only to exclude the most egregious deviations by setting an extreme significance threshold such as 10^{-7} or less, and using exact tests for rare alleles (Weir *et al.*, 2005).

Missing data at individual genotypes is not uncommon, but when missing rates exceed about 5 % at a SNP, or about 10 % for an individual, there is reason to be suspicious of the assay or DNA sample and the most prudent course is to repeat the experiment for that SNP or sample (WTCCC, 2007). For case–control studies, in particular, markers with differences in missing data rates between cases and controls often yield false positives, which are sometimes striking in magnitude (Clayton *et al.*, 2005). Several statistics are available for testing the difference in missing data rates between cases and controls, e.g. one based on the normal approximation to the binomial distribution:

$$z = \frac{m_c - m_t}{\sqrt{m(1 - m)(1/n_c + 1/n_t)}},$$

where m_c and m_t are the proportions of missing genotypes among cases (c) and controls (t), samples sizes (missing + non-missing data) are labelled as n_c and n_t and m is the overall missing genotype rate at the marker.

Graphical displays are useful for identifying suspicious markers or samples. According to the L-shaped distribution of effect sizes, a study of 10^5–10^6 genetic markers genotyped on 1–2K cases and equal numbers of controls should reveal few genuine loci with single-locus p values below 10^{-6}, with many more associated loci lying in the range 10^{-3}–10^{-5}
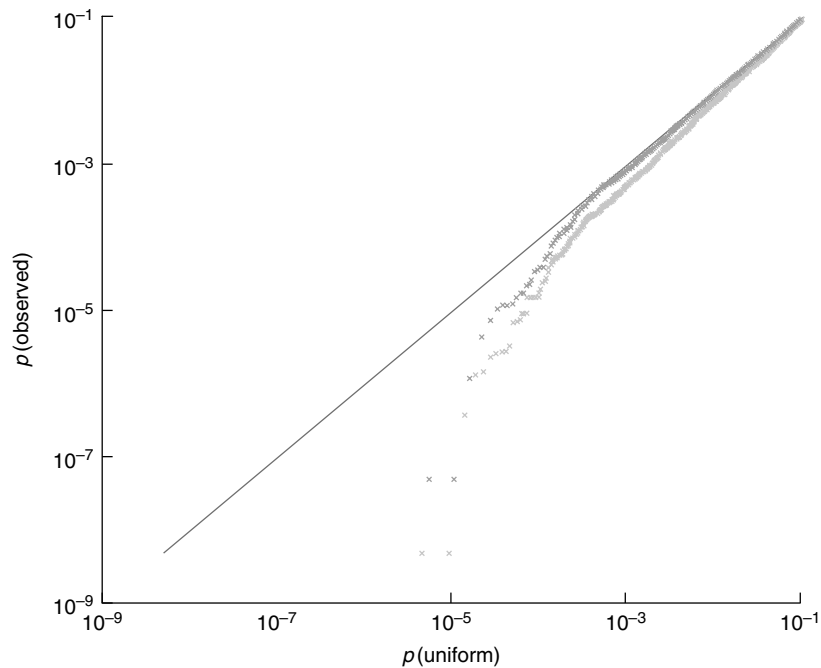
Figure 37.1 Probability–probability plot for association statistics from a WGA study of type 2 diabetes (T2D) genotyped for the Illumina Human1 BeadArray SNPs, from Sladek *et al.* (2007, supplementary Figure 4). Unadjusted p values for the maximum statistic over three tests of association are plotted against the expected uniform distribution of p values under the null hypothesis of no association, genome-wide. Systematic deviations from the $y = x$ line are indicative of the effects of population stratification, and some extreme deviations could reflect genotyping error. Corresponding values after genomic-control adjustment are also shown, and adhere to the $y = x$ line over most of the distribution; some of the very small p values that deviate from this line were subsequently confirmed to be associated with T2D. [Reprinted by permission from Macmillan Publishers Ltd: Sladek R *et al.*, A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 2007; 445: 881–885.]

(Zondervan and Cardon, 2004). Accordingly, if one observes many highly significant loci in a particular study, they may be less likely to reflect real discoveries and more likely to reflect systematic genotype error in some of those markers. This logic has led to the widespread use of quantile–quantile (QQ) plots to examine the overall distribution of test statistics and assess whether there are too many data points in the tail. This approach, used to assess HWE deviation by Weir *et al.* (2005), was shown to be extremely effective in the case–control context by Clayton *et al.* (2005). To construct such plots, the ordered test statistics for association are plotted against the corresponding expected order statistics. For example, Clayton *et al.* (2005) plotted ascending values of the Cochran–Armitage test statistic (see Section 37.3) against $F^{-1}[i/(N + 1)]$, where $F[]$ is the χ_1^2 distribution function. Sladek *et al.* (2007) use a similar approach of plotting the unadjusted p values for the maximum statistic over three specific tests of association for each SNP against the expected uniform distribution (Figure 37.1). In either case, large values deviating from the $y = x$ null can be a symptom of genotyping error, as demonstrated by Clayton

et al. (2005). This subjective screening procedure remains extremely useful to determine whether the data filters (missing data rates, HWE, allele frequency thresholds) in a particular study are sufficient to eliminate most of the problematic markers.

Another useful graphical procedure involves plotting, for each individual, the fraction of all markers that are heterozygous against the proportion of missing data for that individual (WTCCC, 2007). This plotting procedure can identify DNA samples that perform poorly in WGA genotyping, resulting in a high proportion of missing data and/or too few heterozygotes.

Genotyping so many markers in WGA studies enables identification of closely related individuals or duplicate DNA samples. Abecasis *et al.* (2001) showed that comparing the average number of alleles that are identical by state for two individuals with the variance of IBS sharing can identify duplicate samples or MZ twins, full-siblings, parent–offspring and half-siblings. In other words, for two individuals j and k having genotypes g_j and g_k at marker i of N, plot

$$\bar{x}_{\text{ibs}}^{(j,k)} = \frac{1}{2N} \sum_{i=1}^{N} x_i^{(j,k)},$$

against

$$s_{x_{\text{ibs}}^{(j,k)}}^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_i^{(j,k)} - \bar{x}_{\text{ibs}}^{(j,k)} \right)^2,$$

where

$$x_i^{(j,k)} = \begin{cases} 0 & \text{if} & g_j, g_k = mm, MM \\ 1 & \text{if} & g_j, g_k = mm, Mm \text{ or } Mm, MM \\ 2 & \text{if} & g_j, g_k = mm, mm \text{ or } Mm, Mm \text{ or } MM, MM. \end{cases}$$

More distant relatives are difficult to distinguish from unrelated individuals using this approach. See also Devlin and Roeder (1999) and Voight and Pritchard (2005).

37.3 SINGLE-LOCUS ANALYSIS

The current standard practice involves an individual test of each SNP typed in the WGA study, to identify any promising associations. We now review the main tests employed.

Consider a sample of unrelated cases, affected by the disease of interest, and unaffected controls, typed at a SNP with alleles denoted by M and m. We can represent the sample genotype data in a 2×3 contingency array (Table 37.1). Under the null hypothesis of no association with the disease, we expect the relative genotype frequencies to be the same in cases and controls. Thus, as described by Clayton (**Chapter 36**), we can construct a score test for association by calculating the standard Pearson's χ^2 statistic for independence of the rows and columns, given by

$$X_{\text{Gen}}^2 = \sum_{i=0,1,2} \sum_{j=A,U} \frac{(n_{ij} - \mathbf{E}[n_{ij}])^2}{\mathbf{E}[n_{ij}]}, \tag{37.1}$$

Table 37.1 Representation of SNP genotype data for a population-based association study in a 2×3 contingency array. The counts, n_{ij}, denote the observed sample frequency of individuals carrying i copies of allele m, and phenotype j, where $j = A$ corresponds to cases and $j = U$ corresponds to controls.

Genotype	Cases	Controls	Total
MM	n_{0A}	n_{0U}	$n_{0.}$
Mm	n_{1A}	n_{1U}	$n_{1.}$
Mm	n_{2A}	n_{2U}	$n_{2.}$
Total	$n_{.A}$	$n_{.U}$	$n_{..}$

where

$$\mathbf{E}[n_{ij}] = \frac{n_{i.}n_{.j}}{n_{..}}.$$

The test statistic, $X^2{}_{\text{Gen}}$, has an approximate χ^2 distribution with 2 df under the null hypothesis of independence (no association).

It may be of interest to investigate the association further by estimating the OR of disease for each genotype at the SNP. It is customary to calculate the sample OR relative to the most common genotype in controls as a baseline. For example, the OR for genotype *mm* relative to the baseline genotype *MM* is estimated by

$$\psi_{mm|MM} = \frac{n_{2A}n_{0U}}{n_{2U}n_{0A}}.$$

The variance of the log OR is approximately

$$\mathbf{V}[\ln \psi_{mm|MM}] = \frac{1}{n_{2A}} + \frac{1}{n_{2U}} + \frac{1}{n_{0A}} + \frac{1}{n_{0U}},$$

with corresponding 95 % confidence interval

$$\ln \psi_{mm|MM} \pm 1.96 \times \sqrt{\mathbf{V}[\ln \psi_{mm|MM}]}.$$

Case–control studies are retrospective in the sense that subjects are ascertained on the basis of their disease status, and are then genotyped. Since cases are over-sampled, the disease risk cannot be directly estimated. However, assuming the disease under investigation to be rare, the OR gives an approximation to the relative risk: an individual carrying genotype *mm* is approximately $\psi_{mm|MM}$ times more likely to develop the disease than an individual of genotype *MM*.

To reduce the df of the association test, we can focus on *allelic* effects by assuming that alleles at the SNP act independently in terms of disease risk. In other words, we assume a multiplicative model for disease penetrances so that $\psi_{mm|MM} = \psi_{Mm|MM}^2$, and hence a linear (additive) trend in the log-odds of disease with each copy of allele m. Under this assumption, we can test for association between the SNP and disease by means of the Cochran–Armitage trend test, given by

$$X^2_{\text{C-A}} = \frac{n_{..}[n_{..}(n_{1A} + 2n_{2A}) - n_{.A}(n_{1.} + 2n_{2.})]^2}{n_{.A}(n_{..} - n_{.A})[n_{..}(n_{1.} + 4n_{2.}) - (n_{1.} + 2n_{2.})^2]}. \tag{37.2}$$

Under the null hypothesis of no association between the SNP and disease, X^2_{C-A}, has an approximate χ^2_1 null distribution. We can also calculate the OR for allele m, relative to allele M, by

$$\psi_{m|M} = \frac{n_{1A}n_{0U}/(n_{0.} + n_{1.}) + n_{2A}n_{1U}/(n_{1.} + n_{2.}) + 4n_{2A}n_{0U}/(n_{0.} + n_{2.})}{n_{0A}n_{1U}/(n_{0.} + n_{1.}) + n_{1A}n_{2U}/(n_{1.} + n_{2.}) + 4\sqrt{(n_{2A}n_{2U}n_{0A}n_{0U})}/(n_{0.} + n_{2.})}.$$

The OR can be interpreted as follows: an affected individual is $\psi_{m|M}$ times more likely to carry genotype Mm than genotype MM at the SNP, and $\psi^2_{m|M}$ times more likely to carry genotype mm.

As discussed by Clayton (**Chapter 36**), the tests described above rely on asymptotic assumptions that are unlikely to hold when there are small genotype counts or the minor allele frequency is low. By conditioning on the marginal frequencies of the contingency array, exact tests can be constructed for the genotype-based or Cochran–Armitage statistics.

37.3.1 Logistic Regression Modelling Framework

For complex traits, we expect that disease risk might be modified by environmental effects that cannot be easily accommodated by the tests described above. A more flexible framework for modelling the relationship between disease phenotype and SNP genotype makes use of logistic regression techniques, as described by Clayton (**Chapter 36**). Consider, as before, a sample of unrelated cases and controls, yielding genotype data **G**. The logistic regression model is parameterised in terms of the log-odds of disease for each SNP genotype, denoted by $\boldsymbol{\beta}$. The log-likelihood of observed phenotype data, **y**, is given by

$$\ln f(\mathbf{y}|\mathbf{G}, \boldsymbol{\beta}) = \sum_i [y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i)],$$

where y_i denotes the disease phenotype of the ith individual. The probability, π_i, that the ith individual is affected, given their genotype G_i at the SNP, is given by the *logit* link function,

$$\pi_i = \frac{\exp \eta_i}{1 + \exp \eta_i},$$

where the linear predictor is $\eta_i = \beta_{G_i}$.

Under the null hypothesis of no association of the SNP with the disease, we expect each genotype to have equal odds of disease, so that the linear component $\eta_i = \beta_0$. Evidence of association corresponds to deviation from this null model. Assuming that the alleles at the SNP act independently in terms of disease risk, in the same way as for the Cochran–Armitage trend test, there is an additive effect of the two alleles in the log-odds of disease. Under this additive model, treating allele M as baseline, the linear predictor for the ith individual is given by

$$\eta_i = \beta_0 + \beta_A z_{(A)i}, \tag{37.3}$$

where β_A denotes the additive effect of allele m, and $z_{(A)i}$ is an indicator variable representing the additive component of the ith genotype, summarised in Table 37.2.

Table 37.2 Coding of additive and dominance components of SNP genotypes.

Genotype	Additive component $z_{(A)i}$	Dominance component $z_{(D)i}$
MM	-1	0
Mm	0	1
Mm	1	0

Deviation from the additive model is referred to as *dominance*, and can be thought of as an interaction between the pair of alleles within each genotype at the SNP. Under this non-additive model, the linear predictor for the ith individual extends to

$$\eta_i = \beta_0 + \beta_A z_{(A)i} + \beta_D z_{(D)i}, \tag{37.4}$$

where β_D denotes the dominance effect of allele m over allele M, and $z_{(D)i}$ is an indicator variable representing the dominance component of the ith genotype, summarised in Table 37.2.

We can test for association between disease and the SNP by comparing nested submodels of (37.4) with $\beta_A = 0$ and/or $\beta_D = 0$ using analysis of deviance. For example, the log-likelihood ratio statistic,

$$\Lambda_{\text{Gen}} = 2\ln f(\mathbf{y}|\mathbf{G}, \hat{\beta}_0, \hat{\beta}_A, \hat{\beta}_D) - 2\ln f(\mathbf{y}|\mathbf{G}, \hat{\beta}_0, \beta_A = 0, \beta_D = 0),$$

provides a genotype-based test of association, having an approximate χ_2^2 null distribution, and the maximum likelihood estimates, $\hat{\boldsymbol{\beta}}$, under each model are obtained numerically. The model in (37.4) is parameterised in a different way compared to the model used in the 2 df test described by Clayton (**Chapter 36**). However, although model parameter estimates are interpreted in different ways, the statistical tests are equivalent, and both are approximately equal to the Pearson 2 df test defined in (37.1).

Assuming that alleles at the SNP have independent effects on disease risk, we can form an additive test of association,

$$\Lambda_{\text{Add}} = 2\ln f(\mathbf{y}|\mathbf{G}, \hat{\beta}_0, \hat{\beta}_A, \beta_D = 0) - 2\ln f(\mathbf{y}|\mathbf{G}, \hat{\beta}_0, \beta_A = 0, \beta_D = 0),$$

which has an approximate χ_1^2 distribution and is equivalent to the Cochran–Armitage test (37.2). Within this framework, we can also test for deviations from the additive model of alleles at the SNP on the log odds of disease, given by the statistic

$$\Lambda_{\text{Dom}} = 2\ln f(\mathbf{y}|\mathbf{G}, \hat{\beta}_0, \hat{\beta}_A, \hat{\beta}_D) - 2\ln f(\mathbf{y}|\mathbf{G}, \hat{\beta}_0, \hat{\beta}_A, \beta_D = 0),$$

having an approximate χ_1^2 distribution under the null hypothesis of no dominance. We can test other disease models within this framework by imposing constraints in the model parameters. For example, to allow for a pure recessive effect of allele m, we constrain $\beta_D = -\beta_A$, leading to a test of association having an approximate χ_1^2 null.

Within the logistic regression framework, it is straightforward to incorporate covariates, $\mathbf{x}$, in the linear component, to allow for environmental effects. Specifically, the linear

predictor of the ith individual extends to

$$\eta_i = \beta_0 + \sum_j \gamma_j x_{ij} + \beta_A z_{(A)i} + \beta_D z_{(D)i}, \qquad (37.5)$$

where x_{ij} is the response of the ith individual to the jth covariate, and γ_j is the corresponding regression coefficient. Under the null hypothesis of no association of the SNP with disease, each individual is equally likely to be affected in terms of their genotype, with risk modified only by the effects of the covariates. Thus, we can consider the same likelihood ratio tests of association by comparing sub-models of (37.5) subject to the constraints $\beta_A = 0$ and/or $\beta_D = 0$ by analysis of deviance, in the same way as above.

37.3.2   Interpretation of Results and Correction for Multiple Testing

A significant result in a test of association may suggest that the SNP itself is causative, directly influencing disease risk. However, such an assertion would need to be established via further functional studies. In fact, there are a number of possible alternatives.

1. Alleles at the SNP are correlated with alleles at the functional polymorphism, but do not directly influence disease risk. Such an indirect association occurs as a result of background LD between the two loci.

2. Alleles at the SNP and the functional polymorphism are confounded with underlying population structure that is not accounted for in the analysis. In the presence of structure, cases may be ascertained preferentially from one stratum of the population, because of higher disease prevalence and/or biased ascertainment. If SNP genotype frequencies vary between strata, we may detect an apparent association with disease in the population overall, even if there is no association within each of the individual population strata. This problem is addressed in more detail below.

3. For tests with significance level α, we expect $100\alpha\%$ of non-associated SNPs tested to show false positive signals of association.

The choice of significance level α should take account of multiple testing in WGA studies. For example, if we choose $\alpha = 5\%$ and test 20 independent SNPs for association with disease, we expect one of them to show significant evidence of association, even if none is truly associated with disease. It is crucial, therefore, to correct for multiple testing to maintain the type I error rate for the experiment overall (i.e. for all the SNPs tested in the association study).

The simplest approach to allow for multiple testing is to make use of a Bonferroni correction. Under this approach, each test is treated as independent, and the SNP-wise significance level is adjusted to achieve an overall experiment-wise type I error rate of $100\alpha\%$. Specifically, when testing N SNPs, we use a significance level of $100\alpha/N\%$ at each SNP. The disadvantage of this approach is that each test is assumed to be independent, whereas for WGA studies we expect adjacent SNPs to be correlated owing to background patterns of LD throughout the genome, and thus the Bonferroni correction will be conservative. An alternative correction that overcomes the problem of correlated tests makes use of the false discovery rate (FDR), by fixing the expected number of false positives among significant associations (Benjamini and Hochberg, 1995). Specifically, if

we select an uncorrected SNP-wise significance level of α, the FDR is given by $N\alpha/k$, where k is the number of SNPs with a p value of less than α. Therefore, we can fix the SNP-wise significance level so as to obtain an overall FDR of 5 %.

The most appropriate methods for correcting for multiple testing make use of permutation procedures. The null distribution of experiment-wise association statistics is generated by calculating the maximum association statistic over the genome for a large number of permutations of the original phenotype and genotype data. In the simplest procedures, phenotype labels are permuted while keeping genotypes fixed, thus maintaining the LD structure throughout the genome. Corrected p values for each SNP can then be calculated by comparing the observed test statistic with the distribution of the maximum test statistic from each permutation. For WGA studies, this process will be extremely computationally intensive. However, for an empirical experiment-wise significance level of the order of 5 %, as few as 100 permutations of the data may be adequate.

A final approach to take account of multiple testing makes use of Bayesian statistical theory. Under this approach, we assign a prior probability of association to each SNP, reflecting our beliefs about the number of disease associated loci before we look at the observed genotype data. The advantage of the Bayesian framework is that we can allow our prior probability of association to vary across SNPs, reflecting their functional relevance or the results of previous linkage and/or association studies.

37.4   POPULATION STRUCTURE

One of the potential problems of population-based association studies is population structure, which, if not accounted for in the analysis, can inflate the false positive error rate for detecting associations. Consider a population consisting of two underlying strata, where the disease is common in stratum one, but rare in stratum two. If cases and controls are selected at random from the population overall, without regard to the underlying structure, cases will be preferentially selected from stratum one. As a result, SNPs with allele frequency differences between the strata will appear to be associated with disease, even if there is no association within each stratum.

One obvious approach to deal with the problem of structure is to match cases and controls for stratum of the population, e.g. on the basis of self-described ethnicity or geographical location. However, with migration and admixture between ethnic groups, these indicators may not reflect the complex nature of fine-scale structure within populations. Family-based association studies (see **Chapter 38**) provide a design-based solution to the problem, in which, for example, the two alleles of each parent of an affected child are matched, but this solution is expensive in terms of additional genotyping and limited to studies for which trios or other suitable family structures are available. For WGA studies, the problem of population structure can be overcome by statistical methods that account for the underlying structure in analyses. The most important of such methods are briefly reviewed here (see **Chapters 35** and **36** for more detail and further discussion).

One of the simplest methods to identify, and adjust for, structure in population-based WGA studies is genomic control (Devlin and Roeder, 1999). Under the null hypothesis of no disease association, the distribution of Cochran–Armitage test statistics is χ_1^2. However, in a stratified population, we expect that there would be allele frequency differences at many SNPs, genome-wide, and hence an excess of false positive signals of

association. As a result, the observed distribution of association statistics will be inflated, with the magnitude of the inflation reflecting the extent of structure. The genomic control method takes account of structure by a linear rescaling of observed test statistics to approximately restore the χ_1^2 null distribution. Figure 37.1 shows a plot of p values for the WGA of Sladek *et al.* (2007) before and after genomic-control adjustment. This approach is appealing in its simplicity, but is limited to a simple test of association which, for example, do not incorporate environmental effects.

A more complex solution to the problem of population structure has been proposed by Pritchard *et al.* (2000a) and has come to be known as *structured association*. Under an admixture model, the proportion of an individual's genome that descends from each of K specific ancestral strata is treated as unknown. The posterior distribution of ancestry for each sampled individual is then approximated using Bayesian Markov chain Monte Carlo (MCMC) methods, using the STRUCTURE algorithm, based on genotype information from several hundred genome-wide SNPs (see **Chapter 30** for further details). Tests for association then compare allele frequencies between cases and controls within strata, implemented in the companion STRAT software (Pritchard *et al.*, 2000b). Alternatively, population ancestry could be included as covariates in a logistic regression framework to allow for more flexible modelling of genetic and non-genetic risk factors for complex disease. Potential disadvantages of the structured association approach include the following: (1) the number of ancestral subpopulations is unknown, and must be inferred using an *ad hoc* estimation procedure and (2) the MCMC algorithm is computationally intensive, and cannot in practice accommodate the numbers of markers used in WGA studies.

Setakis *et al.* (2006) have proposed accounting for population structure within the logistic regression model by treating the genotypes at genome-wide SNPs directly as covariates. Backward elimination or Bayesian shrinkage model selection techniques are employed to reduce the over-fitting problem arising from including many covariates in the model. Reich *et al.* (2006) have suggested using principal components analysis (PCA) to infer population structure and provide formal significance tests for between-strata differences. The eigenvectors and the corresponding loadings for each individual from the PCA can be used as covariates within a logistic regression framework to account for the underlying structure. A key advantage of this approach is computational efficiency, since PCA can be applied to data sets with $> 10^5$ SNPs.

37.5 MULTI-LOCUS ANALYSIS

One of the most attractive features of SNPs for complex disease gene mapping is their abundance throughout the genome. However, each individual SNP provides little information about disease association unless it is highly correlated with the underlying (unobserved) causal polymorphism. Single-SNP analyses may thus be inefficient for mapping. For high-density panels of SNPs, such as those utilised in WGA studies, we expect that there would be correlations between several SNPs flanking a causal polymorphism, and so simultaneous analyses of multiple SNPs may jointly provide evidence of association for modest gene effects, even when the individual SNPs do not.

The logistic regression model described above provides a natural framework for multi-locus association analysis. Additive and dominance effects of all SNPs in the same

gene or small genomic region can be fitted simultaneously in the logistic regression model. To allow for correlations between SNPs, and to reduce the problem of over-parameterisation, standard statistical model building techniques, such as forward selection and/or backward elimination, can be utilised to identify good combinations of SNPs to describe the association with disease. Such model building techniques tend to over-fit the observed phenotype and genotype data, and so appropriate correction for the selection process should be taken into account to avoid inflation of type 1 error, e.g. by permutation testing.

37.5.1 Haplotype-based Analyses

An alternative approach to multi-locus analysis is to focus on haplotype effects. Haplotypes are particularly attractive because much of diversity within blocks of LD is driven by mutation, rather than recombination. As a result, much of common genetic variation can be structured into haplotypes within blocks that are rarely disturbed by meiosis. Furthermore, Clark (2004) emphasises that the functional properties of a protein are determined by the linear sequence of amino acids, corresponding to DNA variation on a haplotype. For example, there is evidence that a combination of causal variants in cis in the *HPC2/ELAC2* gene increases the risk of prostate cancer (Tavtigian *et al.*, 2001). Finally, a rare causal allele may reside on a specific haplotype background that would not otherwise be identified through single-locus methods.

It is common to assume that each of the pair of haplotypes, H_{i1} and H_{i2}, forming the diplotype, H_i, of the ith individual, contributes independent effects to disease risk, where haplotypes are labelled according to their relative frequency in the population. Under this assumption, we can parameterise the logistic regression model in terms of the log odds of disease for each haplotype. Thus, the linear predictor of the ith individual is given by

$$\eta_i = \beta_0 + \sum_j \gamma_j x_{ij} + \beta_{H_{i1}} + \beta_{H_{i2}}.$$

In this expression, β_k denotes the log-OR of the kth most frequent haplotype, relative to the baseline haplotype, usually taken to be the most common, so that $\beta_1 = 0$. Furthermore, β_0 is the baseline log odds of disease and γ_j denotes the effect of the jth covariate. An affected individual is approximately $\exp[\beta_{H_{i1}} + \beta_{H_{i2}}]$ times more likely to carry diplotype H_i than to carry two copies of the baseline haplotype.

One of the obvious problems of haplotype-based analyses is that we do not observe the diplotypes, **H**, directly from the unphased genotype data. One solution to the problem would be to take a point estimate of the diplotype for each individual, using statistical methodology, such as PHASE (Stephens *et al.*, 2001; Stephens and Donnelly, 2003) or by maximum likelihood via implementation of the expectation–maximisation (E–M) algorithm (Excoffier and Slatkin, 1995), and to treat this estimate as if it were known. However, this does not take account of uncertainty in the haplotype reconstruction process, and as a result, the variances of model parameters are under-estimated, and the false positive error rate is inflated.

A better approach to allow for unknown phase is to consider the distribution of diplotypes consistent with each multi-locus genotype, denoted $f(H_i|G_i, \mathbf{h})$ for the ith individual, given unknown population haplotype frequencies **h**. The likelihood of observed

phenotype data, $\mathbf{y}$, given the unphased genotype data, can then be expressed as

$$f(\mathbf{y}|\mathbf{G}, \mathbf{x}, \mu, \boldsymbol{\beta}, \boldsymbol{\gamma}, h) = \prod_i \sum_{H_i \in G_i} f(y_i|H_i x_i, \mu, \beta, \gamma) f(H_i|G_i h), \qquad (37.6)$$

where $H_i \in G_i$ denotes the set of diplotypes consistent with the observed multi-locus genotype of the ith individual. Under the null hypothesis of no association between SNP haplotypes and disease, the log odds of disease for each haplotype will be the same, and $\boldsymbol{\beta} = 0$. Thus, we can construct a likelihood ratio test of association by considering the deviance,

$$\Lambda_{\text{Hap}} = 2 \ln f(\mathbf{y}|\mathbf{G}, \mathbf{x}, \hat{\mu}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{h}}) - 2 \ln f(\mathbf{y}|\mathbf{G}, \mathbf{x}, \hat{\mu}, \boldsymbol{\beta} = 0, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{h}}),$$

where $f(\mathbf{y}|\mathbf{G}, \mathbf{x}, \hat{\mu}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{h}})$ and $f(\mathbf{y}|\mathbf{G}, \mathbf{x}, \hat{\mu}, \boldsymbol{\beta} = 0, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{h}})$ are obtained by maximising the likelihood (37.6), respectively, over the set of parameters $\{\mu, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{h}\}$, and over $\{\mu, \boldsymbol{\gamma}, \mathbf{h}\}$, subject to the constraint $\boldsymbol{\beta} = 0$. Under the null hypothesis of no association, the deviance Λ_{Hap} has an approximate χ^2 distribution with $d - 1$ df, where d is the number of distinct haplotypes consistent with the observed sample genotypes.

Zaykin *et al.* (2002) propose a two-stage strategy of first obtaining the distribution of diplotypes consistent with the unphased genotype of each individual by application of the E–M algorithm to the complete sample of cases and controls, but other algorithms for haplotype reconstruction, such as PHASE, could be used to obtain estimates of $f(H_i|G_i, \mathbf{h})$. In the subsequent test of association, the likelihood (37.6) is maximised only over the parameters μ, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$. Schaid *et al.* (2002) use the same two-stage approach, but perform a score test of association between SNP haplotypes and disease, which is asymptotically equivalent to the likelihood ratio test of Zaykin *et al.* (2002).

37.5.2 Haplotype Clustering Techniques

One potential problem with haplotype-based analyses is lack of parsimony, since many haplotypes may be consistent with the observed unphased genotype data, some of which may be very rare. In the logistic regression model, a parameter is required for each haplotype, except the baseline, leading to a test of association with disease that, potentially, has many degrees of freedom. The effects of rare haplotypes will be difficult to estimate, and there may be a lack of power to detect association, particularly if only one or two haplotypes are at high or low risk of disease. We could resolve this problem by combining rare haplotypes into a single pooled category. However, a more satisfactory approach to reduce the dimensionality of the problem is to take advantage of the expectation that similar SNP haplotypes in a small genomic region tend to share recent common ancestry, and hence are likely also to share the same alleles at the underlying functional polymorphism(s). Therefore, similar SNP haplotypes are expected to have comparable disease risks, and thus could naturally be combined in the analysis.

A number of methods that group haplotypes according to some similarity measure and then assign the same risk to all haplotypes within the same cluster have been proposed (Templeton *et al.*, 1987; 1988; 1992; Templeton and Sing, 1993; Molitor *et al.*, 2003a; 2003b; Durrant *et al.*, 2004; Morris, 2005; 2006). In this way, we parameterise the logistic regression model in terms of the log ORs for each cluster, which will be less than the number of haplotypes, reducing the degree of freedom of the resulting test of association. Morris (2005) clusters haplotypes according to a Bayesian partition model

(Knorr-Held and Rasser, 2000; Denison and Holmes, 2001). The model is defined by specifying the number of clusters of haplotypes and the centre of each cluster, taken from the set of distinct haplotypes consistent with the observed multi-locus genotype data, without replacement. All remaining haplotypes are then assigned to the nearest cluster centre, where distance between a pair of haplotypes is measured in terms of unweighted SNP allele mismatches. Morris (2005) has developed a Bayesian MCMC algorithm, GENEBPM, to sample from the posterior distribution of haplotype clusters, and the corresponding cluster log ORs, allowing for the inclusion of covariates such as environmental risk factors. Output from the algorithm can be used to (1) estimate the log OR of disease for each distinct haplotype; (2) identify clusters of haplotypes with similar disease risks; and (3) estimate the posterior probability of haplotype association with disease. More recently, Morris (2006) has extended the GENEBPM algorithm to allow for deviations from the assumption of independent haplotype effects on disease risk.

The evolution of a sample of haplotypes within a population can be represented by means of a genealogical tree, most readily modelled by the coalescent process with recombination. A number of methods have been proposed for fine-scale association mapping with samples of unrelated cases and controls that take account of the shared ancestry of their chromosomes explicitly (Morris *et al.*, 2002; 2004; Zollner and Pritchard, 2005; Minichiello and Durbin, 2006). These approaches are computationally intensive, making them impractical for the analysis of large WGA studies. Haplotype similarity can be represented by means of a dendogram, and can be thought of as representing their evolution in the population. Thus, haplotype clustering can be thought of as an approximation to the more complex population genetics processes underlying their evolution, providing computationally efficient algorithms that can be applied on the scale of WGA studies, while maintaining the principal features of the shared ancestry of a sample of chromosomes.

Multi-locus analysis of SNP haplotypes is appropriate within candidate genes or small genomic regions that have been subject to limited ancestral recombination. In the context of WGA studies, haplotype-based analyses are appropriate within blocks of strong LD. Haplotype-based analyses may provide additional information with higher-density genotyping in a follow-up study of associated regions from an initial genome scan. Employing clustering techniques may reveal patterns of haplotype similarity that help to refine the likely location of the underlying causal polymorphisms, and may identify individuals with high probability of carrying high-risk alleles who could be sequenced in functional studies.

37.6   EPISTASIS

The traditional definition of epistasis is the masking, or modification, of the effects of genotypes at one polymorphism by genotypes at a second one. One classical example of epistasis is the coat colour of Labrador dogs. One gene controls hair colour, differentiating between black and brown coats. However, a second gene determines the deposition of hair colour, where dogs carrying one genotype will be golden, irrespective of their black/brown genotype, masking the effects of the first gene. The existence of epistasis is not surprising since we expect the biological mechanisms underlying complex diseases to be extremely intricate, incorporating the effects of multiple genetic risk factors, acting

together in some way. Furthermore, there is increasing evidence from model organisms, including *Drosophila melanogaster* and *Saccharomyces cerevisiae*, that epistasis occurs frequently, involves multiple polymorphisms, and may contribute large effects to the genetic component of phenotypic variation (Mackay, 2001; Brem and Kruglyak, 2005; Brem *et al.*, 2005; Storey *et al.*, 2005).

From a statistical viewpoint, epistasis corresponds to an interaction between genotypes at two or more loci. Thus, the logistic regression modelling framework described above for multi-locus analysis generalises naturally to allow for epistasis. For example, to model interaction between a pair of SNPs, j and k, the linear predictor for the ith individual can be expressed as

$$
\begin{aligned}
\eta_i = \beta_0 &+ \beta_{Aj} z_{(Aj)i} + \beta_{Dj} z_{(Dj)i} + \beta_{Ak} z_{(Ak)i} + \beta_{Dk} z_{(Dk)i} \\
&+ \beta_{AAjk} z_{(Aj)i} z_{(Ak)i} + \beta_{ADjk} z_{(Aj)i} z_{(Dk)i} + \beta_{DAjk} z_{(Dj)i} z_{(Dk)i} \\
&+ \beta_{DDjk} z_{(Dj)i} z_{(Dk)i},
\end{aligned}
\tag{37.7}
$$

where the SNP genotype indicator variables are defined in Table 37.2. The parameters β_{Aj} and β_{Dj}, correspond to the additive and dominance *main effects*, respectively, of SNP j, with β_{Ak} and β_{Dk} interpreted in the same way for SNP k. The four interaction terms, $\beta_{AA}, \beta_{AD}, \beta_{DA}$, and β_{DD}, correspond to additive–additive, additive–dominance, dominance–additive, and dominance–dominance contributions to epistasis between SNPs j and k.

We can test for joint association of SNPs j and k with disease, allowing for epistasis between the two loci, by comparing the interaction model (37.7) and the null model $\eta_i = \beta_0$. Under the null hypothesis of no association of either SNP with disease, the difference in deviances between the two models has an approximate χ^2 distribution with 8 df. Furthermore, we can test for epistasis between SNPs j and k in their association with disease by comparing the interaction model (37.7) with a constrained model $\beta_{AA} = \beta_{AD} = \beta_{DA} = \beta_{DD} = 0$, where the difference in deviances now has 4 df. Alternatively, standard statistical model building techniques can be utilised to infer the best combination of main effects and interactions to describe the association with disease. With more than two SNPs, higher-order interactions could also be incorporated, although these effects are difficult to estimate without very large sample sizes, and are difficult to interpret.

In the presence of epistasis between SNPs, we would expect modelling interaction effects to lead to increased power over tests that include only the corresponding main effects. However, researchers are often reluctant to consider epistasis because of the fear that a less parsimonious model will decrease the power to detect association unless the interaction effects are large. Marchini *et al.* (2005) have shown that this fear is misplaced by demonstrating that testing for association allowing for additive main effects and additive–additive epistasis for each of $n(n-1)/2$ pairs of SNPs has greater power to detect association with disease than n single-locus tests with additive-only effects. This result holds for a range of interaction models, despite the additional burden of multiple testing via Bonferroni correction.

Evans *et al.* (2006) have investigated the use of two-stage approaches for testing for association between pairs of interacting SNPs and disease to reduce the multiple-testing burden of a complete two-dimensional scan of the genome. In the first stage, a single-locus test of association is performed at each SNP. Only those SNPs passing some predetermined threshold of significance are carried forward to the second stage of testing where various

strategies could be employed to test for association of pairs of SNPs with disease, allowing for epistasis. For example, we could test for association between SNPs carried forward to the second stage, or we could consider all possible pairs of SNPs that contain at least one carried forward to the second stage of testing. The choice of a stringent level of significance for the first stage of testing reduces the number of tests performed overall, but also reduces the probability of detecting association of disease with a pair of SNPs made up of a strong epistatic component with minimal main effects. In fact, the results of detailed simulations over a wide range of models of epistasis suggest that two-stage strategies are, in general, less powerful than a two-dimensional scan of the genome, irrespective of the significance threshold at stage one, even while taking account of the additional correction for multiple testing.

An alternative approach to allow for epistasis in association studies is to make use of Bayesian model averaging techniques (Hoeting *et al.*, 1999). This approach can be used to obtain the joint posterior distribution of the main effects of all SNPs and interactions between all pairs of SNPs by considering the space of all possible models of association, conditional on the observed genotypes and disease phenotypes. Conti *et al.* (2003) describe a Bayesian MCMC algorithm to sample from the posterior distribution of models incorporating pair-wise and higher-order epistasis between multiple SNPs in a candidate gene, together with interaction with non-genetic risk factors. They demonstrate the utility of this approach to a population-based association study of colorectal polyps with candidate metabolic genes, allowing for epistasis between polymorphisms and interaction with non-genetic risk factors including smoking and consumption of well-done red meat. The main advantage of Bayesian model averaging approaches is that all SNPs can be considered simultaneously, rather than focusing on each pair of SNPs. However, these methods are likely to be extremely computationally intensive on the scale of WGA studies.

37.7   REPLICATION

Because of the small effect sizes of complex traits, the multiple-testing problem, and limited sample sizes and genotyping resources, most WGA studies have only marginal power to identify most real effects. As a result, many true results may be difficult to distinguish from chance results in the initial study. For example, one of the first WGA studies of Crohn's disease (Duerr *et al.*, 2006) clearly identified a single locus of large effect, but several other loci, which were subsequently shown to be genuinely associated, had significance levels of only 10^{-3}–10^{-5} and were initially missed (Cardon, 2006). These challenges highlight the fact that WGA scans are best considered as hypothesis-generating mechanisms, and replication studies are needed to validate and refine initial evidence from WGA studies.

In principle, replication studies are straightforward to design and implement: given assumptions of effect size and frequency, one can calculate sample sizes needed for a specified power, and then conduct careful validation experiments. In practice, however, replication has proved to be one of the most difficult areas in human complex trait genetics, one with a rich history of confusion and false claims. Surprisingly, this standard of failure has been observed in the simplest of studies, involving only single candidate genes with but a few genetic markers and statistical tests. In WGA studies, the challenge will be even greater, as large numbers of SNPs may need follow up.

For replication to serve as a useful measure of validation, association studies need to follow the standard principles of experimental design; i.e. replication studies should match the conditions of the initial study as closely as possible and test explicitly defined, refutable hypotheses. In the context of association studies, this means that the same SNPs, same statistical tests and same phenotypes should be tested in replication as those used in the initial WGA scans. Historically, however, investigators have often used nearby SNPs and conducted multiple statistical tests (which may or may not be reported) on samples with related but non-identical phenotypes, all under the guise of replication. Unsurprisingly, the results have been highly varied, with 'replication' sometimes claimed in error, sometimes missed, and occasionally, claimed with the opposite allele to that originally identified, a 'flip-flop' phenomenon that stretches biological credibility in the absence of unmeasured confounders (Patterson and Cardon, 2005; Lin *et al.*, 2007).

One of the main problems is that investigators have attempted to combine replication studies (using the same SNPs) with fine-mapping studies (using additional SNPs to try to find the strongest associated variants). This is an attractive strategy economically but creates challenges to the inference process since the former is fundamentally a hypothesis-testing exercise, while the latter assumes that the null hypothesis has already been rejected. In the association context, these two strategies are extremely difficult to disentangle. At present, the most robust strategy is to separate the two and conduct them in their natural sequential order (Clarke *et al.*, 2007).

Estimating the power and required sample size for a study to replicate WGA findings is complicated by inflation of effect sizes: for all but the largest genetic effects, the loci taken forward for replication will have an estimated effect size drawn from the upper tail of their effect-size distribution over replicate studies, a phenomenon known as the *Winner's Curse* (Bazerman and Samuelson, 1983). The process of selecting only the most significant loci for further scrutiny is known to invoke the Winner's Curse in both linkage (Goring *et al.*, 2001) and association studies (Lohmueller *et al.*, 2003). Accordingly, using these biased estimates to design replication studies results in overestimates of statistical power and underestimates of sample size requirements. This poses a power conundrum: replication studies will be underpowered if they use the biased WGA estimates of effect size, but if successful, they can provide unbiased effect-size estimates of the loci initially identified by WGA. Because many factors can influence the degree of bias, it is not easy to correct the effect-size estimates from WGA studies. Empirical evidence from the large number of studies currently under way will be extremely valuable in helping to provide rough brackets on the range of bias that is recoverable in replication studies of practical size and scope.

37.8 PROSPECTS FOR WHOLE-GENOME ASSOCIATION STUDIES

Initial reports of WGA studies began as early as 2002 (Ozaki *et al.*, 2002), but studies involving more comprehensive coverage of common variants began in 2005 (Klein *et al.*, 2005; Duerr *et al.*, 2006; Hampe *et al.*, 2007). The initial reports are promising, with each study identifying and validating several novel loci for different diseases. These studies demonstrate that WGA can be successful in identifying common variants for complex traits in humans. Given the chequered history of human genetic association studies (Cardon and Bell, 2001; Ioannidis *et al.*, 2001), this is a major advance in the field.

A large number of WGA studies, each involving thousands of individuals and $> 10^5$ genetic markers, are under way and likely to be reported over the next few years. If the initial trends continue, a number of new genes will soon be identified for complex traits, thereby uncovering new biological pathways and stimulating a cascade of focused experimental studies. For statistical genetics research, these studies raise a series of new opportunities and challenges. Knowledge of new genes will enable detailed studies of gene–gene and gene–environment interactions, determination of population-specific effects, and identification of further loci by conditioning on the initial findings. Moreover, owing to ongoing Biobank initiatives in a number of countries, in which $> 10^5$ individuals are being studied for a wide range of human conditions, it will be possible to conduct population-based assessments of specific gene or haplotype effects, evaluate longitudinal genetic effects, and study the impact of particular genes or gene combinations across traits (pleiotropy).

Ongoing technology advances suggest that WGA studies will not end with the current paradigm of 300–500K SNPs, or even with the immediate target of million-SNP panels. There is intense ongoing commercial research into resequencing the DNA of entire genomes of individuals, which, if achieved, would comprise the first true WGO design. If this technology eventually is able to be conducted with high accuracy and low cost so that it can be deployed in large samples, it will fill a void in the current generation of genetic studies: detection and annotation of rare genetic variants in humans. Present technologies focus on common genetic variants (typically $> 1\%$ frequency in particular populations), leaving the majority of human genetic variation unexplored. Rare variants are known to be of critical importance in human diseases (e.g. BRCA1 and BRCA2), so the upcoming resequencing information will be of immense biomedical importance. The statistical challenges facing such studies will be of a difficulty level matching the potential importance, requiring new designs, inferential frameworks, and theoretical models to exploit the emerging resource.

REFERENCES

Abecasis, G.R., Noguchi, E., Heinzmann, A., Traherne, J.A., Bhattacharya, S., Leaves, N.I., Anderson, G.G., Zhang, Y., Lench, N.J., Carey, A., Cardon, L.R., Moffatt, M.F. and Cookson, W.O.C. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. *American Journal of Human Genetics* **68**, 191–197.

Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2001). GRR: graphical representation of relationship errors. *Bioinformatics* **17**, 742–743.

Almind, K., Bjorbaek, C., Vestergaard, H., Hansen, T., Echwald, S. and Pedersen, O. (1993). Amino acid polymorphisms of insulin receptor substrate-1 in non-insulin dependent diabetes mellitus. *Lancet* **342**, 828–832.

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**, 781–791.

Barrett, J.C. and Cardon, L.R. (2006). Evaluating coverage of genome-wide association studies. *Nature Genetics* **38**, 659–662.

Bazerman, M.H. and Samuelson, W.F. (1983). I won the auction but don't want the prize. *Journal of Conflict Resolution* **27**, 618–634.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **B57**, 289–300.

Brem, R.B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 1572–1577.

Brem, R.B., Storey, J.D., Whittle, J. and Kruglyak, L. (2005). Genetic interaction between polymorphisms that affect gene expression in yeast. *Nature* **436**, 701–703.

Cardon, L.R., Bell, J.I. (2001). Association study designs for complex diseases. *Nature Reviews Genetics* **2**, 91–99.

Cardon, L.R. (2006). Genetics. Delivering new disease genes. *Science* **314**, 1403–1405.

Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* **74**, 106–120.

Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Buchman, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. and Sing, C.F. (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *American Journal of Human Genetics* **63**, 595–612.

Clark, A.G. (2004). The role of haplotypes in candidate gene studies. *Genetic Epidemiology* **27**, 321–333.

Clarke, G.M., Carter, K.W., Palmer, L.J., Morris, A.P., Cardon, L.R. (2007). Fine-mapping vs replication in whole genome association studies. *American Journal of Human Genetics*. (The revised version of the paper is currently being reviewed, and we expect a response imminently).

Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, J.D., Smink, L.J., Lam, A.C., Ovington, N.R., Stevens, H.E., Nutland, S., Howson, J.M.M., Faharn, M., Moorhead, M., Jones, H.B., Falkowski, M., Hardenbol, P., Willis, T.D. and Todd, J.A. (2005). Population structure, differential bias and genomic control in a large case-control association study. *Nature Genetics* **37**, 1243–1246.

Conti, D.V., Cortessis, V., Molitor, J. and Thomas, D.C. (2003). Bayesian modelling of complex metabolic pathways. *Human Heredity* **56**, 83–93.

van Dam, R.M., Hoebee, B., Seidell, J.C., Schaap, M.M., Blaak, E.E. and Reskens, E.J. (2004). The insulin receptor substrate-1 Gly972Arg polymorphism is not associated with type 2 diabetes mellitus in two population-based studies. *Diabetic Medicine* **21**, 752–758.

Dawson, E., Abecasis, G.R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D.M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lohmussaar, E., Zernant, J., Tonisson, N., Remm, M., Magi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D.R., Cardon, L.R. and Dunham, I. (2002). A first generation linkage disequilibrium map of chromosome 22. *Nature* **418**, 544–548.

Denison, D.G.T. and Holmes, C.C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics* **57**, 143–149.

Devlin, B., Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997–1004.

Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Seinhard, A.H., Abraham, C., Reueiro, M., Griffiths, A., Dassopoulos, T., Bitton, A., Yong, H., Targan, S., Wu Datta, L., Kistner, E.O., Schumm, L.P., Lee, A.T., Gregersen, P.K., Barmada, M.M., Rotter, J.I., Nicolae, D.L. and Cho J.H. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461–1463.

Dunning, A.M., Durocher, F., Healey, C.S., Teare, M.D., McBride, S.E., Carlomango, F., Xu, C.F., Dawson, E., Rhodes, S., Ueda, S., Lai, E., Luben, R.N., van Rensburg, E.J., Mannermaa, A., Kataja, V., Rennart, G., Dunham, I., Purvis, I., Easton, D. and Ponder, B.A. (2000). The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics* **67**, 1544–1554.

Durrant, C., Zondervan, K.T., Cardon, L.R., Hunt, S., Deloukas, P. and Morris, A.P. (2004). Linkage disequilibrium mapping via cladistic analysis of SNP haplotypes. *American Journal of Human Genetics* **75**, 35–43.

Evans, D., Marchini, J., Morris, A.P. and Cardon, L.R. (2006). Two stage two locus models in genome wide association. *PLoS Genetics* **2**, e157.

Excoffier, L. and Slatkin, M. (1995). Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12**, 921–927.

Florez, J.C., Burtt, N., de Bakker, P.I., Almgren, P., Tuomi, T., Holmkvist, J., Gaudet, D., Hudson, T.J., Schaffner, S.F., Daly, M.J., Hirschhorn, J.N., Groop, L., Altshuler, D. (2004). Association testing in 9,000 people fails to confirm the association of the insulin receptor substrate-1 G972R polymorphism with type 2 diabetes. *Diabetes* **53**, 3313–3318.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., de Felice, M., Lochner, A., Faggart, M., Lui-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. and Althshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.

Goring, H.H., Terwilliger, J.D. and Blangero, J. (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics* **69**, 1357–1369.

Grant, S.F.A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., Styrkarsdottir, U., Magnusson, K.P., Walters, G.B., Palsdottir, E., Jonsdottir, T., Gudmundsdottir, T., Gylfason, A., Saemunds-dottir, J., Wilensky, R.L., Reilly, M.P., Rader, D.J., Bagger, Y., Christiansen, C., Gudnason, V., Sigurdsson, G., Thorsteinsdottir, U., Gulcher, J.R., Kong, A. and Stefansson, K. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature Genetics* **38**, 320–323.

Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F.M., Briggs, J., Gunther, S., Prescott, N.J., Onnie, C.M., Hasler, R., Sipos, B., Folsch, U.R., Lengauer, T., Platzer, M., Matthew, C.G., Krawczak, M. and Schreiber, S. (2007). A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nature Genetics* **39**, 207–211.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14**, 382–417.

Hugot, J.P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J.P., Belaiche, J., Almer, S., Tysk, C., O'Moiran, C.A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J.F., Sahbatou, M. and Thomas, G. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603.

Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A. and Contopoulos-Ioannidis, D.G. (2001). Replication validity of genetic association studies. *Nature Genetics* **29**, 306–309.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

Johnson, G.C.L., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., DiGenova, G., Veda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F., Twells, R.C.J., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tunmilehto, J., Gough, S.C.L., Clayton, D.G. and Todd, J.A. (2001). Haplotype tagging for the identification of common disease genes. *Nature Genetics* **29**, 233–237.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGio-vanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferrix, F.L., Ott, J., Barnstable, C. and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389.

Knorr-Held, L. and Rasser, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **46**, 13–21.

Lawrence, R., Evans, D.E., Morris, A.P., Ke, X., Hunt, S., Paolucci, M., Ragoussis, J., Deloukas, P., Bentley, D. and Cardon, L.R. (2005). Genetically indistinguishable SNPs and their influence on inferring the location of disease-associated variants. *Genome Research* **15**, 1503−1510.

Lin, P.I., Vance, J.M., Pericak-Vance, M.A. and Martin, E.R. (2007). No gene is an island: the flip-flop phenomenon. *American Journal of Human Genetics* **80**, 531−538.

Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. and Hirschhorn, J.N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics* **33**, 177−182.

Mackay, T.F. (2001). The genetic architecture of quantitative traits. *Annual Review of Genetics* **35**, 303−339.

Marchini, J., Donnelly, P. and Cardon, L.R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413−417.

Minichiello, M.J. and Durbin, R. (2006). Mapping trait loci by use of inferred ancestral recombination graphs. *American Journal of Human Genetics* **79**, 910−922.

Molitor, J., Marjoram, P. and Thomas, D. (2003a). Application of Bayesian spatial statistical methods to the analysis of haplotype effects and gene mapping. *Genetic Epidemiology* **25**, 95−105.

Molitor, J., Marjoram, P. and Thomas, D. (2003b). Fine scale mapping of disease genes with multiple mutations via spatial clustering techniques. *American Journal of Human Genetics* **73**, 1368−1384.

Morris, A.P. (2005). Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. *Genetic Epidemiology* **29**, 91−107.

Morris, A.P. (2006). A flexible Bayesian framework for modelling haplotype association with disease allowing for dominance effects of the underlying causative variants. *American Journal of Human Genetics*. **79**, 679−694.

Morris, A.P., Whittaker, J.C. and Balding, D.J. (2002). Fine-scale mapping of disease loci via coalescent modelling of genealogies. *American Journal of Human Genetics* **70**, 686−707.

Morris, A.P., Whittaker, J.C. and Balding, D.J. (2004). Little loss of information due to unknown phase for fine-scale linkage disequilibrium mapping with single nucleotide polymorphism genotype data. *American Journal of Human Genetics* **74**, 945−953.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y. and Tanaka, T. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics* **32**, 650−654.

Parikh, H. and Groop, L. (2004). Candidate genes for type 2 diabetes. *Reviews in Endocrine and Metabolic Disorders* **5**, 151−176.

Patil, N., Bem, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Majoribanks, C., McDonough, D.P., Nguyen, B.T.N., Norris, M.C., Sheehan, J.B., Sten, N., Stern, D., Stokowski, R.P., Thomas, D.J., Trulson, M.O., Vyas, K.R., Frazer, K.A., Fodor, S.P.A. and Cox, D.R. (2001). Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science* **294**, 1719−1722.

Patterson, M. and Cardon, L. (2005). Replication publication. *PLoS Biology* **3**, e327.

Phillips, M.S., Lawrence, R., Sachidanandam, R., Morris, A.P., Balding, D.J., Donaldson, M.A., Studebaker, J.F., Ankener, W.M., Alfisi, S.V., Kuo, F.S., Camisa, A.L., Pazorov, V., Scott, K.E., Carey, B.J., Faith, J., Katari, G., Bhatti, H.A., Cyr, J.M., Derohannessian, V., Elousa, C., Forman, A.M., Grecco, N.M., Hoch, C.R., Kuebler, J.M., Lathrop, J.A., Mockler, M.A., Nachtman, E.P., Restine, S.L., Varde, S.A., Hozza, M.J., Gelfand, C.A., Broxholme, J., Abecasis, G.R., Boyce Jacino, M.T. and Cardon, L.R. (2003). Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genetics* **33**, 382−387.

Pritchard, J.K., Stephens, M. and Donnelly, P.J. (2000a). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945−959.

Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P. (2000b). Association mapping in structured populations. *American Journal of Human Genetics* **67**, 170–181.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. and Lander, E.S. (2001). Linkage disequilibrium in the human genome. *Nature* **411**, 199–204.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Rubinsztein, D.C. and Easton, D.F. (1999). Apolipoprotien E genetic variation and Alzheimer's disease: a meta analysis. *Dementia and Geriatric Cognitive Disorders* **10**, 199–209.

Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M. and Poland, G.A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70**, 425–434.

Setakis, E., Stirnadel, H. and Balding, D.J. (2006). Logistic regression protects against population structure in genetic association studies. *Genome Research* **16**, 290–296.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T.J., Montpetit, A., Pshezhetsky, A.V., Prentki, M., Posner, B.I., Balding, D.J., Meyre, D., Polychronakos, C. and Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885.

Stephens, M. and Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genetic data. *American Journal of Human Genetics* **73**, 1162–1169.

Stephens, M., Smith, N.J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978–989.

Storey, J.D., Akey, J.M. and Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology* **3**, e267.

Taillon-Miller, P., Bauer-Sardina, I.B., Saccone, N.L., Pulzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J.P. and Kwok, P.Y. (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genetics* **25**, 324–328.

Tavtigian, S., Simard, J., Teng, D., Abtin, V., Baumgard, M., Beck, A., Camp, N., Carillo AR, Chen Y, Dayananth P, Desrochers M, Dumont M, Farnham JM, Frank D, Frye C, Ghaffari S, Gupte JS, Hu R, Iliev D, Janecki T, Kort EN, Laity KE, Leavitt A, Leblanc G, McArthur-Morrison J, Pederson A, Penn B, Peterson KT, Reid JE, Richards S, Schroeder M, Smith R, Snyder SC, Swedlund B, Swensen J, Thomas A, Tranchant M, Woodland AM, Labrie F, Skolnick MH, Neuhausen S, Rommens J, Cannon-Albright LA. (2001). A candidate prostate cancer susceptibility gene at chromosome 17p. *Nature Genetics* **27**, 172–180.

Templeton, A.R., Boerwinkle, E. and Sing, C.F. (1987). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. *Genetics* **117**, 343–351.

Templeton, A.R., Crandall, K.A. and Sing, C.F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**, 619–633.

Templeton, A.R. and Sing, C.F. (1993). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics* **134**, 659–669.

Templeton, A.R., Sing, C.F., Kessling, A. and Humphries, S. (1988). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. *Genetics* **120**, 1145–1154.

Valdar, W., Solberg, L.C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W.O., Taylor, M.S., Rawlins, J.N.P., Mott, R. and Flint, J. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics* **38**, 879–887.

Voight, B.F. and Pritchard, J.K. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genetics* **1**, e32.

Weir, B.S. (1996). *Genetic Data Analysis II*. Sinauer Associates Inc: Sunderland, Massachusetts, USA, pp 112–133.

Weir, B.S., Cardon, L.R., Anderson, A.D., Neilsen, D.M. and Hill, W.G. (2005). Measures of human population structure show heterogeneity among genomic regions. *Genome Research* **15**, 1468–1476.

The Wellcome Trust Case Control Consortium (2007). A genome wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* (in press).

Wright, A.F. and Hastie, N.D. (2001). Complex genetic diseases: controversy over the Croesus code. *Genome Biology* **2**, COMMENT2007.

Zaykin, D.V., Westfall, P.H., Young, S.S., Karnoub, M.A., Wagner, M.J. and Ehm, M.G. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity* **53**, 79–91.

Zeggini, E., Parkinson, J., Halford, S., Owen, K.R., Frayling, T.M., Walker, M., Hitman, G.A., Levy, J.C., Sampson, M.J., Feskens, E.J.M., Hattersley, A.T. and McCarthy, M.I. (2004). Association studies of insulin receptor substrate 1 gene (IRS1) variants in type 2 diabetes samples enriched for family history and early age of onset. *Diabetes* **53**, 3319–3322.

Zollner, S. and Pritchard, J.K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169**, 1071–1092.

Zondervan, K.T. and Cardon, L.R. (2004). The complex interplay among factors that influence allelic association. *Nature Reviews Genetics* **5**, 89–100.

Family-based Association

F. Dudbridge

MRC Biostatistics Unit, Institute for Public Health, Cambridge, UK

Family-based methods are a useful means to protect against the confounding effects of population stratification and other factors. The transmission/disequilibrium test is a test of linkage and association that is at once the most popular such method and the starting point for numerous extensions and generalizations, the most prominent of which are reviewed in this chapter. A logistic regression formulation leads to natural extensions for multiallelic markers, haplotypes and quantitative traits. When haplotypes are uncertain, special methods are described for testing for association within families. An approach suitable for general pedigree structures, which derives the conditional distribution of a covariance score function given the pedigree structure, is described. For quantitative traits, ordinary linear regression models are adapted for the family-based design and allow the modelling of linkage in the error variance. When testing association in the presence of linkage, some care is required to avoid bias due to correlated transmissions to siblings, and a likelihood-based solution to this problem that allows for uncertain haplotypes and missing parental genotypes is presented. Some ongoing work in the area is summarized.

38.1 INTRODUCTION

Family-based association studies include a broad range of methods that aim to detect association between genetic markers and disease or quantitative phenotypes, using families as the primary sampling unit. Most often, nuclear families consisting of the two parents and a number of full siblings are used, but extended pedigrees may also be used for testing association, as may subsets of nuclear families such as sib pairs or single parent families. As might be expected given the diversity of family structures, there is a large range of methods for family-based association; here only the most prominent ones will be discussed, and these are able to analyse all of the most common family structures.

The appeal of family-based association is due to several reasons. First, in common with all epidemiological studies, there is the need to match cases to controls at the population and, ideally, the individual level. As discussed in **Chapter 36**, inappropriate matching of cases to controls can lead to selection bias and confounding effects if gene frequencies differ between case and control populations. A situation of particular concern is population

stratification (see **Chapter 35**) in which the study population actually consists of a mixture of discrete sub-populations that differ both in gene frequency and disease prevalence. Case-control sampling might then recruit relatively more cases from the sub-populations with higher prevalence, leading to differences in marginal gene frequency between cases and controls even if a gene has no influence on disease.

Secondly, family-based association tests are confounded with linkage (Ott, 1989), in the sense that association can only be detected in the presence of linkage. While such confounding effects can present technical difficulties in interpreting results (Whittaker *et al.*, 2000), they have the positive effect of confirming that truly associated markers are physically close to the causal genetic variant. This property also reinforces the intuitive view that genetic phenotypes are inherited and should therefore be detectable in families.

Another important advantage of family-based designs is that they allow parent of origin studies that allow for imprinting effects and interactions between maternal and foetal genotypes (Weinberg, 1999). Such analyses are simply not possible in population studies.

By choosing controls from within the same families as the cases, the problem of population stratification is largely eliminated, and the confounding effects of other factors can be substantially reduced. For example, members of the same family are likely to share a high proportion of environmental and dietary exposures, so that differences in phenotype are more likely to be due to genetic differences than to unmeasured confounders. Family-based designs are thus attractive from the point of view of identifying true causal effects. Their principal disadvantages arise from practical matters of recruitment and cost. Suitable family members may be unavailable for genotyping, for example in late-onset diseases in which both parents may be deceased. Family controls are over-matched, since they share, in expectation, up to one-half of their genome with the cases, leading to higher genotyping costs per unit of information as compared to unrelated controls (Cardon and Palmer, 2003). For these reasons, unrelated controls are currently favoured for large-scale genome association scans (see **Chapter 37**), but family-based controls remain useful, particularly for replication studies that can confirm previously suggested associations to a higher standard of evidence.

The current use of family-based controls originated with the haplotype relative risk method (Falk and Rubinstein, 1987). This uses a sample of affected cases together with each of their two parents, and for each case, the alleles present in the parents but not transmitted to the case are combined into a control genotype that can be analysed using standard methods for unrelated cases and controls (see **Chapter 36**). A variation, termed *haplotype-based haplotype relative risk*, treats the allele rather than the genotype as the unit of observation, again following standard methods for unrelated subjects (Terwilliger and Ott, 1992). The problem of estimating genotype relative risks in family-based designs was first addressed by Self *et al.* (1991), from which much of the regression methodology described below is derived.

The transmission/disequilibrium test (TDT) (Spielman *et al.*, 1993) is a landmark in the development of family-based association, both as a point of departure for many extensions and generalizations, and as the baseline standard to which new methods are often compared. Its appeal lies in its simplicity, close relation to standard statistical tests and absolute protection from population stratification. In its simplest form, it is similar to the haplotype-based haplotype relative risk, except that it performs a matched analysis of untransmitted alleles versus transmitted alleles, rather than an unmatched analysis. This basic change allows the TDT to be viewed both as a test of association and as a test of

linkage, and also protects against deviations from Hardy–Weinberg equilibrium that could be induced by non-random mating. Because of these properties, the TDT has become a very popular method and has spawned many extensions. Perhaps more than in other areas of statistical genetics, these methods have become identified with the software that implements them, owing to the variety of possible family structures and difficulties in fitting them to standard statistical models. Here, some of the more established methods will be reviewed, with references to the relevant software, and some comparisons will be drawn between these methods.

The structure of the chapter is as follows. Section 38.2 reviews the TDT and some of its properties. Section 38.3 sets out a logistic regression formulation of the TDT and describes how it can be used to include multiple alleles, genotypes and environmental covariates. Section 38.4 discusses the problem of uncertain haplotypes and missing parental genotypes, and describes the popular TRANSMIT program. In Section 38.5 some approaches for analysing general pedigrees are described, starting with the sib-TDT for sibships and progressing to a general approach implemented in the FBAT software. Methods for quantitative traits are discussed in Section 38.6, with emphasis on the popular QTDT software that allows joint modelling of linkage and association. Section 38.7 focuses on testing association in the presence of linkage; some solutions are given, including a likelihood-based program, UNPHASED, that relates closely to the original TDT. Finally, Section 38.8 gives a summary of ongoing work in this area.

38.2 TRANSMISSION/DISEQUILIBRIUM TEST

In its original form, the TDT considers the transmission of the variant allele of a biallelic marker from heterozygous parents to affected children. Table 38.1 shows the four cell counts relating to the transmissions from a parent to an affected child, arranged as a contingency table. The TDT treats the untransmitted allele as a matched control to the transmitted allele, in which case only the heterozygous parents are informative. The null hypothesis is that heterozygous parents transmit the two alleles with equal probability. Let $T = 1$ when the parent transmits the variant allele, and $T = 0$ otherwise; then, under equally likely transmission $E(T) = \frac{1}{2}$, $\text{var}(T) = \frac{1}{4}$, and by applying the central limit theorem over the $n_{12} + n_{21}$ heterozygous parents, we have the TDT statistic

$$TDT = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}, \tag{38.1}$$

which is asymptotically distributed as χ^2 with 1 degree of freedom.

Table 38.1 Counts of transmissions from n parents of affected children, used in calculating the transmission/disequilibrium test.

		Non-transmitted allele		
		Variant	Common	Total
Transmitted	Variant	n_{11}	n_{12}	$n_{1.}$
allele	Common	n_{21}	n_{22}	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	n

It can be formally shown that equal transmission probability occurs either when there is no linkage between marker and disease, or when there is no association (Ewens and Spielman, 1995). This can be intuitively seen by considering the parental chromosomes that carry disease alleles that are penetrant in the child (assume that Mendelian segregation holds on chromosomes carrying non-penetrant alleles, leading to equal transmission probability). Figure 38.1 shows the four situations that are possible when a parent carries a penetrant allele and is heterozygous at the marker. When there is no linkage, either marker allele is transmitted with the disease allele with probability 1/2 regardless of the haplotype distribution in parental chromosomes: scenarios (a) and (b) are equally likely, as are (c) and (d). But when there is no association, there is no information on which marker allele occurs on the same haplotype as the disease allele, given that the parent is heterozygous: scenarios (a) and (c) are equally likely, as are (b) and (d). Therefore, either marker allele occurs on the disease haplotype with probability 1/2 and is then transmitted with probability 1/2 regardless of the recombination fraction. This means that the null hypothesis of the TDT may be taken to be either no linkage or no association.

Although the TDT is often used as a test of association, it was first proposed as a test of linkage (Spielman *et al.*, 1993). This was motivated by studies of the insulin gene in
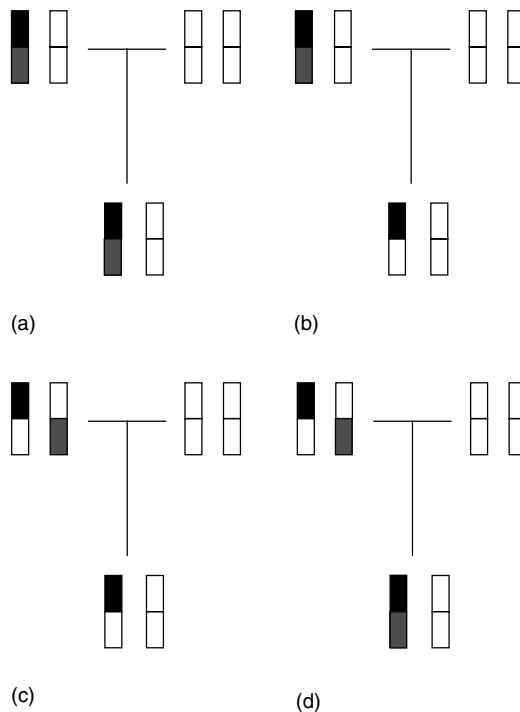
Figure 38.1 Four transmission scenarios for a parent heterozygous at a test marker. The disease locus is shown above the marker locus. Disease risk allele is shown in black, variant marker allele in grey: (a) marker allele on disease chromosome, no recombination; (b) marker allele on disease chromosome, with recombination; (c) marker allele on normal chromosome, no recombination; (d) marker allele on normal chromosome, with recombination.

Table 38.2 Counts of transmissions in n case–parent trios, used in calculating dominant and recessive versions of the TDT. Allele 1 denotes the common allele and allele 2 denotes the variant alleles.

		Offspring genotype			
		1/1	1/2	2/2	Total
	$1/1 \times 1/2$	n_{11}	n_{12}	—	$n_{1.}$
Mating type	$1/2 \times 1/2$	n_{21}	n_{22}	n_{23}	$n_{2.}$
	$1/2 \times 2/2$	—	n_{32}	n_{33}	$n_{3.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

type-1 diabetes, which had shown population association in candidate-gene studies but little evidence for linkage in sib pairs. Thus, the presence of association was established and the TDT was able to show that this association coincided with linkage, as opposed to arising, say, from population stratification. This application has remained useful for candidate-gene studies, but in positional cloning studies that start with genome scans for linkage, it has been more appropriate to treat the TDT as a test of association for fine mapping. Now, with the move towards genome-wide association scans (see **Chapter 37**), the situation will again arise in which population association is demonstrated without evidence for linkage, and the interpretation of the TDT will revert to its original sense as a test of linkage. These considerations have an important bearing on how data are analysed in extended sibships or pedigrees, as discussed in Section 38.7.

The sampling units for the TDT are single parents of affected children, though both parents must be available for each child. This treats the alleles in the affected children as independent, which is true under the null hypothesis. However, when there is linkage and association, the alleles in the children are independent only under the multiplicative model of relative risk (see **Chapter 36**), so the TDT may not be the most powerful test in other models of risk. TDT statistics have been derived for dominant and recessive models that may be preferred in those instances (Schaid and Sommer, 1994). These treat the case–parent trio as the sampling unit, with the relevant cell counts given in Table 38.2. These give the χ^2 statistics

$$TDT_{\text{DOM}} = \frac{\left[n_{23} + n_{22} - \frac{3}{4}n_{2.} + n_{12} - \frac{1}{2}n_{1.}\right]^2}{\frac{3}{16}n_{2.} + \frac{1}{4}n_{1.}} \tag{38.2}$$

and

$$TDT_{\text{REC}} = \frac{\left[n_{33} - \frac{1}{2}n_{3.} + n_{23} - \frac{1}{4}n_{2.}\right]^2}{\frac{1}{4}n_{3.} + \frac{3}{16}n_{2.}} \tag{38.3}$$

38.3 LOGISTIC REGRESSION MODELS

The TDT can be derived from a logistic regression model that offers scope for extending the test in many ways. Recalling that the TDT is equivalent to a matched analysis of transmitted versus untransmitted alleles, the standard approach from epidemiology is to

define a conditional logistic regression model with transmission as the random outcome and alleles as predictors. For a biallelic marker, the log-odds ratio β for transmission is given by

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta, \tag{38.4}$$

where p is the probability that the variant allele is transmitted by a heterozygous parent. The likelihood for n heterozygous parents is

$$L = \prod_{i=1}^{n} \frac{\exp(\delta_i \beta)}{1 + \exp(\beta)}, \tag{38.5}$$

where $\delta_i = 1$ when parent i transmits the variant allele, and is 0 otherwise. Standard likelihood theory yields a score test for $\beta = 0$ that is identical to that obtained using the TDT.

For multiallelic markers, the model can be extended to fit a regression parameter to each pair of transmitted/non-transmitted alleles. However, this can result in a large number of parameters, and a more economical approach is to define marginal odds ratios for the transmission of each allele, and to fit these parameters in a model that conditions on the parental genotype. Thus, for alleles $i = 1, \ldots, k$, we define parameters $\beta_1, \ldots, \beta_k$ and for an i/j parent transmitting allele i we have

$$\text{logit}(p_{ij}) = \beta_i - \beta_j. \tag{38.6}$$

The likelihood contribution for such a parent is $\dfrac{\exp(\beta_i)}{\exp(\beta_i) + \exp(\beta_j)}$. The null hypothesis is that all $\beta_i = 0$, which may be tested using a joint score test, or by using a likelihood ratio test that compares the likelihood under the null to that obtained when it is maximized over all values of $\beta_1, \ldots, \beta_k$ (one parameter must be set to 0 for identifiability). This then yields maximum likelihood estimates of the odds ratios for transmission, which may be regarded as relative risks for disease (Cordell and Clayton, 2002). The model can be fitted using standard software for conditional logistic regression, and the null hypothesis has been shown to correspond to no linkage between marker and disease, or no association between any allele and disease (Sham and Curtis, 1995).

The regression model can be extended in many ways. A useful approach is to regard the analysis not in terms of transmissions from parents, but rather in terms of comparing cases, the affected children, to controls formed from the other combinations of parental alleles. The likelihood contribution from a full case–parent trio is equivalent to that obtained by matching the case to three controls corresponding to each of the other three child genotypes that could be formed from the parental genotypes. The genotypes may now be viewed as predictors and the genotype relative risks estimated, and covariate effects may also be added to the model (Self et $al.$, 1991). Additional markers can be included as covariates, to distinguish the effects of multiple loci in linkage disequilibrium (Cordell and Clayton, 2002).

The regression approach also permits a natural model for parent of origin effects. By defining separate relative risk parameters for the maternal and paternal alleles, tests can be constructed to compare their effects. For example, if a test of equality between maternal and paternal relative risks showed a significant difference, this would provide evidence

of imprinting. If imprinting is present, say in the paternally inherited copy, then a test of the maternal relative risk alone would be more powerful than one of the marginal relative risk. Furthermore, the maternal genotype could be included as an additional covariate, allowing for maternal–foetal genotype interactions at the prenatal stage. Indeed, such an effect could confound a comparison of maternal and paternal risks, if not properly accounted for (Weinberg, 1999).

In parent of origin analyses, it is possible to increase the number of family-based controls by assuming symmetry of parental mating-type probabilities. In other words, the probability of the mother having genotype g_1 and the father having genotype g_2 is assumed to equal that of the mother having g_2 and the father having g_1. Under this assumption, further controls can be constructed by exchanging the genotypes of the parents. For example, suppose the mother has genotype 1/2 and the father 3/4, and the case has genotype 1/3. Then the family-based controls are 1/4, 2/3 and 2/4, and with the symmetry assumption we may also use 3/1, 4/1, 3/2 and 4/2 as controls, where the genotypes are ordered with the maternal allele first. This approach, termed *conditioning on exchangeable parental genotypes* (Cordell *et al.*, 2004) can lead to a substantial increase in effective sample size, compensating for the increased number of parameters that must be estimated in parent of origin models.

Provided the analysis is conditional on the parental genotypes, the TDT is robust to population stratification and will also be a test of linkage. However, it is useful to place the conditional model within an unconditional model, as this will be relevant for the discussion of missing data approaches. Assuming that case–parent trios are ascertained through the disease status of the child, the full likelihood of a trio is

$$L^{(f)} = \Pr(F, M, C|D) = \frac{\Pr(D|F, M, C) \cdot \Pr(F, M, C)}{\Pr(D)}, \tag{38.7}$$

where F, M, C are the genotypes of the father, mother and child, respectively, and D denotes that the child has disease. Assume that (1) conditional on the child genotype, the disease status of the child is independent of the parental genotypes; (2) there is Mendelian transmission, so all children are equally likely from the parents; then this simplifies to

$$L^{(f)} = \frac{\Pr(D|C) \cdot \Pr(F, M)}{\sum\limits_{f, m \in G} \Pr(f, m) \sum\limits_{c \in S(f, m)} \Pr(D|c)}, \tag{38.8}$$

where G is the set of all possible genotypes and $S(f, m)$ is the set of possible child genotypes for parents f and m. To relate this to the conditional likelihood used for the TDT, this can be written as

$$L^{(f)} = \frac{\Pr(D|C)}{\sum\limits_{c \in S(F, M)} \Pr(D|c)} \cdot \frac{\Pr(F, M) \sum\limits_{c \in S(F, M)} \Pr(D|c)}{\sum\limits_{f, m \in G} \Pr(f, m) \sum\limits_{c \in S(f, m)} \Pr(D|c)}$$

$$= L^{(c)} \cdot L^{(p)}. \tag{38.9}$$

In other words, the full likelihood is the product of a conditional likelihood and a parental contribution that gives the probability of observing the parents of an affected child. The parental likelihood includes a mating-type distribution $\Pr(F, M)$ that could

be mis-specified under population stratification. Inference on the conditional likelihood, using a log-odds model for $\Pr(D|C)$, gives the conditional logistic regression TDT (38.6).

38.4 HAPLOTYPE ANALYSIS

The use of haplotypes to identify disease loci has received much attention in recent times (Schaid, 2004) because disease mutations are assumed to have arisen on a single founder chromosome whose haplotype should be more strongly associated to disease than any single marker. Furthermore, haplotype analysis can be used to detect and characterize multiple interacting loci, when haplotypes are constructed from disease alleles themselves.

In principle, haplotypes can be analysed in the same way as any multiallelic marker. There is however a practical difficulty in that haplotypes are usually not directly observed, but must be deduced from genotype data; and then the haplotypes may be ambiguous. For example, a subject that is heterozygous at two loci, that is, with genotypes 1/2 and 1/2, can have two possible pairs of haplotypes: 1−1 and 2−2, or 1−2 and 2−1. When haplotypes are ambiguous, statistical methods can be used to model the possible solutions, for example by maximum likelihood (see **Chapter 36**) or by coalescent modelling (see **Chapter 25**; Stephens *et al.*, 2001). In family data, there are additional complications, including the fact that many haplotypes are deducible after all, and that introducing a model of the haplotype distribution may not retain robustness to population stratification.

When there are complete genotype data available in a family, the haplotypes can often be deduced with certainty. Ambiguity only occurs when there are markers at which both parents are heterozygous for the same alleles (a situation known as *intercross* in experimental genetics) and the child is also heterozygous. Here, we cannot say which parent transmitted which allele, although the transmitted and untransmitted alleles themselves are identifiable. Haplotype ambiguity then occurs if another marker exists at which more than one allele is present in the parents. Because ambiguous haplotypes may be relatively uncommon, particularly for multiallelic markers, it may be possible to perform a standard TDT using just the deduced haplotypes. However, such an approach is biased unless we allow for the fact that not all families could be used (Dudbridge

Figure 38.2 Family in which haplotypes can be deduced, but would not be deducible from other children of the same parents.
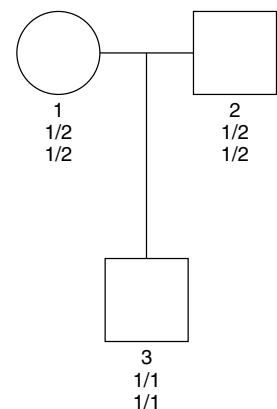
et al., 2000). Figure 38.2 shows a family in which the haplotypes can be deduced, so that we can score two transmissions of the 1−1 haplotype. However, for these parents, the haplotypes can only be deduced in this case and in the case where the 1−1 haplotype is not transmitted twice, so the expected transmission count is 1, with variance 1. We have seen that the usual TDT assumes an expected transmission count of 1 for two parents, with variance 1/2, so scoring just the certain haplotype transmissions would lead to an underestimate of the true variance and an inflated test statistic. Programs are available to adjust the TDT for this situation (Dudbridge *et al.*, 2000; Markianos *et al.*, 2001).

An alternative for analysing only the certain haplotypes is to fit a statistical model to the missing data, using maximum likelihood. A difficulty is that if we assume Hardy−Weinberg equilibrium, which is convenient for limiting the number of parameters, we may mis-specify the distribution of the uncertain haplotypes and lose robustness to population stratification. Clayton (1999) proposed a compromise solution that is implemented in the software TRANSMIT. The method uses a model for haplotype frequencies in a way that reduces its impact on inference on the relative risks.

Recall the factorization of the likelihood contribution from a case−parent trio

$$
L^{(f)} = \frac{\Pr(D|C)}{\sum\limits_{c \in S(F,M)} \Pr(D|c)} \cdot \frac{\Pr(F, M) \sum\limits_{c \in S(F,M)} \Pr(D|c)}{\sum\limits_{f,m \in G} \Pr(f, m) \sum\limits_{c \in S(f,m)} \Pr(D|c)}
$$
$$
= L^{(c)} \cdot L^{(p)}. \tag{38.10}
$$

Let $\boldsymbol{\beta}$ be the vector of haplotype log relative risks. Let $\boldsymbol{\gamma}$ parameterize the haplotype frequencies such that the frequency vector is $\dfrac{\exp(\boldsymbol{\gamma})}{\sum \exp(\gamma_i)}$. Thus, $\boldsymbol{\beta}$ specifies the model for $\Pr(D|C)$ and, assuming Hardy−Weinberg equilibrium, $\boldsymbol{\gamma}$ specifies the model for $\Pr(F, M)$. The parental likelihood depends on both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, whereas the conditional likelihood depends only on $\boldsymbol{\beta}$. The score function for the full likelihood is

$$
\mathbf{u}^{(f)} = \begin{pmatrix} \dfrac{\partial \log L^{(f)}}{\partial \boldsymbol{\beta}} \\[2mm] \dfrac{\partial \log L^{(f)}}{\partial \boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial \log L^{(c)} L^{(p)}}{\partial \boldsymbol{\beta}} \\[2mm] \dfrac{\partial \log L^{(p)}}{\partial \boldsymbol{\gamma}} \end{pmatrix}, \tag{38.11}
$$

which could be used to construct score tests for $\boldsymbol{\beta}$, but this would depend strongly on the model for the parental likelihood. Clayton (1999) proposed instead a *partial score function*

$$
\mathbf{u}^{(*)} = \begin{pmatrix} \dfrac{\partial \log L^{(c)}}{\partial \boldsymbol{\beta}} \\[2mm] \dfrac{\partial \log L^{(p)}}{\partial \boldsymbol{\gamma}} \end{pmatrix}, \tag{38.12}
$$

in which inference about $\boldsymbol{\beta}$ depends only on the conditional likelihood. When there are ambiguous haplotypes, the partial score function becomes a weighted mean over the possible solutions, where the weights are the corresponding full likelihoods. Denoting the

set of possible haplotype solutions by P, the partial score function can be written as

$$\mathbf{u}_P^{(*)} = \frac{\sum_{j \in P} L_{(j)}^{(f)} \mathbf{u}_{(j)}^{(*)}}{\sum_{j \in P} L_{(j)}^{(f)}}. \tag{38.13}$$

This can be used to construct a score test for $\boldsymbol{\beta}$, which depends on the haplotype frequency model only through the weights of the possible solutions. Indeed, when there are no uncertain data, the score test for $\boldsymbol{\beta}$ is the same as in the original TDT. All the extensions available in the logistic regression model can be cast in this framework, although the TRANSMIT program only implements tests of individual haplotypes and a global test of the whole set of haplotypes.

38.5 GENERAL PEDIGREE STRUCTURES

It is often the case that the available family data do not consist of only case–parent trios but include a variety of different pedigree structures. This can be a particular problem for late-onset traits, since parents of affecteds may be deceased at the time of diagnosis, and also for the testing of association in the presence of linkage, because the linkage structure needs to be accounted for in the analysis. A well-known problem affecting the TDT in an ostensibly simple situation arises in the case where one parent is missing (Curtis and Sham, 1995). Suppose the available parent is heterozygous and the child is homozygous. Then we might score the transmitted allele, but this overlooks the fact that, had the child been heterozygous, we would have been unable to identify the transmitted allele and so score the family. But, if we only score homozygous children, they are likely to be homozygous for the more common allele, which would lead to a seeming over-transmission of the common allele even if there were no linkage or association. This is one of many biases that can occur when data are available for some members of a family but not for others (Knapp, 2000).

Two general approaches have emerged to deal with the problems of missing family members. The first is to fit a statistical model to the missing data and conduct an analysis that takes all of the possible completions into account. Here, the approach used by TRANSMIT applies equally well to nuclear families with missing parents. The second approach is to construct test statistics that are unbiased under the null hypothesis while using only the available data. This approach retains complete robustness to population stratification and can be readily applied to arbitrary pedigree structures. Some methods of this type are described in this section.

As noted above, the case of missing parents in late-onset disease is of particular interest. In this case, unaffected siblings may be used as controls, as long as their relationship to affecteds is taken into account. One of the first such methods was the sib-TDT, which uses sibships consisting of at least one affected and one unaffected sib (Spielman and Ewens, 1998). It counts the number of variant alleles in affected sibs, calculating its mean and variance for each family under the assumption that the proportion of variant alleles is the same in affected sibs as it is in unaffecteds. These counts are summed over the set of families to form a z score. Several other sibling-based methods have been

suggested (Curtis, 1997; Boehnke and Langefeld, 1998; Horvath and Laird, 1998), though the sib-TDT remains most closely related to the more recent approaches.

The pedigree disequilibrium test (PDT) combines the principles of the TDT and sib-TDT into a test for general pedigrees (Martin *et al.*, 2000). It splits a pedigree into a list of all case–parent trios and discordant sib pairs with genotype data. For trio j, X_{T_j} is defined as the number of transmissions of the variant allele minus the number of its non-transmissions. For sib pair j, X_{S_j} is defined as the number of copies of the variant allele in the affected sib minus the number in the unaffected sib. A measure of association for the pedigree is then

$$D = \frac{1}{N_T + N_S} \left[\sum_{j=1}^{N_T} X_{T_j} + \sum_{j=1}^{N_S} X_{S_j} \right], \tag{38.14}$$

where N_T and N_S are the total number of trios and discordant sib pairs, respectively. D has expectation 0 in any pedigree. After computing this measure for each of $i = 1, \ldots, N$ pedigrees, the PDT statistic

$$T = \frac{\sum_{i=1}^{N} D_i}{\sum_{i=1}^{N} D_i^2} \tag{38.15}$$

is asymptotically χ^2 with 1 degree of freedom. This gives a valid test of linkage or association in any pedigree structure, although some pedigrees are uninformative, notably the affected sib pair. The PDT has been adapted to test quantitative traits (Monks and Kaplan, 2000), haplotypes (Dudbridge, 2003) and genotypes (Martin *et al.*, 2003a).

A very flexible approach for constructing unbiased tests is implemented in the software FBAT (Lake *et al.*, 2000). The general principle is that under no linkage or no association, the covariance between genotypes and traits is 0. When T_{ij} denotes the trait value of person j in pedigree i and X_{ij} denotes some coding of its genotype, a score is constructed as

$$S = \sum_i S_i = \sum_i \sum_j T_{ij} X_{ij}. \tag{38.16}$$

The mean and variance of each S_i are computed with respect to its conditional distribution given the minimal sufficient statistic for the null hypothesis (Rabinowitz and Laird, 2000). In practice, this means enumerating all possible founder genotypes and considering the transmission patterns that result in the same information structure as in the observed data: an example is given below. After summation over pedigrees, the standardised score statistic

$$T = (S - E(S))[\text{var}(S)]^{-1}(S - E(S)) \tag{38.17}$$

may be referred to the χ^2 distribution with degrees of freedom equal to the rank of var(S), which is 1 for a biallelic marker.

The conditional distribution of S_i is derived separately for each nuclear family configuration, which is tedious but not intractable. A recursive algorithm is available for constructing distributions for general pedigrees. Table 38.3, following Rabinowitz and

Table 38.3 Conditional distributions used by FBAT when testing for linkage with one heterozygous 1/2 parent's genotype available.

Set of genotypes present in children	Conditional distribution
{1/1} or {1/2}	Observed data have conditional probability 1
{1/1, 1/2}	Random assignment of 1/1 and 1/2 that keeps the number of each invariant
{1/1, 2/2} or {1/1, 1/2, 2/2}	Randomly assign 1/1, 1/2 and 2/2 with probabilities 1/4, 1/2, 1/4, independently to each sib, discarding outcome without at least one assignment of 1/1 and one assignment of 2/2

Laird (2000), shows the derivation when testing for linkage of a biallelic marker when one parent's genotype is available. As noted above, when there is only one child, there is no way to score the family without incurring bias. However, when there are several children, their genotypes define a set of configurations whose probabilities can be computed exactly, assuming the null hypothesis. Each configuration gives rise to a value of S_i that can be used to obtain the mean and variance of S_i for that family type.

For example, suppose there is one heterozygous parent with three children with genotypes *AA, AB* and *AA*, respectively. Then, row 2 of Table 38.3 is appropriate, and denoting the value of S_i by $S_i(G_1, G_2, G_3)$ when the ordered genotypes of the children are G_1, G_2 and G_3, the mean of S_i is

$$\frac{1}{3}[S_i(AA, AB, AA) + S_i(AB, AA, AA) + S_i(AA, AA, AB)] \tag{38.18}$$

and its variance is

$$\frac{2}{9}[(S_i(AA, AB, AA))^2 + (S_i(AB, AA, AA))^2 + (S_i(AA, AA, AB))^2$$
$$- S_i(AA, AB, AA)S_i(AB, AA, AA) - S_i(AA, AB, AA)S_i(AA, AA, AB)$$
$$- S_i(AB, AA, AA)S_i(AA, AA, AB)]. \tag{38.19}$$

As described, the FBAT method is appropriate for any trait T_{ij}. For a disease, it is appropriate to code T_{ij} as 1 for affected and 0 for unaffected. Similarly X_{ij} may be any coding of genotypes: for biallelic markers, it is convenient to code X_{ij} as the number of variant alleles carried by the child, which corresponds to the multiplicative model favoured by the TDT. Dominant or recessive models may be coded using 1 for risk genotypes and 0 otherwise. Multivariate calculations are possible for multiallelic markers.

The FBAT approach is more flexible than the PDT, as it can use information from concordant sibships and also distinguish tests of linkage from tests of association. It has been subject to numerous extensions, including analysis of haplotypes (Horvath *et al.*, 2004) and multivariate traits (Lange *et al.*, 2003). Its strengths are the guaranteed protection from population stratification and applicability to any pedigree structure. Nevertheless, the approach has some limitations, including difficulty in estimating relative risk and sub-optimal handling of additional covariates. When these issues are important,

the likelihood-based approaches described in Sections 38.3 and 38.7 may be more appropriate.

38.6   QUANTITATIVE TRAITS

So far, the emphasis has been on association to binary traits, but the attractions of the family-based design apply equally to quantitative traits. In the quantitative setting, population stratification is taken to mean a difference in the trait mean across sub-populations, together with a difference in gene frequency. It is usually assumed that the parametric form of the trait distribution is otherwise unchanged, such as a normal, and also that the variance and higher moments are also unchanged: therefore the effect of stratification can be thought of as a shift in location of the trait distribution.

Two types of regression models have been developed for family-based association of quantitative traits. The first uses logistic regression in a similar way as for discrete traits, treating transmission as the random outcome. The second uses linear regression with the genotype as the independent variable and the quantitative trait as the dependent variable, with adjustment to respect the family-based design. Generally, the methods based on linear regression are more sensitive to the assumption of normality, and may require more nuisance parameters than the methods using logistic regression; but they can be more flexible and powerful when their assumptions are met.

The logistic regression approach treats transmission as the random outcome, as for discrete traits, and treats the quantitative trait as an effect modifier for alleles acting as predictors (Waldman et $al.$, 1999). For a biallelic marker, the transmission probability of the variant allele from a heterozygous parent is given by

$$\text{logit}(p) = \alpha + Y_i \beta, \tag{38.20}$$

where Y_i is the trait value of child i, regression coefficient β is the transmission parameter of the variant allele, and the intercept α is included to account for association to a phenotype for which all the children have been selected. In an unselected sample, the intercept could be omitted.

For multiallelic markers, a conditional logistic regression model can be defined in a manner similar to (38.6). For a j/k parent transmitting allele j to child i,

$$\text{logit}(p_{jk}) = \alpha_j - \alpha_k + Y_i(\beta_j - \beta_k). \tag{38.21}$$

Models of this form are most sensitive to log-linear effects of the trait value on the transmission probability, and this is appropriate for small effects on normally distributed traits (Clayton and Jones, 1999). In this case, $Y\beta$ is the relative trait mean for a subject with trait Y carrying the variant allele, compared to a reference allele. If the traits are not zero-centred, it is prudent to include the intercept term α to implicitly subtract the mean from each trait value. This model can be extended to genotype association by using a polytomous logistic regression for multinomial outcomes (Kistner and Weinberg, 2004). These models can be fitted by standard software.

Linear regression is a natural framework for modelling association to quantitative traits. Here, the genotypes are treated as categorical variables that predict the expected trait value,

and we test whether the mean is different for subjects having different genotypes. This is reminiscent of the one-way analysis of variance, but the main issue is to allow for different trait means in different population strata. Potentially, each family is in its own stratum, so the analysis must allow for the mean to differ between families.

A reasonable approach is to conduct ordinary linear regression with the genotypes normalised within families (Lunetta *et al.*, 2000),

$$Y_i = \alpha + \beta(G_i - \psi_i) + e_i, \tag{38.22}$$

where now α is the overall mean of Y, G_i is a code for the genotype of child i, ψ_i is the expected value of G_i given the mating type of the parents and an assumed genetic model, and e_i is the error. For example, for a biallelic marker, we might let G denote the number of copies of the variant allele. If one parent is homozygous for the variant allele ($G = 2$) and the other is heterozygous ($G = 1$), then we may define $\psi_i = \frac{3}{2}$ for that family. If both parents are heterozygous, we may have $\psi_i = \frac{1}{4} \cdot 0 + \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 = 1$. This coding is equivalent to a model proposed by Rabinowitz (1997), but the framework allows for other codings that are more sensitive to dominant or recessive effects (Lunetta *et al.*, 2000).

A more general model, implemented in the QTDT software, decomposes the total population association into two orthogonal components: the between-family association, and the within-family association (Abecasis *et al.*, 2000). The advantage of this model is that the within-family association is robust to population stratification, but in the absence of stratification, the same model can give a more powerful test of the total association. Furthermore, a difference between the between- and within-family association parameters can be taken as evidence of stratification. The method also allows for covariance between the traits of multiple siblings, in terms of variance components, which is not considered by the other models. Since the covariance between siblings depends upon linkage, the QTDT model allows linkage and association to be modelled and tested separately, in contrast to most other family-based designs (Cardon and Abecasis, 2000).

The model for the trait mean of child j in family i is

$$Y_{ij} = \alpha + \beta_b B_i + \beta_w(G_{ij} - B_i) + e_{ij}, \tag{38.23}$$

where β_b and β_w are between- and within-family coefficients for association, G_{ij} codes for the genotype of child j in family i, and B_i is an expected value of B_{ij} for family i. To allow for multiple children, B_i is defined as

$$B_i = \tfrac{1}{2}(G_{iF} + G_{iM}), \tag{38.24}$$

where G_{iF} and G_{iM} code for the father's and mother's genotypes, respectively, or when parents are not available

$$B_i = \frac{1}{n_i} \sum_j G_{ij}, \tag{38.25}$$

where n_i is the number of children in family i. The likelihood for a nuclear family is the multivariate normal density with the mean vector obtained by this model and the variance–covariance matrix constructed from variance components (see **Chapter 19**).

Likelihood ratios are used to test for within-family association ($\beta_w = 0$), total association ($\beta_w = 0$ and $\beta_b = 0$, with 2 degrees of freedom) and population stratification ($\beta_w = \beta_b$). Furthermore, β_w is a valid estimate of the additive genetic value of the test marker (Fulker *et al.*, 1999; Abecasis *et al.*, 2000).

This model can be regarded as a regression on the within-family component, with the intercept treated as a random effect that is modelled by the between-family component. An alternative approach would be to treat the intercept as a fixed effect that depends upon the parental mating type (Gauderman, 2003), so that the expected trait value within a family is specified directly:

$$Y_{ij} = \alpha_i + \beta_w G_{ij} + e_{ij}, \tag{38.26}$$

where α_i are set to be equal for families having the same mating type. This approach has been shown to have improved power for detecting gene–gene and gene–environment interactions, although it requires more nuisance parameters than the model (38.23): this can be a serious problem for multiallelic markers or haplotypes.

A potential problem with the linear regression models is that they rely on normality of the residual distribution and, in the case of the full likelihood implemented in the QTDT software, on multivariate normality of the trait distribution. These assumptions may be violated if the underlying distribution is non-normal, or if the sample has been selected, for example, to have extreme trait values. A solution to this problem is to adopt a retrospective formulation in which the linear model is used to predict the probability of genotype given the trait, via Bayes' theorem (Liu *et al.*, 2002). With some care, this approach can be implemented in standard software (Gauderman, 2003) but it is worth noting that, for small effects on normal traits, the logistic regression models (38.21) are approximately equivalent to the retrospective likelihood approach.

38.7   ASSOCIATION IN THE PRESENCE OF LINKAGE

It was earlier noted that the TDT and related methods can be regarded either as tests of linkage in the presence of association or tests of association in the presence of linkage. Often, the two interpretations are interchangeable: we are seeking evidence of a genetic effect, realised through linkage and association. In some contexts, however, the distinction is important. A situation that has received much attention occurs in positional cloning, when a genome scan for linkage identifies a broad genomic region and association methods are then used to fine map a gene by exploiting the short-range extent of linkage disequilibrium (see **Chapter 27**). In this case, family-based association methods must allow for the fact that linkage has been established for the tested markers, so that the test is strictly one of association.

To illustrate why this is necessary, consider a nuclear family with two children, typed for a marker that is completely linked to the disease gene (Figure 38.3). If there is no association, we have no information on which marker allele occurs together with the disease allele, given that the parent is heterozygous. But if there is complete linkage, the same marker allele will be transmitted to both children: in other words, the transmission to the second child is not independent of the first. Therefore, a standard TDT treating
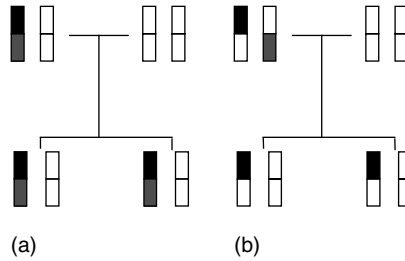
Figure 38.3 Two families that are equally likely under no association, given that the parent is heterozygous for the marker. Under complete linkage, only 2 or 0 transmissions of the variant allele can occur.

all transmissions as independent will inflate the evidence for association and increase the false-positive rate.

This problem only arises when there are multiple offspring in a family and the tested marker is linked. When all families have just one offspring, there is no need to allow for linkage. Most of the previously described methods are susceptible to the problem, except for the QTDT program, which explicitly models the correlation between children's traits in terms of the evidence for linkage (Abecasis *et al.*, 2000).

A common approach to allow for dependent observations is to use a variance estimate that treats the whole family as one sampling unit. When using the logistic regression models, this can be achieved via score or Wald tests of the odds ratios, counting all children independently but using a 'cluster' variance estimate instead of the analytic form (Gould *et al.*, 2006, chapter 1). The same approach is available in the TRANSMIT, PDT and FBAT programs. In general terms, the method involves calculating a score S_i and its expectation for each family, as if the children were independent. The variance is then estimated from the family-wise scores rather than from individual contributions, using the *empirical variance* estimator

$$\mathrm{var}(S) = \sum_{i=1}^{N} [S_i - E(S_i)]^{\mathrm{T}} [S_i - E(S_i)]. \tag{38.27}$$

Because this method is based on treating families as sampling units, the resulting score tests are unbiased.

A limitation of the empirical variance approach is that it does not easily extend to maximum likelihood estimation of missing data and uncertain haplotypes. Essentially, this is because the underlying likelihood still assumes that transmissions are independent, and the adjustment for correlated transmissions only occurs at the testing stage. While the TRANSMIT software implements an empirical variance estimate to allow for correlated transmission, it has been shown to be biased when parents are missing (Martin *et al.*, 2003b). These authors suggested a method to estimate the degree of linkage and thence adjust a score test; however, this method is difficult to implement in nuclear families with more than two children.

