

# Advanced Molecular Biology

A Concise Reference



R. M. Twyman

# **Advanced Molecular Biology**

A Concise Reference



To my parents, Peter and Irene  
and to my children, Emily and Lucy

# Advanced A Concise Reference Molecular Biology

**Richard M. Twyman**

*Neurobiology Division, MRC Laboratory of Molecular Biology, Hills Road,  
Cambridge CB2 2QH, UK*

**Consultant Editor**

**W. Wisden**

*MRC Laboratory of Molecular Biology, Cambridge, UK*

**BIOS**  
**SCIENTIFIC**  
**PUBLISHERS**

---

© BIOS Scientific Publishers Limited, 1998

First published 1998

All rights reserved. No part of this book may be reproduced or transmitted, in any form or by any means, without permission.

A CIP catalogue record for this book is available from the British Library.

ISBN 1 85996 141 X

---

**BIOS Scientific Publishers Ltd**  
9 Newtec Place, Magdalen Road, Oxford OX4 1RE, UK  
Tel: +44 (0)1865 726286. Fax: +44 (0)1865 246823  
World Wide Web home page: <http://www.bios.co.uk/>

**DISTRIBUTORS**

*Australia and New Zealand*

Blackwell Science Asia  
54 University Street  
Carlton, South Victoria 3053

*India*

Viva Books Private Limited  
4325/3 Ansari Road, Daryaganj  
New Delhi 110002

Published in the United States of America, its dependent territories and Canada by Springer-Verlag New York Inc., 175 Fifth Avenue, New York, NY 10010-7858, in association with BIOS Scientific Publishers Ltd

Published in Hong Kong, Taiwan, Singapore, Thailand, Cambodia, Korea, The Phillippines, Indonesia, The People's Republic of China, Brunei, Laos, Malaysia, Macau and Vietnam by Springer-Verlag Singapore Pte. Ltd, 1 Tannery Road, Singapore 347719, in association with BIOS Scientific Publishers Ltd

---

Production Editor: Andrea Boshier.  
Typeset by Poole Typesetting (Wessex) Ltd, Bournemouth, UK.  
Printed by Redwood Books, Trowbridge, UK.

# Contents

<b>Abbreviations</b>	<b>ix</b>
<b>How to use this book</b>	<b>xi</b>
<b>Preface</b>	<b>xii</b>
<b>1. Biological Heredity and Variation</b>	<b>1</b>
Mendelian inheritance	1
Segregation at one locus	2
Segregation at two loci	8
Quantitative inheritance	11
<b>2. The Cell Cycle</b>	<b>21</b>
The bacterial cell cycle	21
The eukaryotic cell cycle	23
The molecular basis of cell cycle regulation	26
Progress through the cell cycle	28
Special cell cycle systems in animals	33
<b>3. Chromatin</b>	<b>35</b>
Nucleosomes	35
Higher order chromatin organization	38
Chromatin and chromosome function	39
Molecular structure of the bacterial nucleoid	42
<b>4. Chromosome Mutation</b>	<b>45</b>
Numerical chromosome mutations	45
Structural chromosome mutations	49
<b>5. Chromosome Structure and Function</b>	<b>57</b>
Normal chromosomes — gross morphology	57
Special chromosome structures	58
Molecular aspects of chromosome structure	60
<b>6. Development, Molecular Aspects</b>	<b>65</b>
Differentiation	65
Pattern formation and positional information	72
The environment in development	75
<b>7. DNA Methylation and Epigenetic Regulation</b>	<b>93</b>
DNA methylation in prokaryotes	93
DNA methylation in eukaryotes	94
Epigenetic gene regulation by DNA methylation in mammals	97
<b>8. The Gene</b>	<b>103</b>
The concept of the gene	103

Units of genetic structure and genetic function	104
Gene-cistron relationship in prokaryotes and eukaryotes	104
Gene structure and architecture	106
<b>9. Gene Expression and Regulation</b>	<b>111</b>
Gene expression	112
Gene regulation	113
Gene expression in prokaryotes and eukaryotes	115
<b>10. Gene Transfer in Bacteria</b>	<b>117</b>
Conjugation	117
Transformation	119
Transduction	120
<b>11. The Genetic Code</b>	<b>127</b>
An overview of the genetic code	127
Translation	127
Special properties of the code	129
<b>12. Genomes and Mapping</b>	<b>133</b>
Genomes, ploidy and chromosome number	134
Physico-chemical properties of the genome	134
Genome size and sequence components	135
Gene structure and higher-order genome organization	136
Repetitive DNA	139
Isochore organization of the mammalian genome	143
Gene mapping	144
Genetic mapping	146
Physical mapping	151
<b>13. Mobile Genetic Elements</b>	<b>165</b>
Mechanisms of transposition	166
Consequences of transposition	172
Transposons	175
Retroelements	180
<b>14. Mutagenesis and DNA Repair</b>	<b>183</b>
Mutagenesis and replication fidelity	183
DNA damage: mutation and killing	185
DNA repair	187
Direct reversal repair	187
Excision repair	191
Mismatch repair	194
Recombination repair	197
The SOS response and mutagenic repair	197
<b>15. Mutation and Selection</b>	<b>201</b>
Structural and functional consequences of mutation	201
Mutant alleles and the molecular basis of phenotype	209
The distribution of mutations and molecular evolution	211
Mutations in Genetic Analysis	213

<b>16. Nucleic Acid Structure</b>	<b>223</b>
Nucleic acid primary structure	223
Nucleic acid secondary structure	226
Nucleic acid tertiary structure	231
<b>17. Nucleic Acid-Binding Properties</b>	<b>235</b>
Nucleic acid recognition by proteins	236
DNA-binding motifs in proteins	237
RNA-binding motifs in proteins	243
Molecular aspects of protein-nucleic acid binding	244
Sequence-specific binding	246
Techniques for the study of protein-nucleic acid interactions	249
<b>18. Oncogenes and Cancer</b>	<b>253</b>
Oncogenes	254
Tumor-suppressor genes	258
<b>19. Organelle Genomes</b>	<b>263</b>
Organelle genetics	263
Organelle genomes	264
<b>20. Plasmids</b>	<b>271</b>
Plasmid classification	271
Plasmid replication and maintenance	273
<b>21. The Polymerase Chain Reaction (PCR)</b>	<b>279</b>
Specificity of the PCR reaction	279
Advances and extensions to basic PCR strategy	283
Alternative methods for <i>in vitro</i> amplification	284
<b>22. Proteins: Structure, Function and Evolution</b>	<b>287</b>
Protein primary structure	288
Higher order protein structure	289
Protein modification	295
Protein families	297
Global analysis of protein function	304
<b>23. Protein Synthesis</b>	<b>313</b>
The components of protein synthesis	313
The mechanism of protein synthesis	315
The regulation of protein synthesis	318
<b>24. Recombinant DNA and Molecular Cloning</b>	<b>323</b>
Molecular cloning	324
Strategies for gene isolation	331
Characterization of cloned DNA	336
Expression of cloned DNA	339
Analysis of gene regulation	342
Analysis of proteins and protein-protein interactions	345

<i>In vitro</i> mutagenesis	346
Transgenesis: gene transfer to animals and plants	348
<b>25. Recombination</b>	<b>369</b>
Homologous recombination	369
Homologous recombination and genetic mapping	373
Random and programmed nonreciprocal recombination	376
Site specific recombination	378
Generation of immunoglobulin and T-cell receptor diversity	379
Illegitimate recombination	382
<b>26. Replication</b>	<b>389</b>
Replication strategy	389
The cellular replisome and the enzymology of elongation	392
Initiation of replication	400
Primers and priming	404
Termination of replication	404
The regulation of replication	406
<b>27. RNA Processing</b>	<b>411</b>
Maturation of untranslated RNAs	411
End-modification and methylation of mRNA	412
RNA splicing	414
RNA editing	421
Post-processing regulation	421
<b>28. Signal Transduction</b>	<b>425</b>
Receptors and signaling pathways	425
Intracellular enzyme cascades	431
Second messengers	434
Signal delivery	440
<b>29. Transcription</b>	<b>443</b>
Principles of transcription	443
Transcriptional initiation in prokaryotes — basal and constitutive components	445
Transcriptional initiation in eukaryotes — basal and constitutive components	447
Transcriptional initiation — regulatory components	450
Strategies for transcriptional regulation in bacteria and eukaryotes	456
Transcriptional elongation and termination	458
<b>30. Viruses and Subviral Agents</b>	<b>467</b>
Viral infection strategy	468
Diversity of replication strategy	469
Strategies for viral gene expression	475
Subviral agents	477
<b>Index</b>	<b>489</b>

# Abbreviations

A	adenine (base), adenosine (nucleoside)	CTF	CAAT transcription factor
AER	apical ectodermal ridge	CTP	cytidine triphosphate
AMP, ADP, ATP	adenosine monophosphate, diphosphate, triphosphate	DAG	diacylglycerol
ANT-C	Antennapedia complex	Dam	DNA adenine methylase
AP site	apurinic/apyrimidinic site	dATP	deoxyadenosine triphosphate
APC	anaphase-promoting complex	Dcm	DNA cytosine methylase
ARS	autonomously replicating sequence	DEAE	diethylaminoethyl
ATPase	adenosine triphosphatase	DIF	differentiation inducing factor
b, bp	base, base pair	DMS	dimethylsulfate
BAC	bacterial artificial chromosome	DNase	deoxyribonuclease
BCR	B-cell receptor	dNTP	deoxynucleotide triphosphate
BER	base excision repair	DSBR	double strand break repair
bHLH	basic helix-loop-helix	dsDNA/	
BMP	bone morphogenetic protein	RNA	double-stranded DNA/RNA
BX-C	Bithorax complex	EGF(R)	epidermal growth factor (receptor)
bZIP	basic leucine zipper	ER	endoplasmic reticulum
C	cytosine (base), cytidine (nucleoside)	ES cell	embryonic stem cell
CAK	CDK-activating kinase	EST	expressed sequence tag
CaM	calmodulin	FGF(R)	fibroblast growth factor (receptor)
CAM	cell adhesion molecule	FISH	fluorescence <i>in situ</i> hybridisation
cAMP	cyclic AMP	G	guanine (base), guanosine (nucleoside)
CAP	catabolite activator protein	GABA	$\gamma$ -aminobutyric acid
CBP	CREB factor binding protein	GAP	GTPase-activating protein
CDK	cyclin-dependent kinase	GMP, GDP,	
cDNA	complementary DNA	GTP	guanosine monophosphate, diphosphate, triphosphate
CDR	complementarity determining region	GNRP	guanine nucleotide releasing protein
cf.	compare	GPCR	G-protein-coupled receptor
cGMP	cyclic guanosine monophosphate	GTF	general transcription factor
CKI	cyclin-dependent kinase inhibitor	GTPase	guanosine triphosphatase
CMV	cauliflower mosaic virus	Hfr	high frequency of recombination
cpDNA	chloroplast DNA	HLH	helix-loop-helix
CREB	cAMP response element binding (factor)	HMG	high mobility group
cRNA	complementary RNA	hnRNA,	
CTD	C-terminal domain	hnRNP	heterogeneous nuclear RNA, RNP



HOM-C	homeotic complex	PDE	phosphodiesterase
HPLC	high pressure/performance liquid chromatography	PDGF(R)	platelet-derived growth factor (receptor)
HSV	herpes simplex virus	PEV	position effect variegation
HTH	helix-turn-helix	PI(3)K	phosphoinositide 3-kinase
ICE	interleukin 1 $\beta$ converting enzyme	PKA, PKC, PKG	protein kinase A, C, G
IFN	interferon	PLA, PLB, PLC, PLD	phospholipase A, B, C, D
Ig	immunoglobulin	Poly(A)	polyadenylate
IL	interleukin	POU	Pit-1/Oct-1,2/Unc-86 HTH module
Ins	inositol	PrP	prion related protein
IRES	internal ribosome entry site	PtdIns	phosphatidylinositol
IS	insertion sequence	q.v.	<i>quod vide</i> (which see)
ITR	inverted terminal repeat	QTL	quantitative trait locus
Jak	Janus kinase	RACE	rapid amplification of cDNA ends
kb, kbp	kilobase, kilobase pairs	RAPD	randomly amplified polymorphic DNA
kDNA	kinetoplast DNA	RF	replicative form
LCR	locus control region	RFLP	restriction fragment length polymorphism
LINE	long interspersed nuclear element	RNase	ribonuclease
Lod	log of the odds ratio	RNP	ribonucleoprotein
LTR	long terminal repeat	rRNA	ribosomal RNA
MAPK	mitogen-activated protein kinase	RSS	recombination signal sequence
MAR	matrix associated region	RT-PCR	reverse transcriptase PCR
5-meC	5-methylcytosine	RTK	receptor tyrosine kinase
MEK	MAPK/Erk kinase	SAM	S-adenosylmethionine
MHC	major histocompatibility complex	SAPK	stress-activated protein kinase
MPF	mitosis/maturation promoting factor	SAR	scaffold attachment region
mRNA	messenger RNA	SCE	sister chromatid exchange
mtDNA	mitochondrial DNA	SCID	severe combined immune deficiency
N-CAM	neural cell adhesion molecule	SDS	sodium dodecylsulfate
NAD	nicotinamide adenine dinucleotide	SH	Src homology domain
NCR	noncoding region	SINE	short interspersed nuclear element
NER	nucleotide excision repair	snRNA	small nuclear RNA
NMP, NDP, NTP	nucleotide monophosphate, diphosphate, triphosphate	SR	sarcoplasmic reticulum
NMR	nuclear magnetic resonance	SRF	serum response factor
NOR	nucleolar organizer region	SRP	signal recognition particle
OD	optical density	SSB	ssDNA-binding protein
ORF	open reading frame	ssDNA/ RNA	single-stranded DNA/RNA
PAC	P1 artificial chromosome		
PCNA	proliferating cell nuclear antigen		
PCR	polymerase chain reaction		

SSLP	short sequence length polymorphism	TSG	tumor suppressor gene
STAT	signal transducer and activator of transcription	U	uracil (base), uridine (nucleoside)
STRP	short tandem repeat polymorphism	UTP	uridine triphosphate
STS	sequence tagged site	UTR	untranslated region
SV40	simian vacuolating virus 40	V(D)J	variable, diversity, junctional gene segments
T	thymine (base), thymidine (nucleoside)	VEGF(R)	vascular endothelial growth factor (receptor)
TAF	TBP-associated factor	VNTR	variable number of tandem repeats
TBP	TATA-binding protein	VSG	variable surface glycoprotein
TCR	T-cell receptor	VSP	very short patch (repair)
TF	transcription factor	XIC	X-inactivation centre
TGF	transforming growth factor	XP	xeroderma pigmentosum
Tn	bacterial transposon	YAC	yeast artificial chromosome
tRNA	transfer RNA	YEpl	yeast episomal plasmid
TSD	target site duplications	ZPA	zone of polarizing activity
TSE	transmissible spongiform encephalopathy		

## How to use this book

The book is divided into chapters concerning different areas of molecular biology. Key terms are printed in **bold** and defined when they are first encountered. The book is also extensively cross-referenced, with q.v. used to direct the reader to other entries of interest, which are shown in *italic* as listed in the index. The index shows page numbers for key terms, section titles and important individual genes and proteins. Page numbers are followed by (f) to indicate a relevant figure, (t) to indicate a table, or (bx) to indicate a quick summary box.

# Preface

This book began life as a set of hastily scrawled lecture notes, later to be neatly transcribed into a series of notebooks for exam revision. The leap to publication was provoked by an innocent comment from a friend, who borrowed the notes to refresh her understanding of some missed lectures, and suggested they were useful enough to be published as a revision aid.

The purpose of the book has evolved since that time, and the aim of the following chapters is to provide a concise overview of important subject areas in molecular biology, but at a level that is suitable for advanced undergraduates, postgraduates and beyond. In writing this book, I have attempted to combat the frustration I and many others have felt when reading papers, reviews and other books, in finding that essential points are often spread over many pages of text and embellished to such an extent that the salient information is difficult to extract. In accordance with these aims, I have presented 30 molecular biology topics in what I hope is a clear and logical fashion, limiting coverage of individual topics to 10–20 pages of text, and dividing each topic into manageable sections. To provide a detailed discussion of each topic in the restricted space available means it has been necessary to assume the reader has a basic understanding of genetics and molecular biology. This book is therefore not intended to be a beginners guide to molecular biology nor a substitute for lectures, reviews and the established text books. It is meant to complement them and assist the reader to extract key information. Throughout the book, there is an emphasis on definitions, with key terms printed in bold and defined when first encountered. Figures are included where necessary for clarity, but their style has been kept deliberately simple so that they can be remembered and reproduced with ease. There is extensive cross-referencing between sections and chapters, which hopefully stresses the point that while the book may be divided into discrete topics (Transcription, Development, Cell Cycle, Signal Transduction, etc.), all these processes are fundamentally interlinked at the molecular level. A list of references is provided at the end of each chapter, but limited mostly to recent reviews and a few classic papers where appropriate. I hope the reader finds *Advanced Molecular Biology* both enjoyable and useful, but any comments or suggestions for improvements in future editions would be gratefully received.

I would like to thank the many people without whose help or influence this book would not have been possible. Thanks to Alison Morris, who first suggested that those hastily scrawled lecture notes should be published. Thanks to Stuart Glover, Liz Jones, Bob Old, Steve Hunt and John O'Brien, who have, in different ways, encouraged the project from its early stages. Many thanks to Steve Hunt, Mary-Anne Starkey, Nigel Unwin and Richard Henderson at the MRC Laboratory of Molecular Biology who have supported this project towards the end. Special thanks to those of greater wisdom than myself who have taken time to read and comment on individual chapters: Derek Gatherer, Gavin Craig, Dylan Sweetman, Phil Gardner, James Palmer, Chris Hodgson, Sarah Lummis, Alison Morris, James Drummond, Roz Friday and especially to Bill Wisden whose comments and advice have been invaluable. Finally, thanks to those whose help in the production of the book has been indispensable: Annette Lenton at the MRC Laboratory of Molecular Biology, and Rachel Offord, Lisa Mansell, Andrea Boshier and Jonathan Ray at BIOS.

Richard M. Twyman

# Chapter 1

## Biological Heredity and Variation

### Fundamental concepts and definitions

- In genetics, a **character** or **characteristic** is any biological property of a living organism which can be described or measured. Within a given population of organisms, characters display two important properties: **heredity** and **variation**. These properties may be simple or complex. The nature of most characters is determined by the combined influence of genes and the environment.
- **Simple characters** display **discontinuous variation**, i.e. phenotypes can be placed into discrete categories, termed **traits**. Such characters are inherited according to simple, predictable rules because **genotype** can be inferred from **phenotype**, either directly or by analysis of crosses or pedigrees (see *Table 1.1* for definitions of commonly used terms in transmission genetics). For the simplest characters, the phenotype depends upon the genotype at a single gene locus. Such characters are not solely controlled by that locus, but different genotypes generate discrete, contrasting phenotypes in a particular genetic background and normal environment. When associated with the nuclear genome of sexually reproducing eukaryotes, such characters are described as **Mendelian** — they follow distinctive patterns of inheritance first studied systematically by Gregor Mendel. Not all simple characters are Mendelian. In eukaryotes, **non-Mendelian characters** are controlled by organelle genes and follow different (although no less simple) rules of inheritance (see *Organelle Genomes*). The characters of, for example, bacteria and viruses are also nonMendelian because these organisms are not diploid and do not reproduce sexually.
- **Complex characters** often display **continuous variation**, i.e. phenotypes vary smoothly between two extremes and are determined quantitatively. The inheritance of such characters is not predictable in Mendelian terms and is studied using statistical methods (**biometrics**). Complex characters may be controlled by many loci (**polygenic theory**), but the fact which distinguishes them from the simple characters is usually not simply the number of interacting genes, but the influence of the environment upon phenotypic variance, which blurs the distinction between different phenotypic trait categories and makes it impossible to infer genotype from phenotype.

### 1.1 Mendelian inheritance

**Principles of Mendelian inheritance.** For genetically amenable organisms (i.e. those which can be kept and bred easily in large numbers), the principles of inheritance can be studied by setting up large-scale crosses (directed matings) and **scoring** (determining the phenotype of) many progeny. Mendel derived his rules of heredity and variation from the results of crosses between pure breeding, contrasting varieties of the garden pea *Pisum sativum* and crosses involving hybrid plants. Although he worked exclusively with one plant species, his conclusions are applicable to all sexually reproducing eukaryotes, including those (e.g. humans) which cannot be studied in the same manner. For these unamenable organisms, heredity and variation are studied by the analysis of pedigrees (*Box 1.1*). Mendel's principles of inheritance can be summarized as follows.

- (1) The heredity and variation of characters are controlled by factors, now called **genes**, which occur in pairs. Mendel called these factors *Formbildelementen* (form-building elements).
- (2) Contrasting traits are specified by different forms of each gene (different **alleles**).
- (3) When two dissimilar alleles are present in the same individual (i.e. in a **heterozygote**), one trait displays **dominance** over the other: the phenotype associated with one allele (the **dominant allele**) is expressed at the expense of that of the other (the **recessive allele**).