A likelihood-based solution to the problem of correlated transmission is based on the factorization of the nuclear family likelihood into conditional and parental contributions (38.9). Allowing for k affected children, the full likelihood for a nuclear family

assuming independent transmissions is

$$L^{(f)} = \frac{\Pr(D_1, \ldots, D_k | C_1 \ldots, C_k)}{\sum\limits_{c_1, \ldots, c_k \in S(F, M)} \Pr(D_1, \ldots, D_k | c_1, \ldots, c_k)}$$

$$\cdot \frac{\Pr(F, M) \sum\limits_{c_1, \ldots, c_k \in S(F, M)} \Pr(D_1, \ldots, D_k | c_1, \ldots, c_k)}{\sum\limits_{f,m \in G} \Pr(f, m) \sum\limits_{c_1, \ldots, c_k \in S(f,m)} \Pr(D_1, \ldots, D_k | c_1, \ldots, c_k)}$$

$$= L^{(c)} \cdot L^{(p)}. \tag{38.28}$$

Inferences should be based on the conditional likelihood in order to retain the properties of the TDT, but the parental likelihood is needed to model the possible completions of missing data. In order to distinguish the two components, separate sets of relative risk parameters are specified in each component, with inference being drawn only on those in the conditional component. Let $\boldsymbol{\beta}$ be the vector of haplotype log relative risks in the conditional component, and $\boldsymbol{\varphi}$ those in the parental component. Let $\boldsymbol{\gamma}$ parameterize the haplotype frequencies as in (38.11), and let (g_1, g_2) be the two haplotypes comprising a genotype g. Assuming a multiplicative model for the joint risk of the children, we have

$$L^{(c)} = \frac{\prod\limits_{i=1}^{k} \exp(\beta_{C_{i1}} + \beta_{C_{i2}})}{\left[\sum\limits_{c \in S(F,M)} \exp(\beta_{c_1} + \beta_{c_2}) \right]^k} \tag{38.29}$$

and

$$L^{(p)} = \frac{\exp(\gamma_{F_1} + \gamma_{F_2} + \gamma_{M_1} + \gamma_{M_2}) \left[\sum\limits_{c \in S(F,M)} \exp(\varphi_{c_1} + \varphi_{c_2}) \right]^k}{\sum\limits_{f,m \in G} \exp(\gamma_{f_1} + \gamma_{f_2} + \gamma_{m_1} + \gamma_{m_2}) \left[\sum\limits_{c \in S(f,m)} \exp(\varphi_{c_1} + \varphi_{c_2}) \right]^k} \tag{38.30}$$

performing inference only on $\boldsymbol{\beta}$. At this point, the children are still assumed to be independent. In order to relax this assumption, we then introduce the additional step of conditioning on an identity-by-descent (IBD) vector. This vector contains one entry for each transmission to each child, indicating whether the phase of that transmission is the same as that of the paternal transmission to the first child. The conditional likelihood can be written down without actually computing the IBD vector, and has the form

$$L^{(c|\text{IBD})} = \frac{\exp\left(\sum\limits_{i=1}^{k} \beta_{C_{i1}} + \beta_{C_{i2}} \right)}{\exp\left(\sum\limits_{i=1}^{k} \beta_{C_{i1}} + \beta_{C_{i2}} \right) + \exp\left(\sum\limits_{i=1}^{k} \beta_{C_{i1}} + \beta_{U_{i2}} \right)}{} \tag{38.31}$$
$$+ \exp\left(\sum\limits_{i=1}^{k} \beta_{U_{i1}} + \beta_{C_{i2}} \right) + \exp\left(\sum\limits_{i=1}^{k} \beta_{U_{i1}} + \beta_{U_{i2}} \right)$$

where U_i denotes the genotype formed from the two alleles not transmitted to child i. The parental contribution becomes

$$L^{(p|\text{IBD})} = \frac{\exp(\gamma_{F_1} + \gamma_{F_2} + \gamma_{M_1} + \gamma_{M_2})}{\displaystyle\sum_{f,\,m \in G} \exp(\gamma_{f_1} + \gamma_{f_2} + \gamma_{m_1} + \gamma_{m_2}) \left[\sum_{c \in S(f,m)} \exp(\varphi_{c_1} + \varphi_{c_2}) \right]^k}$$

$$\cdot \left[\exp\left(\sum_{i=1}^{k} \varphi_{C_{i1}} + \varphi_{C_{i2}} \right) + \exp\left(\sum_{i=1}^{k} \varphi_{C_{i1}} + \varphi_{U_{i2}} \right) \right.$$

$$\left. + \exp\left(\sum_{i=1}^{k} \varphi_{U_{i1}} + \varphi_{C_{i2}} \right) + \exp\left(\sum_{i=1}^{k} \varphi_{U_{i1}} + \varphi_{U_{i2}} \right) \right] \tag{38.32}$$

and the full likelihood $L^{(f|\text{IBD})} = L^{(c|\text{IBD})} \cdot L^{(p|\text{IBD})}$ is then maximised over the full set of parameters $(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\gamma})$. This formulation has the effect of treating the whole family as one sampling unit, resulting in unbiased inference about $\boldsymbol{\beta}$.

When the data are complete, the total likelihood factorises completely into conditional and parental components, yielding inference on $\boldsymbol{\beta}$ that is equivalent to the conditional logistic regression formulation of the TDT. When there are incomplete data, the likelihood contribution of a family is the sum of the likelihoods for the possible completions. Then the parental likelihood has the effect of weighting the conditional analysis, in an analogous manner to the TRANSMIT method. In fact, when $\boldsymbol{\varphi}$ is set to $\mathbf{0}$, the score function for the full likelihood is very similar to the partial score function used by TRANSMIT, with a slight difference in the weights used when the data are incomplete. The advantage of freely estimating $\boldsymbol{\varphi}$ is that the IBD vector can be identified in the parental contribution, which is not possible when $\boldsymbol{\varphi} = \mathbf{0}$, thus eliminating the bias seen in TRANSMIT.

The model in (38.32) has been implemented in the software UNPHASED (Dudbridge, 2006). The likelihood formulation also allows estimation of odds ratios and testing of covariate effects, and extends readily to quantitative traits by treating the traits as effect modifiers as in (38.21).

It might be argued that, despite considerable efforts to devise valid tests of association in the presence of linkage, the problem is not particularly important. The reason is that, in many studies of complex disease, the evidence for linkage is fairly weak, resulting in only moderate dependence between transmissions to multiple children. Furthermore, given the weak evidence from traditional linkage analysis, it may be more worthwhile to seek stronger evidence for linkage, via association methods. Nevertheless, in situations in which the linkage evidence is particularly strong, for example, in some HLA-linked diseases, it is useful to have methods that are specifically designed to detect association.

38.8 CONCLUSIONS

Family-based methods are a useful and important means to protect against confounding effects in association studies. All the methods presented here compare, by various designs, the alleles transmitted to the study subjects to the alleles not transmitted by

their parents. This genetic matching eliminates the possibility of population stratification, which is a concern in genetic studies because allele frequencies vary considerably between populations, owing to several factors including random drift (see **Chapter 31**). Furthermore, the shared environment within families ensures that other confounders, possibly not identified, are automatically controlled for. Together with the fact that family-based association can only be detected in the presence of linkage, these methods provide a benchmark that continues to be widely used in candidate-gene studies, and will serve as an important method for replication as genome-wide scans for population association become more widespread.

From the original TDT method (Spielman *et al.*, 1993), the field has grown to encompass a wide range of applications, of which only the most prominent have been covered here. To a large extent, the field relies on customised software because many of the methods do not employ standard statistical models and data structures. The TRANSMIT (Clayton, 1999) and FBAT (Lake *et al.*, 2000) programs are widely used for hypothesis testing, and the QTDT program (Abecasis *et al.*, 2000) is widely used for testing and estimation of both linkage and association to quantitative traits. UNPHASED is a serious alternative for analysis of general nuclear families, providing estimation as well as hypothesis testing and allowing gene–gene and gene–environment interactions (Dudbridge, 2006). A further alternative is FAMHAP, which is faster but slightly less accurate (Becker and Knapp, 2004). Despite the proliferation of custom software, family-based analysis can be cast into standard models for conditional logistic regression, by regarding each nuclear family as a stratum with control subjects constructed from the other possible combinations of parental alleles. Provided these controls are constructed appropriately, a wide variety of analyses are possible using standard software (Cordell *et al.*, 2004).

A number of topics have not been covered here, and are the subject of ongoing research. These include the combination of family-based and population-based studies (Epstein *et al.*, 2005), and multiple imputation approaches for dealing with missing data (Kistner and Weinberg, 2005). Recent methods for extracting maximal information from incomplete data have attracted much interest (Rabinowitz, 2002; Allen *et al.*, 2005) because they are robust to violations of simplifying assumptions such as Hardy–Weinberg equilibrium. Currently, their computational complexity is the greatest obstacle to more widespread use. Bayesian approaches have not been discussed here, but there is some work in this area (George and Laud, 2002; Bernardinelli *et al.*, 2004; Denham and Whittaker, 2003). These methods may assume greater importance as investigators aim to incorporate external information on gene function and interaction into epidemiological studies.

REFERENCES

Abecasis, G.R., Cardon, L.R. and Cookson, W.O. (2000). A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics* **66**, 279–292.

Allen, A.S., Satten, G.A. and Tsiatis, A.A. (2005). Locally-efficient robust estimation of haplotype-disease association in family-based studies. *Biometrika* **92**, 559–571.

Becker, T. and Knapp, M. (2004). Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genetic Epidemiology* **27**, 21–32.

Bernardinelli, L., Berzuini, C., Seaman, S. and Holmans, P. (2004). Bayesian trio models for association in the presence of genotyping errors. *Genetic Epidemiology* **26**, 70–80.

Boehnke, M. and Langefeld, C.D. (1998). Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *American Journal of Human Genetics* **62**, 950–961.

Cardon, L.R. and Abecasis, G.R. (2000). Some properties of a variance components model for fine-mapping quantitative trait loci. *Behavior Genetics* **30**, 235–243.

Cardon, L.R. and Palmer, L.J. (2003). Population stratification and spurious allelic association. *Lancet* **15**, 598–604.

Clayton, D. (1999). A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *American Journal of Human Genetics* **65**, 1170–1177.

Clayton, D. and Jones, H. (1999). Transmission/disequilibrium tests for extended marker haplotypes. *American Journal of Human Genetics* **65**, 1161–1169.

Cordell, H.J., Barratt, B.J. and Clayton, D.G. (2004). Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genetic Epidemiology* **26**, 167–185.

Cordell, H.J. and Clayton, D.G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *American Journal of Human Genetics* **70**, 124–141.

Curtis, D. (1997). Use of siblings as controls in case-control association studies. *Annals of Human Genetics* **61**, 319–333.

Curtis, D. and Sham, P.C. (1995). A note on the application of the transmission disequilibrium test when a parent is missing. *American Journal of Human Genetics* **56**, 811–812.

Denham, M.C. and Whittaker, J.C. (2003). A Bayesian approach to disease gene location using allelic association. *Biostatistics* **4**, 399–409.

Dudbridge, F. (2003). Pedigree disequilibrium tests for multilocus haplotypes. *Genetic Epidemiology* **25**, 115–121.

Dudbridge, F. (2006). UNPHASED user manual. Technical report 2006/05, MRC Biostatistics Unit, Cambridge.

Dudbridge, F., Koeleman, B.P., Todd, J.A. and Clayton, D.G. (2000). Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *American Journal of Human Genetics* **66**, 2009–2012.

Epstein, M.P., Veal, C.D., Trembath, R.C., Barker, J.N., Li, C. and Satten, G.A. (2005). Genetic association analysis using data from triads and unrelated subjects. *American Journal of Human Genetics* **76**, 592–608.

Ewens, W.J. and Spielman, R.S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. *American Journal of Human Genetics* **57**, 455–464.

Falk, C.T. and Rubinstein, P. (1987). Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics* **51**, 227–233.

Fulker, D.W., Cherny, S.S., Sham, P.C. and Hewitt, J.K. (1999). Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics* **64**, 259–267.

Gauderman, W.J. (2003). Candidate gene association analysis for a quantitative trait, using parent-offspring trios. *Genetic Epidemiology* **25**, 327–338.

George, V. and Laud, P.W. (2002). A Bayesian approach to the transmission/disequilibrium test for binary traits. *Genetic Epidemiology* **22**, 41–51.

Gould, W., Pitblado, J. and Sribney, W. (2006). *Maximum Likelihood Estimation with Stata*, 3rd edition. Stata Press, College Station, TX.

Horvath, S., Xu, X., Lake, S.L., Silverman, E.K., Weiss, S.T. and Laird, N.M. (2004). Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genetic Epidemiology* **26**, 61–69.

Horvath, S. and Laird, N.M. (1998). A discordant-sibship test for disequilibrium and linkage: no need for parental data. *American Journal of Human Genetics* **63**, 1886–1897.

Kistner, E.O. and Weinberg, C.R. (2004). Method for using complete and incomplete trios to identify genes related to a quantitative trait. *Genetic Epidemiology* **27**, 33–42.

Kistner, E.O. and Weinberg, C.R. (2005). A method for identifying genes related to a quantitative trait, incorporating multiple siblings and missing parents. *Genetic Epidemiology* **29**, 155–165.

Knapp, M. (2000). The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. *American Journal of Human Genetics* **64**, 861–870.

Lake, S.L., Blacker, D. and Laird, N.M. (2000). Family-based tests of association in the presence of linkage. *American Journal of Human Genetics* **67**, 1515–1525.

Lange, C., Silverman, E.K., Xu, X., Weiss, S.T. and Laird, N.M. (2003). A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* **4**, 195–206.

Liu, Y., Tritchler, D. and Bull, S.B. (2002). A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits. *Genetic Epidemiology* **22**, 26–40.

Lunetta, K.L., Faraone, S.V., Biederman, J. and Laird, N.M. (2000). Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *American Journal of Human Genetics* **66**, 605–614.

Markianos, K., Daly, M.J. and Kruglyak, L. (2001). Efficient multipoint linkage analysis through reduction of inheritance space. *American Journal of Human Genetics* **68**, 963–977.

Martin, E.R., Bass, M.P., Gilbert, J.R., Pericak-Vance, M.A. and Hauser, E.R. (2003a). Genotype-based association test for general pedigrees: the genotype-PDT. *Genetic Epidemiology* **25**, 203–213.

Martin, E.R., Bass, M.P., Hauser, E.R. and Kaplan, N.L. (2003b). Accounting for linkage in family-based tests of association with missing parental genotypes. *American Journal of Human Genetics* **73**, 1016–1026.

Martin, E.R., Monks, S.A., Warren, L.L. and Kaplan, N.L. (2000). A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *American Journal of Human Genetics* **67**, 146–154.

Monks, S.A. and Kaplan, N.L. (2000). Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. *American Journal of Human Genetics* **66**, 576–592.

Ott, J. (1989). Statistical properties of the haplotype relative risk. *Genetic Epidemiology* **6**, 127–130.

Rabinowitz, D. (1997). A transmission disequilibrium test for quantitative trait loci. *Human Heredity* **47**, 342–350.

Rabinowitz, D. (2002). Adjusting for population heterogeneity and misspecified haplotype frequencies when testing nonparametric null hypotheses in statistical genetics. *Journal of the American Statistical Association* **97**, 742–751.

Rabinowitz, D. and Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity* **50**, 211–223.

Schaid, D.J. (2004). Evaluating associations of haplotypes with traits. *Genetic Epidemiology* **27**, 348–364.

Schaid, D.J. and Sommer, S.S. (1994). Comparison of statistics for candidate-gene association studies using cases and parents. *American Journal of Human Genetics* **55**, 402–409.

Self, S.G., Longton, G., Kopecky, K.J. and Liang, K.Y. (1991). On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics* **47**, 53–61.

Sham, P.C. and Curtis, D. (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Annals of Human Genetics* **59**, 323–336.

Spielman, R.S. and Ewens, W.J. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics* **62**, 450–458.

Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**, 506–516.

Stephens, M., Smith, N.J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978–989.

Terwilliger, J.D. and Ott, J. (1992). A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Human Heredity* **42**, 337–346.

Waldman, I.D., Robinson, B.F. and Rowe, D.C. (1999). A logistic regression based extension of the TDT for continuous and categorical traits. *Annals of Human Genetics* **63**, 329–340.

Weinberg, C.R. (1999). Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *American Journal of Human Genetics* **65**, 229–235.

Whittaker, J.C., Denham, M.C. and Morris, A.P. (2000). The problems of using the transmission/disequilibrium test to infer tight linkage. *American Journal of Human Genetics* **67**, 523–526.

39

Cancer Genetics

M.D. Teare

Mathematical Modelling and Genetic Epidemiology, University of Sheffield Medical School, Sheffield, UK

Cancers result from an accumulation of inherited and somatic mutations yielding cells that have acquired the necessary characteristics for unregulated growth. The development of the tumour can be viewed as an evolutionary process, involving several classes of genes in tumour initiation and progression.

This chapter provides an overview of the development of cancer genetic models. The earlier mathematical and statistical models based on population and cancer family observations, are followed by evolutionary models applied to molecular genetic sequence data.

39.1   INTRODUCTION

Cancer is a genetic disease, in that a normal cell must undergo mutations to lose or gain functions allowing a tumour to develop. The earliest evidence indicating a genetic origin for cancer was reported in 1890. Abnormal configurations of chromosomes were observed in dividing cancer cells (von Hansemann, 1890), though at that time the link between chromosomes and inheritance was not known. This was followed by Theodore Bovari's experiments on the fertilisation of sea urchin eggs, strongly suggesting that individual chromosomes contained different information (Bovari, 1904). Bovari had developed procedures by which he could induce aberrant chromosome segregation during mitosis. He was then able to study the fate of these cells after mitosis. In most cases, the unequal distribution of chromosomes would lead to a detrimental effect on the cell. However, he noted that on rare occasions, some particular configurations of chromosomes would generate a cell with the ability of unlimited growth, and this effect could be passed on to its progeny. This suggested that tumours might arise through abnormal segregation of chromosomes to daughter cells (Bovari, 1914). He then went on to explore and develop this hypothesis, which led him to predict many aspects of genetic mechanisms in cancer development.

39.2 ARMITAGE–DOLL MODELS OF CARCINOGENESIS

39.2.1 The Multistage Model

In the early 1950s two hypotheses, regarding mechanisms for carcinogenesis, were put forward based on the studies of cancer mortality. Fisher and Holloman (1951) and Nordling (1953) studied the mortality rates for several tumour types and found that the logarithm of the death rate due to cancer increased in direct proportion to the logarithm of the age at death. More specifically they found that the death rate increased proportionately to the sixth power of the age at death. This relationship seemed to be limited to mortality rates between the ages of 25 and 75. It could be argued that data above age 75 were unreliable and cancer in children and young adults may be affected by other factors.

Fisher and Holloman proposed that this observed relationship between age and mortality could result if a colony of six or seven cancer cells was the critical mass necessary for independent growth to be sustained. Experimental data on the relationship between the dose of carcinogen and cancer incidence already exists and in general, the relationship between the two was more consistent with arithmetic rather than geometric data. (Berenblum and Shubik, 1949). Based on this apparent contradiction, Armitage and Doll (1954) felt that this made the Fisher and Holloman hypothesis 'untenable'. Nordling proposed an alternative hypothesis that the observed relationship could be explained if a single cancer cell was the result of seven successive and accumulated mutations. This model would explain the relationship if the probability of each mutation remains constant throughout the specified age range.

The development of the tumour can be represented as the following changes occurring in the cell (or its lineal descendants).

$$\text{E}_0 \xrightarrow{\;p_1\;} \text{E}_1 \xrightarrow{\;p_2\;} \text{E}_2 \cdots \xrightarrow{\;p_n\;} \text{E}_n.$$
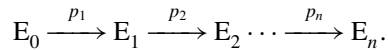
E_0 represents the normal cell and E_n the malignant tumour cell. The p_i represent the rate of change or transition to the next stage. The reciprocal of these rates represents the average sojourn time, spent in years, in state or stage E_i. We denote as t, the time that the cell enters state E_n.

$$p_1 p_2 p_3 p_4 p_5 p_6 p_7 t^7. \tag{39.1}$$

The probability that a mutation of type i has occurred by time t is approximately given by $p_i t$, when each p_i is small. Thus if seven distinct mutations were required, the probability that all mutations have occurred by time t is as given above in (39.1). However, the model requires that the mutations occur in a specific order, and there are $n!$ of these. Therefore, the probability that a cell is in state E_n, at time t, is given by $p_1 p_2 \cdots p_n t^n / n!$

The probability of malignancy by time t can be thought of as equivalent to prevalence. The incidence is given by the derivative of the prevalence with respect to t to yield,

$$\frac{p_1 p_2 \cdots p_n t^{n-1}}{(n-1)!}. \tag{39.2}$$

Thus the double logarithmic plot should show a straight line with slope $(n-1)$. The incidence rate for an individual (as opposed to a cell) is given by (39.2) multiplied by

N_s, which is the average number of cells in the susceptible target tissue. The above result will be approximately true when the mutation probabilities (i.e. $p_i t$) are small.

This simple model has assumed that the specific mutations can be modelled by constant rates. Armitage and Doll (1954) went on to derive expressions allowing for variable mutation rates to illustrate what relationships might be expected in situations where variable exposure to mutagen influenced the mutation rate. Now assume that one, p_c say, varies with time. The probability that this cth change occurs in the time interval $(t_0, t_0 + \mathrm{d}t_0)$, and the nth in interval $(t, t + \delta t)$ is given by the product of the following three probabilities:

1. The probability that $(c - 1)$ changes have occurred in the interval $(0, t_0)$ is given as derived above by, $p_1 \cdots p_{c-1} t_0^{c-1}/(c - 1)!$

2. The probability (conditional that exactly $c - 1$ changes occurred before t_0) that the cth change occurs in the small interval $(t_0, t_0 + \mathrm{d}t_0)$, is given by $p_c(t_0)\mathrm{d}t_0$, where $p_c(t_0)$ represents the mutation rate at time t_0.

3. The probability (conditional on the cth change before t_0) that the $(c + 1)$th, $(c + 2)$th, to nth changes occur in time (t_0, t) is given by

$$\frac{p_{c+1}p_{c+2} \cdots p_{n-1}(t - t_0)^{n-c-1}p_n}{(n - c - 1)!}\mathrm{d}t.$$

This results in the expression for the total probability for the nth event occurring in the time interval $(t, t + \mathrm{d}t)$ as follows:

$$\frac{p_1 \cdots p_{c-1}p_c(t)p_{c+1} \cdots p_{n-1}p_n t^{n-1}}{(n - 1)!}\mathrm{d}t.$$

The $p_c(t)$ is the weighted mean of $p_c(t_0)$ over the whole time period, with the weight at time t_0 being proportional to $t_0^{c-1}(t - t_0)^{n-c-1}$, (which takes its maximum at $t_0 = t(c - 1)/(n - 2)$). Armitage and Doll argue that this slight modification to their constant rate model is enough to explain the deviations from the sixth power law seen in some tumours. Since the weighted mean of the varying mutation rate will depend upon the age, t, the overall incidence will not show a linear relationship with the sixth power of t.

The Nordling model would also explain that the experimentally observed data, i.e. rate of tumour incidence is directly proportional to the concentration of an effective carcinogen, if the effect of the carcinogen is to increase the mutation rate. Armitage and Doll (1954) attempted to test the hypothesis of Nordling by examining cancer mortality for cancer at a variety of different sites and for each sex. In particular they were interested to see if the hypothesis could explain data where the significance of carcinogenic factors was suspected to be variable. They examined the relationship between mortality rate and age for cancers at the commonest sites. These relationships are shown for a selection of common cancers but for more recent time periods (CancerStats, Cancer Research UK) in Figure 39.1. As trends in mortality can be influenced by improvements through treatment, incidence data only is presented. The log (incidence) for cancer of the colon and lung shows an approximately similar linear relationship with log (age). However, the log–log plot for female breast cancer deviates quite strongly from a linear relationship.

By inspection of the log–log plots they found that cancers fell into two main groups. Those that appeared consistent with proportional relationships, (in their study these were
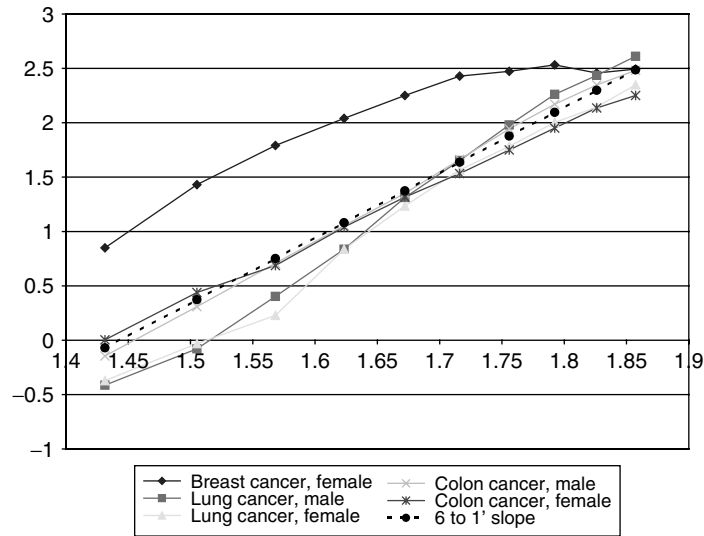
Figure 39.1 log(age) – log(incidence) plots for a selection of common site specific cancers.

cancers in the oesophagus, stomach, pancreas, colon and rectum) and those that did not. In the group that appeared to show linear relationships their regression coefficient for the slope of the line ranged from 5.0 to 6.5. These cancers show a similar relation to those reported by Fisher and Holloman and by Nordling. However, the other group sometimes showed a strong deviation from a linear relationship. These were cancers of the lung, bladder, female breast, ovary, cervix and uterus. Armitage and Doll pointed out that this latter group consists of cancers where the effects of causal factors are known to be variable. Tumours believed to be influenced by hormonal or endocrine secretions, might be expected to show evidence of age-specific effects. Incidence of bladder cancer was strongly associated with occupational hazards which may have both cohort and age-specific effects. A significant proportion of lung cancer was suspected to be due to the increasing rates of cigarette smoking and therefore exposure to this hazard was also not a constant with respect to time. Looking at the observed data in this way gave more support to the original hypothesis of Nordling (1953).

This multistage model appeared to fit the data well and could account for deviations from a constant mutation rate by allowing a variable rate affected by variations in exposure to carcinogens or mutagens. However, at the time, experimental data did not demonstrate evidence for more than two stages in tumour initiation, i.e. an exposure making an impact at an early versus late stage. As a consequence of this argument and suggestions made by Platt (1955), Armitage and Doll went on to offer a two-stage model for carcinogenesis.

39.2.2   The Two-stage Model

In this reduced model (Armitage and Doll, 1957), the first or early stage results in a change that gives one cell (and its descendants) a growth advantage over the adjacent normal cells. The clinical cancer then results from the second or late event occurring in this rapidly dividing 'clone' of cells.

This model assumes that the rates of both the first and second stage are influenced by carcinogenic exposures or agents. Assuming that the exposure is summarised by a 'dose' effect parameter d, the number of mutated cells will be proportional to the initiating dose and will also grow at an exponential rate depending upon the relative growth advantage of the mutated cells over the normal state. So the number of cells at time t will be n_t, which is proportional to $d_1 e^{kt}$, where k is a constant, and d_1 is the dose of the agent influencing the first stage. The resulting cancer incidence will be proportional to the dose effect influencing the second event (d_2) and the number of susceptible cells, i.e. $d_1 d_2 e^{kt}$.

Although individuals are likely to be exposed to the inducing agents throughout life, provided the dose remains constant the resulting incidence will be proportional to both concentrations. This model gives rise to a similar relationship between age and incidence as the multistage model.

This parsimonious two-stage model is able to account for all the features observed in cancer incidence plots. It explains the long latent period often observed after initial exposure to a carcinogen and, when incidence is low, a linear relationship can be seen between the concentration of initiating carcinogen and incidence by age. This two-stage model was also appealing as it reflected the biological process of a cell accumulating two mutated copies of the same gene or the reduction to homozygosity (which is discussed below).

39.2.2.1 The Philadelphia (Ph) Chromosome

Developments in cytogenetic techniques meant that by 1956 it was possible to unambiguously count the number of chromosomes in a normal cell (Tijo and Levan, 1956; Ford and Hamerton, 1956). With further technical improvements it became possible to distinguish and classify chromosome groups facilitating comparisons of karyotype in different cell types. This led to a major milestone in cancer genetics, the discovery of the *Ph* chromosome seen in cells from patients with chronic myelogenous leukaemia (Nowell and Hungerford, 1960). This was first reported as a shortened chromosome in the G chromosome group (this consisted of chromosomes 21 and 22; at the time there was no way to distinguish them). For the first time a specific chromosomal change was associated with a specific tumour type. Many years later it was found that the *Ph* chromosome was in fact a translocation between the long arm of chromosome 9 and a large part of chromosome 22 (Rowley, 1973). Work continued studying this translocation and it could be demonstrated that the translocation brought two genes together resulting in the production of a chimeric protein (Heisterkamp *et al.*, 1983) which stimulates a growth signal pathway. This work confirmed that the genetic translocation was a critical event in the origin and cause of this cancer.

Over the 1950s and 1960s evidence appeared to accumulate that cancer could arise as a result of few mutations, though this argument was blurred by the fact that the early and late stage events could themselves be the result of several 'changes'. Work on childhood tumours seemed to confirm that a combination of inherited and somatic mutations played an essential role in tumour development (Burch, 1962; Falls and Neel, 1951; Crowe *et al.*, 1956).

This research culminated in the key publication by Knudson (1971), where he performed a statistical analysis of retinoblastoma cases and proposed the 'two-mutation' hypothesis for cancer development. Knudson observed that familial cases and sporadic (non-hereditary) cases could arise from the same genetic mechanism. Experimental evidence was accumulating for genes that could suppress tumours (Harris *et al.*, 1969) and Knudson's hypothesis specifically relating to retinoblastoma would lead to the discovery of the first tumour suppressor gene (TSG).

Retinoblastoma is a tumour of the retina and can occur in one (unilateral) or both (bilateral) eyes. Knudson studied the reports of all the cases of retinoblastoma admitted to the MD Anderson Hospital between 1944 and 1969, consisting of 23 bilateral and 25 unilateral cases. In 14 of the bilateral cases, it was possible to estimate the number of distinct tumours in one eye. Knudson used other sources of data to derive population based estimates of the distribution of tumour types among gene carriers (unaffected, unilateral, and bilateral) of cases. The bilateral cases were assumed to be inherited, though the inherited lesion was not assumed to be fully penetrant as bilateral cases do occur in unaffected parents. Assuming that tumours develop according to a Poisson Process, Knudson found that the observed distribution of tumours, including the numbers in each eye, was consistent with a Poisson with rate ($m =$)3.

If n is the total number of target cells (i.e. those cells susceptible to mutation) contained by the two eyes (or retinae), then m/n is the probability that an (inherited) mutant cell develops into a primary tumour. Using prior estimates of n, yields an estimated tumour development rate (in the inherited form) of 0.75×10^{-6}. Assuming that this rate is the same in both the sporadic and hereditary form, this represents the probability of the second event.

Knudson then illustrated that the probability of the first mutation in sporadic tumours is approximately the same as the probability of a new mutation occurring in the germ line, which then passed on to future generations. The germinal mutation rate, μ_g, is approximately equal to 5×10^{-6} per generation. Assuming a generation time of 25 years, this means the sporadic mutation rate (expressed per year) is approximately 2×10^{-7}.

Conditional on the assumption that bilateral cases have all inherited the first mutation, if only one second step is involved in retinoblastoma development, the distribution of bilateral cases should follow an exponential. Therefore, the fraction of total bilateral cases that occurs in a given interval, the hazard rate, should be constant and the proportion of survivors (in the cohort of germinal or hereditary mutation carriers) is given by $S_h = \exp(-k_h t)$, where k_h is a constant. For the non-hereditary form where a single cell must acquire two mutations the proportion of survivors $S_n = \exp(-k_n t^2)$, k_n is a constant representing the non-hereditary mutation rate. Knudson found that the observed distribution of the ages at onset was consistent with his derived models, though the mutation rates looked slightly different in the bilateral and unilateral forms, which he argued could be accounted for by a small proportion of the unilateral cases being hereditary. Though other models could have accounted for such differences, this model suggested that the age distribution for familial and nonfamilial cases was resulting from the accumulation of the **same** genetic changes. Knudson went on to study the two-mutation model with application to other childhood cancers, but his theory was not confirmed through molecular studies until 1983, (see section on mutations resulting in loss of function).

39.2.2.2 Mutations Resulting in Gain of Function

In the 1960s it was known that cells in culture could be transformed by a number of viruses and retroviruses. Cellular genes were found to have similar sequences to those in the transforming retroviruses. These 'normal' genes were involved in regulating growth and differentiation, the inappropriate activation of which could lead to carcinogenesis. This class of genes was termed *proto-oncogenes*. When inappropriately activated they were termed *oncogenes*. Many oncogenes were identified through transfection experiments.

The *Ph* chromosome is an example of a specific balanced translocation giving rise to an activated oncogene (ABL–BCR fusion gene). Cytogenetic studies confirmed that chromosomal breakpoints seen in common translocations in other cancers were near to known proto-oncogenes (Rabbitts, 1994). Oncogenes are not only activated through translocation, some gain function through chromosomal amplification.

39.2.2.3 Mutations Resulting in Loss of Function

Gene-transfer studies seemed to present strong evidence for a single mutational step in carcinogenesis, but this is an artifact of the experimental design. However, evidence from observational studies such as Knudson's was more consistent with a two-stage process. In order for the inherited mutated allele to not have deleterious consequences on the organism, it was likely that this mutation was recessive at the cellular level. Comings (1973) further postulated that the two mutations might affect the same locus, resulting in loss of that gene function.

In 1983, through biochemical and molecular studies Cavenee *et al.* (1983) found that retinoblastoma development required loss of both copies of a specific region of chromosome 13. Thus the Rb1 locus was the first cloned example of a TSG, where, in contrast to the oncogene, loss of function is required for carcinogenesis. This result confirmed Knudson's two hit hypothesis, in that the familial and sporadic form of retinoblastoma were in fact due to the same genetic mechanism.

The success of the Rb story led to two major strategies for identifying further cancer predisposing genes, familial cancer linkage studies, and 'loss of heterozygosity' (LOH) studies. LOH studies stemmed from comparing genotypes observed in constitutional and tumour DNA in cancer patients. It was hypothesised that sections of chromosomes that are lost in tumours are likely to contain TSGs. If a genetic marker was located in a chromosomal region where one copy of a chromosome had been lost, the genotype would appear as homozygous as only the 'non-lost' allele would be detected. This observation would only be informative if the constitutional genotype was heterozygous and hence 'LOH'. There was great enthusiasm for studying familial forms of cancer as these could be expected to lead more quickly to the discovery of the target/predisposing genes. There were also many studies concentrating on patterns of allelic loss in tumours from panels of unrelated individuals. LOH profiles were found to be generally consistent within tumour types, but this approach alone led to few TSG discoveries (Presneau *et al.*, 2003).

However, genetic linkage studies (see **Chapter 33**) did lead to the discovery of many TSGs, though it became clear that single tumour suppressors may account for rare syndromes or a small fraction of the familial effect (Garber and Offit, 2005). In adult onset of disease, it can be very difficult to collect sufficient samples for informative linkage analysis. Often the pedigree comes to the attention of the investigators when

many of the affected individuals are dead. Techniques were developed so that archived tumours could be used to reconstruct the constitutional genotypes. Through this strategy it became apparent that the observations of somatic change in the tumours could be used in a similar way to further informative meioses. This approach was built into a formal framework (Teare *et al.*, 1994; Rohde *et al.*, 1995; 1997), and derives from the linkage approach.

The tumour suppressor model assumes the high cancer risk in selected families is due to a dominantly inherited defective or predisposing allele. However, at the cellular level the defective gene acts recessively, as it requires the wild-type allele to be inactivated or lost. If a genetic marker is in linkage with the disease locus then it is expected to see over-segregation of one marker allele in affected individuals. By a simple extension to this, if the predisposing locus behaves as a tumour suppressor and the wild-type function is lost through loss of chromosomal material during mitosis, then the marker allele in phase with the wild type would also have a tendency to be lost in the tumour, and the marker allele in phase with the disease allele would have a tendency to be retained in the tumour. Besides the parameters used in classic linkage analysis this model requires parameters representing the probability of 'loss of genetic material' and somatic recombination. Features of the model are illustrated in Figure 39.2.

In the most general form we require three probabilities ($\lambda_{DD}, \lambda_{Dd}, \lambda_{dd}$) representing the probability of losing one marker allele in tumour, conditional on the genotype at the disease locus, and three similar probabilities ($\varepsilon_{DD}, \varepsilon_{Dd}, \varepsilon_{dd}$) representing the probability of losing both alleles at the marker locus conditional on the disease locus genotype. In the case where only one allele is lost the parameter ρ represents the probability that the marker allele lost in the tumour is in phase with the predisposing disease allele in the germ line. This parameter can be thought of as similar to θ, the probability of meiotic recombination (Table 39.1).
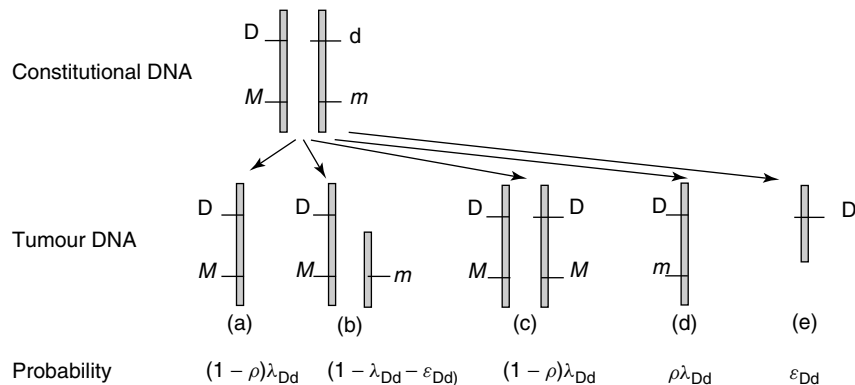
Figure 39.2 Examples of genotype configurations in tumour DNA, when marker is located on the same chromosome as TSG locus. Dd represents genotype at TSG locus, D = defective, d = wildtype; Mm represents heterozygous marker genotype. (a) Hemizygous, (b) retention of both alleles, (c) loss of wildtype chromosome followed by duplication, (d) somatic recombination and loss of chromosome strand bearing wildtype allele, (e) loss of both marker alleles. The model assumes that the wildtype TSG allele is lost.

Table 39.1 Probabilities of loss or retention of marker alleles in the tumour conditional on marker and TSG genotype.

Marker genotype	Constitutional TSG genotype					
	DD		Dd		dd	
MM	*MM* $1 - \lambda_{DD} - \varepsilon_{DD}$ $-M$ $1/2\lambda_{DD}$	*M−* $1/2\lambda_{DD}$ $-$ ε_{DD}	*MM* $1 - \lambda_{Dd} - \varepsilon_{Dd}$ $-M$ $\rho\lambda_{Dd}$	*M−* $(1-\rho)\lambda_{Dd}$ $-$ ε_{Dd}	*MM* $1 - \lambda_{dd} - \varepsilon_{dd}$ $-M$ $1/2\lambda_{dd}$	*M−* $1/2\lambda_{dd}$ $-$ ε_{dd}
Mm	*Mm* $1 - \lambda_{DD} - \varepsilon_{DD}$ $-m$ $1/2\lambda_{DD}$	*M−* $1/2\lambda_{DD}$ $-$ ε_{DD}	*Mm* $1 - \lambda_{Dd} - \varepsilon_{Dd}$ $-m$ $\rho\lambda_{Dd}$	*M−* $(1-\rho)\lambda_{Dd}$ $-$ ε_{Dd}	*MM* $1 - \lambda_{dd} - \varepsilon_{dd}$ $-M$ $1/2\lambda_{dd}$	*M−* $1/2\lambda_{dd}$ $-$ ε_{dd}
mm	*mm* $1 - \lambda_{DD} - \varepsilon_{DD}$ $-m$ $1/2\lambda_{DD}$	*m−* $1/2\lambda_{DD}$ $-$ ε_{DD}	*mm* $1 - \lambda_{Dd} - \varepsilon_{Dd}$ $-m$ $\rho\lambda_{Dd}$	*m−* $(1-\rho)\lambda_{Dd}$ $-$ ε_{Dd}	*mm* $1 - \lambda_{dd} - \varepsilon_{dd}$ $-m$ $1/2\lambda_{dd}$	*m−* $1/2\lambda_{dd}$ $-$ ε_{dd}

The cells of the table display the four possible genotypes in the tumour DNA, and the conditional probability of each genotype.

The original studies of LOH were so-called because it was commonly not feasible to detect loss of one out of two identical alleles; only loss of one out of two different alleles (i.e. reduction to homozygosity) could be detected. Loss of both marker alleles has been reported but we would expect it to be rare when the marker is close to a gene where at least one functional copy is essential. The probabilities in the table relate to phase-known genotypes but as in any linkage study, these can only be inferred from the segregation seen in the family. It is, therefore, apparent that the only informative situation for studying loss of alleles is when the constitutional genotype at the marker is heterozygous.

The benefits of including such observations in classic linkage analysis are mainly to increase power. If the cancer syndrome leads to individuals suffering multiple primary tumours, each tumour can be included as an independent observation. Affected parent–offspring pairs also became informative provided one tumour of the pair was observed (Rohde *et al.*, 1997). The tumour observations are considered in an extension to the penetrance function and the evidence for linkage is summarised in the same manner as classic linkage analysis with the lod score function (see **Chapter 33**) maximised over both θ and ρ. This method has so far had limited application as it is difficult to gain access to sufficient quantities of good quality tumour DNA. With current technological developments, especially with respect to whole genome amplification this approach may still prove to have a wide application. The method can also be extended to allow for other forms of loss of function such as epigenetic silencing (see **Chapter 40**).

39.2.2.4 *Models Attempting to Identify the order of Acquired Mutations and hence Tumour Origin*

As implied above, it is possible to develop cancer at more than one site (this is common in those predisposed to cancer development). However, when tumours are synchronously (clinically) detected these can be two independently arising tumours, or one primary and an associated metastatic tumour. If the tumours are of independent origin you would

expect to see different patterns of mutations in them (even if the mutation targets are the same). However if one tumour is a 'clonal outcrop' related to an original primary, these tumours would be expected to share more mutation patterns. Establishing the true primary tumour may have important treatment implications.

An extension of the LOH linkage model was developed to consider the evidence that two concurrently diagnosed tumours (a and b) shared the same origin. In this setting the objective is not to map the predisposing locus but to assess how similar the LOH pattern is between two tumours from the same individual, and to infer which of the two cancers is the original. The model is formulated as shown in Table 39.2, a contribution to the likelihood is made for each observed (usually microsatellite) marker. The loss parameter index refers to c = common origin, a = occurs in a during metastasis, and b = occurs in b during metastasis. This model was applied to a data set consisting of 62 patients diagnosed with concurrent endometrial and ovarian cancer (Brinkman *et al.*, 2004).

In the familial cancer setting statistical model development has focused on extending the mixed model (Lalouel *et al.*, 1983; Antoniou and Easton, 2006). Such models can allow for effects of known risk genes through mutation screening and linkage data, whilst fitting models to the residual genetic component.

39.2.2.5   A General Theory of Carcinogenesis

Comings (1973) brought together many of the varied theories of carcinogenesis into a single model proposing that cancer arises when a cell has accumulated sufficient mutations to bypass the normal constraints on growth. By studying disease progression in Chronic Myelogenous Leukaemia patients Nowell (1976) observed that additional chromosomal alterations (besides the *Ph* Chromosome) accumulated with disease progression. These observations led him to propose the clonal evolution hypothesis for carcinogenesis. Under this model a cell suffers a first mutation, conferring a growth advantage. This generates a colony of genetically unstable daughter cells, with further mutations accumulating in successive generations. Thus more malignant subclones would evolve driven by selection through growth advantage.

Major support for this hypothesis came from molecular genetic studies on colorectal cancer by Vogelstein and co-workers (reviewed by Vogelstein and Kinzler, 1993). This form of cancer is recognised to develop through several clear stages of malignancy, from pre-malignant lesion, through to metastasis, and offers a means to study the accumulation of genetic changes directly by taking samples of early and late stage lesions. A number of key somatic changes are now well established, namely, the early event of LOH of chromosome 5p (the location of the APC gene), and the late event of TP53 mutation.

Table 39.2   Components of the likelihood for each compound allelic state.

Probability	At each examined locus, for a pair of tumours a and b
$(1 - \alpha)$	Not informative
$\alpha(1 - \lambda c)(1 - \lambda a)(1 - \lambda b)$	Informative, no allele loss detected
$\alpha(1 - \lambda c)\lambda a(1 - \lambda b)$	Informative, allele loss in a, not in b
$\alpha(1 - \lambda c)(1 - \lambda a)\lambda b$	Informative, allele loss in b, not in a
$\alpha\{\lambda c + (1 - \lambda c)\lambda a \lambda b[(1 - \rho)2 + \rho\ 2]\}$	Informative, loss of same allele
$\alpha[2(1 - \lambda c)\lambda a \lambda b(1 - \rho)\rho]$	Informative loss of different alleles

Recently, Luebeck and Moolgavkar (2002) used a multistage epidemiological model to fit the age-specific incidence of colorectal cancer in the US SEER registry (Surveillance, Epidemiology and End Results). Assuming a model of two-stage clonal expansion (TSCE) they used a maximum likelihood approach to estimate the number of mutations necessary in the pre-initiated state before the cell acquires the capacity for increased clonal expansion. They found evidence that the US age-specific incidence data was consistent with two rare events for the initiation stage, followed by a high frequency event resulting in rapid clonal expansion. One further rare event was required for the adenoma to progress to carcinoma.

As can be seen from the example plots in Figure 39.1, the risk of most common cancers increases with age. However, Peto and Mack (2000) reported evidence of high constant risk of breast cancer in twins and relatives of breast cancer patients. By plotting the incidence of breast cancer in relatives from the time that the family index case was diagnosed, their data appeared to suggest that the risk was constant, with respect to time and age of the relative at diagnosis. They showed a similar effect in twins and in studies of cancer incidence in the contra lateral breast. They proposed a 'molecular clock' model, such that breast cancers arise in a subset of susceptible women who, from a predetermined inherited age-point, are at high constant risk of breast cancer. This pattern was not observed in other familial cancers, and a later examination of a larger data set did not confirm the hypothesis, presenting stronger evidence that familial breast cancer risk did vary with age (Hemminki and Granström, 2002).

39.2.2.6 *Mathematical Modelling of Tumour Initiation and Progression*

It is now generally accepted that cancer arises as the result of mutations in three distinct classes of cancer susceptibility genes – gatekeepers, caretakers and landscapers (Vogelstein and Kinzler, 2002). Oncogenes and tumour suppressors belong to the gatekeeper class. Caretakers include DNA repair genes whose function is to ensure genome integrity. Mutations in caretaker genes lead to genetic instability allowing the cell to accumulate further mutations more rapidly. Landscapers do not themselves lead to abnormal cellular growth, but may provide the means for the tumour to resist the immune system and develop further stromal support systems. Mathematical models examining the abstract population dynamics of the evolving tumour have been developed by Nowak and colleagues (reviewed in Michor *et al.*, 2004).

Multicellular organisms are made up of a variety of cell types. Populations of specific cell types form '*compartments*' and these compartments of cells develop to perform or maintain organ specific functions. Homeostatic mechanisms exist to ensure that the total cell number within a compartment remains constant over time. Therefore, a balance between cell birth and cell death must be maintained. Tumourigenesis will follow if the cell birth rate exceeds the death rate.

Michor *et al.* (2004) have explored the population dynamics of cancer progression by using a Moran Process. This stochastic process imposes a constant population size. At time zero, all cells in the compartment are unmutated. The oncogene model requires the cell to acquire only one mutation to alter its fitness. At each time step a cell is randomly selected (proportional to its fitness) for duplication, the daughter cell replaces another randomly selected cell in the compartment. They demonstrate that large compartments (where number of cells is large) accelerate the accumulation of advantageous (for the

cell) mutations, but slow down the accumulation of deleterious mutations. The converse is true for small compartments. This model thus predicts that the size of the compartment may be influential in determining the type of mutations that are observed.

When modelling the TSG element two mutations (each with distinct mutation rate μ_1 and μ_2) are necessary for the cell's fitness to be increased. In addition they assume that $\mu_1 > \mu_2$. They show that the probability that a cell arises with two mutated copies by time t, is again influenced by compartment size, denoted as N. When $N < 1/\sqrt{\mu_2}$, a cell with one mutated copy reaches fixation before a cell with two copies arises. Hence it takes two rate-limiting steps to inactivate a TSG in a small compartment. In moderately sized compartments where, $1/\sqrt{\mu_2} < N < 1/\sqrt{\mu_1}$, a cell with two mutated copies will emerge before the first mutation has reached fixation. In large compartments, where $N > 1/\sqrt{\mu_1}$, cells with one mutated copy will arise immediately and the dynamics is dominated by the waiting time for the second mutational step. Thus as the compartment size increases the TSG function is inactivated by two, one or zero rate-limiting steps.

Michor *et al.* (2004) went on to explore the dynamical effects of chromosomal instability (CIN) and TSG inactivation. The CIN event has an associated cost or fitness and the cell with CIN has an increased rate of LOH. In a small compartment, it takes two rate-limiting steps for a cell to accumulate two mutations in the TSG with or without CIN present. For a wide range of parameter values one or very few neutral CIN genes in the genome suffice to ensure that CIN initiates tumour formation in a pathway where a single TSG must lose function. One or several 'costly' CIN genes may initiate tumour formation in situations where functional loss of two TSGs is necessary. These simple mathematical models offer insightful explanations for some of the similarities and differences observed between cancers.

39.2.2.7 Mutational Analyses of Tumour DNA

The development of a cancer involves somatic mutation and selection, and population genetic approaches introduced in **Chapter 22** can be applied. Phylogenetic relationships (see **Chapter 16**) can be used in studies of many tumours for the accumulation of specific mutations within a single gene. Yang *et al.* (2003) formulated a likelihood based model to analyse the full range of mutations found in a TP53 database. Their approach specifically allows the mutation or codon substitution rate to vary depending upon both tumour type and functional domain. TP53 consists of six distinct functional domains Roemer (1999). Mutations are generally classified into missense (resulting in amino acid change), nonsense (non translation or truncation of protein) or silent (no amino acid change).

Parameters exist to represent relative substitution rates; (1) nonsense vs silent (2) missense vs silent and (3) transition vs transversion. These substitution rates can be decomposed into domain specific and dinucleotide specific rates, yielding likelihood ratio tests comparing nested hypotheses. Through applying this method to a TP53 database (consisting of over 15 000 tumour samples), they found strong evidence that the mutation rates were domain dependent. They also found evidence that the transition vs transversion bias varied across different tumour types.