**Table 1.1:** Definitions of some common terms used in transmission genetics

Term	Definition
Allele	Broadly, a variant form of a gene specifying a particular trait. At the molecular level, a sequence variant of a gene (q.v. <i>wild-type allele</i> , <i>mutant allele</i> , <i>polymorphism</i> )
Character	A biological property of an organism which can be detected or measured
Character mode	A general type of character, e.g. eye color
Character trait, trait, variant	A specific type of character, e.g. <i>blue</i> eye color
Gene	Broadly, a hereditary factor controlling or contributing to the control of a particular character. At the molecular level, a segment of DNA (or RNA in some viruses) which is expressed, i.e. used to synthesize one or more products with particular functions in the cell (q.v. <i>gene</i> , <i>cistron</i> , <i>gene expression</i> )
(Gene) locus	The position of a gene (or other marker or landmark) on a chromosome or physical or genetic map. A useful term because it allows discussion of genes irrespective of genotype or zygosity
Genetic	Pertaining to genes. Of characters, heredity and variation arising from the nucleotide sequence of the gene (c.f. <i>epigenetic</i> , <i>environmental</i> )
Genotype	The genetic nature of an individual, often used to refer to the particular combination of alleles at a given locus
Hemizygous	Containing one allele in a diploid cell, often used to refer to <i>sex-linked</i> genes (q.v.)
Hereditary	Passed from parent to offspring. Has a wider scope than the term <i>genetic</i> : includes genetic inheritance (inheritance of nucleotide sequence) as well as epigenetic inheritance (the inheritance of information in DNA structure) and the inheritance of cytoplasmic or membrane components of the cell at division
Heterozygous	Containing different alleles at a particular locus
Homozygous	Containing identical alleles at a particular locus
Phenotype	The outward nature of an individual, often used to refer to the nature of particular characters
Pleiotropic	Affecting more than one character simultaneously
Variation	The diversity of a particular character in a given population. Variation can be continuous or discontinuous
Zygosity	The nature of alleles at a locus — homozygous, heterozygous or hemizygous

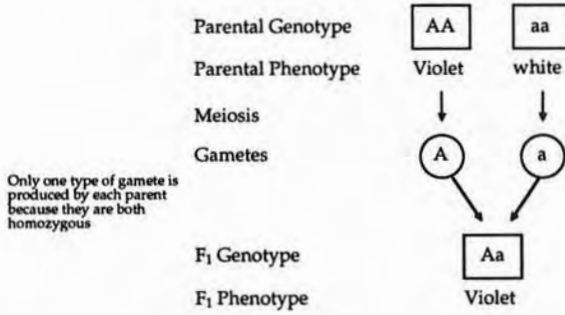
For a more precise structural and functional definition of genes and alleles, see The Gene, and Mutation and Selection.

- (4) Genes do not blend, but remain discrete (**particulate**) as they are transmitted.
- (5) During meiosis, pairs of alleles segregate equally so that equivalent numbers of gametes carrying each allele are formed.
- (6) The segregation of each pair of alleles is independent from that of any other pair.

**1.2 Segregation at one locus**

**Crosses at one locus.** Five of Mendel’s principles can be inferred from the **one-point cross (one-factor cross)**, where a single gene locus is isolated for study. A cross between contrasting pure lines produces hybrid progeny and establishes the **principle of dominance** (Figure 1.1). A **pure line** breeds true for a particular trait when self-crossed or inbred, and from this it can be established that the pure line contains only *one type of allele*, i.e. all individuals are homozygous at the locus of interest. A cross between contrasting pure lines thus produces a generation of uniform hybrids, where each individual is heterozygous, carrying one allele from each pure line. This is the **first filial**



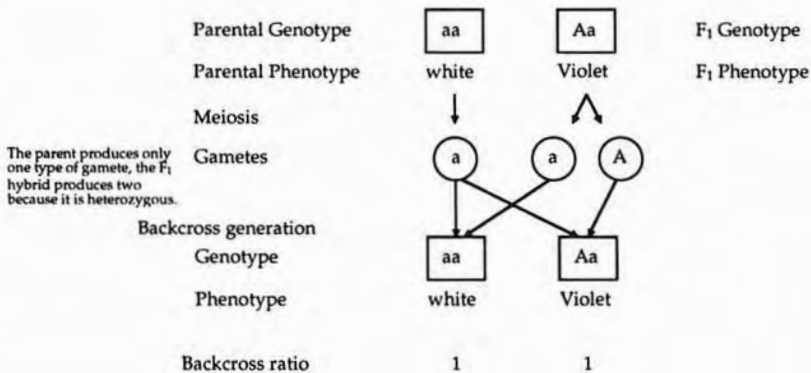


**Figure 1.1:** A cross between pure lines. This generates a hybrid F<sub>1</sub> generation and establishes the principle of dominance. Here the A allele, which in homozygous form specifies violet-colored flowers, is dominant to the a allele, which in homozygous form specifies white-colored flowers. The flower color locus is found on chromosome 1 of the pea plant and is thought to encode an enzyme involved in pigment production; the a allele is thought to be null.

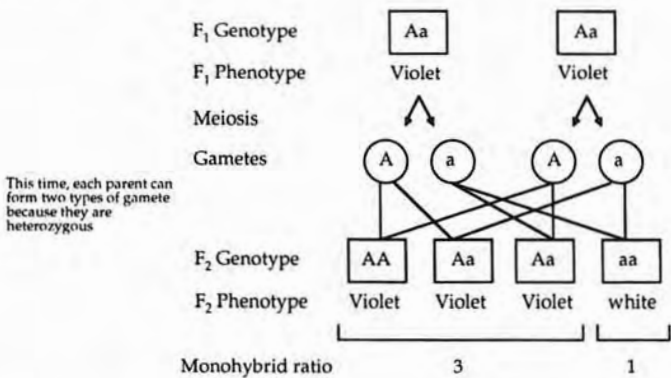
generation (F<sub>1</sub> generation). In each of his crosses, Mendel showed that the phenotype of the F<sub>1</sub> hybrids was identical to one of the parents, i.e. one of the traits was *dominant* to the other.

A **backcross** (a cross involving a filial generation and one of its parents), can confirm that the F<sub>1</sub> generation is heterozygous. If the F<sub>1</sub> generation is crossed to the homozygous parent carrying the recessive allele, the 1:1 ratio of phenotypes in the first backcross generation confirms the F<sub>1</sub> genotype (Figure 1.2). This type of analysis demonstrates the power of genetic crosses involving a **test stock** (which carries recessive alleles at all loci under study) to determine unknown genotypes, and a similar principle can be used in *genetic mapping* (q.v.). The reappearance of the recessive phenotype (i.e. white flowers) in the F<sub>2</sub> generation confirms that pairs of alleles remain particulate during transmission and are neither displaced nor blended in the hybrid to generate the phenotype.

An F<sub>1</sub> **self-cross** (self-fertilization) or, where this is not possible, an **intercross** between F<sub>1</sub> individuals can be termed a **monohybrid cross** because the participants are heterozygous at one particular locus. Such a cross demonstrates the **principle of equal segregation**, which has become known as **Mendel's First Law**. The ratio of phenotypes in the subsequent **second filial generation** (F<sub>2</sub> generation) is 3:1 (Figure 1.3). This is known as the **monohybrid ratio**, and would be expected to



**Figure 1.2:** A backcross between the F<sub>1</sub> hybrid and its recessive parent. Because the recessive phenotype reappears in the progeny, this cross proves that the F<sub>1</sub> generation is heterozygous, i.e. that the recessive allele is still present as a discrete unit.



**Figure 1.3:** A monohybrid cross. The 3:1 monohybrid ratio demonstrates that alleles segregate equally at meiosis. A **mating diagram** is used in this figure to show all possible combinations of gametes at fertilization.

arise only if there was equal segregation of alleles at meiosis (generating equal numbers of gametes carrying each of the two possible alleles).

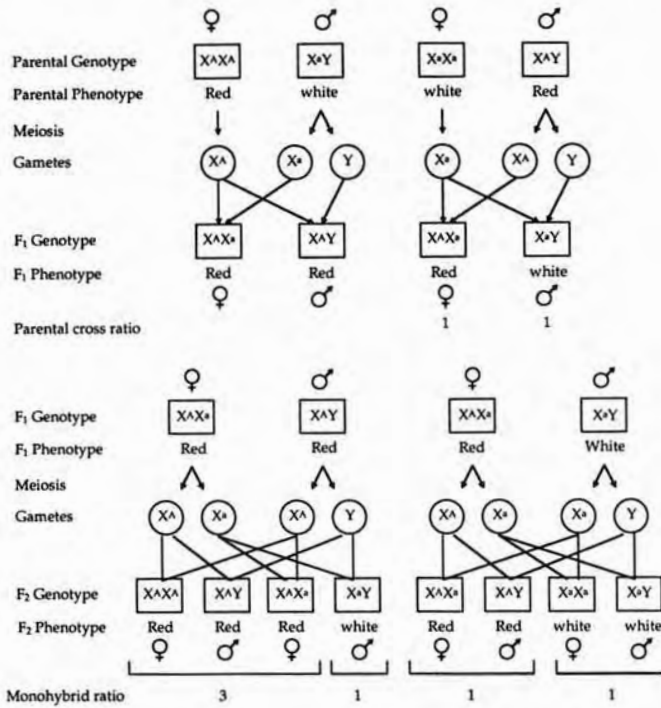
**Deviation from the monohybrid ratio.** Many characters follow broadly Mendelian inheritance patterns but show specific deviations from the monohybrid segregation ratio discussed above. The analysis of such characters has demonstrated that all but one of Mendel’s six principles may be broken, a fact that suggests that Mendel enjoyed a degree of luck in his choice of seven characters which did not suffer from any of the complications discussed below. The rule that is never broken is the principle of particulate inheritance — genes do not blend, but remain as discrete units when they are transmitted.

**Parental nonequivalence.** One of the fundamental conclusions from Mendel’s experiments was that for every locus, one allele is derived from each parent, i.e. their contribution to the zygote is equal. Hence the results of **reciprocal crosses** (pairs of crosses where males of genotype A are crossed to females of genotype B and *vice versa*) are equivalent, and this is the basis of **autosomal inheritance** patterns in pedigrees (Box 1.1). Parental equivalence reflects the fact that diploid eukaryotic cells carry two sets of chromosomes, one derived from each parent. In most cases, both parents contribute the same number of chromosomes and each is equally active. There are two important exceptions: **sex-linked inheritance** (due to structural hemizygosity), and **monoallelic expression** (due to functional hemizygosity).

**Sex-linked inheritance.** The **sex chromosomes** (q.v.) control **sex determination** (q.v.) in animals and are asymmetrically distributed between the sexes (c.f. *autosomes*). There is a **homogametic sex**, which possesses a pair of identical sex-chromosomes and thus produces one type of gamete, and a **heterogametic sex**, which possesses either a pair of nonidentical sex chromosomes or a single, unpaired sex-chromosome and thus produces two types of gamete.

In mammals, females are homogametic: they carry two copies of the X-chromosome whereas males carry one X-chromosome and one Y-chromosome, and are therefore heterogametic. There are two short regions of homology between the X- and Y-chromosomes, the **major and minor pseudoautosomal regions**, which facilitate pairing during meiosis. The major pseudoautosomal region is the site of an obligatory cross-over, and genes located there are inherited in a normal autosomal fashion (**pseudoautosomal inheritance**). Other genes are described as **sex-linked** because their expression depends upon how the sex chromosomes are distributed.

In crosses, **X-linked genes** can be identified because the results of reciprocal monohybrid crosses are not the same (Figure 1.4). If the dominant allele is carried by the female, normal Mendelian



**Figure 1.4:** X-linked inheritance. Because the male is hemizygous, the results of reciprocal crosses are not equivalent. The segregation ratios are linked to the sex-ratios, resulting in sex-specific phenotypes, and the male always transmits his X-linked allele to his daughters.

ratios are observed, but if the dominant allele is carried by the male, specific deviations in both the F<sub>1</sub> and F<sub>2</sub> generations occur because the male is hemizygous. In either case, X-linked genes show phenotypic sex-specificity, whereas for autosomally transmitted traits, the segregation ratios are sex-independent. Furthermore, because males inherit their X-chromosome only from their mothers and transmit it only to their daughters, the sex-phenotype relationship alternates in each generation, a phenomenon termed **criss-cross inheritance**. This is the major characteristic used to distinguish X-linked inheritance patterns in human pedigrees (Box 1.1).