The work by Yang *et al.* (2003), considered only the mutations detected in TP53 in many cancer types. However, the cancer cell is the direct result of the accumulation of mutations at many loci. In terms of the cancer modelling, it is important to be able to distinguish between target genes that are to be mutated (i.e. the mutation is pathogenic) and

passenger mutations. Greenman *et al.* (2006), extend the likelihood approach of Yang *et al.* (2003) also deriving test statistics and a means to estimate the parameters incorporated into their model. The main emphasis of this parameterisation is to distinguish between those mutations which have driven the cell towards the tumour state and mutations which can be classed as hitch-hikers or passengers. They do this by explicitly describing the selection process separately from the mutation process. The set of silent mutations is used to estimate the mutation rates under the null (i.e. no association between cancer development and mutation). This method is in essence similar to epidemiological studies constructed to evaluate risk factors, whilst controlling for confounding factors. The strata in this setting takes account of mutation types (essentially six different types), as these may be dependent upon the cell and its environment.

As they are considering many genes, they classify non-silent mutations into three classes, missense and nonsense, as in Yang *et al.*, but have a further class for non-silent mutations at splice sites. Their approach assumes that a sample of cancers have been examined or screened for mutations within a specified and common set of cancer genes (or a screened genome). They are then able to test for evidence of a drive towards increased selection for specific non-silent mutation types. They apply this model to a series of 518 genes sequenced in 25 breast tumours (Stephens *et al.*, 2005). This set contained a total of 91 base substitutions in 71 genes. They found strong evidence of selection on rates for nonsense and splice-site mutations, and were able to estimate that 29.8 of these base substitutions were pathogenic.

39.2.2.8 *Future Directions*

The statistical methods presented here incorporate only a subset of cancer mutation observations. Methods need to be developed that can handle the full range of genetic alterations. In tumours these include epigenetic effects and DNA sequence level mutations ranging from a single base change to loss or duplication of whole chromosomes. Mathematical models suggest that accounting for the evolutionary dynamics and the size and structure of the tumour target tissue may assist in successfully distinguishing between hitch-hiker and functional change mutations.

ELECTRONIC RESOURCES

CancerStats, Cancer Research UK, http://info.cancerresearchuk.org/cancerstats/

REFERENCES

Antoniou, A.C. and Easton, D.F. (2006). Models of genetic susceptibility to breast cancer. *Oncogene* **25**, 5898–5905.

Armitage, P. and Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* **8**, 1–12.

Armitage, P. and Doll, R. (1957). A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *British Journal of Cancer* **11**, 161–169.

Berenblum, I. and Shubik, P. (1949). An experimental study of the initiating stage of carcinogenesis, and a re-examination of the somatic cell mutation theory of cancer. *British Journal of Cancer* **3**, 109.

Bovari, T. (1904). *Ergebnisse Über die Konstitution der Chromatischen Substanz des Zelkerns*. Gustav Fischer, Jena.

Bovari, T. (1914). *Zur Frage der Entstehung Maligner Tumoren*. Gustav Fischer, Jena, pp. 1–64.

Brinkman, D., Ryan, A., Ayhan, A., McCluggage, W.G., Feakins, R., Santibanez-Koref, M.F., Mein, C.A., Gayther, S.A. and Jacobs, I.J. (2004). A molecular genetic and statistical approach for the diagnosis of dual-site cancers. *Jounal of National Cancer Institute* **96**, 1441–1446.

Burch, P.R.J. (1962). A biological principle and its converse: some implications for carcinogenesis. *Nature* **195**, 241–243.

Cavenee, W.K., Dryja, T.P., Phillips, R.A., Benedict, W.F., Godbout, R., Gallie, B.L., Murphree, A.L., Strong, L.C. and White, R.L. (1983). Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature* **305**, 779–781.

Comings, D. (1973). A general theory of carcinogenesis. *Proceedings of the National Cancer Institute* **70**, 3324–3328.

Crowe, F.W., Schull, W.J. and Neel, J.V. (1956). *A Clinical, Pathological and Genetic Study of Multiple Neurofibromatosis*. Charles C. Thomas, Springfield, Ill.

Falls, H.F. and Neel, J.V. (1951). Genetics of retinoblastoma. *AMA Archives Of Ophthalmology* **46**, 367–389.

Fisher, J.C. and Holloman, J.H. (1951). A hypothesis for the origin of cancer foci. *Cancer* **4**, 916–918.

Ford, C.E. and Hamerton, J.L. (1956). The chromosomes of man. *Nature* **178**, 1020–1023.

Garber, J.E. and Offit, K. (2005). Hereditary cancer predisposition syndromes. *Journal of Clinical Oncology* **23**, 276–292.

Greenman, C., Wooster, R., Futreal, P.A., Stratton, M.R. and Easton, D.F. (2006). Statistical analysis of pathogenocity of somatic mutations in cancer. *Genetics* **173**, 2187–2198.

von Hansemann, D. (1890). Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung. *Virchows Archiv fur Pathologische Anatomie* **119**, 299.

Harris, H., Miller, O.J., Klein, G., Worst, P. and Tachibana, T. (1969). Suppression of malignancy by cell fusion. *Nature* **223**, 363–368.

Heisterkamp, N., Stephenson, J.R., Groffen, J., Hansen, P.F., de Klein, A., Bartram, C.R. and Grosveld, G. (1983). Localisation of the c-ab1 oncogene adjacent to a translocation break point in chronic myelocytic leukemia. *Nature* **306**, 239–242.

Hemminki, K. and Granström, C. (2002). Risk for familial breast cancer increases with age. *Nature Genetics* **32**, 233.

Knudson, A.G. (1971). Mutation and cancer: a statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* **68**, 820–823.

Lalouel, J.M., Rao, D.C., Morton, N.E. and Elston, R.C. (1983). A unified model for complex segregation analysis. *American Journal of Human Genetics* **35**, 816–826.

Luebeck, E.G. and Moolgavkar, S.H. (2002). Multistage carcinogenesis and the incidence of colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 15095–15100.

Michor, F., Iwasa, Y. and Nowak, M.A. (2004). Dynamics of cancer progression. *Nature Reviews Cancer* **4**, 197–205.

Nordling, C.O. (1953). A new theory of the cancer inducing mechanism. *British Journal of Cancer* **7**, 68–72.

Nowell, P.C. (1976). The clonal evolution of tumour cell populations. *Science* **194**, 23–28.

Nowell, P.C. and Hungerford, D. (1960). A minute chromosome in human granulocytic leukemia. *Science* **132**, 1497.

Peto, J. and Mack, T.M. (2000). High constant incidence in twins and other relatives of women with breast cancer. *Nature Genetics* **26**, 411–414.

Platt, R. (1955). Clonal aging and cancer. *Lancet* **265**, 867.

Presneau, N., Manderson, E.N. and Tonin, P.N. (2003). The quest for a tumor suppressor gene phenotype. *Current Molecular Medicine* **3**, 605–629.

Rabbitts, T.H. (1994). Chromosomal translocations in human cancer. *Nature* **372**, 143–149.

Roemer, K. (1999). Mutant p53: gain-of – function oncoproteins and wild-type p53 activators. *Biological Chemistry* **380**, 879–887.

Rohde, K., Teare, M.D. and Santibanez Koref, M.S. (1997). Analysis of genetic linkage and somatic loss of heterozygosity in affected pairs of first-degree relatives. *American Journal of Human Genetics* **61**, 418–422.

Rohde, K., Teare, M.D., Scherneck, S. and Santibanez Koref, M.S. (1995). A program using constitutional loss of heterozygosity data to ascertain the location of predisposing genes in cancer families. *Human Heredity* **45**, 337–345.

Rowley, J.D. (1973). A new consistent chromosomal abnormality in chronic myelogenous leukemia. *Nature* **243**, 290–293.

Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., Hunter, C., Bignell, G., Teague, J., Smith, R., Stevens, C., O'Meara, S., Parker, A., Tarpey, P., Avis, T., Barthorpe, A., Brackenbury, L., Buck, G., Butler, A., Clements, J., Cole, J., Dicks, E., Edwards, K., Forbes, S., Gorton, M., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jones, D., Kosmidou, V., Laman, R., Lugg, R., Menzies, A., Perry, J., Petty, R., Raine, K., Shepherd, R., Small, A., Solomon, H., Stephens, Y., Tofts, C., Varian, J., Webb, A., West, S., Widaa, S., Yates, A., Brasseur, F., Cooper, C.S., Flanagan, A.M., Green, A., Knowles, M., Leung, S.Y., Looijenga, L.H., Malkowicz, B., Pierotti, M.A., Teh, B., Yuen, S.T., Nicholson, A.G., Lakhani, S., Easton, D.F., Weber, B.L., Stratton, M.R., Futreal, P.A. and Wooster, R. (2005). A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nature Genetics* **37**, 590–592.

Teare, M.D., Rohde, K. and Santibanez Koref, M.S. (1994). The use of loss of constitutional heterozygosity data to ascertain the location of cancer predisposing genes in cancer families. *Journal of Medical Genetics* **31**, 449–452.

Tijo, H.J. and Levan, A. (1956). The chromosome numbers of man. *Hereditas* **42**, 1–6.

Vogelstein, B. and Kinzler, K.W. (1993). The multi-step nature of cancer. *Trends in Genetics* **9**, 138–141.

Vogelstein, B. and Kinzler, K.W. (2002). *The Genetic Basis of Human Cancer*. 2nd edition. McGraw-Hill.

Yang, Z., Ro, S. and Rannala, B. (2003). Likelihood models of somatic mutation and codon substitution in cancer genes. *Genetics* **165**, 695–705.

40

Epigenetics

K.D. Siegmund

Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

and

S. Lin

Department of Statistics, Ohio State University, Columbus, OH, USA

In recent years, epigenetic changes have been implicated to be associated with a number of complex human diseases. In particular, evidence is mounting that aberrant DNA methylation in the CpG islands of gene promoters is linked to cancer. These discoveries are in part propelled by high-throughput technologies, allowing one to interrogate specific CpG sites or profiling methylation patterns of the entire genome. This chapter provides a review of statistical treatments to several problems in epigenetics based on DNA-methylation data. The detail of methylation characterization varies from modeling DNA-methylation patterns in individual cells to high-throughput methylation profiling in human tissue. The challenge on how to integrate '-omics'-scale data, both genetic and epigenetic, is discussed.

40.1 A BRIEF INTRODUCTION

In almost every cell of a human's body the DNA content and nucleotide sequence are nearly identical. However, in order to specialize in function, each cell type expresses a characteristic subset of genes. For example, an epithelial cell in the colon expresses a different subset of genes from a ductal epithelial cell in the breast. Epigenetics, derived from Greek to mean 'upon' genetics, refers to the transmission of information regarding expression of genes to daughter cells at cell division. By contrast, genetic information is transmitted by the nucleotide sequence of DNA. Mechanisms conveying epigenetic information in humans are not fully understood, but are known to involve the interrelated processes of DNA methylation, histone modification and chromatin structure. Although it is the combination of these factors and others that results in gene expression or

silencing, DNA methylation, a hallmark of epigenetic information, is the focus of this chapter.

Mammalian DNA methylation occurs when a methyl group is added to the cytosine residue of a CpG dinucleotide (Jaenisch and Bird, 2003). Indeed, normally, a large portion of the genomic DNA (except CpG islands (CGIs) within gene promoters) is methylated at CpG dinucleotides, which plays an important role in normal X-chromosome inactivation and genomic imprinting. X-chromosome inactivation is the phenomenon in which one of the X chromosomes in a female (either the maternally or paternally derived one) is randomly inactivated in an early embryonic cell, and that the same X in all cells descended from that cell are inactivated (Ross *et al.*, 2005). This clearly reflects its epigenetic nature as it is a heritable change in gene function (inactivation) without a change in the sequences on the X chromosome involved. It is noted that, although normal X inactivation is female related, X inactivation is not restricted to females. For example, it occurs in males with Klinefelter syndrome who have more than one X chromosome (Iitsuka *et al.*, 2001).

Genomic imprinting, also known as *parent-of-origin effect*, is another epigenetic factor. More than 1% of all mammalian genes are known to be imprinted and catalogued (Morison *et al.*, 2001). In addition to DNA methylation, histone modification and differential packing density of DNA by histone proteins are other mechanisms known to be involved in the process of imprinting (Bartolomei and Tilghman, 1997; Strauch and Baur, 2005).

Before birth, the basic pattern of DNA methylation is established (Bird, 2002). These patterns are replicated at cell division along with DNA sequence. However, with age, methylated areas may lose methylation and unmethylated areas may gain methylation. For example, early studies of DNA methylation reported a global decrease in 5-methylcytosine content with increasing age (Wilson and Jones, 1983). More recent studies of selected CpG sequences have found age-related increases in DNA methylation (Ahuja *et al.*, 1998; Toyota *et al.*, 1999). Age-related changes are believed to be influenced by a combination of local DNA structure (methylation centers), gene expression level and environmental exposures (Ahuja and Issa, 2000). A recent study of monozygotic twins found that variation in DNA-methylation levels of twins increased with age (Fraga *et al.*, 2005) emphasizing the importance of nongenetic factors.

Many human diseases have an epigenetic component (e.g. cancer (Laird, 2005), Klinefelter syndrome (Iitsuka *et al.*, 2001), systemic lupus (Januchowski *et al.*, 2004; Patole *et al.*, 2005), and Rett syndrome (Kriaucionis and Bird, 2003)). Similar to changes seen in cells with aging, cancer is marked by a global loss of methylation (Ehrlich, 2002) that is seen in conjunction with increased (hyper)methylation of CGIs in the promoter regions of tumor suppressor genes (Jones and Laird, 1999). Although the study of cancer genetics has traditionally focused on identifying heritable DNA variants that increase cancer susceptibility, epigenetic studies have now shown that DNA methylation can be one 'hit' in the pathway to cancer (Jones and Laird, 1999) (see **Chapter 39**). A large number of epigenetic changes are observed in cancer cells, suggesting either a defect in the machinery for maintaining epigenetic signatures, or the clonal expansion of a single cell that has undergone a number of stochastic changes that have accumulated with age. The most well-known epigenetic signature defect in cancer is the CpG island methylator phenotype (CIMP) (Toyota *et al.*, 1999; Weisenberger *et al.*, 2006). CIMP is described by the hypermethylation of a number of CGIs in a subset of cancers. It was first identified

for colorectal cancer but now has been reported for numerous other cancer sites (Issa, 2004).

In general, the epigenetic code is considered to be erased during meiosis (Chong and Whitelaw, 2004; Rakyan *et al.*, 2003). Researchers have argued that the erasing and reprogramming of the epigenetic state provides a 'clean slate' for the fertilized egg to restore its totipotency. However, for certain loci in plants the transmission of the epigenetic state through the germline has been demonstrated (Chandler and Stam, 2004). Now transgenerational epigenetic inheritance has also been observed in mice (Morgan *et al.*, 1999), rats (Anway *et al.*, 2005) and most recently humans (Chan *et al.*, 2006). These studies generate great interest since transgenerational epigenetic inheritance has been suggested as a mechanism for adaptive evolution (Belyaev *et al.*, 1981; Jablonka and Lamb, 1989; Monk, 1995). Although these processes are extremely important, in this chapter, we limit our attention to questions involving mitotic inheritance and aberrant DNA methylation in CGIs that occurs with ageing and cancer.

Implementation of state-of-the-art microarray and other high-throughput technologies has made possible the measurements of methylation signatures of multiple gene promoters simultaneously. Recently, the idea of a Human Epigenome Project has been conceptualized (Jones and Martienssen, 2005). As information from the Human Genome Project has already had a profound impact in both basic and translational sciences (Collins, 2004; Ponder, 2001), it is anticipated that the Human Epigenome Project, combined with high-throughput methylation assays and other types of epigenomic and genomic data, will facilitate our fundamental understanding of aberrant epigenetic mechanisms and propel research in areas such as cancer genetics (Jones and Martienssen, 2005; Rakyan *et al.*, 2004). Mathematical modeling and statistical methods have started to be developed to mine such massive amount of data, which could help contribute to the successful realization of the epigenome project.

In this chapter, we will focus on introducing, describing and discussing several problems and statistical treatments, one related to modeling methylation drift in human cell populations and others involving cancer genomics and high-throughput methylation data, either targeted interrogation or whole genome profiling. One of the overarching themes in many of the statistical treatments is finite mixture modeling, which, to some extent, reflects the heterogeneity nature of the data, the sample, and the underlying tumor progression mechanism.

40.2 TECHNOLOGIES FOR CGI METHYLATION INTERROGATION

There are a variety of platforms for analyzing DNA methylation including Bisulfite genomic sequencing (Frommer *et al.*, 1992), methylation-specific polymerase chain reaction (MSP) (Herman *et al.*, 1996), MethyLight (Eads *et al.*, 2000a), and chip-based technologies (Schumacher *et al.*, 2006; Khalili *et al.*, 2007). Some approaches sequence single DNA clones while others quantify methylation frequency in a DNA sample. Some technologies measure total genomic methylation content while others investigate a single CpG site or a region of linked CpGs. Several technologies are high throughput, making possible the measurements of methylation signatures of multiple genes simultaneously. In what follows, we briefly introduce two high-throughput platforms, namely, MethyLight and methylation microarrays.

40.2.1 MethyLight

MethyLight is high throughput in the number of samples (1000), and can analyze large sets of reactions (10–100). It has been widely used in the analysis of clinical samples (Eads *et al.*, 2000b; Virmani *et al.*, 2002; Weisenberger *et al.*, 2006; Widschwendter *et al.*, 2004) but is also amenable to the analysis of formalin-fixed samples such as those collected in large epidemiological studies (unpublished data). At the same time it is highly sensitive and has been proposed for the study of early detection of cancer (Laird, 2003).

MethyLight utilizes a fluorescence-based real time quantitative polymerase chain reaction (PCR) for measuring DNA methylation. The technology uses three different oligonucleotides, one forward and one reverse PCR primer and one hybridization probe. Quantitative values are determined from a standard curve of defined dilutions of a reference sample plotted as log quantity versus $C(t)$ value, the cycle number at which the fluorescent signal surpasses a detection threshold. The quantitative value for each experimental sample is derived from a linear regression on this standard curve (Eads *et al.*, 2000a). Variation in the DNA quantity and integrity and differences in efficiencies of reactions are controlled by normalizing against methylation-independent and methylated-reference reactions. Briefly, the quantitative measure from the experimental sample is normalized to that using a methylation-independent control reaction by computing the ratio. A reference reaction using an enzymatically methylated sample of SssI-treated sperm DNA is similarly normalized. The ratio of the normalized value for the experimental sample to that of the methylated-reference sample gives the percent of methylated reference (PMR).

MethyLight reactions are designed to detect molecules in which all CpGs (usually $\sim$8) are methylated. Because of this stringent criterion for detection, methylation is not found in some samples. This results in a distribution of the data that is quantitative, but having an 'excess' of zeros. This has motivated the development of novel statistical methods for cluster analysis described in Section 40.4.1.

40.2.2 Methylation Microarrays

One of the genomic strategies for efficient scanning of tumor genome for methylation alterations is CGI microarrays; the first of which is known as *differential methylation hybridization* (*DMH*) (Huang *et al.*, 1999). This first generation DMH arrayed GC-rich tags derived from a human CGI genomic library onto solid supports (e.g. nylon membranes). Then CpG DNA amplicons derived from paired tumor and normal samples (probes) are cohybridized to the microarray. Differentially methylated probes can then be identified, which signify methylation alterations of corresponding sequences in the tumor sample. There are a number of successful applications using this first generation of DMH arrays. For example, DMH was used to analyze DNA from 17 paired tumor and normal breast tissues, and it was observed that approximately 1 % of the CGI loci screened display tumor-specific hypermethylation (Yan *et al.*, 2001). More sophisticated DMH arrays that are CGI library based are now commercially available, including the Toronto human CpG 12K arrays (HCGI12K, University Health Network Microarray Center).

More recently, as new platforms are becoming increasingly available for building custom microarrays, other technologies for CGI (or large genomic regions) methylation profiling have mushroomed. Such works include that using the Affymetrix tiling microarrays on chromosomes 21 and 22 (Schumacher *et al.*, 2006) and the Agilent's custom

CpG promoter Methylation (CpGpM) arrays (Khalili *et al.*, 2007). Another technology that does not rely on the use of methylation sensitive restriction enzymes is that based on DNA immunoprecipitation (Keshet *et al.*, 2006; Weber *et al.*, 2005; Mukhopadhyay *et al.*, 2004). Briefly, sonicated DNA is immunoprecipitated using an anti-5-methylcytosine monoclonal antibody. Precipitated DNA from tumor samples and sonicated DNA from normal controls are then fluorescent labeled and cohybridized to the microarrays. There is a great deal of similarity between methylation microarray experiments and those performed using other types of microarrays, such as gene expression arrays discussed in **Chapter 6–Chapter 9**, but key differences exist, including sample preparation and experimental conditions.

40.3   MODELING HUMAN CELL POPULATIONS

Variable DNA methylation patterns observed in populations of morphologically identical cells can carry information about cell dynamics. Since DNA methylation patterns are copied fairly faithfully from one generation of cells to the next, random errors in methylation accumulated during cell division may record the histories of the cells. Molecular clock approaches, used to compare genomes between viruses, are now being used to make inference on human stem cells (Kim *et al.*, 2005a; 2005b; 2006; Yatabe *et al.*, 2001). The first study of this kind, set out to determine if stem cells in the colon are immortal (Yatabe *et al.*, 2001), is described below. Similar methods are being applied in studies of cancer to determine whether all cancer cells are immortal or whether cancer stem cells exist.

40.3.1   Background

The epithelial tissue from the human colon is sustained by millions of crypts, each containing approximately 2000 cells. Each crypt contains a mixture of stem cells and differentiated cells. The stem cells reside near the bottom of the crypt while their differentiated offspring migrate toward the colon surface. Mature cells are short lived and essentially all differentiated cells are replaced in about a week. Copy errors from cell division are maintained in the only long-lived cells, the stem cells.

The exact number of stem cells and their characteristics are unknown (Kim and Shibata, 2002). Specifically, it is unknown whether stem cells are immortal, each division resulting in two new cells, one stem cell and one cell that will differentiate, or whether they are defined by niches. Niches are regions containing cells that are externally directed to function as stem cells. One approach for inferring which of these two models may be at work in the human colon is to analyze DNA methylation patterns among crypt cell populations.

40.3.2   Methylation Patterns

DNA methylation patterns are determined using bisulfite genomic sequencing of cloned PCR products. The clones are sequenced at a series of 5–8 CpG sites. The data are coded such that '1' denotes methylated and '0' unmethylated CpGs. The code for a string of CpG sites is called a *tag*. An unmethylated tag of eight CpGs is represented by eight 0's

('00 000 000'). When studying N CpGs, there are a total of 2^N possible tags. In Yatabe *et al.* (2001), multiple clones are measured for each crypt (range 5−8) and multiple crypts per subject (range 7−9). The CpG regions studied are CpG-rich and are selected as they should normally be unmethylated in the colon.

Tags are summarized using three statistics: (1) the proportion of methylated sites (percent methylation), (2) the number of unique tags per crypt, and (3) the average Hamming distance, the average number of site differences between any two tags from the same crypt. In general, percent methylation captures information on the numbers of cell divisions since birth. Number of unique tags and average Hamming distance are measures of crypt diversity. The more stem cells or longer-lived stem cell lineages, the greater the number of unique tags and average Hamming distance.

For diploid genomes, a sensible use of average Hamming distance should only evaluate pairs of alleles sharing a common ancestor (within lineage). In Yatabe *et al.* (2001), the average Hamming distance is from a pool of alleles and represents a mixture of within- and between-lineage distance for two of the three loci studied (the third being on the X chromosome and measured only in males). The pure within-lineage average Hamming distance might be estimable from a statistical model or, in future studies, computed directly by using a nearby single nucleotide polymorphism to distinguish the different chromosomal lineages in the cell population.

Measures of cell diversity allow us to asses the histories of cells in the human colon. Under an immortal stem cell model, sequences will become increasingly diverse over time and we should observe similarity in the distances (or number of unique tags) among cells within crypts to those between crypts. For a niche, random loss and replacement of stem cells within a crypt eventually leads to a series of bottlenecks where all cells are related to a new most recent common ancestor. This leads to more closely related cells and would result in smaller distances (or fewer unique tags) among cells within a crypt than among cells between crypts.

40.3.3   Modeling Human Colon Crypts

Populations of cells in a human colon crypt are characterized using mathematical models from population genetics. More on such models can be found in **Chapter 22**. The parameter of primary interest is the probability that a stem cell divides asymmetrically, leaving one stem cell and one nonstem cell at each generation. For a niche, each stem cell division can result in 0, 1, or 2 stem cells as offspring. These occur with probabilities p_0, p_1, and p_2 respectively, where $p_0 + p_1 + p_2 = 1$. For an immortal lineage where the stem cell will never go extinct, $p_1 = 1$ and $p_0 = p_2 = 0$. Yatabe *et al.* (2001) consider a niche with $p_1 < 1$ and $p_0 = p_2 = (1 - p_1)/2$. The number of stem cells is constrained to be constant across all generations, but allows for a dynamic population where some stem cells go extinct and others spread throughout the crypt until their descendents populate the entire crypt. The models used for constraining the population size from one generation to the next arose from (Cannings, 1974; Karlin and McGregor, 1964). For details, the reader is referred to the supplementary material from (Yatabe *et al.*, 2001).

Other parameters in the model include the number of stem cells in the crypt and the frequency of methylation errors. Methylation errors can allow the addition of a novel methylation event or loss of a methyl group. The addition or loss of a methyl group can occur at different frequencies and are assumed to happen independently at each CpG site.

Future work may determine if these sites are indeed methylated independently or whether they depend on the methylation status of neighboring sites.

Quantities fixed in the analysis are the number of generations of cell division and the total number of cells in a crypt. The number of generations of cell division is a function of age of the individual (e.g. one division each day) and possibly exposures. For instance, in addition to age, the number of cell divisions in endometrial glands may depend on the number of live births of the woman and whether she is obese (Kim *et al.*, 2005b). Under the molecular clock hypothesis, a single choice of parameter values should yield the data observed for individuals of all ages and exposures.

As direct likelihood computation is difficult, rejection algorithms, a simulation-based approach, have been used for parameter estimation. In rejection algorithms the goal is to estimate the posterior distribution of the data given the parameters, $P(D|\theta)$. Methylation data are simulated under a proposed phylogenetic model (D') and compared to real methylation data (D). Parameters are selected such that the simulated data reflects the real data. As the chance of the real data being duplicated is extremely rare, typically summary statistics from the simulated and real data are compared (S' and S, respectively). The parameters are estimated by values for which the distance between the summary statistics for the simulated and real data is below some tolerance ($|S' - S| < \varepsilon$). This approach has been named *Approximate Bayesian Computation* (Beaumont *et al.*, 2002).

40.3.4 Summary

Methylation data in the human colon are consistent with the existence of stem cell niches, each niche containing multiple stem cells (Yatabe *et al.*, 2001). The existence of niches is also supported by studies of small intestine (Kim *et al.*, 2005a) and endometrial glands (Kim *et al.*, 2005b). In endometrial glands, an age-related association between mitotic age and methylation was reported; methylation was increasing with age prior to menopause and level thereafter. In addition, methylation was higher in obese women or those with lower parity, important epidemiologic risk factors in cancer. The analysis of random replication errors may provide new methods for studying cell proliferation with age and cancer. However, not all tissues show age-related changes in methylation. Methylation in the brain and heart do not increase with age (unpublished data), as would be expected in organs whose tissues do not divide in adults.

40.4 MIXTURE MODELING

Statistical modeling using finite mixtures of distributions has become one of the staples in analyzing large scale biological data, thanks to their flexibility. It provides a mathematically based approach for modeling a wide variety of random phenomena. Clustering or classification, either of features or of samples, is a class of problems often treated by such an approach. For instance, mixture modeling has been used extensively in gene expression analysis (see McLachlan *et al.*, 2006 and references therein). Analysis of methylation data has also made use of this versatile modeling framework. We briefly introduce the unified framework of finite mixture modeling next, which will be followed by three subsections; each describes a mixture modeling solution to an epigenetic problem.

We let $Y_i, i = 1, \ldots, n$, denote a random sample of size n of variable Y, which can be a scalar or a vector. Without loss of generality, we use $f(y)$ to denote the probability

density function of Y, where $f(y)$ will be viewed as a probability distribution in the case that Y is a discrete random variable. Suppose Y_i is from a heterogeneous population with K subpopulations having component densities denoted by $f_k(y), k = 1, \ldots, K$. Then the density $f(y)$ can be written as

$$f(y|\theta) = \sum_{k=1}^{K} \pi_k f_k(y|\theta_k), \tag{40.1}$$

where the π_k's are referred to as the *mixing proportions* (or weights) that are nonnegative and sum to 1; that is, $0 \leq \pi_k \leq 1, k = 1, \ldots, K$, and $\sum_{k=1}^{K} \pi_k = 1$. The parameter vector $\theta = (\theta_k, k = 1, \ldots, K)$, where θ_k is the parameter of the kth component density.

In the above formulation of the mixture model, the number of mixture components K is fixed. However, in many applications, such as in the case of learning the number of cancer subtypes classified by methylation data as in Section 40.4.1, K is unknown and has to be estimated along with the weights and the parameters of the component densities. Typically, in such situations the number of subgroups is determined using the Bayesian information criterion (BIC $= -2 \times$ log likelihood $+$ number of parameters $\times$ ln(number of observations)). The BIC allows the comparison of nonnested models having different numbers of clusters; the model with the lowest BIC is selected as best. By convention, differences in BIC greater than 2 are considered positive evidence for model differences, differences between 6 and 10 strong evidence and differences greater than 10 very strong evidence (Fraley and Raftery, 1998; Kass and Raftery, 1995).

40.4.1 Cluster Analysis

40.4.1.1 Background

Cancer patients with identical diagnoses show variable response to therapy. This has spawned the use of molecular analysis for the classification of disease subtypes. The use of cluster analysis, an approach to finding novel subgroups in data, has exploded. DNA methylation is amenable to disease classification as DNA-based signatures are more stable than alternate molecular features such as protein or RNA expression. Not only do DNA-methylation profiles differ across tissues in the body, but also among different cancer histologies from the same organ (Model *et al.*, 2001; Virmani *et al.*, 2002). This has motivated the use of DNA methylation for discovering novel disease subtypes.

One approach to clustering samples is to use finite mixture models. The number of subgroups is determined using the BIC and the mixing weights give the probability of belonging to each subgroup. Using these weights, samples can be classified by assigning them to the subgroup to which they have the greatest probability of belonging. Software exists for fitting mixtures of binary data and mixtures of normals. In order to accommodate data showing an excess of zeros as is commonly observed using the MethyLight technology, two novel approaches have been proposed.

40.4.1.2 The Bernoulli–Lognormal Mixture

DNA methylation is the outcome and we use the term *feature* to refer to the CpG regions studied. In the Bernoulli–lognormal mixture model, the outcome is modeled using a mixture of discrete and continuous components. A stringent conditional independence

assumption is made such that within subgroup k, the methylation levels are independent across features. The likelihood for a single sample is the product of the density for each feature across all CpG regions. Suppressing the notation for feature, the distribution for a single measurement is

$$f_k(y|\theta) = (1 - p_k)^{I_{\{y=0\}}} \left(p_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(z - \mu_k)^2}{2\sigma^2} \right) \right)^{I_{\{y>0\}}}, \qquad (40.2)$$

where p_k is the probability of detecting methylation in class k, $I_{\{\cdot\}}$ is an indicator function, which is equal to 1 if the condition in the brackets is satisfied, otherwise it is equal to 0, z is the log-transformed value of y if $y > 0$, μ_k is the mean of the (log-transformed) methylation level among the observations having positive measurements in group k, and σ^2 is the variance of the positive measurements. The variance could be allowed to vary by group or feature, but is assumed constant. For each feature $2 \times K$ parameters are estimated, a cluster-specific mean and probability of positive methylation. In total, the model estimates $F \times 2 \times K + 1$ parameters, where F is the number of features. The model has been fit using the expectation maximization (EM) algorithm and multiple starting values (Siegmund *et al.*, 2004).

40.4.1.3 Truncated Normal

MethyLight data have also been clustered using a truncated normal distribution. This model assumes that measurements of zero are due to the true methylation value falling below a threshold of detection; this threshold depends primarily on the biochemical properties of the reaction. The truncated normal density is

$$f_k(y|\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(z - \mu_k)^2}{2\sigma^2} \right) \right)^{I_{\{y>\tau\}}}, \qquad (40.3)$$

where z, μ_k, σ^2 and $I_{\{\cdot\}}$ are as defined above and τ denotes the detection threshold that is unique to each MethyLight reaction. For each feature, $K + 1$ parameters are estimated, the cluster-specific means and the (common) truncation value. In total $F \times (K + 1) + 1$ parameters are estimated; fewer parameters than the Bernoulli–lognormal. A version of this model has been fit using Markov chain Monte Carlo (Marjoram *et al.*, 2006).

40.4.1.4 Summary

The Bernoulli–lognormal model was applied to a study of DNA methylation in 87 lung cancer cell lines (41 small cells and 46 nonsmall cells). The purpose was to demonstrate whether subtypes of cancer could be identified using DNA-methylation profiles alone. Cluster analysis was performed using a subset of seven CpG regions. Applying the BIC criterion provided very strong evidence for a two-cluster model; however, the classification of the samples to the known disease subtypes yielded high error rates ($\sim$20 %) (Siegmund *et al.*, 2004). This suggested that DNA-methylation profiles could distinguish cancer subtype but that larger numbers of features would be needed to reduce the classification error rate.

The Bernoulli–lognormal and truncated normal mixture models have been evaluated via simulation studies. Not surprisingly, those studies demonstrated that the performance

of both approaches relied on the key properties of the data being analyzed. Standard lessons were reported such as using the (correct) model that required the fewest number of parameters tended to result in the lowest classification error rates. This was true of the truncated normal model performing better than the Bernoulli–lognormal model when the proportion of zeros was due to a detection threshold (Marjoram *et al.*, 2006). However, this superiority was lost when the data included (unmodeled) correlation among genes within a subgroup, perhaps because the more flexible Bernoulli–lognormal model could absorb model misspecification with its added parameters.

40.4.2 Modeling Exposures for Latent Disease Subtypes

Typically novel disease subtypes identified by epigenetic profiles are characterized using external information. Sometimes the external information is a clinical endpoint such as survival or response to treatment. Other times, for instance when studying the etiology of disease, the external information might be an environmental exposure such as smoking history or dietary folate intake. Often the analysis proceeds by a two-step procedure of first, identifying the disease subtypes and second, associating the disease subtypes with outcomes or exposures. Such an analysis is simple and can be accomplished using standard statistical software. However, this two-step approach does not take measurement error of the first step into account. Simple solutions exist for getting unbiased estimates of association when the latent disease subtypes are the exposure for a clinical endpoint. In such situations the errors in variables problem relates to measurement error in the exposure and a quick fix is to substitute the membership probabilities π_k instead of a hard assignment into the most likely category as the exposure (Stram *et al.*, 2003). This method falls within the general class of single-imputation approaches utilized in epidemiological studies. However, this shortcut does not work when the latent disease subtype is the outcome. In that situation one solution is an extension of the finite mixture model.

Let **x** be vector of q exposures. In the extended finite mixture model

$$f_k(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}) f_k(\mathbf{y}|\mathbf{x}; \theta), \tag{40.4}$$

where $\pi_k(\mathbf{x})$ is the probability that a sample belongs in disease subtype k given exposures **x**. The probabilities $\pi(\mathbf{x}) = [\pi_1(\mathbf{x}), \ldots, \pi_k(\mathbf{x})]$ are modeled using polytomous logistic regression, logit $\pi(\mathbf{x}) = \alpha + \beta\mathbf{x}$, where α is a $(K - 1)$-dimensional vector and β is a $(K - 1) \times q$-dimensional matrix. A simplification of this model assumes that the exposure acts directly on the disease subtype and not on any individual feature such that conditional on disease subtype, the distribution of the measurements is independent of the exposure **x**, $(f_k(\mathbf{y}|\theta))$.

The extended finite mixture model is fit by Siegmund *et al.* (2006) using the EM algorithm. The parameter of interest is β, the log–odds ratio measuring the association between exposure and disease subtype. Its standard error is computed using the observed information matrix as described by Louis (1982). Siegmund *et al.* (2006) found that estimates from the extended finite-mixture model were unbiased and had the correct standard error estimates; however, adequate sample sizes were needed in order for the algorithm to converge. These results were compared to the naïve two-step analysis. When there is error in the identification of disease subtype, the estimate of the log–odds ratio was biased toward the null and its standard error underestimated by the two-step analysis.

When disease subtypes were distinct and could be determined from the methylation data with certainty, the two-step analysis was adequate.

The two approaches were compared using a real data set of colorectal adenomas. Researchers hypothesized that low folate availability would be associated with abnormal methylation across a number of CGIs in colorectal adenomas. An analysis of red blood cell folate level and DNA-methylation subtype in colorectal adenomas estimated an odds ratio (OR) of 0.31 (95 % confidence interval (CI) 0.08–1.26) from the extended finite mixture model ($n = 58$ case subjects). The OR estimates using the two-phase approach was 0.44 (95 % CI 0.15–1.28).

40.4.3 Differential Methylation with Single-slide Data

40.4.3.1 Background

Recent discoveries that *de novo* methylation of CGIs can be associated with multiple types of cancer have led to tremendous interests in whole-genome methylation profiling (Gebhard *et al.*, 2006; Ordway *et al.*, 2006; Weber *et al.*, 2005). One of the tasks of making sense of such methylation profiling is to uncover gene promoters that are hypermethylated in tumor samples. Such outcomes, combined with other biological findings and evidence, may lead to the capturing of 'tumor signatures'.

In some circumstances, especially in pilot studies, a single microarray may be used to probe the methylation pattern of a sample without either biological or experimental replicates. For gene expression profiling, there are a number of methods proposed for single-slide analysis. The earliest method in dealing with single-slide data was based on some fold–change criterion (Schena *et al.*, 1996). Several methods have been proposed to provide more formal statistical treatments to the problem, including the gene-pooling method of Chen *et al.* (1997), the Bayesian hierarchical $\gamma - \gamma$ model of Newton *et al.* (2001), and more recently, the normal–uniform (NU) mixture modeling approach of Dean and Raftery (2005) (see **Chapter 6** for more details on single-slide analysis tools.)

Since epigenomic methylation profiling using microarrays is a relatively young field compared to its mature sister of gene expression profiling (see **Chapters 6–8**), there are few methods developed specifically for methylation data. Although methods in the literature for gene expression analysis may be adapted for methylation data, systematic evaluations of their performances in the latter context are currently lacking. Furthermore, features of methylation data may be different from those of gene expression data, and, as such, it should not be taken for granted that relative performances of statistical methods for gene expression data would hold for methylation data. Due to these considerations, a novel mixture modeling approach was proposed (Khalili *et al.*, 2007).

40.4.3.2 A γ-normal-γ Model

CGI loci are to be classified into hypermethylated, hypomethylated, or nondifferentiated according to normalized logarithms of methylation intensity ratios of the tumor sample to the normal sample (y). Normalization methods for cDNA-gene expression arrays (see **Chapters 6–9**) are applicable to data from methylation microarrays.

Loci that are nondifferentiated are those that have equal methylation intensities in the tumor and the normal. On the other hand, loci that are hypermethylated will have positive

log ratios ($y > 0$), whereas those with negative log ratios ($y < 0$) are hypomethylated. Each of the hypomethylated, nondifferentiated, and hypermethylated components can be modeled by a γ, a normal, and a γ distribution, respectively, leading to a gamma−normal−gamma (GNG) mixture as follows:

$$f(y|\theta) = \pi_1 G(-y|\alpha_1, \beta_1) I_{\{-y>0\}} + \pi_2 N(y|\mu, \sigma^2) + \pi_3 G(y|\alpha_2, \beta_2) I_{\{y>0\}}, \qquad (40.5)$$

where the weight parameters $\pi_k \left(\pi_k \geq 0, k = 1, 2, 3; \sum_{k=1}^{3} \pi_k = 1 \right)$ represent the relative proportions of the probes belonging to these three components. Again, I is the indicator function, which is equal to 1 if the condition in the curly brackets is satisfied; otherwise it is equal to 0. The parameter vector $\theta = (\alpha_1, \beta_1, \mu, \sigma^2, \alpha_2, \beta_2, \pi_1, \pi_2, \pi_3; \sum \pi_k = 1)$ can be estimated using the EM algorithm. Once the parameters are estimated, the probability of each loci belonging to each of the hypomethylated, nondifferentiated, and hypermethylated components can be computed, leading to a classification rule based on these probabilities (Khalili *et al.*, 2007).

40.4.3.3 *Normal−uniform Model*

A two-component NU mixture model can be adapted from Dean and Raftery (2005) to classify CGI loci into differentially methylated or not.

The NU modeling philosophy differs from that of the GNG model in that the NU model lumps both the hypermethylated and the hypomethylated loci into a single component, whereas in GNG, these two types of loci are modeled separately to reflect their opposite−extremes feature. The normal component is intended to capture those loci that are nondifferentiated, as in the GNG model, and the uniform component is for those that are differentially methylated, either hyper- or hypomethylated. Thus, the distribution for the normalized log methylation intensity ratio is

$$f(y|\theta) = (1 - \pi)N(y|\mu, \sigma^2) + \pi U(y|a, b), \qquad (40.6)$$

where π represents the proportion of differentially methylated loci, μ and σ are the mean and standard deviation of the normal distribution, and $[a, b]$ is the interval on which the uniform distribution is defined. The parameter vector $\theta = (\mu, \sigma^2, a, b, \pi)$ can also be estimated using the EM algorithm. Bayes rule with 0−1 loss is used to classify loci into one of the two components (Dean and Raftery, 2005).

40.4.3.4 *Further Comments*

Applications of the GNG and NU models to three breast cancer cell lines to profile their individual methylation signature based on whole-genome methylation intensity data generated from the CpGpM arrays demonstrated better goodness of fit of the GNG model over the NU model. High reliability of the predicted hypermethylated loci based on the GNG model was also ascertained through a bootstrap resampling technique (Khalili *et al.*, 2007). The reason for the need of analyzing these three breast cancer cell lines separately as opposed to analyzing them together is that they are inherently heterogeneous biologically, and thus their methylation signatures are expected to be different as well. As discussed earlier, the uniform component of the NU model is designed to capture both

the hyper- and hypomethylated loci, which are the two extremes in the data distribution. However, the uniform distribution inevitably also captures some of the loci that are nondifferentiated, which may affect the performance of the overall model. Nevertheless, it is premature to conclude that the GNG is a better model than the NU model for identifying CpG loci that are differentially methylated based on single-slide data, as the results are only from limited experiments and the comparisons are on a relative basis rather than relying on a 'gold standard'. For that matter, it is worth noting that other single-slide analytic methods in the gene expression literature may prove to be better alternatives for solving this epigenetic problem.

40.5   RECAPITULATION OF TUMOR PROGRESSION PATHWAYS

40.5.1   Background

Recapitulating pathways of tumor progression has significant implications in understanding the disease, in influencing the treatment decision, and in developing novel drug targets. Traditionally, tumor progression pathway studies have largely focused on discerning relationships among various stages, from precancerous to preinvasive/invasive, and to metastasis, using morphological data. For example, in breast cancer studies, several pathways have been proposed to describe the relationship between grades of ductal carcinoma in situ (DCIS) and grades of invasive ductal carcinoma (IDC) (Buerger *et al.*, 1999; Gupta *et al.*, 1997; Roylance *et al.*, 2002). Interpreting pathways as compartmental models, (Sontag and Axelrod, 2005; Subramanian and Axelrod, 2001) used a set of differential equations to describe the flows of contents between compartments. In particular, they studied the relative plausibility of four pathways for describing the observed counts of DCIS and IDC concurrence. It was concluded that breast tumors are heterogeneous in nature, and pathways that depict simple progression patterns, such as linearity, do not provide adequate description of such tumor samples.

The heterogeneous nature of tumors suggests that clinical phenotypic data need to be coupled with molecular signatures in order to lead to a more satisfactory recapitulation of progression pathways. To this end, it is critical to note that recent advances highlight an important role of epigenetically mediated gene silencing in tumorigenesis (Baylin, 2005). The number of hypermethylated genes tends to increase in more malignant cells, and different tumor types may be marked by their unique epigenetic signatures. Thus, the notion of utilizing DNA-methylation profiles together with clinically observable phenotypes to recapitulate tumor progression was conceptualized (Wang *et al.*, 2007). More importantly, the idea is practically feasible as these epigenetic marks are stable and heritable in tumor genomes (Baylin, 2005), and thus methylation data derived from tumors of different patients can be used as surrogates for examining progression patterns of tumors.

40.5.2   Heritable Clustering

A tumor progression pathway can be described by a directed graph with nodes corresponding to tumor stages and each directed edge denoting possible progression from one stage to another. In the current context, a node is depicted by the epigenetic (methylation) signature of the selected genes and the clinical variables (phenotypic data); see Figure 40.1
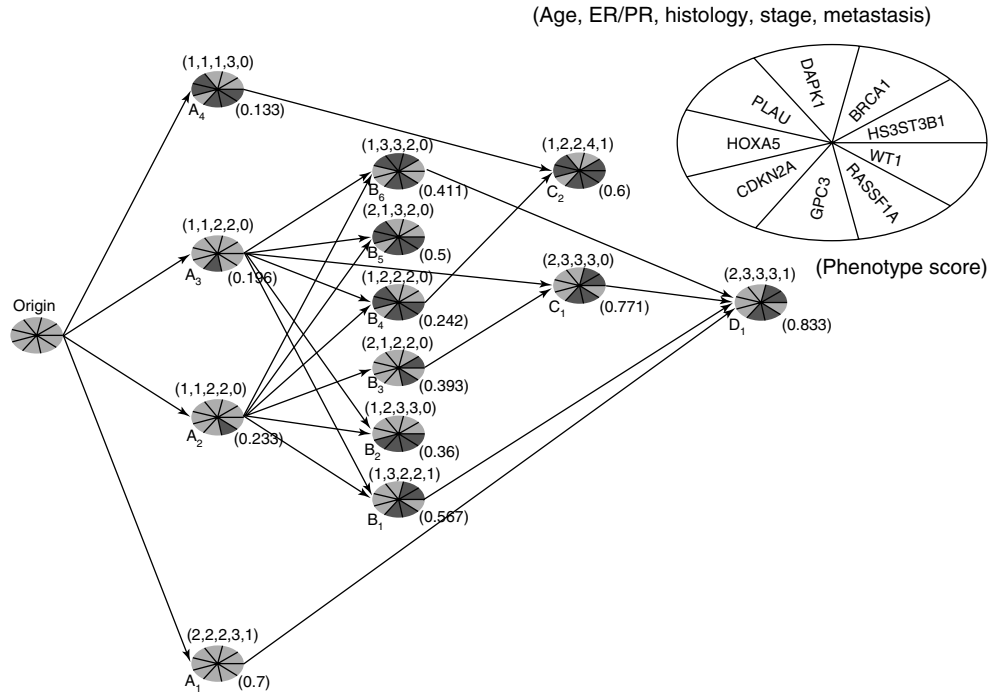
Figure 40.1 Example progression pathway network.

for an example. The heritable clustering algorithm of Wang *et al.* (2007) unfolds in three stages. First, the number of clusters (nodes of the pathway) is determined. Then, the samples are assigned to clusters and the cluster characteristics, both epigenetically and phenotypically, are determined. Finally, the clusters are organized into a pathway network to capture tumor progression to adhere to the notion that the hypermethylated loci acquired at each node are passed on to subsequent nodes and a progeny node is more aggressive phenotypically than its ancestor nodes. For determining the number of clusters, two parameters play a central role. One is the weight parameter (w) that is needed to balance the relative contributions from the methylation signature and the clinical phenotypes. The other is the within cluster similarity parameter (ε) to guarantee a certain degree of homogeneity. For each pair of (w, ε), the number of clusters $K(w, \varepsilon)$ from a clustering algorithm and the resulting total similarity $T_S(w, \varepsilon)$ can be determined. The number of clusters associated with the pair of parameters that maximizes the following objective function

$$f(w, \varepsilon) = \log(T_S(w, \varepsilon)) - \frac{K(w, \varepsilon)}{P + G}, \tag{40.7}$$

is the number of nodes of the pathway to be constructed, where P and G are the number of phenotypes and gene promoters, respectively.

Tumor samples are then classified into these nodes to define their characteristics, both epigenetically and phenotypically. Wang *et al.* (2007) considered four distance-based algorithms, H-clust, K-means, partitioning around medoids (PAM), and simulation-based (SIM). In addition, they also proposed an iterative likelihood based algorithm, which can

account for dependencies among variables as well as incomplete data. Briefly, the idea is to group tumors with similar epigenetic and phenotypic characteristics according to their parameter vectors. That is, one assumes that tumors within a cluster (C_k) share the common distributional parameter vector $\theta^k = \{\theta_G^k, \theta_P^k\}$, which represents the cluster profile and will be updated iteratively. At each iteration, let $I_{k(t)} = 1$ if tumor t is in cluster C_k, otherwise it is 0. Thus, the joint likelihood is

$$L\left(\theta_G^k, \theta_P^k, k = 1, \ldots, K | \mathbf{X}_t, \mathbf{Y}_t, t = 1, \ldots, T\right) = \prod_{k=1}^{K}\left\{\prod_{t=1}^{T}\left[P\left(\mathbf{X}_t, \mathbf{Y}_t | \theta_G^k, \theta_P^k\right)\right]^{I_k(t)}\right\},$$

(40.8)

where $\mathbf{X}_t, \mathbf{Y}_t, t = 1, \ldots, T$ are the observed methylation signature and phenotypes, respectively, T is the number of tumor samples, and K is the number of clusters. A tumor sample that has a larger likelihood in a different cluster than the one it is currently assigned to is a potential candidate for switching class membership.

Nodes with distinctive characteristics are then assembled into a directed graph to represent the pathway of tumorogenesis. The characteristics of a node are described by a center and a score, both epigenetically and phenotypically, signifying the 'severity' of the tumor stage it represents. Directed edges are used to connect any two nodes that have an ancestor–progeny relationship, where a node is an ancestor of another node (progeny node) if the progeny node has more severe characteristics, reflecting a temporal order.

Although strict epigenetic heritability is observed in Wang *et al.* (2007) in building the pathway, one may entertain the idea that methylation is reversible by lifting this stringent requirement.

40.5.3 Further Comments

A key utility of the heritable clustering algorithm is that the resulting pathway offers an opportunity for scientists to visualize the relationships between methylated gene promoters, observed clinical phenotypes, and their progression patterns. By utilizing both epigenetic signature and phenotypic information, pathway of tumor stages can be refined to better reflect its temporal ordering. For a panel of genes that are tumor-associated or tumor suppressor genes, hypermethylation in the promoters may lead to tumorigenesis or tumor progression, and thus their interplay with observable clinical characteristics can lead to better capturing of progression pathways. For instance, tumors with more aggressive phenotypes tend to exhibit higher level of methylation in such genes. In the same vein, other biological data types, such as histone modification, promoter sequence data and protein–DNA-binding data, which interact with methylation in a complex fashion, may also help to further define the pathways and offer a more system approach to recapitulate tumor progression. As the application of the algorithm to a set of primary breast cancer data demonstrates, the likelihood approach offers a more flexible and competitive algorithm for obtaining tight clusters than distance-based clustering algorithms (Wang *et al.*, 2007). Furthermore, the resulting pathway network portrays complex and nonlinear interplays between the methylation signature of the selected genes and the phenotype data, echoing the conclusion drawn by Sontag and Axelrod (2005) regarding nonlinearity and heterogeneity of tumor progression.

40.6 FUTURE CHALLENGES

A major challenge facing researchers studying complex traits is the integration of data from the various '-omics' platforms. Numerous biological and biomedical areas are propelled by recent advances of high-throughput technologies. In addition to whole-genome methylation profiling, a relatively new comer, whole genome profiling of DNA variations (single-nucleotide polymorphism (SNP) arrays), genetic changes in terms of copy number alterations (comparative genomic hybridization (CGH) arrays), transcriptom profiling (gene expression arrays), protein–DNA binding (chromatin immunoprecipitations (ChIP)–chip), and RA-omics (micro-RNA arrays), are platforms that provide large amount and diverse types of genomic and epigenomic data. Each of these contains valuable information on different aspects of the whole biological system, but more importantly, if all information is considered jointly in a truly integrated fashion, then the whole is greater than the sum of its parts. Such data integration is a goal of the new NIH-directed initiative, the Cancer Genome Atlas.

One of the earliest integrated approaches is that of combining DNA variation, gene expression, and phenotypic data to dissect complex traits (Schadt *et al.*, 2005). A popular way of studying the interplay of the diverse data types is through treating the gene expression intensity as a quantitative trait, the expression quantitative trait loci (eQTL) approach (Schadt *et al.*, 2003; Wayne and McIntyre, 2002). Similar eQTL approach was also used to study natural variation in human gene expression and potential regulatory role of SNPs (Morley *et al.*, 2004). Transcription regulation modeling is another area in which diverse data types have been jointly considered. Microarray gene expression data, ChIP–chip protein-DNA-binding data, and promoter sequence data are all believed to play a role in transcription regulation, but their combined information has proved to be much more powerful in inferring regulatory elements and complex regulatory networks (Bar-Joseph *et al.*, 2003; Liu *et al.*, 2002; Sun *et al.*, 2006). Integrating methylation profiling data with other types of epigenomic and genomic data is the natural next step to increase the power for various predictive investigations, including disease diagnostics, prognostics and therapeutics. However, cross-platform integration is by no means a trivial exercise, as different data types are interrelated in a complex manner through a highly sophisticated, yet unknown, biological system. As such, modeling of such a complex system is highly challenging. Such interdisciplinary problem requires knowledge in biology, technology, statistical modeling, and computational skills. From the statistical perspective, hierarchical modeling holds promises to tackling such problems. Large scale capable computational methods such as Markov chain Monte Carlo and approximate Bayesian computation are key to successful implementations of cross-platform models.

In conclusion, the field of epigenetics is in its infancy compared to the field of genetics. Although the impact epigenetic states will have on dissecting complex traits is still unknown, it is clear that genetic and environmental factors are not its only determinants and DNA sequence is not its only heritable unit.

REFERENCES

Ahuja, N. and Issa, J.P. (2000). Aging, methylation and cancer. *Histology and Histopathology* **15**, 835–842.

Ahuja, N., Li, Q., Mohan, A.L., Baylin, S.B. and Issa, J.P. (1998). Aging and DNA methylation in colorectal mucosa and cancer. *Cancer Research* **58**, 5489–5494.

Anway, M.D., Cupp, A.S., Uzumcu, M. and Skinner, M.K. (2005). Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science* **308**, 1466–1469.

Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A. and Gifford, D.K. (2003). Computational discovery of gene modules and regulatory networks. *Nature Biotechnology* **21**, 1337–1342.

Bartolomei, M.S. and Tilghman, S.M. (1997). Genomic imprinting in mammals. *Annual Review of Genetics* **31**, 493–525.

Baylin, S.B. (2005). DNA methylation and gene silencing in cancer. *Nature Clinical Practice Oncology* **2**(Suppl. 1), S4–S11.

Beaumont, M.A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

Belyaev, D.K., Ruvinsky, A.O. and Borodin, P.M. (1981). Inheritance of alternative states of the fused gene in mice. *The Journal of Heredity* **72**, 107–112.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes and Development* **16**, 6–21.

Buerger, H., Otterbach, F., Simon, R., Poremba, C., Diallo, R., Decker, T., Riethdorf, L., Brinkschmidt, C., Dockhorn-Dworniczak, B. and Boecker, W. (1999). Comparative genomic hybridization of ductal carcinoma *in situ* of the breast-evidence of multiple genetic pathways. *The Journal of Pathology* **187**, 396–402.

Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. *Advances in Applied Probability* **6**, 260–290.

Chan, T.L., Yuen, S.T., Kong, C.K., Chan, Y.W., Chan, A.S.Y., Ng, W.F., Tsui, W.Y., Lo, M.W.S., Tam, W.Y., Li, V.S.W. and Leung, S.Y. (2006). Heritable germline epimutation of MSH2 in a family with hereditary nonpolyposis colorectal cancer. *Nature Genetics* **38**(Suppl. 10), 1178–1183, (advanced online publication).

Chandler, V.L. and Stam, M. (2004). Chromatin conversations: mechanisms and implications of paramutation. *Nature Reviews Genetics* **5**, 532–544.

Chen, Y., Dougherty, E.R. and Bittner, M.L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* **2**, 364–374.

Chong, S. and Whitelaw, E. (2004). Epigenetic germline inheritance. *Current Opinion in Genetics and Development* **14**, 692–696.

Collins, F.S. (2004). The case for a US prospective cohort study of genes and environment. *Nature* **429**, 475–477.

Dean, N. and Raftery, A.E. (2005). Normal uniform mixture differential gene expression detection for cDNA microarrays. *BMC Bioinformatics* **6**, 173.

Eads, C.A., Danenberg, K.D., Kawakami, K., Saltz, L.B., Blake, C., Shibata, D., Danenberg, P.V. and Laird, P.W. (2000a). MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Research* **28**, E32.

Eads, C.A., Lord, R.V., Kurumboor, S.K., Wickramasinghe, K., Skinner, M.L., Long, T.I., Peters, J.H., Demeester, T.R., Danenberg, K.D., Danenberg, P.V., Laird, P.W. and Skinner, K.A.

(2000b). Fields of aberrant CpG island hypermethylation in Barrett's esophagus and associated adenocarcinoma. *Cancer Research* **60**, 5021–5026.

Ehrlich, M. (2002). DNA methylation in cancer: too much, but also too little. *Oncogene* **21**, 5400–5413.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T.D., Wu, Y.Z., Plass, C. and Esteller, M. (2005). From the cover: epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10604–10609.

Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal* **41**, 578–588.

Frommer, M., Mcdonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L. and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 1827–1831.

Gebhard, C., Schwarzfischer, L., Pham, T.H., Schilling, E., Klug, M., Andreesen, R. and Rehli, M. (2006). Genome-wide profiling of CpG methylation identifies novel targets of aberrant hypermethylation in myeloid leukemia. *Cancer Research* **66**, 6118–6128.

Gupta, S.K., Douglas-Jones, A.G., Fenn, N., Morgan, J.M. and Mansel, R.E. (1997). The clinical behavior of breast carcinoma is probably determined at the preinvasive stage (ductal carcinoma in situ). *Cancer* **80**, 1740–1745.

Herman, J.G., Graff, J.R., Myohanen, S., Nelkin, B.D. and Baylin, S.B. (1996). Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 9821–9826.

Huang, T.H., Perry, M.R. and Laux, D.E. (1999). Methylation profiling of CpG islands in human breast cancer cells. *Human Molecular Genetics* **8**, 459–470.

Iitsuka, Y., Bock, A., Nguyen, D.D., Samango-Sprouse, C.A., Simpson, J.L. and Bischoff, F.Z. (2001). Evidence of skewed X-chromosome inactivation in 47,XXY and 48,XXYY Klinefelter patients. *American Journal of Medical Genetics* **98**, 25–31.

Issa, J.P. (2004). CpG island methylator phenotype in cancer. *Nature Reviews Cancer* **4**, 988–993.

Jablonka, E. and Lamb, M.J. (1989). The inheritance of acquired epigenetic variations. *Journal of Theoretical Biology* **139**, 69–83.

Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics* **33**, 245–254.

Januchowski, R., Prokop, J. and Jagodzinski, P.P. (2004). Role of epigenetic DNA alterations in the pathogenesis of systemic lupus erythematosus. *Journal of Applied Genetics* **45**, 237–248.

Jones, P.A. and Laird, P.W. (1999). Cancer epigenetics comes of age. *Nature Genetics* **21**, 163–167.

Jones, P.A. and Martienssen, R. (2005). A blueprint for a human epigenome project: the AACR human epigenome workshop. *Cancer Research* **65**, 11241–11246.

Karlin, S. and McGregor, J. (1964). Direct product branching processes and related Markov chains. *Proceedings of the National Academy of Sciences of the United States of America* **51**, 598–602.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R.A., Niveleau, A., Cedar, H. and Simon, I. (2006). Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nature Genetics* **38**, 149–153.

Khalili, A., Potter, D., Yan, P., Li, L., Gray, J., Huang, T.H. and Lin, S. (2007). Gamma-normal-gamma mixture model for detecting differentially methylated loci in three breast cancer cell lines. *Cancer Informatics* **2**, 43–54.

Kim, J.Y., Siegmund, K.D., Tavare, S. and Shibata, D. (2005a). Age-related human small intestine methylation: evidence for stem cell niches. *BMC Medicine* **3**, 10.

Kim, J.Y., Tavare, S. and Shibata, D. (2005b). Counting human somatic cell replications: methylation mirrors endometrial stem cell divisions. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 17739–17744.

Kim, K.M. and Shibata, D. (2002). Methylation reveals a niche: stem cell succession in human colon crypts. *Oncogene* **21**, 5441–5449.

Kim, J.Y., Tavare, S. and Shibata, D. (2006). Human hair genealogies and stem cell latency. *BMC Biology* **4**, 2.

Kriaucionis, S. and Bird, A. (2003). DNA methylation and Rett syndrome. *Human Molecular Genetics* **12**, Spec No. 2, R221–R227.

Laird, P.W. (2003). The power and the promise of DNA methylation markers. *Nature Reviews Cancer* **3**, 253–266.

Laird, P.W. (2005). Cancer epigenetics. *Human Molecular Genetics* **14**, Spec No. 1, R65–R76.

Liu, X.S., Brutlag, D.L. and Liu, J.S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology* **20**, 835–839.

Louis, T.A. (1982). Finding the observed infromation matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 226–233.

Marjoram, P., Chang, J., Laird, P.W. and Siegmund, K.D. (2006). Cluster analysis for DNA methylation profiles having a detection threshold. *BMC Bioinformatics* **7**, 361.

McLachlan, G.J., Bean, R.W. and Jones, L.B. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* **22**, 1608–1615.

Model, F., Adorjan, P., Olek, A. and Piepenbrock, C. (2001). Feature selection for DNA methylation based cancer classification. *Bioinformatics* **17**(Suppl. 1), S157–S164.

Monk, M. (1995). Epigenetic programming of differential gene expression in development and evolution. *Developmental Genetics* **17**, 188–197.

Morgan, H.D., Sutherland, H.G., Martin, D.I. and Whitelaw, E. (1999). Epigenetic inheritance at the agouti locus in the mouse. *Nature Genetics* **23**, 314–318.

Morison, I.M., Paton, C.J. and Cleverley, S.D. (2001). The imprinted gene and parent-of-origin effect database. *Nucleic Acids Research* **29**, 275–276.

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747.

Mukhopadhyay, R., Yu, W., Whitehead, J., Xu, J., Lezcano, M., Pack, S., Kanduri, C., Kanduri, M., Ginjala, V., Vostrov, A., Quitschke, W., Chernukhin, I., Klenova, E., Lobanenkov, V. and Ohlsson, R. (2004). The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Research* **14**, 1594–1602.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

Ordway, J.M., Bedell, J.A., Citek, R.W., Nunberg, A., Garrido, A., Kendall, R., Stevens, J.R., Cao, D., Doerge, R.W., Korshunova, Y., Holemon, H., McPherson, J.D., Lakey, N., Leon, J., Martienssen, R.A. and Jeddeloh, J.A. (2006). Comprehensive DNA methylation profiling in a human cancer genome identifies novel epigenetic targets. *Carcinogenesis* **27**, 2409–2423.

Patole, P.S., Zecher, D., Pawar, R.D., Grone, H.J., Schlondorff, D. and Anders, H.J. (2005). G-rich DNA suppresses systemic lupus. *Journal of the American Society of Nephrology* **16**, 3273–3280.

Ponder, B.A. (2001). Cancer genetics. *Nature* **411**, 336–341.

Rakyan, V.K., Chong, S., Champ, M.E., Cuthbert, P.C., Morgan, H.D., Luu, K.V. and Whitelaw, E. (2003). Transgenerational inheritance of epigenetic states at the murine Axin(Fu) allele occurs after maternal and paternal transmission. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 2538–2543.

Rakyan, V.K., Hildmann, T., Novik, K.L., Lewin, J., Tost, J., Cox, A.V., Andrews, T.D., Howe, K.L., Otto, T., Olek, A., Fischer, J., Gut, I.G., Berlin, K. and Beck, S. (2004). DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biology* **2**, e405.

Ross, M.T., *et al.* (2005). The DNA sequence of the human X chromosome. *Nature* **434**, 325–337.

Roylance, R., Gorman, P., Hanby, A. and Tomlinson, I. (2002). Allelic imbalance analysis of chromosome 16q shows that grade I and grade III invasive ductal breast cancers follow different genetic pathways. *The Journal of Pathology* **196**, 32–36.

Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., Lum, P.Y., Leonardson, A., Thieringer, R., Metzger, J.M., Yang, L., Castle, J., Zhu, H., Kash, S.F., Drake, T.A., Sachs, A. and Lusis, A.J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37**, 710–717.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B. and Friend, S.H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O. and Davis, R.W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 10614–10619.

Schumacher, A., Kapranov, P., Kaminsky, Z., Flanagan, J., Assadzadeh, A., Yau, P., Virtanen, C., Winegarden, N., Cheng, J., Gingeras, T. and Petronis, A. (2006). Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Research* **34**, 528–542.

Siegmund, K.D., Laird, P.W. and Laird-Offringa, I.A. (2004). A comparison of cluster analysis methods using DNA methylation data. *Bioinformatics* **20**, 1896–1904.

Siegmund, K.D., Levine, A.J., Chang, J. and Laird, P.W. (2006). Modeling exposures for DNA methylation profiles. *Cancer Epidemiology, Biomarkers and Prevention* **15**, 567–572.

Sontag, L. and Axelrod, D.E. (2005). Evaluation of pathways for progression of heterogeneous breast tumors. *Journal of Theoretical Biology* **232**, 179–189.

Stram, D.O., Leigh Pearce, C., Bretsky, P., Freedman, M., Hirschhorn, J.N., Altshuler, D., Kolonel, L.N., Henderson, B.E. and Thomas, D.C. (2003). Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity* **55**, 179–190.

Strauch, K. and Baur, M.P. (2005). Parent-of-origin, imprinting, mitochondrial, and X-linked effects in traits related to alcohol dependence: presentation group 18 of genetic analysis workshop 14. *Genetic Epidemiology* **29**(Suppl. 1), S125–S132.

Subramanian, B. and Axelrod, D.E. (2001). Progression of heterogeneous breast tumors. *Journal of Theoretical Biology* **210**, 107–119.

Sun, N., Carroll, R.J. and Zhao, H. (2006). Bayesian error analysis model for reconstructing transcriptional regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 7988–7993.

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J.G., Baylin, S.B. and Issa, J.P. (1999). CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 8681–8686.

Virmani, A.K., Tsou, J.A., Siegmund, K.D., Shen, L.Y., Long, T.I., Laird, P.W., Gazdar, A.F. and Laird-Offringa, I.A. (2002). Hierarchical clustering of lung cancer cell lines using DNA methylation markers. *Cancer Epidemiology, Biomarkers and Prevention* **11**, 291–297.

Wang, Z., Yan, P., Potter, D., Eng, C., Huang, T.H. and Lin, S. (2007). Heritable clustering and pathway discovery in breast cancer integrating epigenetic and phenotypic data. *BMC Bioinformatics* **8**, 38.

Wayne, M.L. and McIntyre, L.M. (2002). Combining mapping and arraying: an approach to candidate gene identification. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 14903–14906.

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L. and Schubeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics* **37**, 853–862.

Weisenberger, D.J., Siegmund, K.D., Campan, M., Young, J., Long, T.I., Faasse, M.A., Kang, G.H., Widschwendter, M., Weener, D., Buchanan, D., Koh, H., Simms, L., Barker, M., Leggett, B., Levine, J., Kim, M., French, A.J., Thibodeau, S.N., Jass, J., Haile, R. and Laird, P.W. (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nature Genetics* **38**, 787–793.

Widschwendter, M., Siegmund, K.D., Muller, H.M., Fiegl, H., Marth, C., Muller-Holzner, E., Jones, P.A. and Laird, P.W. (2004). Association of breast cancer DNA methylation profiles with hormone receptor status and response to tamoxifen. *Cancer Research* **64**, 3807–3813.

Wilson, V.L. and Jones, P.A. (1983). DNA methylation decreases in aging but not in immortal cells. *Science* **220**, 1055–1057.

Yan, P.S., Chen, C.M., Shi, H., Rahmatpanah, F., Wei, S.H., Caldwell, C.W. and Huang, T.H. (2001). Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. *Cancer Research* **61**, 8375–8380.

Yatabe, Y., Tavare, S. and Shibata, D. (2001). Investigating stem cells in human colon by using methylation patterns. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10839–10844.

Part 7

Social and Ethical Aspects

41

Ethics Issues in Statistical Genetics

R.E. Ashcroft

Queen Mary's School of Medicine and Dentistry, University of London, London, UK

The ethical, legal and social issues concerning genetic research (genethics) are extensive and complex. This chapter reviews some of the central issues in current debates. The first part of the chapter considers the scope of genethics and the relationship between ethics, morality, professional conduct and genetics research. It then considers the relationship between risk-control and benefit-maximising models of governance in genetics research. The second part of a chapter, using a case study of the UK Biobank, looks at the main issues in governance of genetic databases, including scientific value of the research, recruitment, consent and mental capacity, voluntariness and incentives to participate, feedback of research results, confidentiality and security. The third part of the chapter looks at issues in scientific conduct of research, concentrating on stewardship and benefit sharing with host communities. The fourth part looks briefly at societal issues, concentrating on two case studies: geneticisation and reductionism, and the issue of race in genetic research.

41.1   INTRODUCTION: SCOPE OF THIS CHAPTER

The ethical, social and legal issues arising in genetic research and its applications are so extensive that they have generated a whole field of research and scholarship, often referred to as ethical, legal and social implications (ELSI) of genetics or 'genethics' (Clarke and Ticehurst, 2006; Sherlock and Morrey, 2002). A few examples of topics discussed in this literature include ethical debates about bio-safety of genetically modified crops, the morality of patenting genetically engineered living creatures, the obligations researchers in human biodiversity may have to share the benefits of any commercialisable discoveries with the donors of samples and the communities of which they are members, the moral limits on modification of the human genome, and the question of whether prenatal or pre-implantation genetic testing amounts to discrimination against the disabled or a new variant eugenics.

These issues are so diverse and complex that to cover them all adequately would require book-length treatment. In the present chapter I concentrates on issues of primary concern to statistical geneticists. This is a necessarily selective survey. I concentrates on an issue that is of considerable practical concern in genetic research – the ethical governance on

research using biobanks. This topic cuts across many other topics in the ethics of human genetic research as well as touching on some aspects of research on non-human organisms. The second part of this chapter discusses the ethics of research using biobanks so far as this relates to the interests of sample donors. The third part of this chapter discusses the ethics of such research so far as this relates to the relationships between researchers. The final part of the chapter discusses the ethics of research so far as this relates to the interests of society.

Before beginning an examination of these issues, it will be useful to review, in this first part of the chapter, what the main sources of ethical argument are in the literature.

41.1.1   What is Ethics?

Ethics can be defined as philosophical inquiry into the values, rules of conduct and character traits, which are involved in right action, doing good and living well. It is often contrasted with morality, which is the commonly shared set of rules and principles shared within a community and taken for granted in assessing one's own behaviour and that of others. As defined, ethics can be thought of as systematic inquiry into the foundations of morality and – where necessary – correction of the principles of morality in the light of reason and evidence (Benn, 1998).

Ethics in this philosophical sense should not be confused with 'professional ethics'. Professionals such as doctors or lawyers may refer to conduct as 'unethical', by which they mean that it violates the formal or informal norms of expected behaviour by members of that profession, as laid out in codes of conduct or as inculcated through professional training. This usage of the term ethics is analogous to the term morality as defined above. Professional ethics (or, as I would prefer to say, professional morality) is referred to as such to distinguish it from common morality, which is the morality shared by most members of a community whether or not they are members of a profession. Philosophical ethics may address questions of professional ethics, but professional ethics need not be philosophical. In recent years there has been a considerable growth in teaching of and research into professional ethics – for example, medical ethics, which is now taught formally in all medical schools in the United Kingdom. Professional ethics is developed in close liaison with the legal and regulatory requirements on professionals.

One important part of philosophical ethics is bioethics. Bioethics can be defined as the application of principles and methods of analysis of philosophical ethics to the analysis of moral and social problems arising in the life sciences and medicine. Genethics is therefore part of bioethics.

The methods of bioethics are generally analytic and philosophical. Nevertheless, in recent years considerable attention has been paid to the need for ethical arguments to make use of the best quality social-scientific evidence: as has been said, good ethics needs good facts. This is particularly important in the area of population and public health genetics, since studies involve large numbers of participants, and are occasions of potential controversy. Empirical research has some potential both to clarify what issues are at stake in the participant community and also to build trust and confidence in the aims and processes of research. Examples of such research include the large scale social survey research using the Euro barometer survey into public attitudes and values concerning biotechnology (Bauer and Gaskell, 2002) and qualitative interview studies into research participants reasons for agreeing or refusing to take part in biobank research (Haimes and Whong-Barr, 2004). The Wellcome Trust has funded a major programme of social research

on genethics aspects of biobank research, summarised in Haimes and Williams (2006) and on specific issues in pharmacogenomics research, summarised in UK Pharmacogenetics Study Group (2006). Recent collections presenting empirical research from the United Kingdom and other European countries are, Tutton and Corrigan (2004) and Ashcroft and Hedgecoe (2006).

In addition to analytical and empirical methods, research in genethics necessarily overlaps with research in law and public policy. There has been a considerable growth in research in medical and biotechnology law over the past twenty years, using both standard case and statute law resources and also, increasingly, the jurisprudence of human rights. In the last ten years there have been a number of significant international human rights declarations concerning genetics, and these are directly relevant to population genetic research.

41.1.2 Models for Analysing the Ethics of Population Genetic Research

There are a number of different ways of conceiving what ethically is at stake in population genetics research, each of which goes some way to shaping the regulatory and ethical framework in which contemporary genetic research is done. We can start by categorising these into two broad groups: benefit-maximisation models and risk control models. Benefit maximisation-models of the ethics of population genetic research focus on the benefits of such research, and seek ways to maximise these. Risk-control models identify specific risks which may be involved in genetic research and seek to control and in many cases minimise these. While these might be seen as complementary in that benefit-maximisation models recognise risk-based constraints, and risk control models try to avoid strong risk aversion, which might dilute any benefits accruing to such research to the point of futility, nevertheless they differ in emphasis and orientation.

41.1.2.1 Risk Control Models

For historical reasons, I thinks it is arguable that risk-control models of genethics have dominated discussion. We can identify a number of factors explaining this. The first is the history of eugenics, especially in the way eugenic ideas were used to justify forced sterilisations, barriers to 'mixed' or 'inappropriate' marriages, racial discrimination, abuse of the mentally ill and learning disabled, up to mass killings of the genetically 'unfit' under the German Third Reich (Paul, 2002). Any genetic research in humans since the Second World War has needed to establish the distance between its aims and those of eugenics. Genetic research and genetic medicine is widely seen as an area in which serious risks of personal harm and social injustice need to be forestalled or overcome. A second factor is the history of the ethics of research on human subjects. Again because of the Nazi experience (and parallel episodes in the Japanese empire, and subsequent morally problematic experiments carried out during the Cold War by NATO and Warsaw Pact countries and elsewhere), research on human beings has been considered intrinsically risky (Rhodes, 2005). To the extent that population genetics for medical or non-medical purposes involves engaging with human subjects, it is seen as a form of research on human subjects and is governed by the research ethics risk control paradigm. Next, because of the scale of modern population genetic research two further types of risk have come to the fore. The first is the role of the state in funding, facilitating and regulating genetics

research. Concerns of modern citizens about state interference in their lives are read across into the field of genetics research. The second is the role of the commercial private sector in developing technologies using genetic research and its applications. Concerns about the self-interested behaviour of corporate actors are similarly read across into the field of genetics research. These types of risk are exemplified by concerns about data protection, confidentiality and privacy, intellectual and real property rights in samples, data and discoveries, benefit sharing with donors of samples or data, exploitation of poor individuals or groups (especially transnationally), state coercion and surveillance, sharing of information with agencies or individuals for non-medical purposes and so on (Bauer and Gaskell, 2002). Finally, there is a general concern with any kind of research into the biological characteristics of human beings, that such research undermines human identity or human dignity. This kind of risk is often invoked in religious contexts, but it has also been influential in the framing of certain kinds of regulation, most notably international patent law, and law and regulations prohibiting germline gene therapy and reproductive cloning (Beyleveld and Brownsword, 2001). Unlike the first four kinds of risk, this is not clearly linked to any specific historical or political experience, but may be linked to a more general historical epoch often referred to as 'modernity', which many intellectual historians associate with secularisation or with the 'scientific revolution' of the sixteenth century. Further, this sort of risk is generally considered qualitatively rather than quantitatively. Notionally it cannot be traded off against other benefits or risks, because undermining human dignity (for instance by undermining the grounds for considering human beings fundamentally equal and members of the same human family) would be absolutely wrong. This sort of risk is held in mind especially in international human rights declarations.

At this stage, my purpose in describing these five kinds of risk is not to argue about their scope, significance and centrality to the ethics of genetics research, but rather to argue that concentrating on these kinds of risk makes sense to many analysts of the ethics of genetics for historical and political reasons. Because they are seen to be salient, they frame and shape much of the discussion of the ethics of genetic research, and much of the regulatory framework for genetic research is fixed by the concern for these risks and solutions developed in other contexts for managing them.

As well as identifying these five kinds of risk by the context in which they arise and the sorts of harm or moral danger involved, they can also be categorised by the level at which they operate: individual, family, social group, society as a whole, humanity as a whole. There is a pronounced tendency in the literature to concentrate in risks of harm to identifiable individuals, and to regulate that risk of harm through a dependency on the informed consent mechanism. As seen, it is harder to regulate risks which operate at group level, and the troubles attending the extension of informed consent models to group protection are well known and difficult.

41.1.2.2 *Benefit Maximisation Models*

Perhaps surprisingly, benefit-maximisation models are rarely articulated formally, but are instead the 'common sense of science'. Benefit maximisation arguments can appear in various forms, from appeals to the future health benefits that will accrue if particular research lines are (successfully) pursued, to appeals to a right to freedom of scientific inquiry, to more bluntly economic arguments about the inefficiency of research

ethics regulations. In the current state of ethical debate it is arguable that benefit-maximisation models function in two ways: first, as a corrective to (excessive) caution in risk-control models, and second, as substantive arguments about the best way to get maximum value out of particular research resources. An example of the latter type of argument is the argument about whether privately funded entrepreneurship and 'pay-per-view' style models of access to genetic databases by researchers, or alternatively open source, publicly funded databases, are more efficient and effective in producing high quality genetic science and technologies (Sulston and Ferry, 2002; Rabinow, 2003).

41.2 A CASE STUDY IN ETHICAL REGULATION OF POPULATION GENETICS RESEARCH: UK BIOBANK'S ETHICS AND GOVERNANCE FRAMEWORK

In order to understand the ways in which the different kinds of research risk frame the governance of population genetic research, while retaining an intention to do research that is maximally beneficial, it is helpful to consider a case study. The UK Biobank is a major initiative in studying the interaction between genes and environment in a health context, using a large cohort study drawn from the UK population. It is described on the project's website as follows:

> UK Biobank is a long-term national project to build the world's largest information resource for medical researchers. It will follow the health of 500 000 volunteers aged 40–69 in the UK for up to 30 years.
>
> The UK Biobank is funded by the Department of Health, the Medical Research Council, the Scottish Executive, and the Wellcome Trust medical research charity. The project will help approved researchers to develop new and better ways of preventing, diagnosing and treating common illnesses such as cancer, heart disease, diabetes and Alzheimer's disease.[1]

On the website, the need for the UK Biobank is explained as follows:

> UK Biobank is a medical research study of the impact on health of lifestyle, environment and genes in 500 000 people currently aged 40–69 from all around the UK. This age group is being studied because it involves people at risk of developing serious diseases – including cancer, heart disease, stroke, diabetes, dementia – over the next few decades.
>
> The UK National Health Service treats the single largest group of people anywhere in the world, and keeps detailed records on all of them from birth to death. Consequently, follow-up of UK Biobank participants through routine medical and other records will allow identification of those who develop a wide range of disabling and life-threatening conditions. This will make UK Biobank a uniquely valuable resource.
>
> Scientists have known for many years that our risks of developing different diseases are due to the complex combination of different factors: our lifestyle and environment; our personal susceptibility (genes); and the play of chance (luck). Because UK Biobank will involve thousands of people who develop any particular disease, it will be able to show more reliably than ever before why some people develop that disease while others do not. This should help to find new ways to prevent death and disability from many different conditions.[2]

Recruitment and participation are described thus:

People to invite into UK Biobank are identified from central registries. The only information provided, in confidence, to UK Biobank is name, address, sex, date of birth, and general practice. These details are processed centrally (in accordance with the Data Protection Act) and an invitation letter sent directly by UK Biobank. General practitioners are advised that people registered with their practices may be invited to take part.

Taking part in UK Biobank involves participants in:

- Attending a local study assessment centre for about 90 minutes to answer some simple questions, to have some standard measurements, and to give small samples of blood (about 2 tablespoons) and urine;
- Agreeing to allow their health to be followed for many years by UK Biobank directly through routine medical and other records;
- Being re-contacted by us to answer some additional questions and/or attend a repeat assessment visit (which would be entirely optional).

A very wide range of tests will be done on the blood and urine samples for approved medical research, and it is impossible to predict all of the uses to which the samples might be put during the next few decades. But, since none of these individual test results will be fed back to participants, their doctors or anyone else, taking part in UK Biobank should not have any adverse effects on participants (including their employment status or ability to get insurance).

By analysing the answers, measurements and samples collected from participants, medical researchers will be able to work out why some people develop particular diseases while others do not. Although taking part in UK Biobank may not help participants directly, it should give future generations a much better chance of living their lives free of diseases that disable and kill.[3]

The UK Biobank has a detailed governance framework, which I will discuss and describe in detail below. For the moment, let us examine the issues this initial presentation of the UK Biobank highlights as being ethically important or of interest to the general reader of its website.

41.2.1   The Scientific and Clinical Value of the Research

The first thing emphasised in the public presentation of the Biobank is the *expected value* of the research. Participants are invited to take part, and the public (and scientific community) are invited to support the project, on the basis that the Biobank will produce important new knowledge, which will make a significant contribution to the understanding, prevention, diagnosis and treatment of major diseases.

In this regard, UK Biobank is typical of most human subjects research involving large numbers of people (Tutton and Corrigan, 2004). To succeed, it must recruit a large number of people, and secure their consent to a variety of investigations, some of which may be painful or inconvenient, and some of which may require long-term contact. Biobank is presented as a moral enterprise. It is thought not only to be scientifically interesting, but also to be potentially useful and helpful in advancing a vital interest that all share, their interest in being healthy and in receiving good medical care. A potential participant might ask him or herself, 'why should I volunteer?' One answer might be: to help scientists do

something they think is really interesting. For some research participants, this would be a sufficient reason to take part. Many people do think that science is intrinsically valuable, and that it is exciting and honourable to play a part in its advance. Yet in fact this appeal to shared scientific curiosity is relatively rare in modern biomedical research. A risk of selling science to the public on the grounds of utility is that scientific projects may fail, or may fail to produce the expected or hoped for results. In clinical trials research, researchers are aware of the 'therapeutic misconception' held by many patients, who believe that they are receiving the treatment judged to be best for them, when in fact they are participating in a trial which may or may not benefit them personally. This issue is less pertinent in epidemiological research, but something similar may apply at the level of society at large: social participation in research like Biobank is not guaranteed to benefit anyone individually or as a collective. Yet researchers have an incentive to emphasise the chances of success, and to play down the chances of failure.

Participants in research like Biobank who take part on the grounds of its expected usefulness may perceive this utility in one of three ways. First, they may perceive participation as useful to them personally, here and now. Second, they may perceive participation as generating benefits, which will be useful to people like them (possibly including themselves personally) in future. Third, they may perceive participation as generating benefits, which may benefit people in future, without consideration of their own personal interests. It is only rarely the case that there are direct personal benefits to participating in epidemiological research, but see the discussion of feedback of results to participants below. On the other hand, researchers and research participants alike frequently emphasise the second and third kind of reason to take part as good reasons.

Participation on the basis of benefits to people like myself (possibly including me) in future could be called *solidary participation*: I take part out of solidarity with people I recognise as my peers and whose suffering I want to ease, on the basis that they would do the same for me, or that I hope that they would. Participation on the basis of solidarity is a theme that has long been emphasised in epidemiology, at least in the post-war period in which epidemiology in the United Kingdom and other welfare states was linked institutionally to social medicine, to socialised healthcare, and to social movements such as the Trades Union movement (Ashcroft *et al.*, 2000). Solidarity may be presented in a more or less moralised form. On the one hand, many commentators would see solidarity as the basis of any moral relationship with others, since what motivates us to help others is an understanding of the plight of others based on empathy. On the other hand, solidarity could be a much more pragmatic motive, where people reason on a quasi-contractual basis that they are obliged to put into a social relationship more or less what they get out of it. Some recent work argues that people have a duty to participate in research, because they have benefited to date from research in which others have participated. Solidary participation on this understanding is no more than enlightened self-interest.

Participation on the basis of benefits to others without consideration of the relationship they may have to people can be described as altruistic participation. Appeals to pure altruism are relatively rare, just as it was noted that appeals to shared curiosity are rare. In practice, most appeals for research participation appeal either to direct self-interest (through payment to participants in healthy volunteer research, or through an improved chance of receiving the treatment that is most effective or safest in a clinical trial) or to solidarity. The reasons for this are complex and are suspected to be not very well understood, but part of the explanation may be the dominance of a 'risk control' approach

to research ethics. The influential Declaration of Helsinki and most other statements of research ethics emphasise that the welfare of the individual subject shall take precedence over the interests of science and society. A consequence of this emphasis is that any approach to research recruitment emphasises the interests of the participant rather than the benefits to science or society, even where the interests of the participant are only minimally affected, or where those interests are entirely consistent with maximising the scientific and social benefits.

So far i have considered the ways in which appeal to scientific value figure in arguments persuading participants to take part in Biobank, and in persuading the public to accept Biobank as a national scientific project using public resources. It is important also to emphasise the other side to this argument, which is that for these reasons to hold, it must actually be the case that Biobank represents good value for money and a worthwhile investment of scientific resources of time, talent and facilities. In general terms, it is accepted that scientific projects must be well founded in terms of posing a well-defined question, to which the answer is not already known, and which, if answered, would generate new knowledge of scientific importance. The twin mechanisms of systematic review (to establish what is already known, and to what degree of methodological confidence) and peer review (to establish the credentials of the research team, the likely scientific value of the research and the value for money the project offers) are intended to ensure that scientific projects put to the public to invite participation are well founded scientifically, and thus that the claims made about the utility of the research are as well supported as is possible.

An important issue in genetic research, however, is that in many cases genetic sample collections are built up not in the context of a specific research project, with tightly designed objectives and research questions, but as resources for use in future unspecified research (Gibbons *et al.*, 2007). Sample collections may be made of those collected for clinical or other purposes, and are turned into research resources retrospectively; or they may be collected for use in one research project, and then reused in other future research projects for other purposes; or they may directly be collected as a sample bank, with the research uses to be specified later. For example, a researcher may build a collection of samples taken from the brains of people with Parkinson's disease, for use by Parkinson's disease researchers, without any specific research project in mind at the point of collection. In a commercial clinical trial, the pharmaceutical company sponsoring the research may build a collection of samples from participants in the trial for pharmacogenetics research either prospectively or retrospectively, to help interpret their trial results and assist in licensing applications. Here, the questions of utility may be more diffuse, and peer review, if practised at all, would focus more on the governance of the collections, and on the research projects, which are subsequently proposed that intend to use the collections.

We have seen how Biobank and other research projects start their approach to ethical governance and recruitment by emphasising the scientific and public value of their projects. The next topic addressed is the methodology of recruitment.

41.2.2 Recruitment of Participants

Biobank are at pains to emphasise how individuals are identified as possible participants. This has two elements: selection and identification. Potential participants are selected on the basis of the inclusion and exclusion criteria, which define the study population in light of the research objectives of the study. Biobank intends to enrol 500 000 individuals,

between 40 and 69 years of age, from all over the United Kingdom. Those excluded from the study are those who are unable to give informed consent (for example, because of diminished mental capacity), those too ill to take part in data collection and those who study recruiters deem uncomfortable with any of the conditions of participation. The latter exclusion essentially is intended to exclude people whom it is felt would prefer not to consent, or seem not to understand what is required of them, but who nonetheless seem for one reason or another to be giving their consent. These are very broad inclusion criteria, as is appropriate for a study, which intends a rather comprehensive analysis of the relationship between physical and mental health, lifestyle and environmental factors and genotype. Other studies might have more restrictive inclusion criteria. Case-control studies, for instance, where 'cases' must meet defined clinical or other criteria for inclusion and where 'controls' must be comparable to selected cases, will have much more specific inclusion and exclusion criteria. In the context of controlled clinical trials there are important issues concerning the fairness of inclusion and exclusion criteria in terms of the distribution of the risks and benefits of research to (non-)participants. This is less of a significant issue in epidemiological research, but does arise in the context of the interpretation of findings, which will be described below.

A more important question concerns how individuals eligible for participation are identified as suitable for recruitment, and how they are then approached. Some studies simply ask people to volunteer, through placing advertisements. It is up to each individual to decide whether they are eligible and interested, and then to contact researchers. For most purposes this is not adequate in terms of constructing an unbiased sample for research, so the practical issue is how to identify potential research participants, and then to contact them, without breaching norms of confidentiality and law relating to data protection. If potential participants are identified through public information (such as electoral rolls or telephone directories), this is rarely a significant issue. However, where participants may be identified through information not in the public domain, such as personal medical records, the situation may be complicated. For a researcher who would not have access to this information routinely (for instance, in the case of medical records, because he or she does not have clinical care of these patients), individual consent might be required for researcher access to these records; but the researcher can only know which records (or which patients) he or she would like to see, and thus whose consent is needed, when he or she has had the chance to look at the records. In practice, pragmatic solutions to this Catch-22 need to be found: either someone who has right of access to the set of records preselects the individuals for the researcher and makes the first approach to them to obtain consent to the researcher approaching them, or actually to enrol them into the study, or the researcher is given a contract with the record-holding organisation that allows them to see records and places them under a contractual duty not to misuse the records outside the terms of the contract. Under some circumstances, it might be that a research project is of such public importance that enrolment of participants may take place without their consent (if this involves data extraction from records only). Many countries have regulations in place to permit this. However, in the case of genetics research, while this method might be used to allow data extraction to identify individuals who are potential research participant, actual recruitment into the research would almost always involve a direct approach to the individuals and the seeking of their consent.

The principal exception to these rules about recruitment of participants are where the samples to be used are de-identified, in such a way that re-identification of individuals is impossible. Normally this would occur when a sample bank is used, rather than samples being collected for a specific research project.

41.2.3 Consent

The central element to the ethics of genetic sample based research is – inevitably – consent. Consent in epidemiological and genetic research has been much debated in recent years (O'Neill, 2002; McHale, 2004; Gibbons *et al.*, 2005). This is in part because genetic research poses new issues, and in part because of the social and economic trends providing the context against which genetic research takes place now.

The standard requirements for a valid informed consent concern (a) the mental capacity of the individual to give consent to the research participation; (b) the voluntariness of the decision; and (c) the information necessary to make a decision (Jackson, 2006).

Mental capacity is a complex topic, and could be treated at length, but for present a few simple points are sufficient. Mental capacity is the ability to understand and retain the information one is given to make a decision, to believe it, and to weigh it and make a choice. A consequence of this definition is that capacity is relative to the nature of the decision being made: it is much easier to show that someone has the capacity to buy a newspaper than it is to show that that they have the capacity to consent to a heart transplant. So one does not simply have 'mental capacity', but rather, under a particular set of circumstances, one can be said to have the capacity to make this sort of decision. In a medical context, adults are presumed to have mental capacity to make any decision that they may be asked to make, and considerable evidence and inquiry may be needed to show that they lack mental capacity to make that decision. This is true even of people with a psychiatric disorder. In the case of children, for the most part, the opposite is true: a child under 18 (or, for many purposes, under 16) has to be shown to have the ability to make a decision, and is presumed to lack that capacity. For medical research purposes, these assumptions about adults, children and mental capacity are controversial. Should it, for instance, be assumed that research is necessarily harder to understand than ordinary clinical treatment? Surely not. But some research is highly complex. Moreover, what may be difficult to understand is that enrolment in research is voluntary, that according to all research guidance it is clear that patients are not to be compelled to take part in research, or threatened with poor treatment if they refuse, or forbidden to leave research once it is started. Patients (in particular) may believe that participation is either obligatory, or necessarily in their best interests, or that they 'owe' their doctor something, when none of these need be true. (This, again, is the 'therapeutic misconception' mentioned above.) So what may be the stumbling block in determining capacity is not the complexity of the technical information, but the changed relationship between researcher and participant (where it was doctor–patient, it is now a rather different relationship between researcher and participant).

We can debate whether this is an issue of capacity, or rather one of voluntariness. In practice, since most epidemiological research is non-therapeutic, people who lack capacity are simply excluded from research (unless the research links specifically to the reasons for their incapacity – as in some psychiatric genetic research), since participation in research can rarely be shown to be in the individual's best interest. This would be the test of whether someone can be enrolled in research when they lack mental capacity. A finer

grained interpretation of this condition can be developed, as in guidelines from the UK Medical Research Council and other similar bodies (Medical Research Council, 1991). On this interpretation, a person lacking capacity to consent to a research project can be included in the research if it is in their best interests, or if the risks are minimal and this research cannot be conducted in people with capacity and this research would benefit others in future with a similar condition and participation is not *against* the individual's interests. Under some circumstances research, which is not minimal risk can be permitted, but there would have to be a compelling case made about why this research was essential for the welfare of people suffering from this condition under investigation and there was no other way to carry it out or to resolve the problem under study.

A related difficult issue concerns the achievement at a later date of capacity by someone enrolled in a study while lacking capacity. For example, a child might be enrolled in a study at birth. When the child reaches majority, it is sometimes argued (as in the Icelandic DNA database, for example) that the now adult individual should have the power to withdraw the consent given on their behalf, or to ratify it. This raises complex issues concerning what it actually means to leave a study in this context, as it will discussed below, but also the wider question of how far 'proxy consent' can really be considered valid, and how far parents (or carers) can choose authoritatively for their children (or incapacitated relative) (Ross, 1998; Archard, 2004).

The voluntariness condition is less controversial. In practice, people may be under a variety of pressures to participate, but these would not normally amount to coercion. It is well known that some people are more successful at recruiting research participants than others, other things being equal, and how far this is due to personal charm or to conveying the importance of the research or a low tolerance to taking no for an answer is often difficult to discern! Ethical concern tends to be directed at formal obstacles to voluntariness. Since the Nuremberg Code of 1947, all research ethics guidelines have insisted that people who take part in research should be free to leave it at any time, in the interests of allowing people to change their minds or go back on decisions they regret or – in the worst case – were suborned into making (Marks, 2006). This poses a difficult challenge in epidemiological research. Since such research depends on the collection of large data-sets and the analysis of such sets as collections of data, often longitudinally, it is not always clear what 'freedom to leave' means. Clearly it can mean that once a person leaves the study, no new data on them can be collected, but existing data can continue to be used. Or it might mean, more permissively, that once they leave, further data on them may be collected from routine data without further contact with the person. Or it might mean, more restrictively, that once they leave existing data on that person should be removed from the data-set (UK Biobank Ethics and Governance Council, 2006). The standard view is that the existing data can be used once the person has decided to leave the study, and that it cannot be removed. To remove the data would compromise the integrity of the data-set, and may actually be impossible for technical reasons, if data once added are de-identified. In more principled terms, it is not clear that the person has a right to request that their data should be removed: firstly, that they agreed to that data being used and cannot retrospectively revoke that agreement, and secondly, that although the data may be derived from them, they are not owned by them, but by the researcher in whom intellectual property in that data resides. What is being negotiated here is the meaning of 'leaving the study'. The normal approach here is to be reasonably explicit about what agreeing to take part in the study amounts to, and about what leaving the study

amounts to. Nonetheless, it is possible to imagine circumstances under which someone loses their trust in the researcher or the research project to the extent that they feel that they have been misled. They might then reasonably say that the data was collected under false pretences, and that it should be deleted. This might arise in a study of race/ethnic differences, for instance, where a person feels that the study undermined the dignity or reputation of their race/ethnic group in a way they could not foresee and were not warned about. Similarly, it might arise where a child was recruited into a study by his or her parents in early childhood, and may at adulthood wish to remove his or her data from the study, which were collected without his or her consent.

A second issue of voluntariness concerns incentives to participate. These may be both formal and informal. Formal incentives, in the form of payments or offers of services in exchange for agreement to participate, are controversial for two main reasons. Firstly, many epidemiologists would feel that people should participate in research for solidary or altruistic reasons, and offering payments, gifts or in-kind exchanges devalues such reasons for participation. They would argue that this discourages people from participating in research unless there is something in it for them. Secondly, ethicists tend to worry that people who take part when there is an incentive scheme are doing so in order to get the incentive, rather than with full understanding of the risks and benefits of participation. The greater the incentive, the greater the risk that this may occur, up to the point where people put themselves knowingly in danger merely because they want or need the incentive. Of course, this is the situation with most phase I drug research and many kinds of employment. But this is felt to be regrettable, if necessary, and not something to be expanded. One hard issue here is that it is almost, if not actually, impossible to draw a line separating reasonable from coercive levels of inducement.

Informal incentives to participate comprise reasons to participate in the research, which are formally part of the research protocol itself, but which are attractive to participants for reasons unconnected with the research. For example, in many epidemiological studies, such as Biobank, participation in the research can involve the collection of vital statistics and medical examinations for the purposes of collecting baseline and study time-point data, but which are also (potentially) useful health screening data for the research participants. Many studies are attractive to participants because they are getting free, or more than usually convenient, 'health checks'. Sometimes, participants may also believe that they are getting access to tests or checks on their health that are not normally available. This is particularly the case in some genetic studies investigating risk factors for common diseases. A patient in a study of the genetic risks for colon cancer, for example, may believe that they will be given individual feedback on their genotype, which would allow them to know whether they were at risk of developing colon cancer.

These informal incentives are highly problematic, and much discussed in the literature (Haimes and Williams, 2006). It is generally accepted that data that can be interpreted 'at the bedside' and which are collected on an identifiable basis should be returned to the participant on request, although if the participant does not want them, there is no need to force these data on them. In most epidemiological research, the researcher is not seeing the participant as a patient, and the researcher has no specific duty of care to the participant, such that if the participant has a high blood glucose level they should be counselled to see their general practitioner about possible diabetes mellitus. Indeed, for many of the tests and measurements that may be done in a research project, the participant would need to be counselled and consent obtained *to the medical test* rather than to research

data collection only if this were done. Nonetheless, if data is collected, which incidentally suggests that the participant would be wise to see a doctor, it is generally good practice to pass this on. More controversy surrounds the feedback of research findings to patients, which only arise out of the analysis of the research data itself. So should it appear that a particular genotype is at greater risk of colon cancer than normal, there is not thought to be any obligation to tell individual research participants with this genotype that they need to see their doctor for further counselling and possible investigations. In part, this is because the quality of such genetic testing in research is not at clinical grade. Participants should normally be kept up to date with the progress of the research, in terms of key findings for instance. But this level of feedback about the group as a whole would not have implications for specific individuals. Or rather, it does not unless the individual were to know that the genotype demonstrated to be predictive of colon cancer happened to be the one he or she carried (Johnstone and Kaye, 2004).

One way to resolve all of this is to be quite explicit up front about what individuals will be fed back and what they will not be fed back, and what participants should do if they have further (medical) questions. However, this does not entirely address the point of principle: do individuals have a right to the data that concerns their health and genetic constitution? Do the researchers have a duty to give it to them? Within a medical relationship, doctors are entitled to withhold information from patients if it may be misinterpreted or if it would be psychologically damaging to the patient to receive a piece of information, which they are ill-prepared to cope with. This entitlement is somewhat controversial even within medicine (it is the so-called 'therapeutic privilege'). Where the researcher has no medical relationship with the participant it is even less clear what the answer is. Current practice is not to feed back individual data directly to the patient, but this may change. In the present context, what is clear is that participants who think they have an informal incentive to participate in a research project, which is that they will get early access to innovative tests, need to be clearly told what they can and cannot expect to receive and why.

This takes us to the most complex issue of all, what is the nature of the information participants should receive in order that they can give a valid consent. The importance of information is twofold. First, the quality of the information and style in which the information is presented has a major influence on the potential participant's ability to understand what is being proposed, and consequently on the likelihood that they agree to take part. Second, the information given defines what it is the participant is consenting to, and hence what that consent authorises the researcher to do in terms of investigations, and how data and samples may be used in research. The latter issue has given rise to heated debate.