In crosses and pedigrees, **Y-linked genes** can be identified because the characters they control are expressed only in males and passed solely through the male line (**holandric**). Few Y-linked traits have been identified in humans.

**Monoallelic expression.** Some autosomal genes are inherited from both parents, but only one allele is active. This is termed **monoallelic expression**, and the locus is functionally, but not structurally, hemizygous. There are two types of monoallelic expression: **parental imprinting**, where the gene inherited from one parent is specifically repressed, and **random inactivation**, where the parental allele to be repressed is chosen randomly. There are two forms of random inactivation in mammals — **X-chromosome inactivation** and **allelic exclusion** of immunoglobulin gene expression. Monoallelic expression is not discussed further in this chapter — see DNA Methylation and Epigenetic Regulation for further discussion of parental imprinting and X-chromosome inactivation, and Recombination for discussion of allelic exclusion.

**Maternal effect and maternal inheritance.** Reciprocal crosses are nonequivalent under several other circumstances. One example is the **maternal effect**, where the phenotype of an individual depends entirely on the genotype of the mother, and the paternal genotype is irrelevant. The maternal effect is observed for genes which function early in development, and reflects the fact that the products of



these genes are placed into the egg by the mother, having been synthesized in her cells, using her genome. Genes which display a maternal effect are actually inherited in a normal Mendelian fashion, but the phenotype is not observed until the following generation (see Figure 6.1) and thus depends on the (equivalent) contributions of the embryo's maternal grandparents. Reciprocal crosses carried out in the grandparental generation would thus be equivalent with respect to the embryonic phenotype. In this way, the maternal effect differs from *maternal inheritance* (q.v.), a form of non-Mendelian inheritance where genes are transmitted solely through the female line because they are located in organelle genomes in the cytoplasm, rather than in the nuclear genome. Maternally inherited genes are not specifically linked to development (i.e. they are expressed throughout the life of the individual) and there is no male contribution in any generation. Thus reciprocal crosses in all generations would be nonequivalent. For further discussion of maternal inheritance and other forms of non-Mendelian inheritance see *Organelle Genomes*.

**Allelic variation and interaction.** The characters described by Mendel occurred in two forms, i.e. they were **diallelic**. For many characters, however, a greater degree of **allelic variation** is apparent. The human ABO blood group locus, for instance, has three physiologically distinct alleles, and in the extreme example of the self-incompatibility loci of clover and tobacco, over 200 different alleles may be detected in a given population. The observed allelic variation also depends upon the level at which the phenotype is determined. At the molecular level, there is often more diversity than is apparent at the morphological level because many of the alleles identified as sequence variants or protein polymorphisms (see *Mutation and Selection*) are neutral with respect to their effect on the morphological phenotype; these are termed **isoalleles**. No matter how many alleles can be distinguished for a particular locus in a population, only two are present in the same diploid individual at any one time. Morphologically distinct alleles can often be arranged in order of dominance, a so-called **allelic series**.

In each of Mendel's crosses, the trait associated with one allele was fully dominant over the other, so that the phenotype of the heterozygote was identical to that of the dominant homozygote. At the biochemical level, such **complete dominance** often reflects the presence of a (recessive) *null allele* (q.v.), which is totally compensated by the presence of a (dominant) normal functional allele; this often occurs where the locus encodes an enzyme, because most enzymes are active at low concentrations — the enzyme for violet petal pigmentation in the pea is one example, but in other plants this is not necessarily the case, leading to incomplete dominance. There are a number of alternative **dominance relationships** and other **allelic interactions**, each with a specific biochemical basis; these are discussed in *Table 1.2*.

The concept of dominance is often applied to alleles, but dominance is a property of characters themselves, not the alleles that control them (only in the case of *paramutation* (q.v.) does a heritable change occur in the allele itself). Dominance also depends on the level at which the phenotype is observed: sickle-cell trait is a partially dominant disease because the effect of the allele is manifest in heterozygotes for normal and sickle-cell variant  $\beta$ -globin production (albeit under extreme circumstances), but when observed at the protein level as bands migrating on an electrophoretic gel, the variant form of  $\beta$ -globin is codominant with the normal protein (i.e. both 'traits' can be observed at the same time).

**Distortion of segregation ratios.** The principle of equal segregation is one of the more robust of Mendel's rules and is inferred from the observation that contributing alleles are represented equally in the progeny of a monohybrid cross. However, there are several ways in which equal representation can be prevented, resulting in distortion of the Mendelian ratios — i.e. a bias in the recovery of a particular allele in the offspring. Such mechanisms fall into two major classes: those acting before and those acting after fertilization.

**Segregation distortion** occurs before fertilization (i.e. so that there is a disproportionate repre-

**Table 1.2:** Dominance relationships and other allelic interactions (interactions at a single locus), with biochemical basis and examples

Allelic interaction	Description
Complete dominance	The dominant allele fully masks the effect of the recessive allele. The phenotype of the heterozygote is identical to that of one of the homozygotes, and the monohybrid ratio is 3:1. This is the classical dominance effect described by Mendel, and often occurs where the recessive allele is null. Examples include violet color pigment in the pea plant, and cystic fibrosis in mammals. Loss of one allele encoding the pigmentation enzyme or transmembrane receptor is compensated by a second, wild-type allele. Alternatively, the dominant allele may be null (q.v. <i>dominant negative</i> ), e.g. in Hirschsprung's disease, which is caused by dominant negative loss of c-RET tyrosine kinase activity — the mutant form of the enzyme sequesters the wild-type enzyme into an inactive heterodimer
No dominance and partial dominance	Neither allele is fully dominant over the other. The phenotype of the heterozygote is somewhere in between those of the homozygotes, and the monohybrid ratio is 1:2:1. If the heterozygous phenotype is exactly intermediate between the two homozygotes, there is no dominance. If the phenotype is closer to one homozygote than the other, there is partial dominance. These dominance relationships occur where there is competition between the products of two alleles (e.g. in sickle-cell trait, where different forms of $\beta$ -globin react differently to low oxygen tension), or if a gene locus is <i>haploinsufficient</i> (e.g. in type I Waardenburg syndrome, which is due to 50% reduction in the synthesis of PAX3 protein)
Overdominance and underdominance	The phenotype of the heterozygote lies outside the range delineated by those of the homozygotes. If the heterozygous phenotype is greater than either homozygous phenotype, the locus shows overdominance; if less, it shows underdominance. The monohybrid ratio is 1:2:1. These relationships occur where there is synergy or antagonism between the products of particular alleles. Overdominance is often observed when considering the combined effects of multiple loci, leading to <b>hybrid vigor (heterosis)</b> , an increase in fitness due to heterozygosity at many loci or <b>inbreeding depression</b> , a decrease in fitness due to homozygosity for many deleterious alleles
Codominance	The phenotype associated with each allele is expressed independently of that of the other. Codominance occurs when there is no competition between alleles, e.g. in the ABO blood group system, where alleles A and B specify different glycoproteins presented on the surface of red blood cells. Both A and B are dominant over O as the latter is a null allele (i.e. the protein remains unglycosylated). However, if an individual carries both A and B alleles, both molecules are presented and the resulting blood group is AB. The monohybrid ratio is 1:2:1
Pseudodominance	An allele appears dominant because the locus is hemizygous. This is applicable to sex-linked loci in the heterogametic sex, e.g. in male mammals (q.v. <i>sex-linkage</i> ) and to individuals with chromosome deletions or chromosome loss (see Chromosome Mutation)
Paramutation	An allelic interaction occurring in the heterozygous state where one allele causes a transiently heritable but epigenetic change in the other, a process often involving methylation of repetitive DNA. This is the only example of an allelic interaction where the DNA itself is the target. For further discussion, see DNA Methylation
Allelic complementation	A phenomenon where two loss-of-function, recessive to wild-type alleles can generate a functional gene product in combination, because they compensate for each other's defects. The principle example of allelic

Continued

	complementation is $\alpha$ -complementation in the expression of <i>E. coli</i> $\beta$ -galactosidase (q.v. <i>recombinant selection</i> )
Trans-sensing	An interaction between alleles which is synapsis-dependent and occurs only in organisms where homologous chromosomes are associated even in mitotic cells (e.g. <i>Drosophila</i> ), or where such association occurs by chance. Examples include <i>transvection</i> (q.v.). For further discussion, see Gene Expression and Regulation

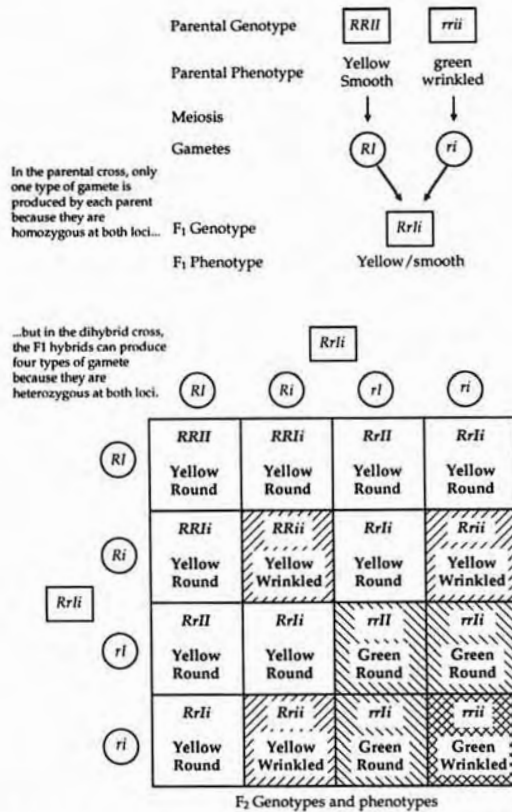
sentation of gametes carrying each allele) and is termed **meiotic drive**. There are two types of drive mechanism, which occur predominantly in the different sexes. **Genic drive** usually occurs in males and involves selective inactivation of sperm of a particular haplotype. Two loci are involved in this type of system, a *trans*-acting **driver** or **distorter** and a *cis*-acting **target**. In the *SD* (segregation distorter) system of *Drosophila*, the target allele is a repetitive DNA element whose copy number correlates to the distortion ratio. The drive locus encodes a product which is thought act at the target allele to perturb chromatin structure, leading to gametic dysfunction. Heterochromatin elements are thought to be involved in many of the characterized genic drive systems, so modulation of DNA structure could be used as a universal mechanism of gamete inactivation. Genic drive is uncommon in females because, as they produce far fewer gametes than males, they would be placed at a selective disadvantage by large-scale gamete inactivation. Drive in females often occurs earlier than in males by a process termed **chromosomal drive**, where the property of a given bivalent at meiosis causes it to adopt a particular orientation in the spindle and thus undergo preferential segregation into either the egg or the polar body (the latter being discarded). Chromosomal drive would not work in males because of the equality of the meiotic products.

Where distortion occurs after fertilization, it reflects differing viabilities of zygotes with alternative genotypes. In its most extreme form, distortion results in the total absence of a particular genotype, indicating the presence of **lethal alleles** (which cause death when they are expressed) whose effects are manifest early in development. The presence of a dominant lethal results in the recovery of only one genotype, the homozygous recessive. The presence of a recessive lethal generates a characteristic 2:1 segregation ratio of dominant homozygotes to heterozygotes, because the homozygous recessive class is not represented. Lethal alleles usually represent the *loss of function* (q.v.) of an essential gene product; thus leaky lethal alleles may be **sublethal** (q.v. *penetrance, expressivity, leaky mutation*).

**Penetrance and expressivity.** Penetrance and expressivity are terms often used to describe the non-specific effects of genetic background, environment and noise on the expression of simple characters (*Box 1.2*). **Penetrance** describes the proportion of individuals of a particular genotype who display the corresponding phenotype. Complete penetrance occurs when there is a 100% correspondence between genotype and phenotype. **Expressivity** reflects the degree to which a particular genotype is expressed, i.e. where the phenotype can be measured in terms of severity, the strongest effects have the greatest expressivity. Incomplete penetrance and variable expressivity often complicate the interpretation of human pedigrees because of the small number of individuals involved. Where incomplete penetrance and variable expressivity affect a character to the degree where it is no longer possible to reliably determine genotype from phenotype, the character can effectively be described as complex (see below).

1.3 Segregation at two loci

**Crosses at two loci.** Mendel's final postulate, which is expressed as the **principle of independent assortment**, can be inferred from a **two-factor cross** (a cross where two loci are studied simultaneously). Two lines which breed true for two contrasting traits are crossed to produce an *F*<sub>1</sub> generation of uniform dihybrids (heterozygous at two loci). If these are self-crossed or intercrossed (a **dihybrid cross**) the *F*<sub>2</sub> generation shows a 9:3:3:1 ratio of the four possible phenotypes. This is termed



**Figure 1.5:** The principle of independent assortment. Two parental lines are chosen which breed true for two contrasting character traits (in this case yellow vs green and smooth vs wrinkled seeds, which are thought to be encoded by the  $I$  locus on chromosome 1 and the  $R$  locus on chromosome 7 of the pea, respectively). The parental cross produces a generation of uniform dihybrids displaying the dominant trait at each locus (in this case smooth yellow seeds). Each F<sub>1</sub> individual produces four types of gamete, which can combine to form nine different genotypes and four different phenotypes in the F<sub>2</sub> generation. The observed 9:3:3:1 ratio would only arise if each pair of alleles segregated independently from each other. The grid shown above to plot the mating information in the dihybrid cross is known as a **Punnett square**, and is more useful than a mating diagram for the simultaneous analysis of two loci.

the **dihybrid ratio**, and is derived as shown in Figure 1.5. The dihybrid ratio could only arise if the segregation of one pair of alleles had no effect on that of the other, i.e. the two allele pairs show independent assortment at meiosis. The principle of independent assortment has become known as **Mendel's Second Law**.

However, where two gene loci are found close together on the same chromosome, coupled alleles tend to segregate together. The 9:3:3:1 dihybrid ratio is thus replaced by a ratio in which two phenotypic classes are common and two are rare. The common class represents the *parental combination* of alleles, i.e. the alleles coupled together on each chromosome, whilst the rare class represents a *recombinant combination* of alleles generated by crossing-over between paired chromosomes at meiosis. This phenomenon is termed **linkage**, and can be exploited to map eukaryotic genomes: see Recombination, and Genomes and Mapping.

**Nonallelic interactions.** Linkage disrupts the dihybrid ratio because independent segregation is prevented. The dihybrid ratio may also be modified in the absence of linkage if the two loci are functionally interdependent, i.e. if both loci contribute to the control of the same character. Various types of **nonallelic interactions** occur which generate specific deviations from the normal 9:3:3:1 dihybrid ratio (Table 1.3, Figure 1.6).



**Table 1.3:** Some different types of nonallelic interactions which generate modified dihybrid ratios. Such interactions are the basis of 'genetic background'

Interaction	Description
None	Two segregating loci which are unlinked and control distinct characters. Dihybrid ratio 9:3:3:1
Additive effects	Two segregating loci contribute to the same character in an additive fashion. Alleles are generally regarded as positive or neutral in effect and the resulting phenotype is determined by the net contribution of all additive loci. The dihybrid ratio would be 9:3:3:1 if two loci were considered in isolation (q.v. <i>quantitative inheritance</i> )
Complementary action of genes	A situation where specific alleles at two gene loci are required to generate a particular phenotype. Thus, if either locus lacks a suitable allele, the phenotype would not be generated. The dihybrid ratio would be 9:7 if dominant alleles were required at both loci, 1:15 if recessive alleles were required, and 4:12 if dominant alleles were required at one locus and recessive alleles at the other
Epistasis	The situation where an allele at one locus prevents or masks the effect of gene expression at a second locus, which is said to be <b>hypostatic</b> . Epistasis may be dominant or recessive, i.e. one or two specific alleles may be required at the epistatic locus for the effect to occur. The dihybrid ratios would be 12:3:1 and 9:3:4, respectively. Epistasis is often observed in genetic pathways or hierarchies where there is an order of gene action, early-acting genes being epistatic to later-acting ones. <b>Phenotypic modification</b> is similar to epistasis except that the <b>modifying allele</b> alters rather than masks the phenotype of the downstream genes
Redundancy/dosage	Alleles of identical genes which have arisen by duplication may interact with each other in a manner which superficially appears allelic. However, they may be separated by recombination, showing that they occupy discrete loci, and such interaction is termed <b>pseudoallelic</b> . The dihybrid ratio would be 15:1. Multiple redundancy is often seen in <i>transgenic</i> organisms (q.v.) where many copies of a transgene have integrated. Transgenic animals may show <b>copy-number-dependent gene expression</b> (i.e. levels of gene product correspond to copy number), or there may be interaction between repetitive copies resulting in <i>homology dependent gene silencing</i> (q.v.)
Suppression	An allele at one locus (the <b>suppressor allele</b> ) prevents the expression of a specific (and usually deleterious) allele at a second locus by compensating in some way for its effect. The term is usually used to describe an interaction where a mutation compensates for a second mutation found at a different locus and restores the wild-type phenotype (q.v. <i>second-site mutations</i> ). Suppression may be dominant or recessive, depending upon whether one or two alleles are required at the suppressor locus. The dihybrid ratios would be 12:3:1 and 9:3:4, respectively. Epistasis differs from suppression in that the former is gene-specific but not allele-specific (i.e. epistasis prevents the expression of <i>all</i> alleles at a particular locus), whereas the latter is gene-specific and allele-specific (i.e. suppression compensates for the effect of a particular allele at the second locus)
Synergism, enhancement	Similar to the suppression mechanism except that the <b>enhancer allele</b> specifically increases the effect of a second mutation, instead of suppressing it. As with suppressor alleles, enhancer alleles may be dominant or recessive. Synergic enhancement differs from additive effects because the enhancer locus alone does not contribute to the phenotype associated with the target (enhanced) locus, i.e. if the target allele is not present, the enhancer locus has no effect on the phenotype. Enhancer alleles should not be confused with <i>enhancer</i> regulatory elements (q.v.)

Normal dihybrid ratio	9 A_B_	3 A_bb	3 aaB_	1 aabb
Dominant epistasis (of allele A over locus B) i.e. A quashes all B alleles making B equivalent to b	12 A_bb		3 aaB_	1 aabb
Recessive epistasis (of genotype aa over locus B) i.e. aa quashes all B alleles making B equivalent to b	9 A_B_	3 A_bb	4 aabb	
Dominant suppression (of allele A over allele b) i.e. A suppresses all b alleles making bb equivalent to B_	12 A_B_		3 aaB_	1 aabb
Recessive suppression (of genotype aa over allele b) i.e. aa suppresses all b alleles making bb equivalent to B_	9 A_B_	3 A_bb	4 aaB_	
Redundant genes i.e. A=B	15 A__			1 aaaa
Dominant complementary genes, i.e. both A and B necessary for phenotype therefore A-bb and aaB- equivalent to aabb	9 A_B_		7 aabb	
Recessive complementary genes i.e. both aa and bb necessary for phenotype therefore A-bb and A-B_ are equivalent	15 A_B_			1 aabb

**Figure 1.6:** The effects of nonallelic interaction on Mendelian dihybrid ratios. Two hypothetical loci, A and B, each comprise a pair of alleles, one of which, denoted by the capital letter, is fully dominant over the other. For normal independent assortment, four phenotypes would be generated corresponding to the generic genotypes A\_B\_, A\_bb, aaB\_ and aabb in the ratio (9:3:3:1). The effects of different types of nonallelic interaction are shown by the modulation of the ratio by changing the phenotypes associated with particular alleles.

## 1.4 Quantitative inheritance

**Types of complex character.** Many characters show continuous variation, i.e. phenotypes are measured in quantitative terms and cannot be placed into discrete traits. The phenotypes often show a normal distribution about a mean value. Such **quantitative characters** are inherited in a complex manner: genotype cannot be deduced from phenotype and no simple rules of heredity can be used to predict the outcome of a cross. The inheritance of such characters is studied using statistical methods (**biometrics**). However, some characters which appear superficially Mendelian are also inherited quantitatively. The discipline of **quantitative inheritance** thus embraces three types of character (Figure 1.7). **Continuous characters** demonstrate true continuous phenotypic variation (i.e. no boundaries between different phenotypes). **Meristic characters** vary in a similar manner to continuous characters, but the intrinsic nature of the character itself demands that phenotypes are placed into discrete categories, usually because the value of the phenotype is determined by counting (hence such characters may also be termed **countable characters**). Finally, **threshold (dichotomous) characters** have two phenotypes — a certain condition can be either present or absent. Such characters thus appear very much like Mendelian diallelic traits, but in this case, an underlying quantitative mechanism controls liability to display the phenotype, which is manifest once some triggering level has been exceeded.