One important feature of much epidemiological research is that data and samples hold much of their value because they can be reused and reanalysed, either directly or in combination with data collected for other purposes. Thus, a tissue sample collection created for research on the consumption of salt and cardiovascular disease could be useful to a researcher interested in the vascular aspects of Alzheimer's disease. Yet the consent taken for building the collection for the first purpose would not necessarily authorise the use of the samples for the second purpose. A sample collection obtained and managed by one research group could be useful to a different research group, possibly in another country. A sample collection built up with public money could be transferred to a spin-out company created by publicly funded researchers, and the samples become a valuable

commercial resource; yet the samples were collected on the basis that these would be a public resource rather than the capital for a commercial venture. In each case, the consent volunteered by research participants may not authorise these changes of use, yet each change might be seen by the researchers (and indeed by ethics committees and research funders) as useful ways of getting the best value out of their investment in building up the sample resource.

The problem here is that consent, to be valid in law, needs to be quite specific, whereas what researchers typically need is a consent that is quite broad and durable in terms of what it authorises in the short and long term. Different approaches have been tried to get around this problem. One approach is to try to create 'broad' consents, which allow participants to consent to a class of uses of their data and sample, rather than to only a narrow and specific use within a tightly defined protocol. The broad consent is supported by independent oversight by a research ethics committee and a governance process, which protects the interests of sample donors or data subjects in place of the protection, which a series of narrow and specific consents would give. This is largely the approach taken by UK Biobank. Another approach is to require a new consent each time a new use of the samples is proposed. In practice this would normally require each individual to be recontacted by the research team. This is advantageous in terms of ensuring that each new investigation is formally authorised by each participant, and it provides a mechanism for the participant to leave, by reminding them that their continued participation is optional. On the other hand, it is cumbersome and expensive, may cause attrition in the study population, and may even be burdensome on participants who are willing to continue as long as they are not bothered too often. This approach is taken in longitudinal studies where recontact would take place in the normal course of events, for new questionnaires or invasive investigations to take place. But in most epidemiological research, where direct contact with participants is unusual after initial recruitment, this approach is unpopular for pragmatic reasons.

The broad consent approach seems popular with the research community, and in many ways is supported by empirical evidence about what research participants want to know. What it does not address is another issue, viz. the ways consent addresses the issues of most concern to research participants. Consent to participate in research, being developed on the basis of the risk-control model appropriate to clinical trials, tends to focus on the risks and benefits of research participation, the nature of the investigations a participant will undergo, and other issues concerned with the personal safety and integrity of the individual. However, for many research participants these are not the only important issues. At least some research participants are interested in issues such as the possible commercialisation of research findings, the possible uses of their data or samples in research of which they may not approve. For instance, while a participant might be entirely happy for their samples to be used in genetic research in cardiovascular disease, they might disapprove of research in the genetic basis of intelligence, and thus the reuse of their samples donated for the former by researchers working on the latter. Now, it is reasonable to say that although participants are the donors of the samples, their donation of the samples involves ceding control of those samples (ownership, if you like, although notions of property in biological samples are controversial too), and beyond certain limits they have no further say in what is done with those samples. Nevertheless, protecting and promoting trust in the research enterprise and in specific researchers may involve giving assurances to sample donors about the kinds of use that are foreseen in the long term,

beyond the terms of narrow and specific consent in the short term and what kinds of use are excluded. UK Biobank, for example, supplements specific informed consent with a commitment to long-term engagement with participants through public announcements, websites, newsletters and other methods.

41.2.4 Confidentiality and Security

One of the more practically important issues of concern to participants, which can have major consequences for the governance of research sample collections, is confidentiality (Laurie, 2002). A general principle of the management of confidential information, such as medical records and genetic information, is that those who have access to it should have access to it only on a need to know basis. This protects the subject of the information from disclosure of information to parties who should not have access to particular items of information. In the context of research, this means that personal information collected or extracted for research purposes should be recorded in a way which minimises the extent to which the information allows identification of the individual data subjects. There is a tension in epidemiological research between the protection of individual privacy and confidentiality by deidentification measures up to full anonymisation, and retention of sufficient information that would allow informative analysis of data-sets, linkage of different items of information, and (re)contact of individuals (for research purposes, or to disclose clinically relevant information to the individual).

Two approaches are popular. One is to protect privacy through recording information in minimally identifying form, relative to the kind of information required for the research project (or sample bank). The advantage of this approach is that it builds in privacy protection in a once-for-all way. The disadvantage is that it is relatively inflexible to changes of protocol or for reuse or reanalysis. The other approach, often combined to some extent with the first, is to protect privacy through access controls and through coding so that linkage between records can be managed on a limited basis but database queries are not permitted, which would allow identification of individuals. This is a technology-based approach, although it is usual for senior management of the project, or, if it exists, the external governance board, to 'hold' the key, which allows deanonymised relinkage, and that this is done only in accordance with a defined protocol. One concern with this approach is that technology-based solutions may be subject to attack, and that in some ways they provide false reassurance about the security of a system which can always be subject to human intervention or human error.[4]

41.3 STEWARDSHIP

Moving away from research participant related ethical issues, we turn to the value of the research resource created in research, which could be a data-set, a sample collection or indeed a research protocol, which partners can sign up to in whole or in part. Much debate has centred on the question of whether and how intellectual property rights should be vested in such resources, and on how the value of these resources can be maintained in the long term. There is no straightforward answer to these questions, but there is consensus on one point, which is that resources created with public or charitable money should normally be regarded as open access resources, and that licensing fees (if any)

should only be levied to cover the costs of processing the requests and upkeep of the research resource, on a non-profit basis. Open resources such as these typically also stipulate that exclusive access licenses will not be granted, and often some requirement is made that research findings and sometimes any additional data or samples collected in the course of licensed research using the resource should be deposited with the resource for future users to access. Some of the challenges here then turn not on use of and access to the research resource, but in finding ways to guarantee its long-term viability. Samples need to be stored, which costs money; and access systems also need maintenance and oversight. At the end of a project lifetime, or when key personnel leave or retire, sample or data collections can fall into disuse or disrepair. Most public research funders (such as the UK Medical Research Council) now require detailed plans to be made for the stewardship of research resources beyond the lifetime of their established funding or identified management.

This issue of the stewardship of a collection relates to wider issues of scientific research integrity: scientists are generally expected to share their data and teach their techniques. This is for three reasons. Firstly, this provides external researchers with a way of checking that research results are valid and non-fraudulent. Second, it allows for the sharing of best practice, so as best to allow the rapid development of science. Third, it encourages the sense of there being a 'community of science' engaged in a common endeavour to advance humanity. Many critics of the commercialisation of science, especially the use of restrictive contracts and intellectual property rights, are generally concerned that this process undermines the three objectives stated here. On the other hand, commercialisation provides its own incentives for entrepreneurship and ingenuity, and can provide wider social benefits in terms of stimulating the technological application of scientific knowledge.

41.3.1 Benefit Sharing

As well as the value of the research resource to researchers, and in many cases to commercial companies, a database or sample collection represents an investment of time and effort on the part of the participants. In many settings, particularly in the developing world, there is a strong sense that researchers owe a duty of reciprocity to their participants to share any financial or clinical benefits of research with their host communities. This may be negotiated as part of 'community consent'. In developed world settings, researchers normally argue that they are dealing with individuals, that the contribution of any given individual on an identifiable basis to a particular project is minimal, and that it is the collected research resource, which has value as a composite. In addition, the research may have been hosted using public infrastructure, and any commercial benefits deriving from the infrastructure are taxed in such a way as to ensure the reinvestment of part of the proceeds in the State. However, these arguments carry less weight where researchers are doing research in a resource poor setting in another country. In such situations it is arguable that the ratio between benefit accruing to the researcher and sponsor and that accruing to research participants is far out of balance, and to expect participants to donate samples altruistically when this may be the only exploitable resource they have is unreasonable and unfair. Benefit sharing agreements may be quite complex and difficult to enforce, especially when they are made between groups and researchers rather than between individuals and researchers. Sometimes benefit sharing may involve money

payments, but more often involve benefits in kind, such as the provision of hospital facilities to a community, or other medical services (Parry, 2004).

41.3.2   Community Involvement

Community involvement has been advocated increasingly in recent years (Hansson, 2005; Haimes and Whong-Barr, 2004). In some developing world settings, community consultation (sometimes involving 'community consent') has been seen as essential to the 'license to do research' in a setting, partly in view of the history of colonial exploitation of poor or vulnerable communities. In the developed world, community consultation has been seen as a useful method for building support for a project, encouraging recruitment, and allowing feedback to participants. On occasion, researchers allow community consultation to play a part in the governance of a project, although a more common approach is to have one or two community members as members of the project steering committee or ethics governance board.

41.4   WIDER SOCIAL ISSUES

Aside from the impact of the research on individual participants or on host communities, there are wider social issues raised by genetics research. As noted at the beginning of this chapter, genetics has many social and ethical issues associated with it. For present purposes, I will concentrate on just two: geneticisation and race/ethnicity in genetics research.

41.4.1   Geneticisation

Geneticisation is the process of transforming diseases or other physiological or psychological traits into traits defined by their genetic causes or risk factors. This is a complex social process, rarely involving straightforward genetic reductionism, but often involving a focussing of attention on the genetic factors influencing a trait and away from other social or biological factors. Genetic reductionism is, roughly, the attempt to explain any physiological or psychological trait exclusively or principally on the basis of genetic factors (Moss, 2003; Sarkar, 1998). For example, starting from the observation that most crimes of violence are committed by men, one could seek to show that the explanation of this is genetic (possibly invoking evolutionary mechanisms) and then to identify specific genes (or gene variants), which account for this, and then proceed to identify individuals with these genes (or gene variants) as being potentially violent criminals, and managing them accordingly. Geneticisation of violent crime need not take as straightforward an approach as this (Wasserman and Wachbroit, 2001). It might well recognise that many factors other than genes are relevant in the causal pathway to violent criminal acts, but nonetheless emphasise that genes are possibly the factor most mutable and controllable in a societal response to violent crime. To take a more medical example, there are many risk factors for cardiovascular disease, from diet, to stress, to genetic constitution. A geneticised approach to cardioprevention would focus on identifying alleles creating higher than average risk of cardiovascular disease, and screen for the presence of these alleles in a population, and tailor prevention strategies accordingly. This would contrast with a more classical public

health approach of changing the behaviour of an entire population in order to reduce the number of cases of heart disease (Khoury *et al.*, 2000).

Geneticisation is clearly a process of some importance, and to the extent that it is more sophisticated than simple genetic reductionism, and allows for non-deterministic causality between gene and phenotype, a language of risk rather than certainty, and the role of other social and biological mediating factors, it causes less concern than genetic reductionism. Genetic reductionism can be criticised on many grounds, but socially the principle risk is that by focussing on heritable characters, it causes increased attention to be paid to reproductive policy and less attention to be paid to social policy. Geneticisation need not have this consequence. Nevertheless, attention to genetic factors over social factors may have the consequences of increased emphasis on individual health and behaviour over social conditions, increased emphasis on personal responsibility for health rather than on collective responsibility for health services and environmental conditions, emphasis on high technology interventions over low-cost social interventions and so on (Novas and Rose, 2000; Hedgecoe, 2001; ten Have, 2001).

While these criticisms have merit in many cases, they risk downplaying the utility that genetic research can have even in social public health. In some respects it is better to think of the geneticisation thesis as a descriptive claim about an ongoing social process rather than as a direct normative critique of such a process. That said, it does capture a worry that is frequently expressed about an individualistic trend in modern healthcare. Interestingly, this contrasts with the arguments frequently raised by genetics researchers about the social basis of their research in trust, community cooperation, benefit sharing and open source data collection and so on.

41.4.2   Race, Ethnicity and Genetics

Any scientific research on racial or ethnic difference is inherently controversial, but genetic research, because of the history of eugenics and biological racism, is especially so (Macbeth and Shetty, 2001; Ellison and Goodman, 2006; see also **Chapter 31**). Central elements of the controversy include the following. Firstly, there is controversy over whether any biological sense can be attributed to socially prevalent conceptions of race, and hence whether scientific inquiry into purported biological concepts of race can have any rational justification (Marks, 1995; Nature Genetics, 2004). Second, even allowing that some biological concept of human variation along race-like lines can be defended, there is a vexed question over whether the biological concepts map onto the concepts used in ordinary social life (Smart *et al.*, 2006; Royal, 2006). If they do, can science be seen as supporting social attitudes to racial difference which may be morally and politically problematic, and if they do not, does using a language so open to misinterpretation not confuse issues in a dangerous way (Ashcroft, 2006). Third, assuming some rigorous and value-neutral concept of racial or ethnic difference can be established, which is biologically useful, do the ways in which these findings are then applied in medicine and applied science make sense? And so on. One of the lessons of this complex debate is quite general across human genetics research, which is that genetic research into the biological bases of traits which are complex in their structure and in their meanings in society is very difficult to carry out with 'clean hands'. This raises large issues about whether certain questions should not be investigated at all, or only with great care, and

what 'approaching an issue with great care' means in terms of the social responsibilities of scientists.

41.5 CONCLUSIONS

This chapter has necessarily been highly selective and rather discursive. It intends to give a summary overview of some of the more important practical and social issues in the conduct of statistical genetic and genetic sample based research. Although some of the issues discussed are complex and confusing, three things remain clear. First, that public trust in medical and biological research remains relatively high, in part because of the care taken to engage with them about science and to maintain high ethical standards. Although many of the issues in this chapter may seem frustratingly unresolved at the level of theory, in practice there is considerable consensus about many of them. For example, the UK Biobank Ethics and Governance Framework has been published for about a year at the time of writing, following extensive consultation with the academic community, commissioned surveys and focus groups, and direct public consultation, and seems to have general support. Second, there is now an extensive and growing literature of philosophical, ethical, legal and empirical research which can help frame and illuminate the issues and help policy-makers, scientists and the public resolve them. Third, that it remains crucial that scientists remain engaged with these debates, both in informing them and in steering them in directions which will both control risks to participants and society, and maximise the benefits that genetic research will generate.

REFERENCES

Archard, D. (2004). *Children: Rights and Childhood*. Routledge, London.

Ashcroft, R.E. (2006). In *Race in Medicine: From Probability to Categorical Practice*, T.H. Ellison and A.H. Goodman, eds, *The Nature of Difference: Science, Society and Human Biology* CRC Press/Taylor & Francis, Boca Raton, FL, 135–153.

Ashcroft, R.E., Jones, S. and Campbell, A.V. (2000). Solidarity in the UK welfare state reforms. *Health Care Analysis* **8**, 377–394.

Ashcroft, R.E. and Hedgecoe, A. (eds) (2006). Genetic databases and pharmacogenetics: social, policy and ethical issues. Symposium. *Studies in History and Philosophy of Biological and Biomedical Sciences* **37**, 499–601.

Bauer, M.W. and Gaskell, G. (eds) (2002). *Biotechnology: The Making of a Global Controversy* Cambridge University Press, Cambridge.

Benn, P. (1998). *Ethics*. UCL Press, London.

Beyleveld, D. and Brownsword, R. (2001). *Human Dignity in Bioethics and Biolaw*. Oxford University Press, Oxford.

Clarke, A. and Ticehurst, F. (eds) (2006). *Living with the Genome: Ethical and Social Aspects of Human Genetics*, Palgrave MacMillan, Basingstoke and New York.

Ellison, G.T.H. and Goodman, A.H. (eds) (2006). *The Nature of Difference: Science, Society and Human Biology*, CRC Press/Taylor & Francis, Boca Raton, FL.

ten Have, H.A. (2001). Genetics and culture: the geneticization thesis. *Medicine, Health Care and Philosophy* **4**, 295–304.

Gibbons, S.M.C., Helgason, H.H., Kaye, J., Nõmper, A. and Wendel, L. (2005). Lessons from European population genetic databases: comparing the law in Estonia, Iceland, Sweden and the United Kingdom. *European Journal of Health Law* **12**, 103–133.

Gibbons, S.M.C., Kaye, J., Smart, A., Heeney, C. and Parker, M. (2007). Governing genetic databases: challenges facing research regulation and practice. *Journal of Law and Society* **34**(2), 163–189.

Haimes, E. and Whong-Barr, M.T. (2004). Key issues in genetic epidemiology: lessons from a UK based empirical study. *TRAMES* **8**, 150–163.

Haimes, E. and Williams, R. (2006). *Review of Research on Human Biological Sample Collections*. Wellcome Trust, London.

Hansson, M.G. (2005). Building on relationships of trust in biobank research. *Journal of Medical Ethics* **31**, 415–418.

Hedgecoe, A. (2001). Ethical boundary work: geneticization, philosophy and the social sciences. *Medicine, Health Care and Philosophy* **4**, 305–309.

Indridason, A. (2004). *Tainted Blood*. Harvill, London.

Jackson, E. (2006). *Medical Law: Text, Cases and Materials*. Oxford University Press, Oxford.

Johnstone, C. and Kaye, J. (2004). Does the UK Biobank have a legal obligation to feedback individual findings to participants? *Medical Law Review* **12**, 239–267.

Khoury, M.J., Burke, W. and Thomson, E.J. (eds) (2000). *Genetics and Public Health in the 21*st *Century: Using Genetic Information to Improve Health and Prevent Disease*, Oxford University Press, Oxford.

Laurie, G. (2002). *Genetic Privacy: A Challenge to Medico-Legal Norms*. Cambridge University Press, Cambridge.

Macbeth, H. and Shetty, P. (eds) (2001). *Health and Ethnicity*, Routledge, London.

Marks, J. (1995). *Human Biodiversity: Genes, Race, and History*. Aldine de Gruyter, New York.

Marks, S.P. (ed) (2006). *Health and Human Rights: Basic International Documents*, Harvard School of Public Health and Harvard University Press, Cambridge, Mass.

McHale, J.V. (2004). Regulating genetic databases: some legal and ethical issues. *Medical Law Review* **12**, 70–96.

Medical Research Council (1991). *The Ethical Conduct of Research on the Mentally in Capacitated*. Medical Research Council, London.

Moss, L. (2003). *What Genes Can't Do?* MIT Press, Boston.

Nature Genetics (2004). *Genetics for the Human Race* **36**(Suppl. 1), S1–S60.

Novas, C. and Rose, N. (2000). Genetic risk and the birth of the somatic individual. *Economy and Society* **29**, 485–513.

O'Neill, O. (2002). *Autonomy and Trust in Bioethics*. Cambridge University Press, Cambridge.

Parry, B. (2004). *Trading the Genome: Investigating the Commodification of Bioinformation*. Columbia University Press, New York.

Paul, D.B. (2002). Is human genetics disguised eugenics? In *The Philosophy of Biology*, D. Hull and M. Ruse, eds, Oxford University Press, Oxford. pp. 536–551.

Rabinow, P. (2003). *French DNA: Trouble in Purgatory*. University of Chicago Press, Chicago.

Rhodes, R. (2005). Rethinking research ethics. *American Journal of Bioethics* **5**(1), 7–28.

Ross, L.F. (1998). *Children, Families and Health Care Decision-Making*. Oxford University Press, Oxford.

Royal, C.D.M. (2006). 'Race' and ethnicity in science, measurement and society. *Biosocieties* **1**, 325–328.

Sarkar, S. (1998). *Genetics and Reductionism*. Cambridge University Press, Cambridge.

Sherlock, R. and Morrey, J.D. (eds) (2002). *Ethical Issues in Biotechnology*, Rowman and Littlefield, Lanham, MD.

Smart, A., Tutton, R., Ashcroft, R.E., Martin, P.A. and Ellison, G.T.H. (2006). Can science alone improve the measurement and communication of race and ethnicity in genetic research? Exploring the strategies proposed by *Nature Genetics*. *Biosocieties* **1**, 313–324.

Sulston, J. and Ferry, G. (2002). *The Common Thread: Science, Politics, Ethics and the Human Genome*. Corgi, London.

Tutton, R. and Corrigan, O. (eds) (2004). *Genetic Databases: Socio-Ethical Issues in the Collection and Use of DNA*, Routledge, London.

UK Biobank Ethics and Governance Council (2006). UK biobank ethics and governance framework, version 2.0, `http://www.ukbiobank.ac.uk/ethics/efg.php` (accessed 22-01-2007).

UK Pharmacogenetics Study Group (2006). Policy issues in pharmacogenetics `http://www.york.ac.uk/res/pgx/publications/` (accessed 22-01-2007).

Wasserman, D. and Wachbroit, R. (eds) (2001). *Genetics and Criminal Behavior*, Cambridge University Press, Cambridge.

42

Insurance

A.S. Macdonald

Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, UK

As soon as DNA-based genetic testing became available, questions were asked about its use by insurers to discriminate against carriers of deleterious alleles. The statistical interest in these questions lies mainly in the application of genetic epidemiology to the actuarial models used to price life and health insurance. This chapter surveys the relevant research, including early work mostly focussed on single-gene disorders, and more recent attempts to predict the relevance of multifactorial genetics to insurance practice.

42.1 PRINCIPLES OF INSURANCE

42.1.1 Long-term Insurance Pricing

Insurance is an unusual branch of commerce because the basis of its sound operation is mathematical in nature. Just as airlines (say) must take the mathematics of aerodynamics as they find it, insurance companies may be in peril if they ignore the actuarial mathematics and statistics that govern their businesses. But actuaries do not deal with precise and impersonal qualities like airflow over a wing; they deal with people and their personal attributes, that lead to an assessment of the risk of making a claim under an insurance contract. Some personal attributes (sex, race, disability) give rise to stronger sensitivities than others (age, weight, smoking habits) and the experience of the last 10 years suggests that any kind of personal genetic information falls in the first group.

The mathematical nature of insurance, and life insurance in particular, is straightforward. The simplest contract, called *whole-of-life insurance*, pays an agreed sum (the sum assured) immediately on the death of the insured person. Suppose a person wishes to buy whole-of-life insurance with sum assured S. The problem is to determine the premium P to be paid at outset (in practice, premiums are usually payable monthly but we ignore this complication). Death being certain, so the insurer will pay out S at some future time with certainty. Intuitively the 'fair' premium before allowing for expenses, profit, etc. appears

to be S. However, this takes no notice of the insurer's ability to invest the premium P at interest. Suppose the insurer can earn interest at rate i per year on invested funds. If the insured person dies after exactly T years, the insurer will then have assets of $P(1 + i)^T$, and will have to pay out S. The 'fair' premium that balances the books is clearly:

$$P = S(1 + i)^{-T}. \tag{42.1}$$

Future cashflows discounted to allow for interest, such as $S(1 + i)^{-T}$, are called *present values*. Equation (42.1) is an example of the actuarial 'principle of equivalence', in which the present values of future income and future outgo are equated, giving an equation to be solved for the unknown P. The problem here is that, except in very unusual circumstances, T is not known in advance.

The problem is resolved by regarding T as a random variable, taking non-negative values, possibly with an upper limit to represent a feasible lifespan. Then $(1 + i)^{-T}$ is also a random variable, and the insurer's books will balance 'on average' if the premium charged is the expected value $E[S(1 + i)^{-T}]$. In other words, if the timing of any future cashflows is uncertain, we equate the expected present values (EPVs) of income and outgo.

This simple principle is widely applied in practice to insurances extending over very long terms, typically covering the event of death (life insurance), onset of a serious disease (critical illness insurance), inability to work (income protection or disability insurance) and the need for long-term care (long-term care insurance). In all cases, the times or ages at which the insured events may occur is a crucial element of the pricing problem. This is what distinguishes life and related insurances from other familiar contracts such as home or motor insurance, which usually do not extend cover beyond a year or two. In the following, we will use the generic term *life insurance* to include all long-term insurance covering illness, disability or death.

The opportunity to invest funds at interest is only one reason why the times of insured events are important. Most life insurance contracts do not run for life, but expire at some advanced age, often an anticipated retirement age. This introduces the possibility that no payment will be made at all. For example, Table 42.1 shows the probabilities that a healthy man of selected ages will die within 35 years, according to the AM92 life table (based on the observation of men with certain types of life insurance contracts in the United Kingdom in 1991–1994). This shows that many insurances will cover quite rare events, whose occurrence or non-occurrence depends on the random future lifetime T, or its analogues. It is evident that good estimates of the distribution of T (denoted $F_T(t)$) and its analogues are needed. Table 42.1 also shows that the financial consequences of inaccurate estimation of $F_T(t)$ could be severe; e.g. attributing, to a person of any given

Table 42.1 Probability that a healthy man will die within 35 years, based on the AM92 life table.

Age	Probability
20	0.042
25	0.067
30	0.111
35	0.186

age, the mortality of a person 5 years younger, would erroneously reduce the probability of their death within 35 years by about one-third, with a corresponding effect on premiums.

42.1.2 Life Insurance Underwriting

Estimating $F_T(t)$ is the basic problem of survival analysis; $F_T(t)$ itself is described in a life table. Although life tables are often based on very large data sets (such as national registries or insurance companies' records) it is evident at once that heterogeneity is an issue. The following are some of the factors known to influence mortality and morbidity (some may be mutually confounding): age, sex, socio-economic class, nationality, education, occupation, diet, smoking and drinking habits, exercise, obesity, medical history, relatives' medical histories, marital status, whether insured or not. For clarity, we suppose individuals are labelled $i = 1, 2, \ldots$, and with the ith person we associate a vector $\mathbf{z}_i = (z_i^1, z_i^2, \ldots, z_i^r)$ representing the values of these or other covariates. The distribution of T is now $F_{T|\mathbf{z}}(t|\mathbf{z})$, depending on the covariates.

Underwriting is the process of obtaining information about risk factors (covariates) and adjusting premiums accordingly. Faced with a multitude of possible risk factors, the underwriter will usually focus on a small number, chosen because of known major effect, correlation with other risk factors (hence suitability as a proxy), and availability of data. Using more risk factors than necessary leads to extra expense. Suppose the first u risk factors are chosen as the basis for underwriting, and define $\mathbf{z}_i' = (z_i^1, \ldots, z_i^u)$ and $\mathbf{z}_i'' = (z_i^{u+1}, \ldots, z_i^r)$. Let the distribution of $\mathbf{z}''$ given $\mathbf{z}'$ be $F_{\mathbf{z}''|\mathbf{z}'}(\mathbf{z}''|\mathbf{z}')$, among those assumed to buy insurance. Suppose the present value of some insurance cashflow depending on T is $\pi(T)$. Then the premium that the ith person ought to be charged under the equivalence principle is the EPV:

$$P(\mathbf{z}_i') = \int_{\mathbf{z}''} \int_t \pi(t) \, \mathrm{d}F_{T|\mathbf{z}}(t|(\mathbf{z}_i', \mathbf{z}'')) \, \mathrm{d}F_{\mathbf{z}''|\mathbf{z}'}(\mathbf{z}''|\mathbf{z}_i'). \tag{42.2}$$

The actual calculation does not go quite like this, since $\mathbf{z}''$ is generally unobserved. Define:

$$G_{T|\mathbf{z}'}(t|\mathbf{z}') = \int_{\mathbf{z}''} F_{T|\mathbf{z}}(t|\mathbf{z}) \, \mathrm{d}F_{\mathbf{z}''|\mathbf{z}'}(\mathbf{z}''|\mathbf{z}'). \tag{42.3}$$

Then the calculated premium ought to be:

$$P(\mathbf{z}_i') = \int_t \pi(t) \, \mathrm{d}G_{T|\mathbf{z}'}(t|\mathbf{z}_i'). \tag{42.4}$$

It is possible that more than u risk factors were modelled, given past data, and as a result of the usual model selection process $\mathbf{z}'$ was chosen as having adequate explanatory power. Then the reduction step in (42.3) is, in principle, carried out explicitly and could be revisited if needed. It is equally possible that only the risk factors $\mathbf{z}'$ were ever investigated, so $G_T(t; \mathbf{z}')$ was estimated directly and (42.3) represents completely unobserved averaging.

Viewed as a prediction problem, (42.2) and (42.4) give correct answers only if the distributions that appear in them are appropriate. Specifically: (1) the distribution that *has* been estimated, $G_{T|\mathbf{z}'}(t|\mathbf{z}')$, is the same in the future as it was in the past, or changes in a predictable way; and (2) the density that averages out the unobserved heterogeneity, $F_{\mathbf{z}''|\mathbf{z}'}(\mathbf{z}''|\mathbf{z}')$, is unchanging over time. We know that (1) is far from true, because there have

been great changes in mortality rates over calendar time. Fortunately these have mostly been decreases (longer lifetimes than predicted) so errors have been profitable for sellers of life insurance. (Not so for sellers of pensions and annuities.) The truth, or otherwise, of (2), is impossible to establish. It may be summed up as follows: premiums will be charged correctly, if the underlying characteristics of the population who actually buy insurance in future, are the same as those of the population used to estimate $G_{T|\mathbf{z}'}(t|\mathbf{z}')$. This explains why insurers often prefer to estimate life tables based on their own experience, rather than external data, even if the latter is much more extensive. The insurer may not know (for example) that heavy smokers are overrepresented among their customers. But as long as they use life tables based on their own data (or adjust industry standard tables appropriately) and as long as this behaviour does not change, they do not need to know. However, ignorance of this kind can be uncomfortable.

If, for some reason, the distribution of the unobserved risk factors among those who actually do buy insurance is $F^*_{\mathbf{z}''|\mathbf{z}'}(\mathbf{z}''|\mathbf{z}')$ rather than $F_{\mathbf{z}''|\mathbf{z}'}(\mathbf{z}''|\mathbf{z}')$, leading to a different theoretical premium $P^*(\mathbf{z}'_i)$, a profit or loss will arise as a result of actually charging $P(\mathbf{z}'_i)$.

42.1.3 Familial and Genetic Risk Factors

Some of the most important risk factors used in life and health underwriting arise from an applicant's personal history, which we interpret broadly to include lifestyle and habits as well as medical history. The most common approach is to charge different premiums based on a small number of well-understood risk factors such as age, sex and smoking habits, but otherwise to limit closer investigation of cases to a small proportion indicated by significantly adverse responses to questions about health and medical history. However, in more recent times 'preferred lives' underwriting has spread quite widely, in which insurers more actively seek out persons with an advantageous risk profile. Leigh (1990) gives a review of underwriting practice, while Brackenridge and Elder (1998) is a standard reference on life insurance underwriting.

Since nearly all life and health insurance is bought by economically active adults to protect dependents or to secure financial transactions, the genetic disorders that are relevant are those with late onset, such that a person may be completely healthy at ages when insurance is purchased. Thus most interest is focused on a small set of relatively rare single-gene dominantly inherited disorders. In 1996, the genetics advisor to the Association of British Insurers listed eight disorders as relevant to insurance, namely Huntington's disease, familial adenomatous polyposis, hereditary breast cancer, early-onset Alzheimer's disease, myotonic dystrophy, multiple endocrine neoplasia, hereditary motor and sensory neuropathy and adult polycystic kidney disease. The last of these was soon removed, since it is normally diagnosed by ultrasonography rather than by a DNA test.

42.1.4 Adverse Selection

Losses arising through unobserved heterogeneity are called *adverse selection*, a term in widespread use in economics. It can arise accidentally, or as a result of one party to a contract deliberately withholding information from the other, or through the actions of competitors in the market. For example, a motor dealer may conceal a car's poor service record, or their customer may conceal their own poor credit record. Similarly, a life insurer

may conceal high charges behind complicated terms and conditions, or the buyer of life insurance may fail to mention a poor health record. Sellers try to protect themselves by demanding information – e.g. credit checks and insurance underwriting – with sanctions if it is not given honestly and accurately. Buyers, who are often relatively powerless individuals, look to laws, regulations and professional advisers for their protection.

A classic example of the insurance industry's reaction to the mere possibility of adverse selection arose in the United Kingdom, when in 1981 one life insurer introduced separate premium rates for smokers and non-smokers. Those for non-smokers were lower than any competitors' premium rates (which was the point of the exercise) and those for smokers were correspondingly higher. Other companies faced the prospect of losing all their non-smoking customers to the innovator, and attracting an all-smoking clientele, while charging premiums that assumed a 'normal' mix of smoking habits. Within a few years, almost all companies charged different rates to smokers and non-smokers. There is no evidence that adverse selection actually did appear, but the threat was enough.

The question of why someone wants to purchase insurance lies behind insurers' concerns about adverse selection. By its nature, insurance covers reasonably rare events, so a modest premium can secure a large amount of cover. Someone who knows they are at high risk – either because they possess information and realise its implications, or because they intend to commit fraud – undermines the sound business model. Not surprisingly, insurers tend to resist suggestions that they should eschew the use of any risk factor they think might be relevant, or that they be instructed to move a covariate from $\mathbf{z}'$ over to $\mathbf{z}''$. In reality, this only matters if the risk factor in question is likely to influence the desire for insurance and motivate the informed individual to purchase it. Sometimes this possibility is very plausible: intimations of mortality as we grow old mean that life insurance without age discrimination would hardly be a feasible proposition. Sometimes it is rather implausible: most smokers are aware that, as a group, they die earlier than non-smokers but: (1) does that knowledge impel them to buy life insurance? and (2) if not, would the sound basis of the business be threatened if we disallowed discrimination against smokers?

The question that has fuelled the genetics and insurance debate is the following: where does personal medical information in general, and genetic information in particular, lie on this spectrum?

42.1.5 Family Medical Histories

For a long time before DNA-based genetic tests were developed, insurers had routinely used family medical history in underwriting. The chief source of information was (and is) a question on the proposal form which typically will ask about: (1) the ages at death and causes of death of the applicant's parents; and (2) whether a parent or sibling has suffered from a disorder that might be inherited. Inquiries rarely if ever extend beyond first-degree relatives, partly because of the difficulty of verifying and interpreting extended pedigrees.

Such family medical histories may disclose a risk of a Mendelian disorder, but they are also strongly predictive of the risk of common diseases, such as coronary heart disease (CHD). In the latter case, the familial link need not be genetic at all, but could lie in shared environment, habits or socio-economic class.

42.1.6 Legislation and Regulation

Sex and disability are qualities that have, in the past, been associated with discrimination that is now deprecated in most modern societies, to the extent of being outlawed in many jurisdictions. However, it is common for insurance companies to be allowed to use sex and disability as risk factors, provided they can show evidence of a statistical or actuarial nature that confirms their relevance. In some countries with such provisions, a commission or tribunal may assess the evidence; in others the courts.

The question of whether various types of genetic information should be given similar protection has been widely considered. Nys *et al.* (2002) give a recent survey of practice in the European Union. The cases of Australia, Sweden and the United Kingdom are perhaps of unusual interest, because in each country the task has fallen to a governmental or legal commission, with some important differences of approach.

1. In the United Kingdom, the Human Genetics Commission (HGC) advises the government on all aspects of genetics, including insurance questions. The insurance industry has, since 1996, agreed to a moratorium on the use of DNA-based test results in underwriting. Its current form, running from 2005 to 2011, has the following key features: (1) insurers will not ask anyone to take a genetic test; (2) insurers will not ask about test results acquired in the course of research trials; and (3) insurers will not ask about existing test results unless a life insurance policy exceeds £500 000 sum assured, or other forms of insurance exceed £300 000 sum assured or equivalent. However, the use of family medical history is unrestricted.

   The UK government has set up the Genetics and Insurance Committee (GAIC) which is responsible for deciding what genetic test results may be used for underwriting, when sums assured exceed the limits mentioned above. It will consider applications submitted to it by the insurance industry, against three criteria: (1) *technical relevance*: does the test accurately measure the genetic information? (2) *clinical relevance*: does a positive result in the test have likely future adverse implications for the health of the individual? (3) *actuarial relevance*: does a positive result justify increased premiums? Thus for the first time in the United Kingdom, discriminatory pricing has to be justified in advance, rather than defended in court if challenged. To date GAIC has approved only one test for use, that is for Huntington's disease in the case of life insurance.

2. In Sweden, similar arrangements are in place but family medical history may not be used. In addition, long-term health insurance for children is sold in Sweden, and genetic information may not be used at all in respect of these contracts. This is a rare example of insurance possibly covering early-onset and recessive disorders.

3. The Australian Law Reform Commission (ALRC) reviewed all aspects of human genetics and their reports (ALRC, 2001; 2002; 2003) are particularly useful references. In respect of insurance, they recommended a system quite similar to that in the United Kingdom.

Many of the social policy aspects of genetics and insurance are discussed in Daykin *et al.* (2003) and Doble (2001) and official reports such as ALRC (2001; 2002; 2003), HCSTC (1995; 2001), HGAC (1997) and HGC (2002).

42.1.7   Quantitative Questions

By its nature, insurance practice seeks to quantify risk, and for that purpose to build models of individual life histories. In this respect, actuaries have much the same interests as epidemiologists. Two quantitative questions address the debate that has grown up around genetics and insurance:

1. If insurers were allowed to use genetic information (including family history) to underwrite insurance premiums, by how much would those premiums increase?

2. If insurers were not allowed to use genetic information to underwrite insurance premiums, what additional costs might they face as a result of adverse selection?

42.2   ACTUARIAL MODELLING

42.2.1   Actuarial Models for Life and Health Insurance

Life and health insurances pay benefits either on the occurrence of an event, or during the continuation of some status. The simplest example is death: life insurance will pay a lump sum upon death, while an annuity or pension will continue to be paid until death. Other events or states that can lead to benefits being paid are suffering one of a specified list of severe illnesses (critical illness insurance); being unable to work (disability or income protection insurance) and inability to care for oneself (long-term care insurance). In each case, the actuary needs a model of an insured person's life history, including all relevant events.

The simplest such model was introduced in Section 42.1.1 to illustrate the statistical principle of insurance, namely that a person's remaining lifetime is represented by a positive random variable T. A traditional life table is simply a tabulation of the survival function $S_T(t)$ associated with T, defined by $S_T(t) = \text{P}[T > t]$. Clearly this is the complement of the distribution function $F_T(t) = \text{P}[T \leq t]$. Assuming T to have a density $f_T(t)$ on a suitable age range, the hazard rate $\lambda(t)$ is defined by $\lambda(t) = f_T(t)/S_T(t)$. The hazard rate has the following intuitive interpretation: conditional on being alive at time (or age) t, the probability of dying by time $t + \text{d}t$ is approximately $\lambda(t)\text{d}t$, for small $\text{d}t$. Figure 42.1 illustrates this model, representing death as a transition between an 'alive' state and a 'dead' state, governed by the hazard rate $\lambda(t)$.

Faced with more complicated types of insurance, other events have to be modelled and a useful approach is to represent events in a life history as transitions between suitably defined states. For example, in Figure 42.2 the 'able' state represents fitness to work, and the 'ill' state represents inability to work. Under the simplest disability insurance policy, the insured person would pay premiums while 'able' and receive regular benefits to replace lost earnings while 'ill'. It quickly becomes impractical to specify such extended models using random times between transitions – analogues of T – as the basic quantities. For
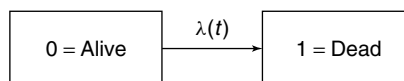
Figure 42.1   A two state model of mortality.
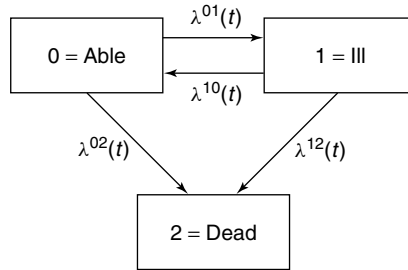
Figure 42.2 A model of illness and death.

example, the number of events (transitions) that may occur in one person's lifetime in the model of Figure 42.2 is not even bounded. Instead it is much more convenient to specify the transition intensities between each pair of states, here denoted $\lambda^{ij}(t)$ between states i and j. These are natural quantities to estimate using occurrence-exposure rates, and if they depend only on the insured person's age the resulting model is Markov and most quantities of actuarial interest can be computed numerically as the solutions of linear ordinary differential equations. We mention the two most important examples (see Hoem (1988) and Norberg (1995) for details).

1. Let $p^{ij}(s, t)$ be the probability of being in state j at time t, conditional on being in state i at time s ($s \leq t$). Then Kolmogorov's forward equations are:

$$\frac{\partial}{\partial t} p^{ij}(s, t) = \sum_{k \neq j} p^{ik}(s, t)\lambda^{kj}(t) - \sum_{k \neq j} p^{ij}(s, t)\lambda^{jk}(t). \qquad (42.5)$$

2. Suppose a lump sum $b^{ij}(t)$ is payable on transition from state i to state j at time t, and a continuous payment at rate $b^i(t)$ per year is made if the insured person is in state i at time t (whether these cashflows are payable to or from the insured person just depends on their signs; we assume positive cashflows are paid to the insured person). Let $V^i(t)$ be the amount the insurance company must hold in reserve at time t, to honour its future obligations, if the insured person is then in state i. (This is called the *prospective policy value*, and may be thought of as the expected value of the insurer's future outgo, net of premiums received, allowing for interest.) Suppose interest can be earned at the instantaneous rate δ *per annum*. Then Thiele's equations are:

$$\frac{\mathrm{d}}{\mathrm{d}t} V^i(t) = \delta V^i(t) - b^i(t) - \sum_{j \neq i} \lambda^{ij}(t)\big(b^{ij}(t) + V^j(t) - V^i(t)\big). \qquad (42.6)$$

In words, the reserve that must be held earns interest, but is depleted by the continuous payments made as long as the insured person remains in state i, and also by the lump sum payable on transition into any other state j. However, when such a transition occurs the reserve that was held, $V^i(t)$, is released, but the reserve $V^j(t)$ required by presence in the new state must now be held instead.
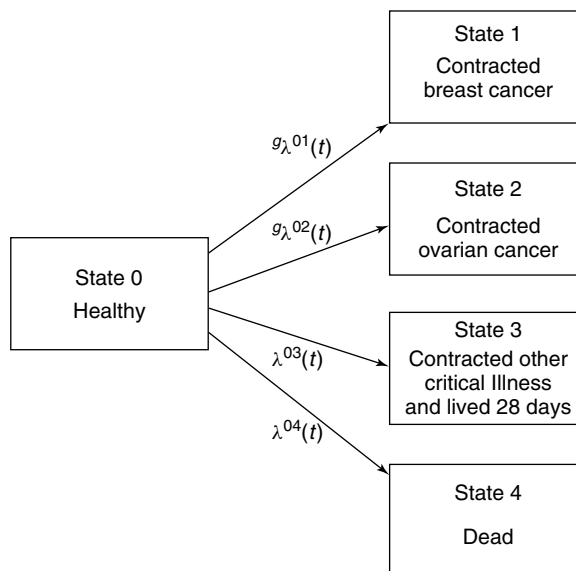
Figure 42.3   A model suitable for pricing critical illness insurance, with breast and ovarian cancer selected as particular causes of claims. The superscript 'g' represents genotype.

For completeness, to illustrate the other insurance contract of practical importance, Figure 42.3 shows a model suitable for pricing and reserving for critical illness insurance. Here, breast and ovarian cancers are picked out as particular causes of claiming; all other causes are combined into a transition into a single state. By letting the transition intensities that govern onset of breast and ovarian cancer depend on genotype g, we can measure the impact of known deleterious genotypes on insurance prices. Similar models can be set up in respect of any genetic disorder, in respect of which the epidemiology is advanced enough to furnish reasonable estimates of the onset rates. What might be regarded as 'reasonable' is a question that could have a bearing on the decisions made by GAIC in the United Kingdom.

42.2.2   Parameterising Actuarial Models

Actuarial researchers almost never have access to pedigree data. Rare exceptions occur when epidemiologists publish details of the pedigrees they have used (Gui and Macdonald, 2002a; Espinosa, 2006). Therefore, they rely on the epidemiological literature. Given the nature of the models described previously, the parameters needed are: (1) age-dependent rates of onset (equivalently, penetrance estimates); (2) duration-dependent rates of mortality post-onset (age dependence may have to be considered too); and (3) mutation frequencies.

Questions concerning known mutation carriers do not require knowledge of the mutation frequencies, but questions concerning incomplete information do. This arises in two circumstances:

1. The cost of adverse selection depends on the size of the group who may gain access to insurance below cost.

2. When a genetic disorder occurs as a rare subset of a common disorder (such as breast cancer) then a family history does not identify the presence of a mutation, it just alters the probability that a mutation is present.

Onset rates and mortality rates may be extracted from the literature by whatever means present themselves, including when necessary being read from graphs (Macdonald (2003b) discusses many of these issues). Occasionally point estimates of onset rates are given (e.g. Ford *et al.*, 1998) which can be used directly, but relative risks, Kaplan-Meier estimates and odds ratios are also common. In many cases ascertainment bias is likely to be present, that even the authors of the articles in question were unable to control, in which case the premium rates obtained from the parameterised model are overstated, and some form of sensitivity analysis is advisable, e.g. by assuming onset rates to be 50 % or less of those estimated.

42.2.3   Market Models and Missing Information

Models such as those illustrated in Figures 42.1 to 42.3 may be used to represent the life history of a person who possesses an insurance contract of the appropriate type. They are adequate for determining premiums and reserves, provided the model has been fully parameterised. However, they have to be extended to address questions of adverse selection or underwriting based on family history.

Adverse selection depends on whether a person decides to buy insurance at a given price, on the basis of the information they possess. Both the information and the decision, therefore, have to be represented in the model. Figure 42.4 illustrates such a model, for the simplest case of critical illness insurance. It represents a person's life history in an insurance market. They are assumed to have a known genotype g_i, and the label i indexes the states and the transition intensities. They start in state $i0$ at (say) age 20. As time passes, they may decide to buy insurance in the normal way (state $i1$) or they may decide to have a genetic test (state $i2$) and, once they know that their genotype is g_i, then decide to buy insurance (state $i3$). Once insured, they remain insured until some suitable age, say 60 or 65, during which time they will receive the benefit if they avoid death (state $i5$) and suffer a critical illness (state $i4$).
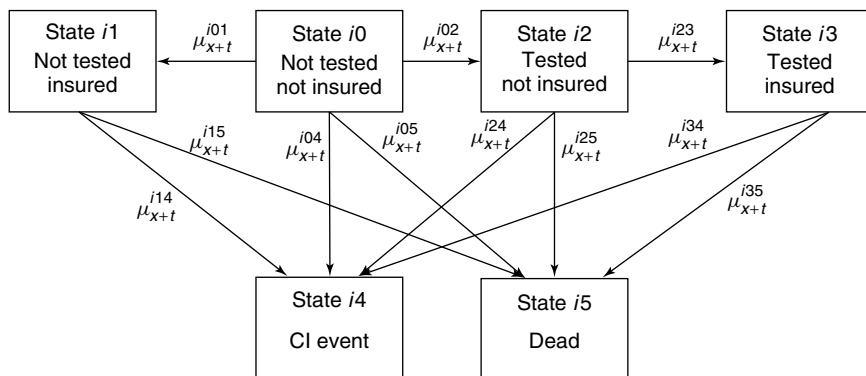
Figure 42.4   A Markov model of insurance purchase and critical illness (CI) insurance events for a person with genotype g_i.

This model assumes the genotype g_i to be fixed, even if not known to the individual at outset. Each possible genotype defines a subset of the population of people who might be active in the insurance market, and we set up such a model for each of them, with transition intensities adjusted to represent each genotype. The probabilities of being in any of the subgroups are just the population frequencies of the genotypes.

In fact we need to refine the partition of the population beyond that defined by genotype, for two reasons: (1) we wish to allow for underwriting based on family history; and (2) genetic testing is most likely to be taken up by people who have a family history of the relevant disorder. An example of such a refinement is illustrated in Figure 42.5, in the case of a dominantly inherited disorder of purely genetic origin (such as Huntington's disease). In such a case, we assume that 'family history' means an affected parent or sibling, and that a healthy person with a family history is 'at risk' but inherited the mutation with probability 1/2. The proportions assumed to be in each subgroup depend on the population mutation frequency, bearing in mind that members of both 'at risk' subgroups have an affected parent.

This model is flexible enough to answer many questions relating to single-gene disorders.

1. The rate at which insurance is purchased in the 'not at risk' subgroup determines the size of the insurance market.

2. The rate at which insurance is purchased by persons who have a family history, or an adverse test result, and the average amount purchased, determines the extent of adverse selection.

3. The rate of genetic testing, given a family history, can be adjusted to suit the clinical approach to the disorder.

4. Underwriting classes are defined by collections of 'insured' states, within each of which the same rate of premium is payable. For example, if insurers underwrite on the basis of family history and there is no genetic testing, states 21, 23, 31 and 33
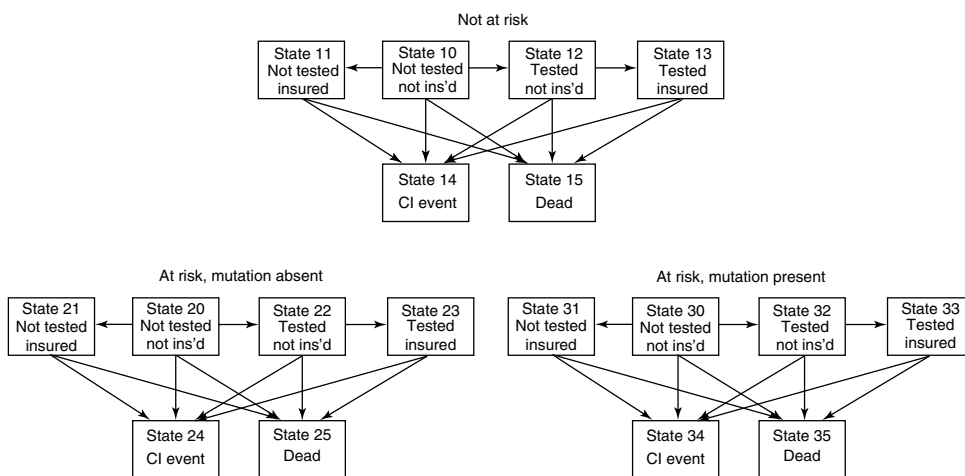
Figure 42.5 A Markov model allowing for a family history of a Mendelian disorder.

would form the 'family history' underwriting class. The effect of introducing genetic testing, and moratoria on the use of genetic information, can then be assessed by changing the composition of the underwriting classes. See Macdonald (2003a) for details.

The algorithm applied to the model is straightforward: the frequency p determines the occupancy probabilities in states $10, 20$ and 30 at birth. Kolmogorov's equations solved forward yield the occupancy probabilities in all states at all relevant ages, then Thiele's equations solved backwards from the age at which insurance expires yield EPVs of future cashflows at all ages back to the age at which insurance purchasing commences. In fact, since Thiele's equations give the EPVs in all states, including the insured states, this model subsumes that of Figure 42.3.