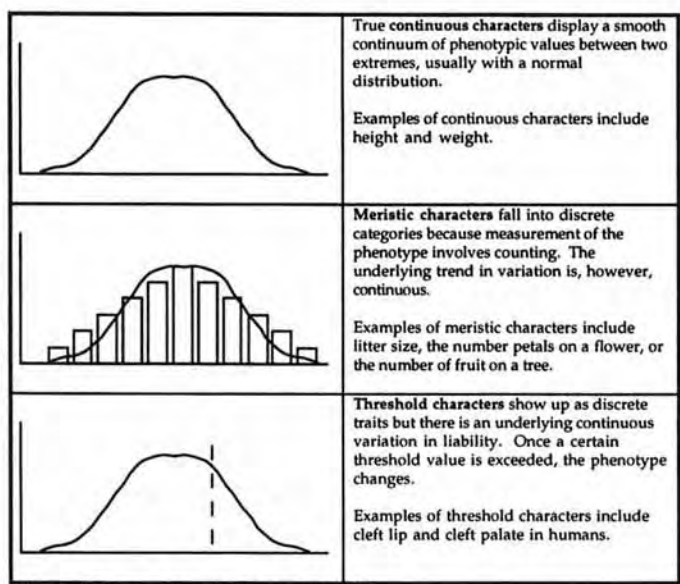


Figure 1.7: Classes of quantitative character

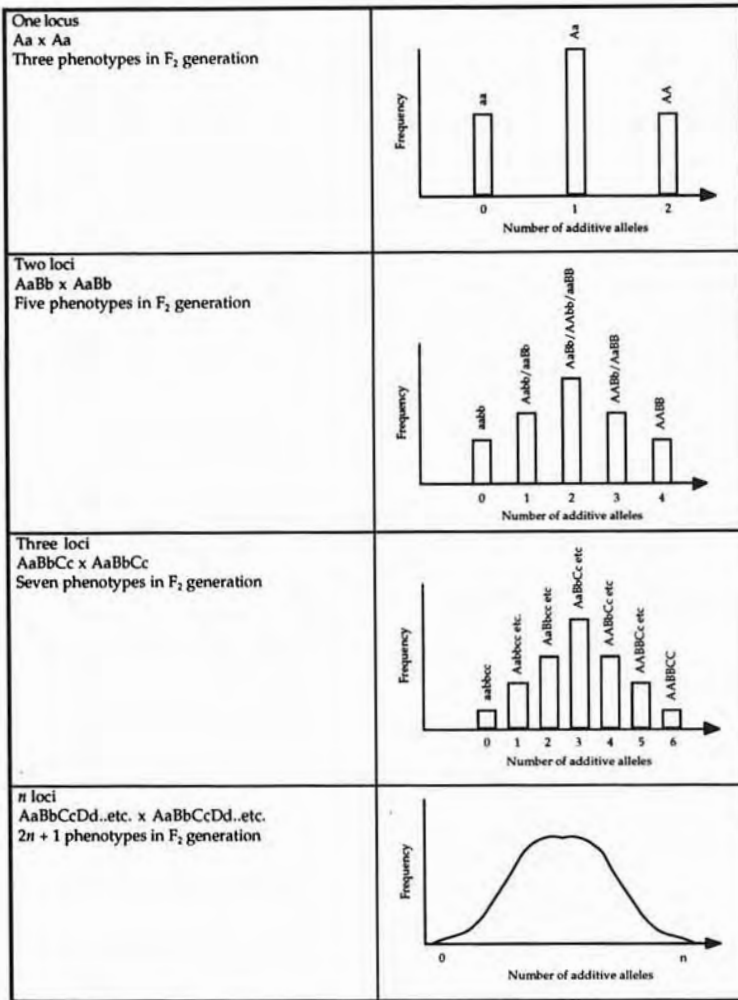
Figure 1.7: Classes of quantitative character.

**The polygenic theory of quantitative inheritance.** Since quantitative characters are unsuitable for Mendelian analysis, they were originally considered to be controlled by factors which were fundamentally different from the genes that control Mendelian characters.

The **polygenic theory** of Fisher established a common basis for the transmission of Mendelian and quantitative characters. The theory proposed that quantitative characters may be controlled by a large number of segregating loci (**polygenes**), each of which contributes to the character in a small but additive manner. In the simplest form of the model, each contributing locus is diallelic, and one allele contributes to the phenotype whilst the other has no effect. The value of each allele is the same, and the phenotype thus depends upon the number of contributing alleles, i.e. the phenotype is specified by a totting up procedure. As the number of segregating loci increases, so does the number of phenotypes, and the distribution approaches that of a normal curve (Figure 1.8). This is because, in the middle range of phenotypic values, many different genotypes will generate the same phenotype (**locus heterogeneity**).

Whilst this provides a very simple model for quantitative inheritance, real populations would be expected to encounter some of the complications discussed in the previous sections. Thus observed patterns of inheritance might be more complex for some of the following reasons:

- (1) the relative contribution of each gene to the phenotype would be different — some loci would have strong effects and others only weak effects (these are termed **major genes** and **minor genes**, respectively);
- (2) the nature of segregation would be complicated by linkage, which would increase the frequency of combinations of alleles found on the same chromosome;
- (3) there would be more than two alleles at some of the contributing loci and these would have differing strengths;
- (4) there would be dominance relationships between alleles;
- (5) there would be nonallelic interactions other than additive effects between contributing loci;
- (6) in natural populations, different alleles would be present at different frequencies, so some genotypes would be relatively common and others rare.

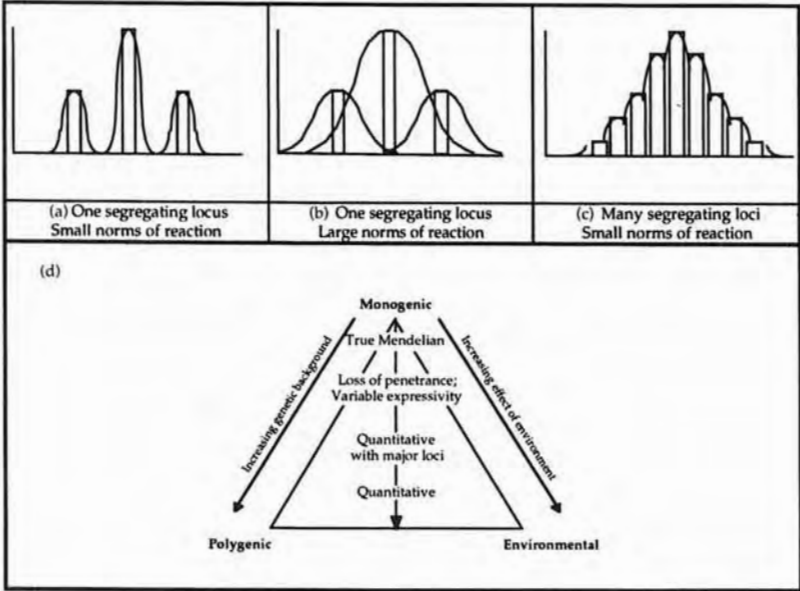


**Figure 1.8:** The polygenic model of quantitative inheritance. As the number of additive loci affecting a given character increases, the number of phenotypes in the  $F_2$  generation also increases. Each locus can contribute a maximum of two additive alleles to the phenotype, so that for  $n$  loci there are  $2n + 1$  phenotypes. In the middle range, many different genotypes generate the same phenotype. The distribution thus approaches that of a normal curve and phenotypic variation appears continuous.

**Environmental influence and the norm of reaction.** Despite the attractive simplicity and widespread use of the polygenic model, there is no conclusive proof that quantitative characters are controlled by polygenes. Where the genes which control quantitative characters have been sought systematically, a small number of major genes and a variable number of minor genes have often been identified, suggesting that a few loci may be sufficient. For human diseases inherited in a complex manner, the major genes are termed **major susceptibility loci** (q.v. *quantitative trait loci*, *QTL mapping*).

In fact, a large number of gene loci is not necessary for continuous variation because all biological characters are influenced to some degree by the **environment** as well as by genotype. The environment in which an organism lives will interact with the genotype to produce the phenotype. Thus, if a single genotype is exposed to a range of environments, a range of phenotypes is observed which is described as the **norm of reaction**. This explains much of the phenotypic variance in





**Figure 1.9:** The effect of environment on the phenotypic variance of a character. A single segregating heterozygous locus generates three genotypes. (a) The phenotypes will fall into discrete traits if the norm of reaction is small, but (b) variation will appear continuous if the norm of reaction is large. (c) If many segregating loci are involved, the polygenic model predicts that the distinctions between genotypes will be small; thus even small norms of reaction smooth the distinctions between individual phenotypes, resulting in continuous variation. (d) Few characters are truly Mendelian, truly polygenic or completely determined by the environment. Most lie between these extremes, somewhere within the triangle. Increasing both the number of genes and the effect of the environment makes a character less Mendelian and more quantitative.

**isogenic populations** (populations where each individual has the same genotype) as all individuals are not exposed to identical environments. [Other, nonenvironmental ways in which isogenic individuals may differ include the presence of somatic mutations (and in vertebrates, the manner in which somatic recombination has rearranged the germline immunoglobulin and T-cell receptor genes), and in female mammals, the distribution of active and inactive X-chromosomes.] The degree to which a phenotype can be shaped by the environment is described as its **phenotypic plasticity**.

Simple characters thus exist because the differences between the mean phenotypic values of each genotype are larger than the norm of reaction for each genotype (put another way, the variance between genotypes is greater than the variance within genotypes). For continuous characters, the opposite is true: the differences between the mean phenotypic value of each genotype is smaller than the norm of reaction for each genotype, so that the latter overlap. This overlap means that the genotype cannot be predicted from phenotype and Mendelian analysis is impossible — the character is quantitative. This effect can occur even if the character is controlled by one segregating locus, but for a polygenic character, as the number of loci increases, the number of genotypes becomes larger and the distinction between them becomes smaller, thus less environmental influence is required to blur the boundaries completely (*Figure 1.9*). By controlling the environment so that the norms of reaction are small, continuous characters controlled by few loci can be resolved into discrete traits and their transmission can be dissected in terms Mendelian inheritance. Characters which do not respond to such experiments are likely to be truly polygenic.

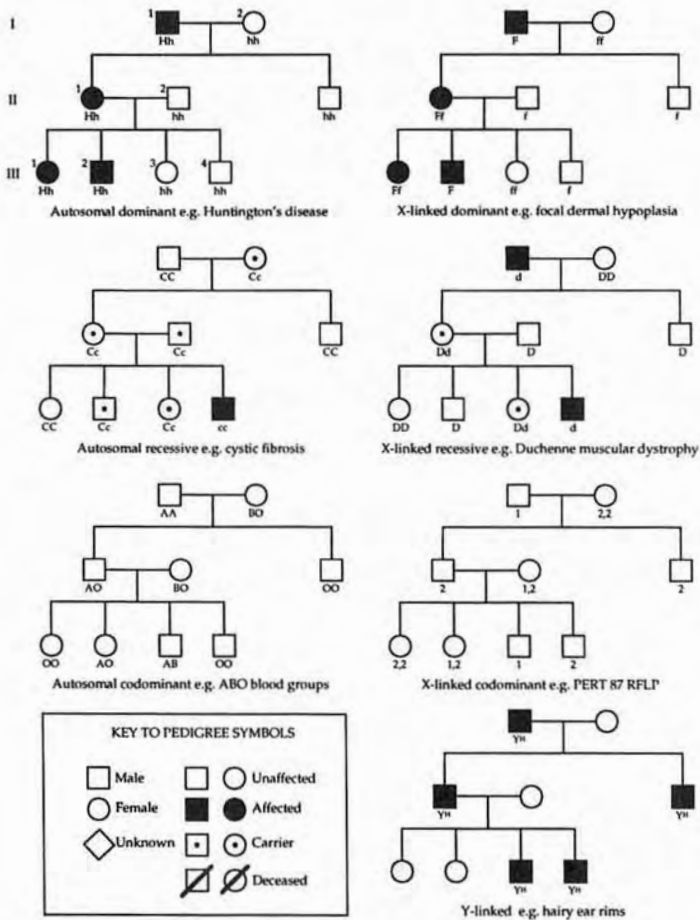
There are relatively few characters which are truly Mendelian, truly polygenic or totally determined by the environment. Most lie somewhere between those three extremes. Mendelian charac-

ters can be regarded as the peak of a triangle, suffering the effects of neither genetic background nor the environment. As the contribution of other genes and the environment increases, the character will begin to show incomplete penetrance and/or variable expressivity and will eventually become quantitative (Figure 1.9).