One slightly technical point arises in connection with premium rates. If level premiums were charged depending on the age at which insurance was purchased, then at any age x, the rate of premium cashflow in each insured state would be an average over the rates of premium charged at all earlier ages, involving the age-dependent rate of purchase and the probabilities of remaining in the insured state since purchase. This is feasible, but can be avoided simply by charging a rate of premium equal to the weighted average intensity from the insured states comprising an underwriting class, into the corresponding 'critical illness event' states, the weights being the occupancy probabilities and the amounts assured. The resulting cashflows are adapted to the Markov framework.

The model can be extended in several ways to meet more demanding problems. Some examples are as follows (Gui, 2003; Gui *et al.*, 2006; Gutiérrez and Macdonald, 2004; 2007; Lu, 2006).

1. Heterogeneous genetic disorders, where different mutations have different penetrances, can be handled by partitioning the population into more subgroups, adding an extra pair of 'at risk' subgroups for each distinguishable mutation.

2. Genetic disorders that account for a small proportion of a common disease, such as breast cancer, cause the problem that family history does not identify mutation-carrying families. However, if the definition of 'family history' is precise enough, e.g. 'two or more first-degree relatives with breast or ovarian cancer before age 50', the onset of a family history can be treated as an event in the insured person's life history, and represented by transition into a 'family history' state.

3. By adding a transition from onset of the disorder to the 'dead' state, life insurance can be modelled. This usually introduces duration dependence, therefore a semi-Markov model.

4. In the case of progressive disorders, such as Huntington's disease, a critical illness insurance claim will usually come much later than onset of the first symptoms, since it will depend on some advanced level of disability. If post-onset survival is modelled, as in (3) above, then an accelerated lifetime model derived from this may be used to model the timing of the insurance claim.

42.2.4 Modelling Strategies

The rarity of single-gene disorders of high relevance to insurance makes it implausible that adverse selection in a well-developed insurance market will be so great as to disrupt it.

This can be tested quite simply by choosing parameters for the model that are guaranteed to be extreme, e.g. that 2 % of the population carry high-penetrance mutations and that anyone with an adverse genetic test result will be quite likely to purchase insurance quickly. If the cost of adverse selection, represented for example by premium rate increases, is small, we can be fairly confident that adverse selection is manageable. The advantage of this 'top-down' strategy is that it does not require detailed studies of individual disorders to be undertaken, and for that reason it was the approach used first. Macdonald (1997; 1999) suggested that 10 % was a very conservative upper limit for premium rate increases in life insurance. Since improving longevity has resulted in life insurance premiums falling by considerably more than this over two decades, that particular market seems reasonably safe.

Some caveats must be entered. A similar outcome may be unlikely in a small, less mature insurance market. But if extreme assumptions suggest that adverse selection could be expensive, we learn nothing from the 'top-down' approach and must turn to the more laborious 'bottom-up' strategy, in which the major relevant disorders are modelled individually, to try to obtain a realistic estimate of likely costs. In the long run this is a better basis for discussing genetics and insurance anyway, so it is a program worth undertaking.

42.2.5 Statistical Issues

As mentioned above, the actuarial researcher rarely has access to the data underlying published estimates of penetrance or onset rates. This makes it difficult to estimate sample variances or confidence intervals for any of the actuarial quantities derived from an epidemiological study, which might have been based on a relatively small number of subjects. Perhaps because they spring from different traditions in actuarial science, premium rates for non-life insurance (house, motor, etc.) have long been treated as estimates in a fully statistical framework, but premiums for life insurance have not. It seems appropriate to do so when considering medical underwriting on the basis of epidemiological data.

Premium rates either are or are based upon EPVs, which are the solutions of Thiele's equations, so it is clear that they are very complicated functions of whatever data underlie the transition intensities of the model. Two approaches, both based on bootstrapping, have been suggested.

1. If a fully parametric model is used with pedigree data (Elston, 1973) the information matrix will be estimated and can be used to bootstrap values of the parameters, assumed to be multivariate normal.

2. If a Kaplan-Meier survival curve is obtained, and the underlying event times, numbers of events and numbers at risk are available, various methods of simulation may be employed. Lu (2006) and Lu *et al.* (2007) found that simple resampling and the so-called weird bootstrap (Andersen *et al.*, 1993) gave similar results, based on studies of polycystic kidney disease.

It is clear that the absence of information about the sampling distribution of premium rates is one of the least satisfactory aspects of actuarial modelling of genetic disorders, but it is equally hard to overcome this disadvantage.

42.2.6 Economics Issues

An important parameter in the models shown in Figures 42.4 and 42.5 is the rate at which insurance is purchased normally, in the absence of any genetic or familial risk. In particular, it influences the severity of any adverse selection. Very few studies exist from which this quantity could be estimated, so in most applications of the model a plausible but hypothetical value has been used. This is flawed from the point of view of economic theory, because if adverse selection increases the price of insurance for everybody, the demand for insurance on the part of low-risk individuals should decrease, hence the 'normal' rate of insurance purchase in the model. Since this increases the proportion of high-risk individuals purchasing insurance, the premium should increase yet again, and there is a risk of entering an 'adverse selection spiral', whose worst outcome would be that the price of insurance should rise to the level of the highest-risk individuals.

Several authors have formulated economic market models to explore the nature of any equilibria that emerge if genetic information is withheld from insurers, see Doherty and Thistle (1996), Hoy and Polborn (2000) and Hoy and Witt (2005). The last of these considered a specific disorder, namely breast cancer and the *BRCA1/2* genes. They modelled a life insurance market in which family background, disclosed to the insurer, produced 13 categories of risk, but other genetic information (test results) was known only by the applicant. In the presence of a high-risk group, the equilibrium insurance premium was up to 297 % of the population weighted probability of death. De Jong and Ferris (2006) suggested a very simple statistical adverse selection model based on the correlation between a risk factor and the random amount of insurance sought, and used it to model adverse selection. In the absence of data relating demand for insurance to genetic risk their main example was the imposition of unisex pricing in the market for annuities. Pauly *et al.* (2003) did obtain estimates of how demand for life insurance might vary with BRCA1/2 genotype, applied in Viswanathan *et al.* (2007). Macdonald and Tapadar (2006) used a simple utility model to map out risk thresholds below which adverse selection was unlikely to appear, given a uniform insurance contract driven by fixed need. They suggested that multifactorial disorders would not lead to an adverse selection spiral.

42.3 EXAMPLES AND CONCLUSIONS

42.3.1 Single-gene Disorders

Tables 42.2 and 42.3 show some examples of critical illness insurance premium rates given by models for selected single-gene disorders. They are expressed as percentages of the standard premiums a healthy person would pay for the same contract. In a sense these are worst-case examples, because premium increases in respect of life insurance are often ameliorated by the significant chances of survival post-onset, but modelling of life insurance is less complete.

In Table 42.2 we assume an adverse genotype to be known; this therefore represents the hypothetical consequences of information that insurers in many countries may not use. In each case a range is shown because of uncertainties within the models, of a non-statistical nature. For example, different epidemiological studies may be used to parameterise the

Table 42.2 Examples of premium ratings for CI insurance for female mutation carriers, as percentages of standard rates..

Disorder	Gene	Age 30 Term 30 years (%)	Age 40 Term 20 years (%)	Age 50 Term 10 years (%)
APKD	APKD1/2	335–435	332–468	297–422
BC/OC	BRCA1	381–1110	355–1112	259–740
BC/OC	BRCA2	252–578	296–768	352–1056
EOAD	PSEN1	866–2040	1032–3022	1076–3714
HD	IT15 (40 CAG)	165–268	169–298	133–247
HD	IT15 (45 CAG)	635–1181	455–1018	224–630
HD	IT15 (50 CAG)	1002–2007	591–1372	276–840

Sources: APKD, Gutiérrez and Macdonald (2003); BC/OC, Macdonald *et al.* (2003); EOAD, Gui and Macdonald (2002b); HD, Gutiérrez and Macdonald (2004).

Table 42.3 Examples of premium ratings for CI insurance for female applicants with a family history, as percentages of standard rates.

Disorder	Gene	Age 30 Term 30 years (%)	Age 40 Term 20 years (%)	Age 50 Term 10 years (%)
BC/OC	BRCA1/2	103–184	–	102–158
EOAD	PSEN1	432–769	363–605	153–198
HD	IT15	203–296	142–202	107–128

Sources: BC/OC, Macdonald *et al.* (2003); EOAD, Gui and Macdonald (2002b); HD, Gutiérrez and Macdonald (2004).

model (APKD), onset rates may be reduced if ascertainment bias is likely to be present (breast/ovarian cancer, EOAD) or it may be uncertain at what stage in the progression of a disease an insurance claim would succeed (Huntington's disease). Since insurers would decline to offer cover once the risk indicates a premium greater than about 350 % of the standard, it is clear that large numbers of mutation carriers would be denied cover. (But note that life insurance cover would generally be offered up to about 500 % of standard premiums, so access to life insurance would be less restricted, albeit at a price.)

Table 42.3 assumes that only the presence of a family history (usually an affected parent) is known; this therefore represents the consequences of information that insurers in many countries may use. Access to critical illness cover is greatly eased, especially at older ages. This is because premium rates are averaged over carriers and non-carriers of mutations, and as persons at risk get older and remain healthy, the chance that they are a mutation carrier falls to well below 1/2. There may also be an element of averaging over more and less severe variants of a disease, since the precise causative mutations are assumed to be unknown.

Table 42.4 shows examples of percentage premium increases that might fall upon all policyholders, if severe adverse selection resulted from limiting access to certain genetic test results. The cost is in two parts if family history as well as genetic test results may not be used, because just adding high-risk individuals to the 'standard' insurance pool will raise premiums, even if nobody changes their behaviour. The meaning of 'severe'

Table 42.4 Examples of percentage premium increases ratings for females in a large CI insurance market, caused by severe adverse selection.

Disorder	Gene	'Lenient' Moratorium on genetic tests (%)	Moratorium on family history	
			Pooling cost (%)	Adverse selection cost(%)
APKD	APKD1/2	0.051–0.072	0.203–0.273	0.111–0.126
EOAD	PSEN1	0.014–0.021	0.077–0.118	0.039–0.084
HD	IT15	0.014–0.025	0.038–0.069	0.041–0.066

Sources: APKD, Gutiérrez and Macdonald (2003); EOAD, Gui and Macdonald (2002b); HD, Gutiérrez and Macdonald (2004).

adverse selection is described in the studies cited, but it perhaps implies a level of rational economic behaviour, not to mention ability to afford insurance, that a sceptic might doubt would happen in practice. Hence these very small percentages argue against adverse selection being a serious threat to the critical illness insurance market, let alone the life insurance market. However, this does depend on: (1) the rarity of relevant disorders; and (2) the significant size of these insurance markets: it might not hold in other circumstances. The basic statistical laws that govern insurance, mentioned in Section 42.1.1, have not altered; it just happens that the numbers are small.

42.3.2 Multifactorial Disorders

The models described above represent single-gene disorders well, because genotype and family history partition the population into a small number of subgroups. They are in principle capable of representing multifactorial disorders, but practical limitations are imposed by increasing complexity, and the fact that current epidemiology does not so often lead directly to usable onset rates. Therefore, such actuarial research as has been done has always incorporated some hypothetical features.

1. Macdonald *et al.* (2005a; 2005b) modelled CHD using three risk factors assumed to be static (sex, smoking status and body mass index) and three assumed to be dynamic (i.e. changing through life: diabetes, two levels of hypercholesterolaemia and three levels of hypertension). The model was parameterised using the Framingham data, which does not include any genetic covariates. The hypothetical genetic component of the model was to assume that some genotype would increase the intensities of transition through the worsening risk factors by a factor of 5 or 50. Table 42.5 shows some examples of the results using a factor of 5 in the form of the percentage *extra* premium that would be charged, above the standard rate (therefore 110 % in Table 42.2 or 42.3 would be +10 % here). The first column shows risk factors already present in the person applying for insurance. What is most remarkable is the negligible increase in premiums for someone presenting no risk factors. Other extra premiums are reasonably consistent with many underwriting guidelines. Table 42.6, by contrast, shows the effect of increasing by five times the direct risks of stroke or heart attack (represented by the transition intensities into those states). The extra premiums are markedly increased. The conclusion is that multifactorial disorders that

Table 42.5 Premium ratings for males, non-smokers, normal body mass index (BMI) aged 35 at entry with policy term 10 years, under hypothetical assumptions of genetic influence increasing the incidence of risk factors five times.

Risk factors	None (%)	H'chol (%)	H'tension (%)	Type 1 diabetes (%)	Type 2 diabetes (%)
				Premium rating factors with five times the incidence rate of	
No risk factors	+0	+5	+13	+2	+6
H'chol Cat 1	+3	+12	+17	+5	+9
Type 1 diabetes	+298	+304	+314	–	–
Type 2 diabetes	+67	+73	+84	–	–
H'tension Cat 1	+6	+12	+29	+8	+12
H'chol Cat 2	+25	–	+45	+28	+32
H'chol Cat 1 and type 1 diabetes	+302	+313	+320	–	–
H'chol Cat 1 and type 2 diabetes	+71	+82	+89	–	–
H'chol Cat 1 and H'tension Cat 1	+9	+19	+34	+11	+15
H'tension Cat 1 and type 1 diabetes	+305	+313	+336	–	–
H'tension Cat 1 and type 2 diabetes	+74	+82	+105	–	–
H'tension Cat 2	+34	+44	+53	+37	+41
H'chol Cat 2 and type 1 diabetes	+330	–	+356	–	–
H'chol Cat 2 and type 2 diabetes	+99	–	+125	–	–
H'chol Cat 2 and H'tension Cat 1	+34	–	+70	+37	+41
H'tension Cat 1, H'chol Cat 1, type 1 diabetes	+309	+322	+342	–	–
H'tension Cat 1, H'chol Cat 1, type 2 diabetes	+78	+91	+112	–	–
H'tension Cat 2 and H'chol Cat 1	+40	+56	+60	+42	+47
H'tension Cat 2 and type 1 diabetes	+342	+354	+368	–	–
H'tension Cat 2 and type 2 diabetes	+111	+124	+137	–	–
H'tension Cat 3	+81	+95	–	+84	+90
H'chol Cat 2, H'tension Cat 1, type 1 diabetes	+342	–	+388	–	–
H'chol Cat 2, H'tension Cat 1, type 2 diabetes	+111	–	+158	–	–
H'chol Cat 2 and H'tension Cat 2	+81	–	+108	+84	+89
H'tension Cat 2, H'chol Cat 1 and type 1 diabetes	+349	+370	+376	–	–
H'tension Cat 2, H'chol Cat 1 and type 2 diabetes	+118	+139	+146	–	–
H'tension Cat 3 and H'chol Cat 1	+89	+111	–	+91	+97
H'tension Cat 3 and type 1 diabetes	+407	+423	–	–	–
H'tension Cat 3 and type 2 diabetes	+176	+193	–	–	–
H'tension Cat 2, and H'chol Cat 2 and type 1 diabetes	+402	–	+438	–	–
H'tension Cat 2, and H'chol Cat 2 and type 2 diabetes	+172	–	+207	–	–
H'tension Cat 3 and H'chol Cat 2	+147	–	–	+150	+157
H'tension Cat 3, H'chol Cat 1, type 1 diabetes	+416	+445	–	–	–
H'tension Cat 3, H'chol Cat 1, type 2 diabetes	+185	+214	–	–	–

Source: Macdonald *et al.* (2005b).

Table 42.6 Premium ratings for males, non-smokers, normal BMI aged 35 at entry with policy term 10 years, under hypothetical assumptions of genetic influence increasing the incidence of CHD and stroke five times.

| Risk factors | Premium rating factors with five times the incidence rate of | | | | | | |
| | None (%) | CHD (%) | Stroke (%) | CHD modified by the presence of | | | |
				H'chol (%)	H'tension (%)	Type 1 diab (%)	Type 2 diab (%)
No risk factors	+0	+142	+37	+26	+29	+1	+5
H'chol Cat 1	+3	+156	+40	+157	+35	+4	+8
Type 1 diabetes	+298	+481	+351	+331	+334	+481	–
Type 2 diabetes	+67	+250	+121	+100	+104	–	+250
H'tension Cat 1	+6	+168	+44	+36	+168	–	+11
H'chol Cat 2	+25	+267	+62	+267	+74	+26	+33
H'chol Cat 1 and type 1 diabetes	+302	+499	+355	+499	+342	+499	–
H'chol Cat 1 and type 2 diabetes	+71	+268	+124	+268	+111	–	+268
H'chol Cat 1 and H'tension Cat 1	+9	+184	+47	+184	+184	+10	+15
H'tension Cat 1 and type 1 diabetes	+305	+513	+361	+343	+513	+513	–
H'tension Cat 1 and type 2 diabetes	+74	+283	+130	+113	+283	–	+283
H'tension Cat 2	+34	+298	+83	+82	+298	+35	+43
H'chol Cat 2 and type 1 diabetes	+330	+641	+384	+641	+392	+641	–
H'chol Cat 2 and type 2 diabetes	+99	+411	+153	+411	+161	–	+411
H'chol Cat 2 and H'tension Cat 1	+34	+309	+73	+309	+309	+35	+44
H'tension Cat 1, H'chol Cat 1, type 1 diabetes	+309	+534	+365	+534	+534	+534	–
H'tension Cat 1, H'chol Cat 1, type 2 diabetes	+78	+304	+134	+304	+304	–	+304
H'tension Cat 2 and H'chol Cat 1	+40	+324	+89	+324	+324	–	+49
H'tension Cat 2 and type 1 diabetes	+342	+681	+414	+403	+681	+681	–
H'tension Cat 2 and type 2 diabetes	+111	+451	+183	+173	+451	–	+451
H'tension Cat 3	+81	+450	+213	+147	+450	+82	+93
H'chol Cat 2, H'tension Cat 1, type 1 diabetes	+342	+694	+398	+694	+694	+694	–

Table 42.6 (*continued*).

H'chol Cat 2, H'tension Cat 1, type 2 diabetes	+111	+465	+167	+465	+465	–	+465
H'chol Cat 2 and H'tension Cat 2	+81	+528	+130	+528	+528	+82	+96
H'tension Cat 2, H'chol Cat 1 and type 1 diabetes	+349	+714	+421	+714	+714	+714	–
H'tension Cat 2, H'chol Cat 1 and type 2 diabetes	+118	+485	+190	+485	+485	–	+485
H'tension Cat 3 and H'chol Cat 1	+89	+485	+220	+485	+485	+90	+102
H'tension Cat 3 and type 1 diabetes	+407	+881	+598	+489	+881	+881	
H'tension Cat 3 and type 2 diabetes	+176	+651	+368	+259	+651	–	+651
H'tension Cat 2, and H'chol Cat 2 and type 1 diabetes	+402	+976	+474	+976	+976	+976	–
H'tension Cat 2, and H'chol Cat 2 and type 2 diabetes	+172	+747	+243	+747	+747	–	+747
H'tension Cat 3 and H'chol Cat 2	+147	+771	+278	+771	+771	+148	+166
H'tension Cat 3, H'chol Cat 1, type 1 diabetes	+416	+925	+608	+925	+925	+925	–
H'tension Cat 3, H'chol Cat 1, type 2 diabetes	+185	+696	+377	+696	+696	–	+696
H'tension Cat 3, H'chol Cat 2 and type 1 diabetes	+491	+1293	+682	+1293	+1293	+1293	–
H'tension Cat 3, H'chol Cat 2 and type 2 diabetes	+260	+1064	+452	+1064	+1064	–	+1064

Source: Macdonald *et al.* (2005b).

act by modifying risk factors for a disease, rather than the disease outcome itself, present significantly more manageable insurance risks. In particular, the kinds of simple models of disease onset applicable in the study of single-gene disorders may be misleading if they are naively applied to the study of multifactorial disorders.

2. Great interest currently centres on large-scale prospective cohort studies in several countries. In the United Kingdom, the Biobank project will recruit 500 000 people aged 40–70, and track them for 10 years, with linkage to national health registers. The resulting data will be made available for analysis in later studies, presumed to be of nested case-control type. Macdonald *et al.* (2006) introduced a hypothetical 2×2 gene–environment model of heart attack risk, and simulated outcomes of the UK Biobank study. The question of interest was as follows: will the kind of case-control

studies likely to follow UK Biobank identify different levels of risk that both insurers and GAIC (in the United Kingdom) might regard as relevant and reliable? The answer was 'doubtful'. Only very large studies had sufficient power to pick out the modestly elevated risk, and that was in the context of a simplified genetic model more akin to a single-gene disorder than to a genuinely multifactorial disorder. This and other studies have found no evidence to suggest that increasing knowledge of multifactorial disorders will lead to changes in insurance practice.

REFERENCES

ALRC (2001). *Protection of Human Genetic Information*. Issues Paper No. 26. Australian Law Reform Commission. `www.alrc.gov.au`.

ALRC (2002). *Protection of Human Genetic Information*. Discussion Paper No. 66. Australian Law Reform Commission. `www.alrc.gov.au`.

ALRC (2003). Essentially yours: the protection of human genetic information in Australia. Report No. 96, Australian Law Reform Commission. `www.alrc.gov.au`.

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Brackenridge, R. and Elder, J. (1998). *Medical Selection of Life Risks*. Macmillan.

Daykin, C.D., Akers, D.A., Macdonald, A.S., McGleenan, T., Paul, D. and Turvey, P.J. (2003). Genetics and insurance – some social policy issues (with discussions). *British Actuarial Journal* **9**, 787–874.

De Jong, P. and Ferris, S. (2006). Adverse selection spirals. *ASTIN Bulletin* **36**, 589–628.

Doble, A. (2001). *Genetics in Society*. Institute of Actuaries in Australia, Sydney.

Doherty, N.A. and Thistle, P.D. (1996). Adverse selection with endogeneous information in insurance markets. *Journal of Public Economics* **63**, 83–102.

Elston, R.C. (1973). Ascertainment and age at onset in pedigree analysis. *Human Heredity* **23**, 105–112.

Espinosa, C. (2006). Ascertainment bias in estimating rates of onset of early-onset Alzheimer's disease: a critical illness and life insurance application Ph.D. dissertation, Heriot-Watt University, Edinburgh.

Ford, D., Easton, D.F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D.T., Weber, B., Lenoir, G., Chang-Claude, J., Sobol, H., Teare, M.D., Struewing, J., Arason, A., Scherneck, S., Peto, J., Rebbeck, T.R., Tonin, P., Neuhausen, S., Barkardottir, R., Eyfjord, J., Lynch, H., Ponder, B.A.J., Gayther, S.A., Birch, J.M., Lindblom, A., Stoppa-Lyonnet, D., Bignon, Y., Borg, A., Hamann, U., Haites, N., Scott, R.J., Maugard, C.M., Vasen, H., Seitz, S., Cannon-Albright, L.A., Schofield, A., Zelada-Hedman, M. and The Breast Cancer Linkage Consortium (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *American Journal of Human Genetics* **62**, 676–689.

Gui, E.H. (2003). Modelling the impact of genetic testing on insurance – early-onset Alzheimer's disease and other single-gene disorders Ph.D. dissertation, Heriot-Watt University, Edinburgh.

Gui, E.H. and Macdonald, A.S. (2002a). A Nelson-Aalen estimate of the incidence rates of early-onset Alzheimer's disease associated with the Presenilin-1 gene. *ASTIN Bulletin* **32**, 1–42.

Gui, E.H. and Macdonald, A.S. (2002b). Early-onset Alzheimer's disease, critical illness insurance and life insurance. Genetics and Insurance Research Centre Research Report 02/2, Heriot-Watt University, Edinburgh.

Gui, E.H., Lu, B., Macdonald, A.S., Waters, H.R. and Wekwete, C.T. (2006). The genetics of breast and ovarian cancer III: a new model of family history with applications. *Scandinavian Actuarial Journal* 338–367.

Gutiérrez, M.C. and Macdonald, A.S. (2003). Adult polycystic kidney disease and critical illness insurance. *North American Actuarial Journal* **7**(2), 93–115.

Gutiérrez, M.C. and Macdonald, A.S. (2004). Huntington's disease, critical illness insurance and life insurance. *Scandinavian Actuarial Journal* 279–313.

Gutiérrez, M.C. and Macdonald, A.S. (2007). Adult polycystic kidney disease and insurance: a case study in genetic heterogeneity. *North American Actuarial Journal* **11**(1), 90–118.

Hoem, J.M. (1988). The versatility of the Markov chain as a tool in the mathematics of life insurance. In *Transactions of the 23rd International Congress of Actuaries*, Helsinki, 171–202.

HCSTC (1995). *House of Commons Science and Technology Committee, Third Report: Human genetics: The Science and its Consequences*. H.M.S.O., London.

HCSTC (2001). House of Commons Science and Technology Committee, Fifth Report: Genetics and insurance. `www.publications.parliament.uk/pa/cm200001/cmselect/cmsctech/174/17402.htm`.

Hoy, M. and Polborn, M. (2000). The value of genetic information in the life insurance market. *Journal of Public Economics* **78**, 235–252.

Hoy, M. and Witt, J. (2005). *Welfare effects of banning genetic information in the life insurance market: the case of the BRCA1/2 genes*. University of Guelph, Discussion Paper 2005-5.

HGAC (1997). *The Implications of Genetic Testing for Insurance*. Human Genetics Advisory Commission, London.

HGC (2002). *Inside information: Balancing interests in the Use of Personal Genetic Data*. The Human Genetics Commission, London.

Leigh, T.S. (1990). Underwriting – a dying art? (with discussion). *Journal of the Institute of Actuaries* **117**, 443–531.

Lu, L. (2006). Some actuarial and statistical investigations into topics on genetics and insurance. Ph.D. dissertation, Heriot-Watt University, Edinburgh.

Lu, L., Macdonald, A.S. and Wekwete, C.T. (2007). Premium rates based on genetic studies: how reliable are they? *Insurance: Mathematics and Economics*.

Macdonald, A.S. (1997). How will improved forecasts of individual lifetimes affect underwriting? *Philosophical Transactions of the Royal Society Series B* **352**, 1067–1075 and (with discussion) *British Actuarial Journal* **3**, 1009–1025 and 1044–1058.

Macdonald, A.S. (1999). Modeling the impact of genetics on insurance. *North American Actuarial Journal* **3**(1), 83–101.

Macdonald, A.S. (2003a). Moratoria on the use of genetic tests and family history for mortgage-related life insurance. *British Actuarial Journal* **9**, 217–237.

Macdonald, A.S. (2003b). Genetics and insurance: What have we learned so far? *Scandinavian Actuarial Journal* 324–348.

Macdonald, A.S., Pritchard, D.J. and Tapadar, P. (2006). The impact of multifactorial genetic disorders on critical illness insurance: a simulation study based on UK Biobank. *ASTIN Bulletin* **36**, 311–346.

Macdonald, A.S. and Tapadar, P. (2006). Multifactorial genetic disorders and adverse selection: Epidemiology meets economics. Submitted.

Macdonald, A.S., Waters, H.R. and Wekwete, C.T. (2003). The genetics of breast and ovarian cancer II: a model of critical illness insurance. *Scandinavian Actuarial Journal* 28–50.

Macdonald, A.S., Waters, H.R. and Wekwete, C.T. (2005a). A model for coronary heart disease and stroke, with applications to critical illness insurance underwriting I: the model. *North American Actuarial Journal* **9**(1), 13–40.

Macdonald, A.S., Waters, H.R. and Wekwete, C.T. (2005b). A model for coronary heart disease and stroke, with applications to critical illness insurance underwriting II: applications. *North American Actuarial Journal* **9**(1), 41–56.

Norberg, R. (1995). Differential equations for moments of present values in life insurance. *Insurance: Mathematics and Economics* **17**, 171–180.

Nys, H., Dreezen, I., Vinck, I., Dierickx, K., Dequeker, E. and Cassiman, J.-J. (2002). *Genetic Testing*. European Commission, Brussels.

Pauly, M., Withers, K., Viswanathan, K.S., Lemaire, J., Hershey, J., Armstrong, K. and Asch, D.A. (2003). *Price Elasticity of Demand for Term Life Insurance and Adverse Selection*. National Bureau of Economic Research, Working Paper 9925.

Viswanathan, K.S., Lemaire, J., Withers, K., Armstrong, K., Baumritter, A., Hershey, J., Pauly, M. and Asch, D.A. (2007). Adverse selection in life insurance purchasing, due to the BRCA 1/2 genetic test and elastic demand. *Journal of Risk and Insurance*.

43

Forensics

B.S. Weir

Department of Biostatistics, University of Washington, Seattle, WA, USA

The use of DNA profiles for human identification often requires statistical genetic calculations. The probabilities for a matching DNA profile can be evaluated under alternative hypotheses about the contributor(s) to the profile, and presented as likelihood ratios. It is conditional probabilities that are needed: the probabilities of profiles given that they have already been seen, and these depend on the relationships between known and unknown people. The algebraic treatment is greatly simplified when it can be assumed that allelic frequencies have Dirichlet distributions over populations. The growing size of DNA profile databases has led to empirical verification of the probabilities of finding pairs of people with the same profile.

43.1   INTRODUCTION

Human individualisation based on the genome exploits the fact that everyone except for identical twins is genetically distinguishable. Moreover, human genetic material is found in every nucleated cell in the body and can be recovered from samples as diverse as bone, blood stains, saliva residues, nasal secretions and even fingerprints. DNA may be recovered from very old samples that have been well preserved, and DNA signatures may even be preserved over successive generations.

Genetic markers have been used for human individualisation since the discovery of blood groups, and statistical genetic arguments have long played a large part in parentage dispute cases. The role of statistical genetics in forensic science increased sharply in the late 1980s when DNA markers began to be used and an emphasis shifted from excluding specific people as being the sources of evidentiary stains to making probability statements about genetic profiles if these people were not the sources. As more markers have been developed, it has become less likely that two people would share the same DNA profile and therefore more likely that people will be convicted on the basis of DNA evidence. Given the serious nature of the charges of many crimes where genetic markers are used, and the serious consequences of conviction for these crimes, there has been considerable scrutiny paid to the statistical genetic arguments

upon which forensic probabilities are based. These arguments are reviewed in this chapter.

A common situation is where DNA is recovered from a biological sample left at the scene of a crime, and there is reason to believe the sample is from the perpetrator of the crime. DNA is also extracted from a blood or saliva sample from a suspect in the crime and is found to have the same profile as the crime sample. An immediate question is how much evidence against the suspect is provided by this matching, and a naive answer might be that the probability of the suspect having the evidentiary profile, if he was not the perpetrator, is the population proportion of that profile. High values for this proportion would favour the suspect and low values would not. How is the proportion to be estimated?

For genetic profiles based on single loci, it is quite feasible to estimate the population proportion of any genotype on the basis of a moderate-sized sample of profiles from that population. The size of the sample would need to be greater for highly variable loci where there are many different genotypes. There are issues of how the population is to be defined, and then how it is to be sampled. As the number of loci used for identification increases, the numbers of genotypes increases substantially and no sample can hope to capture all genotypes. Current practice in the United States is to use a set of 13 short tandem repeat (STR) loci, each with at least 9 alleles and 45 genotypes. The number of 13-locus profiles is therefore considerably more than 10^{21}, so that less than one profile in a trillion of all possible profiles exists anywhere in the world. Although there may not, therefore, be much value in declaring that a particular profile is rare, a first attempt to attach a frequency to a 13-locus profile might be to multiply together the frequencies of the 26 constituent alleles, along with a factor of 2 for every heterozygote. The implied assumption of allelic independence, within and between loci, is a statistical genetic issue.

A more satisfactory approach is to ask the question: What is the probability of an unknown person chosen from the population having a particular genetic profile given that the profile has been seen already for the suspect? Calculating this conditional probability needs to take into account the relationship between the known suspect and the unknown person. This relationship may be due to close family membership or to shared evolutionary history. Once again, these are statistical genetic issues.

43.2 PRINCIPLES OF INTERPRETATION

Evett and Weir (1998) suggested that genetic evidence be interpreted according to three principles:

- *First Principle*: To evaluate the uncertainty of any given proposition it is necessary to consider at least one alternative proposition.

- *Second Principle*: Interpretation is based on questions of the kind 'What is the probability of the evidence given the proposition?'

- *Third Principle*: Interpretation is conditioned not only by the competing propositions, but also by the framework of circumstances within which they are to be evaluated.

The first of these principles leads to the use of likelihood ratios LRs, as will soon be shown. The second is meant to draw a distinction from the very common 'prosecutor's fallacy' (Thompson and Schumann, 1987) of quoting probabilities of a proposition given the evidence. The third principle recognises the difference in the strength of evidence of a blood stain at the scene of a crime having a profile matching that of a suspect, and the evidence of bloodstain in the clothing of a suspect away from the crime scene with a profile matching that of a victim.

The propositions mentioned in the principles refer, in this context, to the source of the genetic profile in the evidentiary stain. For the immediate discussion these will be taken to be

H_p: The profile is from the suspect.
H_d: The profile is from some other person.

Suppose G_s and G_c are the profile types from the suspect and the crime stain, and they are found to match. Then the evidence E is these two profiles: $E = (G_s, G_c)$. Consideration of alternative propositions is carried out by comparing the probabilities of E under these propositions by means of the likelihood ratio

$$\text{LR} = \frac{\Pr(E|H_p)}{\Pr(E|H_d)} = \frac{\Pr(G_s, G_c|H_p)}{\Pr(G_s, G_c|H_d)}$$

$$= \frac{\Pr(G_c|G_s, H_p)}{\Pr(G_c|G_s, H_d)} \times \frac{\Pr(G_s|H_p)}{\Pr(G_s|H_d)}$$

$$= \frac{1}{\Pr(G_c|G_s, H_d)},$$

with the last step depending on the assumption that the profiles must be found to match when they have a common source, and on recognition that $\Pr(G_s)$ does not depend on H_p or H_d.

The forensic question of attaching weight to matching genetic profiles has therefore reduced to the statistical genetic question of determining the probability of a profile given that (the same) profile has been seen already. This conditional probability will be referred to as the *match probability*. It would be a substantial simplification to assume the two profiles were independent and work only with the probability of the crime stain profile:

$$\text{LR} = \frac{1}{\Pr(G_c|H_d)} = \frac{1}{\Pr(G_c)}.$$

The use of LRs for DNA evidence has been described in several textbooks: Aitken (1995), Robertson and Vignaux (1995), Royall (1997), Schum (1994), Evett and Weir (1998), Balding (2005), Buckleton *et al.* (2005) and Lucy (2005).

The odds form of Bayes' theorem relates the probabilities of the propositions after the evidence to the probabilities prior to the evidence:

$$\frac{\Pr(H_p|E)}{\Pr(H_d|E)} = \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \times \frac{\Pr(H_p)}{\Pr(H_d)},$$

which can be stated as

$$\text{Posterior odds on} H_p = \text{LR} \times \text{prior odds on} H_p.$$

It needs to be stressed that the results in this section apply only to the situation where the DNA evidence refers to material left at the crime scene by the perpetrator, and there is no DNA evidence at the scene that does not provide a match to the profile of the suspect. If the evidence refers to a bloodstain found on the clothing of the suspect, e.g. and the stain has a DNA profile matching that of the victim then additional factors need to be considered: What is the probability that the victim's blood would be transferred during the crime? What is the probability that the suspect would have non-self blood on his or her clothing? What is the probability that non-self blood on the suspect's clothing would match that of the victim?

43.3 PROFILE PROBABILITIES

Although it is conditional profile probabilities (match probabilities) that are needed, these are most directly calculated as joint divided by marginal probabilities. Dropping the H_d symbol for convenience:

$$\Pr(G_c|G_s) = \frac{\Pr(G_c, G_s)}{\Pr(G_s)}.$$

Profile probabilities are therefore needed to calculate match probabilities.

At a single locus A with alleles A_u in proportion p_u, the probabilities P_{uv} for genotypes $A_u A_v$ within a single population may be parameterised as

$$P_{uv} = \begin{cases} p_u^2 + fp_u(1 - p_u), & u = v, \\ 2p_u p_v - 2fp_u p_v, & u \neq v. \end{cases} \tag{43.1}$$

As the loci used for individualisation are unlikely to be under the influence of selection, the use of a single inbreeding coefficient f for all genotypes may be thought reasonable, although genotype-specific coefficients should strictly be used for loci with allele-specific mutation processes (Graham *et al.*, 2000). Expressing genotype proportions as functions of allele proportions is necessary for highly variable loci, when many genotypes are not seen in a sample and their population frequencies are difficult to estimate.

43.3.1 Allelic Independence

Considerable attention has been paid to the issue of whether f can be assumed small enough to ignore. Standard testing procedures rarely find significant departures from the hypothesis that f is zero (Weir, 1992; Maiste and Weir, 1995; Zaykin *et al.*, 1995), although this may not be very surprising. The goodness-of-fit test has the statistic

$$X^2 = \sum_u \frac{n(\tilde{P}_{uu} - \tilde{p}_u^2)^2}{\tilde{p}_u^2} + \sum_{u \neq v} \frac{n(\tilde{P}_{uv} - 2\tilde{p}_u \tilde{p}_v)^2}{2\tilde{p}_u \tilde{p}_v},$$

where $\tilde{P}_{uv}$ and $\tilde{p}_u$ are the sample genotype and allele proportions in a sample of n individuals. For a locus with m alleles it is distributed as χ^2 with $m(m-1)/2$ df when

f is zero. When f is not zero, X^2 has non-centrality parameter

$$\lambda = \sum_u [nf^2(1 - p_u)^2] + \sum_{u \neq v} [n2f^2 p_u p_v]$$

$$= (m - 1)nf^2.$$

and this allows the power of the test to be calculated. For a locus with $m = 2$ alleles, the test has one degree of freedom and λ needs to be at least 10.5 for the power to be at least 90 % when the significance level is 0.05. In other words, the sample size must be at least $10.5/f^2$. A sample of 105 000 would therefore be needed to have 90 % chance of detecting an f as small as 0.01. Even an f value of 0.03 would require a sample of over 11 000. In Table 43.1, the sample sizes needed for 90 % power to detect $f = 0.05$ are shown for a range of values of m, along with the powers for that f value when $n = 1000$. It is not likely that Hardy–Weinberg disequilibrium, at the level usually thought to exist in human populations, will be detected with samples of 1000 or less.

What should be done when a test does cause rejection of the null hypothesis? Forensic agencies routinely test for allelic independence at six or more loci in three or more samples. A conventional 5 % significance-level test would be expected to give at least one significant result even if there was independence at all loci in all populations. However, to ignore single rejections on that basis calls into question the logic of performing the test in the first place. Requiring each of t tests to meet a $0.05/t$ significance level in order to declare rejection, the Bonferroni correction, seems unduly conservative. Zaykin *et al.* (2002) examined multiple-testing issues, and showed that Fisher's procedure for combining p values may be more informative. When a hypothesis is true, the p value (the probability of the data or of data with at least as much departure from the hypothesis) has a uniform distribution and minus twice its logarithm has a χ^2 distribution with 2 df. A set of t independent tests can be said to indicate that at least one of the t hypotheses is false if the sum of the $-2\ln(p)$ values exceeds the critical value of the χ^2 distribution with $2t$ df. As an example, LR tests for allelic independence at seven loci were performed on data in three samples (Scholl *et al.*, 1996) and the p values are shown in Table 43.2. For a 5 % significance level, ignoring the multiple-testing issue would lead to rejection of independence at GYPA in the Navajo sample and maybe at D7S8 in that sample. Applying

Table 43.1 Hardy–Weinberg test properties when $f = 0.05$.

m	df	For 90 % power		Power if $n = 1000$
		λ	n	
2	1	10.5	4200	0.35
3	3	14.2	2840	0.44
4	6	17.4	2320	0.50
5	10	20.5	2050	0.54
6	15	23.6	1888	0.57
7	21	26.6	1773	0.60
8	28	29.6	1691	0.63
9	36	32.6	1630	0.64
10	45	35.6	1582	0.65

the Bonferroni correction for the seven tests conducted for the Navajo data would lead to no rejections in that sample; applying the correction to the three tests conducted for GYPA would lead to rejection at that locus; applying the correction to all 21 test would cause no rejections in the whole set. Fisher's procedure, however, suggests dependence in the Navajo sample, but not at the GYPA locus.

An alternative to testing hypotheses about f is to make statements about its value in the relevant population, by means of either a point estimate or a posterior probability distribution (Ayres and Balding, 1998; Shoemaker *et al.*, 1998). Given that human populations do have non-zero values of f, there is some appeal to making probability statements about f lying in a certain range rather than simply failing to reject the false hypothesis that it is zero. Other comments about conventional hypothesis tests were made by Evett and Weir (1998). Although there have been concerns in the past about values of f in forensic calculations, the concerns may have been overstated. As is discussed below, the quantity of interest is the match probability or the probability of a profile given that it as already been seen, and this is a statement about four alleles per locus, whereas f is used in profile probabilities which are statements about two alleles per locus. The distinction was made by Ayres and Overall (1999).

43.3.2 Allele Frequencies

Apart from some uncertainty about f values, the difficulty in using 43.1 is that the allele frequencies are also unknown. This problem is due to uncertainty about the population to which these frequencies apply. The population is sometimes referred to as being that of 'possible perpetrators' as (under H_d) the unknown perpetrator belongs to the population. Under H_p, the suspect must also belong to this population. Usually, however, the population is not defined with sufficient detail to suggest a sampling strategy for estimating allele frequencies from that group. Even an eyewitness description of the appearance of the perpetrator may not delineate the population very precisely. There is also the practical issue of collecting a new sample from which to estimate allele frequencies for every new crime.

Instead, it is customary to base calculations on samples collected from some population that is larger than the relevant population, and which may well have substructure. Typically forensic agencies work with samples of people described by broad racial labels such as 'Caucasian' even when these labels refer to admixed groups, as is the case with 'African

Table 43.2 p values for tests of allelic independence.

Locus	Sample			$-2\sum \ln(p)$
	Navajo	Peublo	Sioux	
LDLR	0.377	0.397	0.599	0.566
GYPA	0.014	0.470	1.000	0.122
HBGG	0.136	0.168	0.790	0.235
D7S8	0.052	1.000	0.804	0.385
GC	0.259	0.124	0.213	0.125
HLA-DQA1	0.438	0.368	0.562	0.569
D1S80	0.750	0.559	0.211	0.563
$-2\sum \ln(p)$	0.031	0.430	0.812	0.216

American'. Assignment of a person to such groups is generally by self reporting, and the samples tend to be drawn from the geographic area served by the agency. Equation 43.1 is now to be understood to apply to a subpopulation, i, of some larger sampled population:

$$P_{uvi} = \begin{cases} p_{ui}^2 + f_i p_{ui}(1 - p_{ui}), & u = v, \\ 2p_{ui}p_{vi} - 2f_i p_{ui}p_{vi}, & u \neq v. \end{cases} \tag{43.2}$$

A simplifying assumption is that all f_i values are equal (and maybe equal to zero), and that all p_{ui} values have the same expected value p_u. Taking expectations over subpopulations:

$$\mathrm{E}(p_{ui}^2) = p_u^2 + p_u(1 - p_u)\theta,$$

where θ or F_{ST} is a measure of population structure. The genotype proportions in the whole population are the expected values of the subpopulation values:

$$P_{uv} = \mathrm{E}(P_{uvi}) = \begin{cases} p_u^2 + Fp_u(1 - p_u), & u = v, \\ 2p_u p_v - 2Fp_u p_v, & u \neq v, \end{cases} \tag{43.3}$$

where F or F_{IT} is the total inbreeding coefficient, in contrast to the within-population inbreeding coefficient $f = (F - \theta)/(1 - \theta)$. If the subpopulations are not inbred, then $f = 0$ and θ could be used in place of F. The population-wide genotype frequencies in 43.1 may be taken to apply, on average, to any subpopulation.

Although 43.3 appears to circumvent the need to know subpopulation allele frequencies, it does raise other issues. It is not possible to estimate F or θ from data only at the whole-population level. Estimation would require observations at the subpopulation level – in which case this overall formulation would not be needed. In practice, a numerical value is assigned to F or θ.

Secondly, it is not possible to estimate p_u^2 as the square of an estimate of p_u as that would ignore the variation in p_u that is being described by θ. Suppose the ith subpopulation has allele A_u with frequency p_{ui}, and it forms a proportion w_i of the whole population ($\sum_i w_i = 1$). A sample of size n individuals from the whole population has n_i individuals from the ith subpopulation. With random sampling, it might be supposed that $n_i = nw_i$. The sampling properties of $\tilde{p}_{ui}$, the proportion of A_u alleles among the $2n_i$ alleles from the ith subpopulation, are (Weir, 1996),

$$\mathrm{E}(\tilde{p}_{ui}) = p_u$$

$$\mathrm{Var}(\tilde{p}_{ui}) = p_u(1 - p_u)\left(\theta + \frac{1 - \theta}{2n_i}\right)$$

$$\mathrm{Cov}(\tilde{p}_{ui}, \tilde{p}_{vi}) = -p_u p_v \left(\theta + \frac{1 - \theta}{2n_i}\right).$$

Assuming the subpopulations to be independent, therefore,

$$\mathrm{E}(\tilde{p}_u) = p_u$$

$$\mathrm{E}(\tilde{p}_u^2) = p_u^2 + p_u(1 - p_u)\left(\theta \sum_i w_i^2 + \frac{1 - \theta}{2n}\right)$$

$$E(2\tilde{p}_u\tilde{p}_v) = 2p_up_v - 2p_up_v\left(\theta\sum_i w_i^2 + \frac{1-\theta}{2n}\right), \quad u \ne v,$$

so that estimating squares or products of allele frequencies by squares or products of estimated allele frequencies causes overestimation for homozygotes and is therefore conservative, but causes underestimation for heterozygotes and is therefore prejudicial. The effects will be small for small θ or for highly structured populations (small w_i).

43.3.3   Joint Profile Probabilities

The joint genotype probabilities needed to calculate match probabilities require account to be taken of the relationships among sets of four alleles, two per individual, in the same way that single genotype probabilities require information about the relationship of pairs of alleles. In the random-mating situation, where allelic relationships do not depend on the arrangements of alleles within genotypes, there are four measures of allelic relationship: θ, γ, δ and Δ for pairs, triples, quadruples and two pairs of alleles, respectively (Cockerham, 1971).

The probability that four alleles are all of type A_u is just the allele frequency p_u if they are all identical by descent, and this identity situation has probability δ. At the other extreme, four alleles are all A_u with probability p_u^4 if there is no identity among them, and this situation has probability $(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)$. Such arguments lead to the joint genotype probabilities

$$\begin{aligned}
\Pr(A_uA_u, A_uA_u) = {} & (1 - 6\theta + 8\gamma + 3\Delta - 6\delta)p_u^4 \\
& + 6(\theta - 2\gamma - \Delta + 2\delta)p_u^3 \\
& + (4\gamma + 3\Delta - 7\delta)p_u^2 + \delta p_u, \quad (43.4) \\
\Pr(A_uA_v, A_uA_v) = {} & 4(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)p_u^2 p_v^2 \\
& + 4(\theta - 2\gamma - \Delta + 2\delta)p_up_v(p_u + p_v) \\
& + 4(\gamma - \delta)p_up_v, \quad u \ne v. \quad (43.5)
\end{aligned}$$

These expressions are greatly simplified under the assumption, made in coalescent approaches, of evolutionary stationarity. Under that assumption, the set of allele frequencies has a Dirichlet distribution over populations with the consequence that the probability of drawing allele A_u from a population given that n previously drawn alleles contained n_u of that type is

$$\Pr(A_u | n_u \text{ among } n) = \frac{n_u\theta + (1-\theta)p_u}{1 + (n-1)\theta}. \quad (43.6)$$

The Dirichlet assumption also implies that $\gamma = 2\theta^2/(1+\theta)$, $\delta = 6\theta^3/(1+\theta)(1+2\theta)$, and $\Delta = \theta^2(1+5\theta)/(1+\theta)(1+2\theta)$ and it leads to the match probabilities given by Balding and Nichols (1994):

$$\Pr(A_uA_v | A_uA_v) = \begin{cases} \dfrac{[2\theta + (1-\theta)p_u][3\theta + (1-\theta)p_u]}{(1+\theta)(1+2\theta)}, & u = v, \\[2ex] \dfrac{2[\theta + (1-\theta)p_u][\theta + (1-\theta)p_v]}{(1+\theta)(1+2\theta)}, & u \ne v. \end{cases} \quad (43.7)$$

Table 43.3 Effects of population structure on match probabilities.

		Reciprocal of match probability			
		$\theta = 0$	$\theta = 0.001$	$\theta = 0.01$	$\theta = 0.03$
$p = 0.01$	Heterozygote	5000	4152	1295	346
	Homozygote	10 000	6 439	863	157
$p = 0.05$	Heterozygote	200	193	145	89
	Homozygote	400	364	186	73
$p = 0.10$	Heterozygote	50	49	43	34
	Homozygote	100	96	67	37

A feature of the Dirichlet distribution is that the frequencies of all pairs of different allelic types have negative correlations. This is trivially true for biallelic single nucleotide polymorphism (SNP) loci but it cannot be true for STR loci affected by stepwise mutation (Graham *et al.*, 2000) although the consequences on match probabilities may not be large for rare genotypes.

The match probabilities in 43.7 are greater than the profile probabilities in 43.3 for allele frequencies less than 0.5. Although a profile is rare, as soon as it has been seen there is an increased probability that there is another copy of it in the population. It needs to be stressed that the equations apply on average for any subpopulation within the population. Some numerical consequences of allowing for population structure are shown in Table 43.3, in the case where all alleles at a locus have the same frequency. The table shows that even small values of θ can have an appreciable effect when allele frequencies are small.

A larger dependency between DNA profiles occurs when two people are in the same family. Brothers, e.g. have at least a 25 % probability of sharing the same genotype at any locus. If only family relatedness is considered and neither relative is inbred, then 43.4 and 43.5 are replaced by

$$\Pr(A_u A_u, A_u A_u) = k_0 p_u^4 + k_1 p_u^3 + k_2 p_u^2 \tag{43.8}$$

$$\Pr(A_u A_v, A_u A_v) = 4 k_0 p_u^2 p_v^2 + k_1 p_u p_v (p_u + p_v) + 2 k_2 p_u p_v, \quad u \neq v. \tag{43.9}$$

Here k_0, k_1, k_2 are the probabilities that the relatives share 0, 1, 2 pairs of alleles identical by descent. Values for these coefficients for some common types of relationship are given in Table 43.4 and the match probabilities are

$$\Pr(A_u A_v | A_u A_v) = \begin{cases} k_0 p_u^2 + k_1 p_u + k_2, & u = v, \\ 2 k_0 p_u p_v + \frac{1}{2} k_1 (p_u + p_v) + k_2, & u \neq v. \end{cases}$$

To complete this section, consider the situation of non-inbred relatives in a population with population structure parameter θ where allele frequencies follow the Dirichlet distribution. Equations 43.4 and 43.5 are modified to

$$\Pr(A_u A_u, A_u A_u) = k_0 \Pr(A_u A_u A_u A_u) + k_1 \Pr(A_u A_u A_u) + k_2 \Pr(A_u A_u)$$

$$\Pr(A_u A_v, A_u A_v) = 4 k_0 \Pr(A_u A_u A_v A_v) + k_1 [\Pr(A_u A_u A_v) + \Pr(A_u A_v A_v)]$$
$$+ 2 k_2 \Pr(A_u A_v), \quad u \neq v.$$

Table 43.4 Identity by descent probabilities for common non-inbred relatives.

Relationship	k_0	k_1	k_2
Identical twins	0	0	1
Full-sibs	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Parent–child	0	1	0
Double first cousins	$\frac{9}{16}$	$\frac{3}{8}$	$\frac{1}{16}$
Half-sibs*	$\frac{1}{2}$	$\frac{1}{2}$	0
First cousins	$\frac{3}{4}$	$\frac{1}{4}$	0
Unrelated	1	0	0

*Also grandparent–grandchild and uncle–nephew.

as given by Fung *et al.* (2003). It needs to be stressed that these results hold only for non-inbred relatives –siblings whose parents are first cousins, e.g. can both be homozygous because their four alleles all descend from a single allele carried by their parents' grandparents. The allelic-set probabilities in these equations refer to the generation to which the relatives' most recent common ancestors belong. The match probabilities become

$$\Pr(A_u A_v | A_u A_v) = \begin{cases} k_0 \dfrac{[2\theta + (1-\theta)p_u][3\theta + (1-\theta)p_u]}{(1+\theta)(1+2\theta)} \\ \quad + k_1 \dfrac{2\theta + (1-\theta)p_u}{1+\theta} + k_2, & u = v, \\ k_0 \dfrac{2[\theta + (1-\theta)p_u][\theta + (1-\theta)p_v]}{(1+\theta)(1+2\theta)} \\ \quad + k_1 \dfrac{2\theta + (1-\theta)(p_u + p_v)}{2(1+\theta)} + k_2, & u \neq v. \end{cases}$$

Parameters p_u and θ are assumed to have the same value in successive generations, so that the same level of approximation holds as in 43.7.

43.4 PARENTAGE ISSUES

The previous section considered relationships among pairs of people caused by evolutionary processes and/or by family membership. Family relatedness lies at the heart of issues concerning parentage, whether these arise in civil paternity suits or in forensic situations such as incest. The genetic evidence often consists of three genetic profiles: those of the mother (G_m), her child (G_c) and the alleged father (G_a) although the child's genotype may be replaced by its paternal allele (A_p). The child's maternal allele is either known from a comparison of G_m and G_c or is assumed to be either of the mother's alleles with equal probability when mother and child are both heterozygotes for the same allele – in either case the paternal allele is deduced by subtraction from G_c. The LR, or paternity index (PI), can be written in several ways, such as

$$\text{PI} = \frac{\Pr(A_p | G_m, G_a, \text{ A is the father of C})}{\Pr(A_p | G_m, G_a, \text{ A is not the father of C})}.$$

The numerator of this expression follows immediately from Mendelian laws, but the denominator depends on the relationship between the three people: mother, alleged father and actual father.

If there is no immediate family relatedness among mother, alleged father and actual father, but they can all be considered to be random members of a population for which the Dirichlet distribution of the last section holds, then the various values of the PI are shown in Table 43.5. The first scenario in that table, e.g. has PI of $1/\Pr(A_u|A_u A_u A_u A_u)$, and the case where mother, child and alleged father are all $A_u A_v$ has a PI of $1/[\Pr(A_u|A_u A_v A_u A_v) + \Pr(A_v|A_u A_v A_u A_v)]$. These can be found from 43.6. Setting $\theta = 0$ leads to the classic results of $1/p_{A_p}$ or $1/2p_{A_p}$, according to whether the alleged father is homozygous or heterozygous for the parental allele A_p. When $\theta \neq 0$, however, it is necessary to consider the genotype of the mother as well as that of the alleged father.

Another extension to classic theory is to allow the alleged father to be related to the actual father for the PI denominator calculations. Provided the alleged father is not inbred and population structure can be ignored the relatedness between these two men can be described by the identity by descent coefficients shown in Table 43.4. If the alleged father is homozygous for the paternal allele A_p, the PI is $2/[(2k_2 + k_1) + (k_1 + 2k_0)p_{A_p}]$, and if he is heterozygous for the paternal allele the PI is half that value. The classic PI values result when $k_2 = k_1 = 0$. The quantity $(k_2/2 + k_1/4)$ is called the *coancestry* or *kinship coefficient of the two relatives*.

Table 43.5   Paternity index for population described by θ.

G_m	G_c	G_a	PI
$A_u A_u$	$A_u A_u$	$A_u A_u$	$\dfrac{1 + 3\theta}{4\theta + (1 - \theta)p_u}$
		$A_u A_w, w \neq u$	$\dfrac{1 + 3\theta}{2[3\theta + (1 - \theta)p_u]}$
	$A_u A_v$	$A_v A_v$	$\dfrac{1 + 3\theta}{2\theta + (1 - \theta)p_v}$
		$A_v A_w, w \neq v$	$\dfrac{1 + 3\theta}{2[\theta + (1 - \theta)p_v]}$
$A_u A_v$	$A_u A_u$	$A_u A_u$	$\dfrac{1 + 3\theta}{3\theta + (1 - \theta)p_u}$
		$A_u A_w, w \neq u$	$\dfrac{1 + 3\theta}{2[2\theta + (1 - \theta)p_u]}$
	$A_u A_v$	$A_u A_u$	$\dfrac{1 + 3\theta}{4\theta + (1 - \theta)(p_u + p_v)}$
		$A_u A_v$	$\dfrac{1 + 3\theta}{4\theta + (1 - \theta)(p_u + p_v)}$
		$A_u A_w, w \neq u, v$	$\dfrac{1 + 3\theta}{2[3\theta + (1 - \theta)(p_u + p_v)]}$

43.5   IDENTIFICATION OF REMAINS

Family relatedness may also be involved for identification of remains. A simple forensic situation is where a bloodstain is found in the car of a person suspected of having murdered a victim, but there is no body. The DNA profile of the bloodstain could be compared to that from material known to be from the victim or it could be compared to profiles determined for the victim's relatives. In that case, calculations similar to those for parentage issues are made. The need to identify bones has received attention in recent years with efforts to identify repatriated Korean and Vietnam War remains. STR profiles were used to identify 26 of 61 bodies following the Waco, Texas fire disaster (Clayton *et al.*, 1995) and 139 of 141 individuals in the 1996 Spitsbergen air accident (Olaisen *et al.*, 1997). There was complete success in identifying tissue recovered from the sites of the 1996 TWA 800 (Ballantyne, 1997) and the 1998 Swissair 111 (Robb, 1999) airplane disasters. A very large task confronted forensic scientists associated with the 2001 World Trade Center disaster, and a personal account is given by C.J. Brenner at `http://www.dna-view.com`. By the end of April 2005, remains of 1592 of the 2749 missing people had been identified and DNA profiling was the sole means of identification in 86 % of those identifications.

As an example of the type of LR calculation that might be undertaken in such cases, suppose that a DNA profile G_x is produced from remains thought to belong to a man for whom profiles are available from his wife G_w, his child G_c and his two parents G_m and G_f. The competing hypotheses, H_p and H_d, are that the remains either are or are not from the missing man. The LR is

$$\mathrm{LR} = \frac{\Pr(G_c, G_x, G_w, G_m, G_f | H_p)}{\Pr(G_c, G_x, G_w, G_m, G_f | H_d)}$$

$$= \frac{\Pr(G_c | G_x, G_m, H_p)\Pr(G_x | G_m, G_f, H_p)}{\Pr(G_c | G_w, G_m, G_f, H_d)\Pr(G_x)}.$$

The probabilities of profiles G_w, G_m, G_f do not depend on the two hypotheses and so cancel out of the ratio. The probability $\Pr(G_x)$ depends on allele frequencies and the remaining three probabilities follow from simple Mendelian calculations. Algebraic expressions in more complex cases are tedious to derive, but Brenner (1997) has developed software that will perform the appropriate symbolic manipulations. There is a growing use of Bayesian networks, or Probabilistic Expert Systems, to provide numerical solutions following the construction of a graph linking all profiles (Dawid and Evett, 1997; Evett *et al.*, 2002; Mortera, 2003).

43.6   MIXTURES

Forensic samples may contain material from more than one contributor. A common situation is for evidence collected in rape cases where material from the victim, possible consensual partners, and the perpetrator(s) may all be present. Even if some of these people contributed only a small proportion of the DNA in the sample, improved technology has made it easier to detect their alleles in the mixed profile. The genetic evidence E for

mixed-stain cases is the set of alleles found among all the people who have either been typed directly or whose type is inferred because they are considered to have contributed to the stain. Consideration of the alleles from people who may have been typed even though they are excluded and/or are hypothesised not to have contributed to the stain is necessary to allow for the effects of population structure.

In line with the principles of evidence interpretation, there needs to be alternative propositions that specify the numbers and profiles of contributors to the evidentiary sample. Some of these contributors will be known and typed people, and some will be unknown people. Those contributors, together with any typed people who are known (under the proposition) not to be contributors, contain among them a set of alleles whose probability depends on the separate allele proportions and the population structure parameter θ according to results such as those in 43.6. There is also a factor of 2 for each known heterozygote, and a term for the number of ways of arranging all $2x$ alleles from x unknown people into pairs. There may be different sets of alleles from unknown people under some propositions, and the probabilities for these sets must be added together. The LR is the ratio of probabilities under alternative propositions.

Much of the complexity in dealing with mixtures can be removed by a mnemonic notation, as laid out in Table 43.6 (Curran et al., 1999; Evett and Weir, 1998). There are sets of alleles (not necessarily distinct) that occur in the crime sample ($\mathcal{C}$). For a particular proposition there may be alleles ($\mathcal{T}$) carried by typed people declared to be contributors, alleles ($\mathcal{W}$) carried by unknown contributors to the sample, as well as alleles ($\mathcal{V}$) carried by people declared not to have contributed to the sample. There are corresponding sets of distinct alleles, and these sets are indicated by a g subscript. Note that the same person may be declared to be a contributor to the sample under one proposition, and declared not to be contributor under another proposition. Note also that the word 'known' in Table 43.6 refers to a value specified by the proposition under consideration.

The alleles in the evidence profile are carried either by typed people declared to be contributors or by unknown people, so that $\mathcal{C}$ is the combination (union) of sets $\mathcal{T}$ and $\mathcal{W}$. For a given proposition, the probability of the evidence profile depends also on the alleles carried by people who have been typed but are declared by that proposition not to have contributed to the profile. For a proposition in which there are x unknown contributors, the probability is $P_x(\mathcal{T}, \mathcal{W}, \mathcal{V}|\mathcal{C}_g)$. This probability is for all $2n_C + 2n_V = 2n_T + 2n_W + 2n_V$ alleles in the sets $\mathcal{T}, \mathcal{W}, \mathcal{V}$, among which allele A_u occurs $c_u + v_u = t_u + w_u + v_u$ times. The probabilities are added over all possible $n_x = (c + r - 1)!/[(c - 1)!r!]$ distinct sets of w_u. As listed in Table 43.7, c is the number of distinct alleles in $\mathcal{C}_g$ and r is the number of alleles carried by unknown people that can be any one of these c alleles.

Generating the n_x sets $\mathcal{W}$ is a two-stage process. Some of the alleles in each set must be present: these are the alleles in the set $\mathcal{C}_g$ that are not in set $\mathcal{T}_g$. Other alleles are not under this constraint because they already occur in $\mathcal{T}_g$, and there are r_u copies of A_u alleles in this unconstrained set. It is a straightforward computing task to let r_1 range over the integers $0, 1, \ldots, r$, then let r_2 range over the integers $0, 1, \ldots, r - r_1$, then let r_3 range over the integers $0, 1, \ldots, r - r_1 - r_2$, etc. The final count r_c is obtained by subtracting the sum of $r_1, r_2, \ldots, r_{c-1}$ from r. The total number of A_u alleles in set $\mathcal{W}$ is $\sum_{u=1}^{c} w_u = 2x$ where $w_u = r_u$ for those alleles in both $\mathcal{C}_g$ and $\mathcal{T}_g$, and $w_u = r_u + 1$ for alleles in $\mathcal{C}_g$ but not in $\mathcal{T}_g$.

For any ordering of the $2x = \sum_u w_u$ alleles in $\mathcal{W}$, successive pairs of alleles can be taken to represent genotypes and there are $(2x)!/(\prod_{u=1}^{c} w_u!)$ possible orderings. This is

Table 43.6 Notation for mixture calculations.

Alleles in the profile of the evidence sample

$\mathcal{C}$	The set of alleles in the evidence profile
$\mathcal{C}_g$	The set of distinct alleles in the evidence profile
n_C	The known number of contributors to $\mathcal{C}$
h_C	The unknown number of heterozygous contributors
c	The known number of distinct alleles in $\mathcal{C}_g$
c_u	The unknown number of copies of allele A_u in $\mathcal{C}$
	$1 \leq c_u \leq 2n_C, \sum_{u=1}^{c} c_u = 2n_C$

Alleles from typed people that H declares to be contributors

$\mathcal{T}$	The set of alleles carried by the declared contributors to $\mathcal{C}$
$\mathcal{T}_g$	The set of distinct alleles carried by the declared contributors
n_T	The known number of declared contributors to $\mathcal{C}$
h_T	The known number of heterozygous declared contributors
t	The known number of distinct alleles in $\mathcal{T}_g$ carried by n_T declared contributors
t_u	The known number of copies of allele A_u in $\mathcal{T}$
	$0 \leq t_u \leq 2n_T, \sum_{u=1}^{c} t_u = 2n_T$

Alleles from unknown people that H declares to be contributors

$\mathcal{W}$	The sets of alleles carried by the unknown contributors to $\mathcal{C}$
x	The specified number of unknown contributors to $\mathcal{C}$: $n_C = n_T + x$
$c - t$	The known number of alleles that are required to be in $\mathcal{W}$
r	The known number of alleles in $\mathcal{W}$ that can be any allele in $\mathcal{C}_g$, $r = 2x - (c - t)$
n_x	The number of different sets of alleles $\mathcal{W}$, $n_x = (c + r - 1)!/[(c - 1)!r!]$
r_u	The unknown number of copies of A_u among the r unconstrained alleles in $\mathcal{W}$
	$0 \leq r_i \leq r, \sum_{i=1}^{c} r_i = r$
w_u	The unknown number of copies of A_u in $\mathcal{W}$: $c_u = t_i + u_u, \sum_{u=1}^{c} u_u = 2x$
	If A_u is in $\mathcal{C}_g$ but not in $\mathcal{T}_g$: $u_u = r_u + 1$. If A_u is in $\mathcal{C}_g$ and also in $\mathcal{T}_g$: $w_u = r_u$

Alleles from typed people that H declares to be non-contributors

$\mathcal{V}$	The set of alleles carried by typed people declared not to be contributors to $\mathcal{C}$
n_V	The known number of people declared not to be contributors to $\mathcal{C}$
h_V	The known number of heterozygous declared non-contributors
v_i	The known number of copies of A_i in $\mathcal{V}$: $\sum_i v_i = 2n_V$

the number of possible sets of unknown genotypes that have each allelic set $\mathcal{W}$. Although it is the genotypes that correspond to the x unknown people, it is the set of $2x$ alleles that determine the probability, in combination with the $2n_T + 2n_V$ alleles among the known people. Because the n_T typed people all have specified genotypes, there is just a factor of 2 for each heterozygote, and there is a factor of 2 for each heterozygote among the set of n_V non-contributors.

Using 43.7, the probabilities of the set of alleles in the evidence is

$$P_x(\mathcal{T}, \mathcal{W}, \mathcal{V}|\mathcal{C}_g) = \sum_{r_1=0}^{r} \sum_{r_2=0}^{r-r_1} \cdots \sum_{r_{c-1}=0}^{r-r_1- \ldots -r_{c-2}}$$

$$\times \frac{(2x)!2^{H_T+H_V}}{\prod_{u=1}^{c} w_u!} \frac{\prod_{u=1}^{c} \prod_{j=0}^{t_u+w_u+v_u-1}[(1-\theta)p_u + j\theta]}{\prod_{j=0}^{2x+2n_T+2n_V-1}[(1-\theta) + j\theta]}. \quad (43.10)$$

Table 43.7 Effects of family relatedness on match probabilities.

Relationship	Match Probability
Homozygotes $A_u A_u$	
Full-sibs	$\dfrac{(1 + p_u)^2 + (7 + 7p_u - 2p_u^2)\theta + (16 - 9p_u + p_u^2)\theta^2}{4(1 + \theta)(1 + 2\theta)}$
Parent and child	$\dfrac{2\theta + (1 - \theta)p_u}{(1 + \theta)}$
Half-sibs	$\dfrac{[2\theta + (1 - \theta)p_u][2 + 4\theta + (1 - \theta)p_u]}{2(1 + \theta)(1 + 2\theta)}$
First cousins	$\dfrac{[2\theta + (1 - \theta)p_u][1 + 11\theta + 3(1 - \theta)p_u]}{4(1 + \theta)(1 + 2\theta)}$
Unrelated	$\dfrac{[2\theta + (1 - \theta)p_u][3\theta + (1 - \theta)p_u]}{(1 + \theta)(1 + 2\theta)}$
Heterozygotes $A_u A_v$	
Full-sibs	$\dfrac{\begin{array}{c}(1 + p_u + p_v + 2p_u p_v) \\ +(5 + 3p_u + 3p_v - 4p_u p_v)\theta + 2(4 - 2p_u - 2p_v + p_u p_v)\theta^2\end{array}}{4(1 + \theta)(1 + 2\theta)}$
Parent and child	$\dfrac{2\theta + (1 - \theta)(p_u + p_v)}{2(1 + \theta)}$
Half-sibs	$\dfrac{\begin{array}{c}(p_u + p_v + 4p_u p_v) \\ +(2 + 5p_u + 5p_v - 8p_u p_v)\theta + (8 - 6p_u - 6p_v + 4p_u p_v)\theta^2\end{array}}{4(1 + \theta)(1 + 2\theta)}$
First cousins	$\dfrac{\begin{array}{c}(p_u + p_v + 12p_u p_v) \\ +(2 + 13p_u + 13p_v - 24p_u p_v)\theta + 2(8 - 7p_u - 7p_v + 6p_u p_v)\theta^2\end{array}}{8(1 + \theta)(1 + 2\theta)}$
Unrelated	$\dfrac{2[\theta + (1 - \theta)p_u][\theta + (1 - \theta)p_v]}{(1 + \theta)(1 + 2\theta)}$

Likelihood ratios are formed as the ratios of two such probabilities.

Every person typed is declared to be either a contributor or a non-contributor. The number of people typed, and the alleles they carry among them, are the same for every proposition. For this reason, $n_T + n_V$, $H_T + H_V$ and $w_u + v_u$ will be the same in the probabilities for each proposition.

If population structure is ignored, and θ is set to zero, 43.10 reduces to

$$P_x(\mathcal{T}, \mathcal{W}, \mathcal{V} | \mathcal{C}_g) = \sum_{r_1=0}^{r} \sum_{r_2=0}^{r-r_1} \cdots \sum_{r_{c-1}=0}^{r-r_1-\ldots-r_{c-2}} \frac{(2x)! 2^{H_T + H_V}}{\prod_{u=1}^{c} w_u!} \prod_{u=1}^{c} p_u^{t_u + w_u + v_u},$$

as an alternative to the expression given by Weir *et al.* (1997). The value of LR now depends only on the numbers and frequencies of the alleles carried by unknown contributors. There is no need to consider the genotypes of typed people, whether or not they contribute to the evidence sample. This is different to the situation where population structure is taken into account – then the genotypes of all typed people are needed.

The arguments made for incorporating non-contributors can be extended. Several people may be typed during the course of an investigation. Even if they are excluded from being contributors, they provide information for the probability calculations when they can be considered to belong to the same subpopulation as (some of) people not excluded. They make their contribution to the calculation via allelic set $\mathcal{V}$.

Gill *et al.* (2006) discussed the complications for interpreting mixtures when some of the alleles in the mixed profile may be masked by typing artefacts such as stutter or may have dropped out completely and are not detected. A complete analysis needs to take into account the relative amounts of DNA inferred to be present at each of the alleles observed to be in the mixture. Having to allow for unseen alleles reduces the possibility of being able to exclude a potential contributor to the mixture simply because that person's alleles are not detected. Great care needs to be taken to avoid prejudicial conclusions if it is decided to ignore those loci in a profile for which interpretation is difficult or alleles are suspected of not being detected.

43.7 SAMPLING ISSUES

43.7.1 Allele Probabilities

In place of the unknown allele probabilities, it is usual to use the allele proportions obtained from a sample of individuals from the population (not from the specific relevant subpopulation). This leads to sampling variation in calculated match probabilities. Suppose the probability P_l at locus l is estimated as $\tilde{P}_l$ and that loci can be treated as being independent. Then the multilocus match probability $P = \prod_l P_l$ is estimated as $\tilde{P} = \prod_l \tilde{P}_l$ and the central limit theorem allows $\ln(\tilde{P})$ to be regarded as having a normal distribution. A 95 % confidence interval for P is, therefore, $(\tilde{P}/C, C\tilde{P})$ where $\ln(C) = 1.96\sqrt{\text{Var}[\ln(\tilde{P})]}$. The task is to estimate the variance of $\ln(\tilde{P})$. From the assumed independence of loci,

$$\text{Var}[\ln(\tilde{P})] = \text{Var}[\sum_l \ln(\tilde{P}_l)] \approx \sum_l \text{Var}(\tilde{P}_l)/P_l^2.$$

As θ is generally assigned a numerical value such as 0.03, rather than being estimated from sample data, it will be assumed constant. For a profile that is homozygous for allele u at locus l:

$$\text{Var}(\tilde{P}_l) \approx \left(\frac{\partial \tilde{P}_l}{\partial \tilde{p}_{lu}}\right)^2 \text{Var}(\tilde{p}_{lu})$$

and for a profile heterozygous for alleles u and v:

$$\text{Var}(\tilde{P}_l) \approx \left(\frac{\partial \tilde{P}_l}{\partial \tilde{p}_{lu}}\right)^2 \text{Var}(\tilde{p}_{lu}) + \left(\frac{\partial \tilde{P}_l}{\partial \tilde{p}_{lv}}\right)^2 \text{Var}(\tilde{p}_{lv}) + 2\left(\frac{\partial \tilde{P}_l}{\partial \tilde{p}_{lu}}\right)\left(\frac{\partial \tilde{P}_l}{\partial \tilde{p}_{lv}}\right) \text{Cov}(\tilde{p}_{lu}, \tilde{p}_{lv}).$$

The variances and covariances of for allele proportions are from Section 43.3.2

$$\text{Var}(\tilde{p}_u) = p_{lu}(1 - p_{lu})\left(\theta \sum_i w_i^2 + \frac{1-\theta}{2n_l}\right),$$

$$\text{Cov}(\tilde{p}_{lu}, \tilde{p}_{lv}) = -p_{lu}p_{lv}\left(\theta \sum_i w_i^2 + \frac{1-\theta}{2n_l}\right),$$

where n_l individuals are scored at locus l. Note that these include the binomial sampling variance within a population as well as the Dirichlet variance between populations. This second term seems necessary when the probabilities $\tilde{P}_l$ are estimated by substituting sample allele proportions into 43.7, since those equations are designed to incorporate evolutionary variation among populations.

The required variance for profile $A_{lu}A_{lv}$ is

$$
\mathrm{Var}[\ln(\tilde{P}_l)] = \begin{cases} p_{lu}(1 - p_{lu})(1 - \theta)^2 \left(\theta \sum_i w_i^2 + \dfrac{1 - \theta}{2n_l} \right) \\ \quad \times \left(\dfrac{1}{3\theta + (1 - \theta)p_{lu}} + \dfrac{1}{2\theta + (1 - \theta)p_{lu}} \right)^2, & u = v, \\[2ex] (1 - \theta)^2 \left(\theta \sum_i w_i^2 + \dfrac{1 - \theta}{2n_l} \right) \left(\dfrac{p_{lu}(1 - p_{lu})}{[\theta + (1 - \theta)p_{lu}]^2} \right. \\ \quad \left. - \dfrac{2p_{lu}p_{lv}}{[\theta + (1 - \theta)p_{lu}][\theta + (1 - \theta)p_{lv}]} + \dfrac{p_{lv}(1 - p_{lv})}{[\theta + (1 - \theta)p_{lv}]^2} \right), & u \neq v. \end{cases}
$$

In practice, the allele probabilities p_{lu} are replaced by sample values $\tilde{p}_{lu}$.

When $\theta = 0$, the variances reduce to

$$
\mathrm{Var}[\ln(\tilde{P}_l)] = \begin{cases} \dfrac{2(1 - p_{lu})}{n_l p_{lu}}, & u = v, \\[2ex] \dfrac{(p_{lu} + p_{lv} - 4p_{lu}p_{lv})}{2n_l p_{lu} p_{lv}}, & u \neq v. \end{cases}
$$

43.7.2 Coancestry

As mentioned earlier, the parameter θ that features so prominently in many of the genetic forensic calculations is generally assigned a numerical value rather than being estimated from forensic databases. Part of the reason for this is the difficulty in obtaining data from subpopulations to allow estimation of θ, the very quantity introduced to avoid having to collect such data in the first place. If observations were available at the crime-relevant subpopulation level, there would not be a need for θ in 43.7.

There are occasions, however, when multiple samples are available from the same major racial classification. Population geneticists have long estimated θ or F_{ST} from such data (e.g. Cavalli-Sforza *et al.*, (1994); **Chapter 29**). A classical procedure for estimation rests on the method of moments (Weir and Cockerham, 1984), generally within a hierarchical sampling framework such as alleles within individuals, individuals within subpopulations, and subpopulations within populations. Recent work by Weir and Hill (2002) has re-examined this classical approach, and has considered the effects of assuming that allele frequencies are normally distributed over populations. For large sample sizes, θ can be estimated for locus l as

$$
\hat{\theta} = \frac{1}{(r - 1)(m_l - 1)} \sum_{i=1}^{r} \sum_{u=1}^{m} \frac{(\tilde{p}_{liu} - \bar{p}_{lu})^2}{\bar{p}_{lu}}.
$$

In this equation, $\tilde{p}_{liu}$ is the sample frequency of allele A_{lu} at locus l in the ith of r samples, and u ranges from 1 to m_l. The quantity $\bar{p}_{lu}$ is the average frequency of A_{lu} over the whole dataset. Estimates can be averaged over loci under the assumption that θ

is the same for each locus. The average estimate is distributed as θ times a χ^2 distribution with $d = (r-1)\sum_l(m_l - 1)$ df, so that a 95 % confidence interval for θ is

$$\left(\frac{d\hat{\theta}}{X_{0.975}}, \frac{d\hat{\theta}}{X_{0.025}}\right),$$

where $X_{0.025}$, $X_{0.975}$ are the 2.5th and 97.5th percentiles of the χ^2 distribution with d df. This asymmetric confidence interval is analogous to the asymmetric posterior probability intervals found by Ayres and Balding (1998) from a Bayesian perspective. It would be possible to incorporate this variation into expressions for the sampling variation of estimated profile probabilities.

43.8 OTHER FORENSIC ISSUES

There are several issues in addition to those already covered that affect the interpretation of DNA evidence. The various problems described in this section could all be avoided by starting with the principles of evidence interpretation and proceeding to an evaluation of conditional profile probabilities. Future directions are indicated by efforts to assign individuals to phenotypic classes or ethnic groups, and to place DNA profile statements in a hierarchy of propositions.

43.8.1 Common Fallacies

By far the most common fallacy associated with the forensic use of DNA is the Prosecutor's fallacy mentioned earlier. Virtually every media account of courtroom testimony contains statements of the form 'Based on the DNA evidence, the forensic scientist testified that there was only one chance in a million that someone other than the defendant left the crime sample' whereas the scientist most probably stated the result correctly as 'If the defendant did not leave the crime sample, there is only one chance in a million that it would match his type.' Weir (2000) has described the frustration in trying to have journalists write carefully, and even writers for *Science* are not immune from the fallacy (*Science* 278:1407, 1997).

The defense attorney's fallacy (Thompson and Schumann, 1987) is less common but may appear compelling to a jury. If the profile in question has a probability of 1 in 1 million, and the defendant lives in a country of 100 million, the defense may claim that he is merely 1 person among the 100 who have that profile. They may even suggest there is a 99 % chance he is not the source of the crime sample. The fallacy arises from a confusion between the probability of the profile under H_d (and hence the LR) and the posterior probability of H_d. It implies the assignment of equal prior probabilities to everyone in the country, and it shows a lack of appreciation for the distribution of the number of matches about the expected value. A full discussion of the related 'island problem' has been given by Dawid (1994) and Balding and Donnelly (1995).

43.8.2 Relevant Population

Estimated profile or match probabilities refer to a population. They are not proportions of people in that population with the profile but instead refer to the probability with which a

person in the population either has the profile or matches the profile from someone already typed. Which population is meant? As the calculations usually refer to the alternative proposition H_d, it is clear that the suspect's ethnicity, e.g. does not define the population. That person did not contribute the crime profile, so his characteristics do not define the population to which the actual contributor belongs (Lewontin, 1993; Weir and Evett, 1992; 1993). However, there is a tendency to concentrate on calculations based on allelic frequencies in a sample of people with the same racial designation as the suspect.

It is proper to define the population by the circumstances of the crime, and these may point to a geographic region or an ethnic group. Usually the conditional probability $\Pr(G_c|G_s, H_d)$ will be greater when the actual contributor and the suspect belong to the same (sub)population, so that focusing on the suspect's racial group has an element of conservativeness. Balding (1999) has presented a more satisfactory approach of allowing the unknown contributor under H_d to belong to one of very many different sets of people defined by the relationship to the suspect: the suspect's siblings, his other relatives, other members of his subpopulation, other members of his racial group, or anyone outside his racial group.

43.8.3 Database Searches

One of the few remaining statistical areas of debate in the forensic uses of DNA concerns the effects of database searches. The treatment so far in this chapter has considered the situation where a suspect has been identified by an investigation and is then typed and found to have a DNA profile matching that of a crime sample. For crimes with no suspect, it is now possible to search large databases of profiles from known individuals such as previously convicted offenders. Does the evidence have any less value against a person who becomes a suspect because his profile was identified after such a search, than when the person was a suspect before being found to match? The confusion has arisen because of the observation that the expected number of matches increases with the size of the database, and there has been a recommendation (National Research Council, 1996) that a profile probability of P should be modified to NP when a suspect is identified by searching a database of size N.

Balding and Donnelly (1996) and Donnelly and Friedman (1999) present careful analyses to show that the LR for the database search case cannot be greater than for the single suspect case. The evidence against the one person who goes to trial is essentially unaltered by the means by which he was identified. It is not the database that is on trial. In responding to Stockmarr (1999), Evett *et al.* (2000) show how the confusion can be avoided by adhering to the Principles of Evidence Interpretation and recognising the difference between LRs and posterior odds. There could well be an argument made that the prior odds against the suspect are less if he was identified by a database search, so that the posterior odds will also be less since the LR is unaltered.

43.8.4 Uniqueness of Profiles

As predicted by Stigler (1995), DNA profiles are now regarded as being as valuable and reliable as fingerprints. Part of this acceptance rests on the very successful use of DNA profiling to identify remains after mass disasters such as airplane crashes, and was not hindered by the anomalous verdict in the criminal trial of O.J. Simpson (Weir, 1995). With this growing acceptance has come less of a need to present statistical arguments.

Moreover, the increasing number of loci used for forensic DNA profiles has made any statistics appear to be beyond credibility. Maybe for these reasons, there has been a move by the Federal Bureau of Investigation (FBI) to dispense with numbers (reported in Science 278:1407, 1997).

Some relevant discussion was given by Kingston (1965) long before the advent of DNA profiling. If a particular item of evidence has a probability P, then he assumed that the unknown number x of occurrences of the profile in a large population of N people is Poisson with parameter $\lambda = NP$: $\Pr(x) = \lambda^x e^{-\lambda}/x!$. If λ is large the binomial distribution would be needed. Both Poisson and binomial require profiles from different people to be independent, so neither can be strictly true for DNA profiles. Suppose a person with the particular profile commits a crime, leaves evidence with that profile at the scene, and then rejoins the population. A person with the profile is subsequently found in the population. A simple model says that the probability that this person is the perpetrator is $1/x$. Although x is not known, it must be at least one, so the probability that the correct person has been identified is the expected value of $1/x$ given that $x \geq 1$:

$$\Pr(\text{correct}) = \frac{\sum_{x \geq 1} \frac{1}{x} \Pr(x)}{\sum_{x \geq 1} \Pr(x)} \approx 1 - \frac{\lambda}{4}.$$

For the United States, with a population of about 3×10^8, a profile with a probability of 10^{-10} would give $\lambda = 0.03$ and a probability that the correct person had been identified of 0.9925. Kingston (1965) went on to determine the probability that at least two people in the population have a particular profile given that at least one person is known to have it. This probability is

$$\Pr(\text{not unique}) = \frac{\Pr(x \geq 2)}{\Pr(x \geq 1)} \approx \frac{\lambda}{2}.$$

For the USA example, the probability that the profile is unique, $(1 - \lambda/2)$, is 0.985.

An alternative calculation was provided by Balding (1999). He supposed that a person (the perpetrator) is sampled anonymously and randomly from a population of $(N + 1)$ people and found to have a certain DNA profile. Each other person has the same probability P of having that profile. A second person (the suspect) is sampled from the population and (event E) is found to have the same profile as the first. Event U is that the second person matches the first and that there is no other person in the population with the same profile. It is possible (event G), with probability $1/(N + 1)$ that the second person is actually the first person. The conditional probability $\Pr(G|E)$, from Bayes' theorem, is $1/(1 + NP)$ since the probability of E given that G did not occur is just P. The probability of U given both G and E is the probability that none of N people have the profile: $(1 - P)^N$, so

$$\Pr(U|E) = \Pr(U|G, E)\Pr(G|E) = \frac{(1 - P)^N}{1 + NP} > 1 - 2\lambda.$$

For the USA example, this lower bound is 0.94.

It is not clear which of the three probabilities (the two of Kingston or that of Balding) the FBI wished to determine, but they did want a value of 0.01 which requires $P = 1/25N$ or $P = 1/50N$ for Kingston and $P = 1/200N$ for Balding. The FBI then reduced P by a factor of 10 to account for uncertainty in estimating its value. There are two problems

with this approach (Weir, 1999; 2001). In the first place, all N profiles in a population are not independent and the Poisson/binomial calculation ignores all the issues of conditional probabilities, population structure and relatedness discussed above. The more serious problem may be that of perception – there is quite a difference between telling a jury that the suspect has been identified as the perpetrator by his DNA profile and telling them that there is a 1 % chance someone else has this profile. The absoluteness implied by statements of identity is not a statistical concept.

A related set of calculations has to do with multiple occurrences of *any* profile, not a particular profile, in a database. This is the so-called birthday problem. The probability that at least two of a sample of n people have the same unspecified birthday (or profile) for the case where every birthday (or profile) has the same probability P, is

$$\text{Pr(at least one match)} = 1 - \text{Pr(no matches)}$$

$$= 1 - \{1(1-P)(1-2P)\cdots[1-(n-1)p]\}$$

$$\approx 1 - \prod_{i=0}^{n-1} e^{-iP} \approx 1 - e^{-n\lambda/2}.$$

In the classic birthday problem, $P = 1/365$ and the probability exceeds 50 % once n gets as large as 23. For the USA example of $P = 10^{-10}$ the chance of some profile being replicated in the population of $N = 3 \times 10^8$ is essentially 100 %. The Arizona Department of Public Safety reported a 'near match' in a database of 65 493 for a profile that had an estimated probability of 1 in 3.1 billion. Using that probability, the chance of finding two matching profiles in the database would be 50 % so it would not have been surprising if they had found a complete match. There are about two billion pairs of people in the Arizona database, so the occurrence of one matching pair is seen to still be a rare event. It is important to stress that the birthday problem requires two or more copies of some birthday (profile), not two or more copies of a particular profile.

There is an interesting twist to the notion expressed above that DNA profiling has become as well-accepted as has fingerprinting. Publications have begun to appear that apply LR approaches to fingerprint minutiæ(Neumann *et al.*, 2006). Such work is needed to assess the evidential contribution of fingerprints that may be partial, distorted or with poor signal to noise ratio.

43.8.5   Assigning Individuals to Phenotypes, Populations or Families

An individual's phenotype depends ultimately on his genetic makeup, and there have been hopes that DNA profiles could be used to identify the ethnicity or physical appearance for the unknown donor of an evidentiary stain. There is the trivial example of sex-linked markers being used to determine gender, or at least X,Y chromosomal composition. There has also been some progress in predicting phenotype when that is limited to red hair (Grimes *et al.*, 2001) because many cases of red hair can be associated with the genotype at a single locus, the melanocortin-1 receptor. Grimes *et al.* conclude 'Given that between 5.3 and 11 % of the population of the British Isles are reported to have red hair, a positive test result could have a major impact on the course of an investigation.'

Much less certainty can be attached to inferring ethnic origin. Lowe *et al.* (2001) used the UK Forensic Science Service panel of six STR loci to suggest a means of reducing the number of interviews needed to resolve a case. They first estimated the

probability of the crime scene profile G for each ethnic group r for which they have allele frequencies. If I indicates an investigator's prior information and $\Pr(r|I)$ his or her prior probabilities for ethnic origin of the crime-scene stain donor, Bayes' theorem provides posterior probabilities of

$$\Pr(r|G, I) = \frac{\Pr(G|r, I)}{\sum_s \Pr(G|s, I)\Pr(s|I)}.$$

If an investigator faces the task of interviewing and/or DNA testing all members of a population until the true donor of the stain is found, then the expected number of interviews may be reduced if the posterior probabilities are used. For example, if the posterior probability is highest for ethnic group R, then preferentially interviewing members of that group will be efficient if the probability is greater than the proportion of people in the population who are of origin R.

Shriver *et al.* (1997) adopted an alternative procedure of looking for DNA loci that are particularly discriminating. They claimed evidence for a set of 10 loci, selected from a set of 1000, that would provide good discrimination between African Americans and European Americans. Their methods and conclusions were challenged by Brenner (1998), who points to the danger of selecting the most discriminating loci from a large number of loci based on the performances in small samples. There is the possibility that 'From among 1000 loci, one could similarly find a set of 10 loci that differentiate the 9-year olds from the 10-year olds in the local playground.' Nevertheless, Brenner supports the work of Lowe *et al.* (2001) that the STR loci used by forensic scientists have good potential for ethnic assignment.

Genetic markers can also be used to assign individuals to families and this is of particular use following mass disasters. An alternative use of markers for making inferences about relatedness is that of familial searching. Bieber *et al.* (2006) describe the situation where an evidentiary profile is compared to every profile in a database and the LR for the two profiles being from (specified) relatives or from unrelated people is calculated. High LRs are suggestive of relatedness, Bieber *et al.* show by simulation when the database did contain one true brother or father of the donor of the evidence sample that, about half the time, the highest LR identified the true relative. To achieve 90 % probability of identifying a subset of the database that included the true relative, it may be necessary to consider at least the 100 highest LRs. Other issues in the estimation of relatedness of relationship from genotypes were reviewed by Weir *et al.* (2006).

43.8.6 Hierarchy of Propositions

This chapter has been concerned mainly with the statistical procedures to quantify the strength of the evidence provided by sets of DNA profiles, especially in the case of matching profiles. These statistical methods are generally straightforward, but they do not capture all the forensic issues. Evett *et al.* (2002) have introduced the concept of a 'hierarchy of propositions' The highest-level propositions have to do with the guilt or innocence of a defendant, whereas the lowest-level propositions have to do with whether a DNA sample is from the defendant. The authors conclude with the substantial point that 'even if the inference with regard to whether or not the source of a sample of DNA is effectively indisputable, the inference with regard to whether the defendant is the offender may be subject to considerable uncertainty'.

43.9   CONCLUSIONS

Quantifying the evidentiary value of matching genetic profiles requires a comparison of the probabilities of the evidence under alternative propositions. These probabilities, in turn, require the probability that an unknown person has the profile given that a known person has the profile, and this is a statistical genetic quantity. The match probability depends on the relationship between the two people, whether this is due to joint family membership, or simply joint membership in some population.

Even though the propositions concerning the origins of a crime stain may list only one person, other people may have been typed as part of the investigation. Their profiles affect the match probabilities under a general model that allows for population structure. All expressions are considerably simplified when the population can be assumed to be have reached an equilibrium stage under the action of evolutionary forces, but this theory is presently in place only for single loci. An assumption is made that different loci are independent, so that match probabilities may be multiplied over loci.

There are many advantages to a general and coherent approach to the interpretation of genetic evidence. This approach, expounded upon by the authors already cited is finding acceptance by forensic scientists in many countries, but it requires a willingness to abandon simplistic approaches that appeal to expedience. The interests of justice, in the long run, will best be served by sound statistical theories.

REFERENCES

Aitken, C.G.G. (1995). *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & sons, New York.

Ayres, K.L. and Balding, D.J. (1998). Measuring departures from Hardy-Weinberg: a Markov Chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity* **80**, 769–770.

Ayres, K.L. and Overall, A.D.J. (1999). Allowing for within-subpopulation inbreeding in forensic match probabilities. *Forensic Science International* **103**, 207–216.

Balding, D.J. (1999). When can a DNA profile be regarded as unique? *Science and Justice* **39**, 257–260.

Balding, D.J. (2005). *Weight-of-evidence for DNA Profiles*. John Wiley & Sons, Chichester.

Balding, D.J. and Donnelly, P. (1995). Inference in forensic identification. *Journal of the Royal Statistical Society Series A* **158**, 21–53.

Balding, D.J. and Donnelly, P. (1996). Evaluating DNA profile evidence when the suspect is identified through a database search. *Journal of Forensic Sciences* **41**, 603–607.

Balding, D.J. and Nichols, R.A. (1994). DNA match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International* **64**, 125–140.

Ballantyne, J. (1997). Mass disaster genetics. *Nature Genetics* **15**, 329–331.

Bieber, F.R., Brenner, C.H. and Lazer, D. (2006). Finding criminals through DNA of their relatives. *Science* **312**, 1315–1316.

Brenner, C.H. (1997). Symbolic kinship program. *Genetics* **145**, 535–542.

Brenner, C.H. (1998). Difficulties in the estimation of ethnic affiliation. *American Journal of Human Genetics* **62**, 1558–1560.

Buckleton, J.S., Triggs, C.M. and Walsh, S.J. (2005). *Forensic DNA Evidence Interpretation*. CRC Press, Boca Raton, FL.

Cavalli-Sforza, L.L., Menozzi, P. and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.

Clayton, T.M., Whitaker, J.P. and Maguire, C.N. (1995). Identification of bodies from the scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci. *Forensic Science International* **76**, 7–15.

Cockerham, C.C. (1971). Higher order probability functions of identity of alleles by descent. *Genetics* **69**, 235–246.

Curran, J., Triggs, C.M., Buckleton, J. and Weir, B.S. (1999). Interpreting DNA mixtures in structured populations. *Journal of Forensic Sciences* **44**, 987–995.

Dawid, A.P. (1994). *Aspects of uncertainty: a tribute to D.V. Lindley*, P.R. Freeman and A.F.M. Smith, eds. John Wiley & Sons, Chichester, pp. 159–170.

Dawid, A.P. and Evett, I.W. (1997). Using a graphical method to assist the evaluation of complicated patterns of evidence. *Journal of Forensic Sciences* **42**, 226–231.

Donnelly, P. and Friedman, D. (1999). DNA database searches and the legal consumption of scientific evidence. *Michigan Law Review* **97**, 931–984.

Evett, I.W., Foreman, L.A. and Weir, B.S. (2000). Letter to the Editor concerning a paper by A. Stockmarr. *Biometrics* **56**, 1274–1275. Response to Devlin, *Biometrics* **56**, 1277.

Evett, I.W., Gill, P.D., Jackson, G., Whitaker, J. and Champod, C. (2002). Interpreting small quantities of DNA: the hierarchy of propositions and the use of Bayesian networks. *Journal of Forensic Sciences* **47**, 520–530.

Evett, I.W. and Weir, B.S. (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Science*. Sinauer, Sunderland, MA.

Fung, W.K., Carracedo, A. and Hu, Y.-Q. (2003). Testing for kinship in a subdivided population. *Forensic Science International* **135**, 105–109.

Gill, P., Brenner, C.H., Buckleton, J.S., Carracedo, A., Krawczak, M., Mayr, W.M., Morling, N., Prinz, M., Schneider, P.M. and Weir, B.S. (2006). DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International* **160**, 90–101.

Graham, J., Curran, J. and Weir, B.S. (2000). Conditional genotypic probabilities for microsatellite loci. *Genetics* **155**, 1973–1980.

Grimes, E.A., Noake, P.J., Dixon, L. and Urquhart, A. (2001). Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype. *Forensic Science International* **122**, 124–129.

Kingston, C.R. (1965). Applications of probability theory in criminalistics. *Journal of the American Statistical Association* **60**, 70–80.

Lewontin, R.C. (1993). Which population?. *American Journal of Human Genetics* **52**, 205.

Lowe, A.L., Urquhart, A., Foreman, L.A. and Evett, I.W. (2001). Inferring ethnic origin by means of an STR profile. *Forensic Science International* **119**, 17–22.

Lucy, D. (2005). *Introduction to Statistics for Forensic Scientists*. John Wiley & Sons, Chichester.

Maiste, P.J. and Weir, B.S. (1995). A comparison of tests for independence in the FBI RFLP databases. *Genetica* **96**, 125–138.

Mortera, J. (2003). *Highly Structured Stochastic Systems*, P.J. Green, N.L. Hjort and S. Richardson, eds. Oxford University Press, Oxford.

National Research Council (1996). *The Evaluation of Forensic DNA Evidence*. National Academy Press, Washington, DC.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., Meuwly, D. and Bromage-Griffiths, A. (2006). Computation of likelihood ratios in fingerprint identification for configurations of three minutiæ. *Journal of Forensic Sciences* **51**, 1255–1266.

Olaisen, B., Sternersen, M. and Mevåg, B. (1997). Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Nature Genetics* **15**, 404–405.

Robb, N. (1999). 229 people, 15,000 body parts: pathologists help solve Swissair 111's grisly puzzles. *Canadian Medical Association Journal* **160**, 241–243.

Robertson, B. and Vignaux, G.A. (1995). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. John Wiley & Sons, Chichester.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall, London.

Scholl, S., Budowle, B., Radecki, K. and Salvo, M. (1996). Navajo, Pueblo, and Sioux population data on the loci HLA-DQA1, LDLR, GYPA, HBGG, Gc, and D1S80. *Journal of Forensic Sciences* **41**, 47–51.

Schum, D.A. (1994). *Evidential Foundations of Probabilistic Reasoning*. John Wiley & Sons, New York.

Shoemaker, J., Painter, I.S. and Weir, B.S. (1998). A Bayesian characterization of Hardy-Weinberg disequilibrium. *Biometrics* **25**, 235–254.

Shriver, M.D., Smith, M.W., Jin, L., Marcini, A., Akey, J.M., Deka, R. and Ferrell, R.E. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics* **60**, 957–964.

Stigler, S.M. (1995). Galton and identification by fingerprints. *Genetics* **140**, 857–860.

Stockmarr, A. (1999). Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. *Biometrics* **55**, 671–677.

Thompson, W.C. and Schumann, E.L. (1987). Interpretation of statistical evidence in criminal trials–the prosecutors fallacy and the defense attorneys fallacy. *Law and Human Behavior* **11**, 167–187.

Weir, B.S. (1992). Independence of VNTR alleles defined as fixed bins. *Genetics* **130**, 873–887.

Weir, B.S. (1995). DNA statistics in the Simpson matter. *Nature Genetics* **11**, 365–368.

Weir, B.S. (1996). *Genetic Data Analysis II*. Sinauer, Sunderland, MA.

Weir, B.S. (1999). Are DNA profiles unique? In Proceedings of the 9th International Symposium on Human Identification, Orlando, Florida, pp. 114–117. `http://www.promega.com/geneticidentity/proceed.html`.

Weir, B.S. (2000). In *Statistical Science in the Courtroom*, J. Gastwirth, ed. Springer-Verlag, New York.

Weir, B.S. (2001). DNA match and profile probabilities: Comment on Budowle (2000) and Fung and Hu (2000). Forensic Science Communications 3. `http://www.fbi.gov/hq/lab/fsc/backissu/jan2001/weir.htm`.

Weir, B.S., Anderson, A.B. and Hepler, A.M. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics* **7**, 771–780.

Weir, B.S. and Cockerham, C.C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.

Weir, B.S. and Evett, I.W. (1992). Whose DNA? *American Journal of Human Genetics* **50**, 869.

Weir, B.S. and Evett, I.W. (1993). Reply to Lewontin. *American Journal of Human Genetics* **52**, 206.

Weir, B.S. and Hill, W.G. (2002). Estimating *F*-statistics. *Annual Review of Genetics* **36**, 721–750.

Weir, B.S., Triggs, C.M., Starling, L., Stowell, L.I., Walsh, K.A.J. and Buckleton, J.S. (1997). Interpreting DNA mixtures. *Journal of Forensic Sciences* **42**, 113–122.

Zaykin, D., Zhivotovsky, L.A. and Weir, B.S. (1995). Exact tests for association between alleles at arbitrary numbers of loci. *Genetica* **96**, 169–178.

Zaykin, D., Zhivotovsky, L.A., Westfall, P.H. and Weir, B.S. (2002). Truncated product method for combining *p*-values. *Genetic Epidemiology* **22**, 170–185.

Reference Author Index

Subject Index