**Box 1.1: Pedigree patterns for Mendelian traits**

**Mendelian pedigree patterns for human traits.** In organisms, such as humans, where large-scale matings are not possible, modes of inheritance cannot be established by offspring ratios. Instead, pedigrees are used, and the mode of inheritance must be assessed by statistical analysis (because of the small size of most human families, it is sometimes difficult to establish an unambiguous inheritance pattern, especially when comparing autosomal and X-linked dominant traits). There are seven basic

pedigree patterns for human traits: autosomal dominant, recessive and codominant, X-linked dominant, recessive and codominant, and Y-linked. The pedigree patterns are shown below and their major characteristic features are listed in the table. Loci on the region of homology shared by the X- and Y-chromosomes are inherited in a normal autosomal fashion because an allele is inherited from each parent — this is known as **pseudoautosomal inheritance**.



Continued

Mode of inheritance	Essential features
Autosomal dominant	Transmission and manifestation by either sex Affected individuals usually have at least one affected parent
Autosomal recessive	Transmission and manifestation by either sex Affected individuals usually have unaffected parents who are <b>carriers</b> (asymptomatic individuals carrying recessive alleles) Increased incidence if there is consanguinity between parents
Autosomal codominant	Transmission and manifestation by either sex Individuals inherit one allele from each parent
X-linked dominant	Transmission and manifestation by either sex, but more common in females Affected males transmit trait to all daughters Affected females usually transmit trait to 50% of sons and daughters
X-linked recessive	Transmission and manifestation in either sex, but much more common in males due to hemizyosity Affected males usually have unaffected parents, but the mother is a carrier Affected females usually have an affected father and a carrier mother, but occasionally the mother is also affected (i.e. homozygous)
X-linked codominant	Transmission and manifestation by either sex Paternal X-linked allele is always passed to daughters, never sons Maternal alleles may be inherited by daughters or sons
Y-linked	Transmission and manifestation in males only (holandric) Affected males have affected fathers and affected sons

Note that in pedigrees, the generations are labeled with roman numerals and the individuals within each generation are numbered from the left. This is shown only for the first pedigree.

**Complications to basic inheritance patterns.** Even if the mode of inheritance for a particular trait is unambiguous, the interpretation of pedigree patterns is complicated by small sample sizes. Further complications arise through factors which reduce the penetrance or vary the expressivity of a given trait. These can often be tolerated in large-scale crosses, but present a serious limitation to the usefulness of many pedigrees because a single case may cause the entire pedigree to be inter-

preted falsely. Some of these complications reflect purely genetic mechanisms (e.g. random clonal X-chromosome inactivation, imprinting, the appearance of new mutations, X-linked male lethality and germline mosaicism), while others may be due to both genetic and environmental factors (i.e. genetic background, environmental influence and developmental noise). One of the most puzzling pedigree complications is **anticipation** — the tendency for some traits to increase in severity and/or show reduced age of onset in successive generations. Recently, a molecular basis for anticipation in several human diseases has been described, reflecting the behavior of pathogenic intergenic *triplet repeats* (Box 15.2).

**Box 1.2:** Causal components of genetic and environmental variance

**The breakdown of phenotypic variance.** Phenotypic variance ( $V_P$ ) is the total observed variance for a given biological character in a given population. It can be broken down into its two major causal components, **genetic variance** ( $V_G$ ), which is the variance contributed by different genotypes (i.e. variation between genotypes), and **environmental variance** ( $V_E$ ), which reflects all external effects and generates the norm of reaction for each genotype (i.e. variation within genotypes). A further component, **gene-environment interactive variance** ( $V_{GE}$ ), reflects the proportion of variance which remains when both genetic and environmental variances have been calculated and subtracted from the total phenotypic variance.  $V_{GE}$  can be thought of as resulting from interaction between the two other components, but in practice it is difficult to measure directly and is often ignored. This relationship can be expressed by the following equation:

$$V_P = V_G + V_E (+ V_{GE})$$

Both genetic and environmental variance can be broken down into several subcomponents.

**Genetic variance.** Genetic variance ( $V_G$ ) is the part of phenotypic variance which arises from differences in genotype between individuals. It can be divided into three further components.

- (1)  $V_A$  is **additive variance** (also known as **genic variance** or the **breeding value**). This reflects the effects of substituting different alleles at loci contributing additively towards a given character. Additive variance is the principle component of phenotypic variance exploited for selective breeding.
- (2)  $V_D$  is **dominance variance**. This reflects the effects caused by allelic interactions at each locus.
- (3)  $V_I$  is **interactive variance**. This reflects the effects caused by nonallelic interactions other than additive effects (e.g. epistasis, suppression, enhancement).

The last two components are usually grouped together as 'nonadditive variance' because they are difficult to isolate with any accuracy. The relative amounts of additive and nonadditive variance for a given character are of particular interest to animal and plant breeders who want to choose the most successful form of artificial selection. The partition of genetic variance can thus be expressed using the formula

$$V_G = V_A + V_D + V_I$$

**Genetic background.** Genetic background is a term used to describe nonspecific genetic effects which alter the expression of a given gene. The effects of genetic background on simple characters lead to variable penetrance and expressivity, and include nonallelic interactions (Table 1.3) as well as position effects, which frequently affect the expression of integrated transgenes and genes involved in large-scale genomic rearrangements. Any variation in a quantitative character caused by genetic background would be described by the component  $V_I$ .

**Environmental variance.** Like genetic variance, environmental variance can also be divided into several subcomponents.

- (1)  $V_{E(g)}$  is **general environmental variance**, and reflects factors to which all members of a given population are exposed.
- (2)  $V_{E(s)}$  is **special environmental variance**, and reflects factors to which only specific groups of individuals are exposed, e.g. the maternal environment during pregnancy in mammals and the common family environment. It is the special environmental variance which makes familial and heritable characters difficult to discriminate (see below).

A further component, which may be considered part of the environment or as a separate source of variance in itself is **developmental noise**. This reflects purely stochastic events which, at the molecular level, may influence gene expression in different cells. It is often difficult to discriminate between developmental noise and variance caused by the environment, but if a character can be scored on each side of the body, both genetic and environmental variance are cancelled and noise is all that remains. The partition of environmental variance can thus be expressed using the formula

$$V_E = V_{E(g)} + V_{E(s)} + \text{Developmental noise}$$

The effects of the environment and noise, as well as genetic background, influence the penetrance and expressivity of simple characters.

**Phenocopies.** A phenocopy is a trait generated purely by modifying the environment. For example, the phenotype of the *sonic hedgehog* knockout mouse is loss of head and midline structures. A phenocopy can be made by starving pregnant rodents of cholesterol, which is normally conjugated to Shh protein and is required for its function. In this example, the effect of mutating the gene can be mimicked by removing a component from the environment which is essential for the function of the gene product.

**Familiality and heritability.** The term **heritability** was coined to express the genetic contribution to phenotypic variance. In experimental organisms, the heritability of a given quantitative character is simple to demonstrate. Individuals are taken from the extremes of a population so that the mean phenotypic value for the character in each subpopulation is far removed from the population mean. Each subpopulation is then interbred and the progeny are scored. If the character is heritable, the means of the progeny will be similar to those of their parents (i.e. at the extremes of the source population), whilst if the observed variance in the source population was entirely environmental in nature, the mean phenotypic value of the new populations would be the same as each other, and the same as that of the source population.

A way to determine heritability without breeding is by looking for resemblance between relatives. Relatives share more genes than random individuals in a population, and phenotypic covariance should reflect underlying genetic similarity. However, relatives tend to share a common environment as well as common genes, so it is important to determine whether the environment contributes significantly to the observed variance. A trait which is shared by relatives is described a **familial**, but not all familial traits are heritable, e.g. children tend to speak the same language as their relatives and language is therefore a trait that runs in families, but it is not heritable: a child born to English parents but raised in a French family would speak French. The way to discriminate between heritability and familiality is to observe phenotypic performance in a number of

different environments. This is easy for a repeatable character (e.g. fleece weight in sheep, milk production in cattle), but is more difficult for a character which is expressed only once (e.g. yield in a cereal crop, human intelligence), and such tests must be carried out on highly related individuals. This is often not possible in humans; thus the genetic basis of human quantitative traits has been difficult to demonstrate. However, twin adoption studies (where identical twins separated at birth and raised in different environments are studied) have been useful.

The term heritability is commonly used in two senses. **Heritability in the broad sense** (designated  $H^2$ ) is also known as the **heritability index (H-statistic)** or the **degree of genetic determination**, and is expressed as  $H^2 = V_G/V_P$ . Broad heritability thus measures the ratio of genetic variance to total phenotypic variance in a given environment. It does not measure the overall importance of genes to the development of a particular character, and assumptions that it does have led to great misuse of the term, especially in its application to human social issues. **Heritability in the narrow sense** (designated  $h^2$ ) is expressed as  $h^2 = V_A/V_P$ , and thus measures the ratio of additive genetic variance to total phenotypic variance. This estimates the degree to which observed phenotypic variance can be influenced by selective breeding. Artificial selection is carried out in defined populations in defined environments to improve commercially valuable characters, so the limitations of broad heritability values, i.e. that they cannot be extrapolated across populations, are not important for the purpose of breeding.

## References

- Connor, J.M. and Ferguson-Smith, M.A. (1994) *Essential Medical Genetics*. 4th edn. Blackwell Science, Oxford.
- Falconer, D.S. and Mackay T.F.C. (1996) *Introduction to Quantitative Genetics*. 4th edn. Longman Group, Harlow.
- Fincham, J.R.S. (1994) *Genetic Analysis*. Blackwell Science, Oxford.
- McKusick, V.A. (ed.) (1996) *Mendelian Inheritance in Man*. 12th edn. Johns Hopkins University Press, Baltimore, MD.
- Avery, L. and Wasserman, S. (1992) Ordering gene function — the interpretation of epistasis in regulatory hierarchies. *Trends Genet.* 8: 312–316.
- Frankel, W.N. (1995) Taking stock of complex trait genetics in mice. *Trends Genet.* 11: 471–477.
- Guarente, L. (1993) Synthetic enhancement in gene activation — a genetic tool come of age. *Trends Genet.* 9: 362–366.
- Hodgkin, J. (1993) Fluxes, doses and poisons — molecular perspectives on dominance. *Trends Genet.* 9: 1–2.
- Lyttle, T.W. (1993) Cheaters sometimes prosper — distortion of Mendelian segregation by meiotic drive. *Trends Genet.* 9: 205–210.
- Mackay, T.F.C. (1995) The genetic basis of quantitative variation — numbers of sensory bristles of

## Further reading



- Drosophila melanogaster* as a model system. *Trends Genet.* 11: 464–470.
- Weeks, D.E. and Lathrop, G.M. (1996) Polygenic disease — methods for mapping complex disease traits. *Trends Genet.* 11: 513–519.
- Wilkie, A.O.M. (1994) The molecular basis of genetic dominance. *J. Med. Genet.* 31: 89–98.
- Wolf, U. (1995) The genetic contribution to phenotype. *Hum. Genet.* 95: 127–148.

### Website

On-line Mendelian Inheritance in Man (OMIM):  
<http://gdbwww.gdb.org/omimdoc/omimtop.html>

**This Page Intentionally Left Blank**

## Chapter 2

# The Cell Cycle

### Fundamental concepts and definitions

- The **cell cycle** is the sequence of events between successive cell divisions.
- Many different processes must be coordinated during the cell cycle, some of which occur continuously (e.g. cell growth) and some discontinuously, as events or landmarks (e.g. cell division). Cell division must be coordinated with growth and DNA replication so that cell size and DNA content remain constant.
- The cell cycle comprises a **nuclear or chromosomal cycle** (DNA replication and partition) and a **cytoplasmic or cell division cycle** (doubling and division of cytoplasmic components, which in eukaryotes includes the organelles). The DNA is considered separately from other cell contents because it is usually present in only one or two copies per vegetative cell, and its replication and segregation must therefore be precisely controlled. Most of the remainder of the cell contents are synthesized continuously and in sufficient quantity to be distributed equally into the daughter cells when the parental cell is big enough to divide. An exception is the **centrosome**, an organelle that is pivotal in the process of chromosome segregation itself, which is duplicated prior to mitosis and segregated into the daughter cells with the chromosomes (the **centrosome cycle**).
- In eukaryotes, the two major events of the chromosomal cycle, replication and mitosis, are controlled so that they can never occur simultaneously. Conversely, in bacteria the analogous processes, replication and partition, are coordinated so that partially replicated chromosomes can segregate during rapid growth. The eukaryotic cell cycle is divided into discrete phases which proceed in a particular order, whereas the stages of the bacterial cell cycle may overlap.
- The progress of the eukaryotic cell cycle is controlled at checkpoints where regulatory proteins receive input from monitors of the cell cycle itself (intrinsic information) and monitors of the environment (extrinsic information). Intrinsic monitoring insures that the stages of the cell cycle proceed in the correct order and that one stage is completed before the next begins. Extrinsic monitoring coordinates cell division with cell growth and arrests the cell cycle if the environment is unsuitable.
- The cell cycle is controlled by protein kinases. Cell cycle transitions involve positive feedback loops which cause sudden bursts of kinase activity, allowing switches in the states of phosphorylation of batteries of effector proteins. Cell cycle checkpoints are regulatory systems which inhibit those kinases if the internal or external environment is unsuitable. The alternation of DNA replication and mitosis is controlled by negative feedback — mitosis is inhibited by unfinished DNA replication, and DNA replication is prevented during mitosis by the phosphorylation and inactivation of a protein required for replication. The cell cycle is the result of a complex network of information, in which kinases are controlled by the integration of multiple positive and negative signals.

### 2.1 The bacterial cell cycle

**DNA replication and growth coordination.** The Helmstetter–Cooper model (or I + C + D model) divides the bacterial chromosome cycle into three phases, the **interval phase**, the **chromosome replication phase** and the **division phase**, represented by the letters I, C and D, respectively.

DNA replication occurs during the C phase; its duration is fixed (about 40 min in *E. coli*), reflecting the time taken to replicate the whole chromosome (see Replication). The D phase begins



when replication is complete, and culminates in cell division. The duration of the D phase is also fixed (about 20 min in *E. coli*), and can be regarded as the time required to synthesize the cellular components required for cell division. The minimum duration of the chromosome cycle in *E. coli* is thus 1 h.

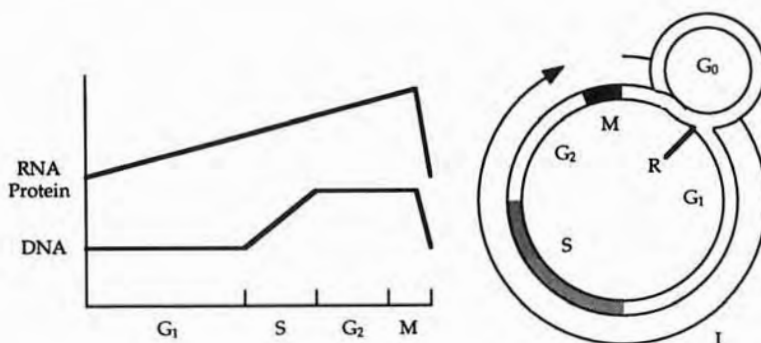
Because  $C + D$  is fixed, any change in the cell doubling time must reflect a change in the duration of I, the interval between successive initiations of replication. The doubling time of *E. coli* can be as long as 3 h or as short as 20 min. During slow growth,  $I > C + D$  and replication is completed before cell division. During rapid growth, however, the doubling time is shorter than the time taken to complete a round of replication and cell division. The only way the cell can accommodate its fixed chromosome cycle into the accelerated cell division cycle is to make  $I < C + D$ , i.e. new rounds of replication must begin before the previous round is complete. Therefore, during rapid growth, daughter cells inherit chromosomes which are already partially replicated (**multiforked chromosomes**) so that replication can be completed before the next round of cell division.

The frequency of initiation is thought to be controlled by a positive regulator which must be present at a certain critical concentration per *origin of replication* (q.v.) for initiation to be successful. During rapid growth, the regulator accumulates more quickly, allowing more frequent initiation. Once initiation has occurred, the number of origins in the cell doubles and the effective concentration of the regulator is halved so that it must accumulate again for another round of initiation to occur. The existence of a positive regulator is predicted because *de novo* protein synthesis is required for initiation; however, the nature of this putative molecule is unknown. The replication initiator protein DnaA is a possible candidate, and factors which control methylation at *oriC* (and the *dnaA* promoter) could also be involved (q.v. *origin of replication*, *Dam* methylation).

**Partition and cytokinesis.** The partition of the replicated chromosome marks the culmination of the chromosome cycle and is followed by cell division. A **septum** forms at the midpoint of the parental cell, which is identified by a **periseptal annulus**, a region of modified cell envelope where the inner and outer membranes are joined together around the circumference of the cell. Additional annuli form by duplication and migrate to positions equivalent to one-quarter and three-quarter cell lengths, and these are the sites of septation in daughter cells during the next round of cell division. Once the septum has formed, the cell undergoes **cytokinesis** — it divides by binary fission.

The identification of mutants which disrupt cell division or partitioning has shown that the two processes can be unhitched, and such mutants fall into several categories. *fts* mutants are deficient in septum formation and thus form filaments that are temperature sensitive (hence the name). The filaments often contain regularly spaced nucleoids, indicating that replication and partitioning mechanisms are still functioning normally. *min* mutants generate septa too frequently, resulting in the formation of **minicells**, which are small cells which contain no chromosomal DNA (although they may contain plasmids). Finally, *par* mutants form normal sized cells but fail to partition the chromosomes properly, so that diploid and anucleate cells arise with high frequency.

The pathway controlling cell division and partition has yet to be determined in full, but several key players have been identified. A good candidate for the initiator of cell division is FtsZ. This protein is structurally and functionally similar to tubulin, which forms the contractile ring in eukaryotic cells. It is distributed ubiquitously during most of the cell cycle, but is localized around the annulus at the beginning of the D phase as a **Z ring**. Its abundance appears to correlate exactly with the frequency of cell division, thus *ftsZ* mutants fail to form septa (and generate filaments) whereas overexpression causes the production of too many septa, and hence minicells. The ZipA protein may be important for the localization of FtsZ because its N terminus is membrane-associated and its C terminus interacts with FtsZ. It is unclear how the septum is positioned in the cell, although **nucleation sites** probably exist because filaments resulting from temperature-sensitive *ftsZ* mutations rapidly form contractile rings at regular intervals when shifted to the permissive temperature. Genes of the *minB* locus limit septation to the central annulus and suppress the process at the terminal



**Figure 2.1:** The standard eukaryotic cell cycle. The chromosome cycle is divided into four stages: **G<sub>1</sub>**, **S** and **G<sub>2</sub>** which constitute the **interphase (I)**, and **M** which is **mitosis**. The left panel shows typical relative durations of the cell cycle stages, although this varies in different species and depending on cell type and growth conditions. Animal cells can withdraw to the quiescent state, **G<sub>0</sub>**, if growth factors are withdrawn in early **G<sub>1</sub>**, but once the cell passes the **restriction point (R)** it becomes committed to a further round of DNA replication and division. The graph compares the accumulation of 'continuous' components with the discontinuous synthesis of DNA. The quantity of all cell components is halved at the end of the **M** phase when cell division occurs.

annuli remaining from previous cell divisions. MinC and MinD are septation inhibitors, whereas MinE antagonizes MinCD activity. The correct balance of all three products thus inhibits terminal septation but protects the central annulus from inhibition.

The partition of bacterial chromosomes proceeds similarly to *plasmid partition* (q.v.). Partitioning may involve the association of the chromosome with the cell membrane, and may be regulated by replication: the origin and terminus of replication, as well as active replication forks, are membrane-associated. Bacterial mutants which affect partition fall into two categories: those which interfere with the separation of interlocked replicated chromosomes (these include topoisomerase and Xer site-specific recombinase mutants) and those which affect the partition process itself. In the latter category no *cis*-acting sites have been found on the chromosome, but several *trans*-acting factors have been identified (e.g. the membrane protein MukA and the microtubule-associated protein MukB; mutations in both genes generate anucleate cells).

## 2.2 The eukaryotic cell cycle

**The standard eukaryotic cell cycle.** The standard eukaryotic cell cycle is divided into four nonoverlapping phases. The discrete events of the chromosome cycle (DNA synthesis and mitosis) occur during the **S phase** and the **M phase**, respectively, and in most cell cycles these landmarks are separated by **G<sub>1</sub>** and **G<sub>2</sub>** gap phases, during which mRNAs and proteins accumulate continuously (Figure 2.1). The process of crossing from one phase of the cell cycle to the next is a **cell cycle transition**. Whereas mitosis is a dramatic event that involves visible reorganization of cell structure, the rest of the cell cycle is unremarkable to the eye and is termed the **interphase**.

Variations on the theme (Table 2.1) include cell cycles where one or both gap phases are omitted, or where either the S phase or the M phase is omitted, leading to halving or doubling of the DNA content, respectively. In addition, a cell may be **arrested** (indefinitely or permanently delayed) at any stage of the cell cycle, as occurs during oocyte maturation and in postmitotic cells such as neurons. Animal cells may withdraw from the cell cycle altogether, entering a quiescent state termed **G<sub>0</sub>**, where both growth and division are repressed. This reflects a continuing requirement for growth factors and other signaling molecules in the environment, and imposes an extra level of regulation on the cell cycle so that the growth and division of individual cells can be coordinated in

**Table 2.1:** Variations on the theme of the four-stage eukaryotic cell cycle

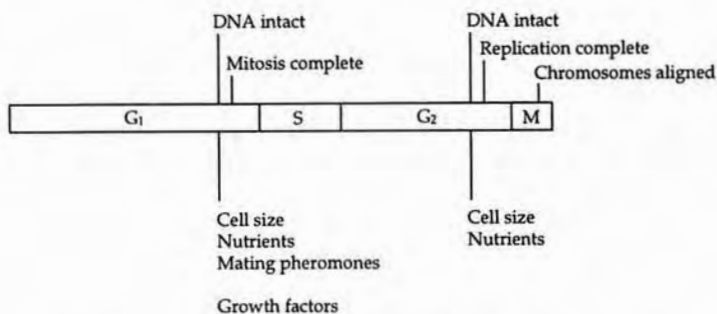
Modification	Circumstances
<i>Stages omitted</i>	
No gap phases	Rapid alternation between M and S phases is characteristic of early development in animals with large eggs because there is enough material in the egg for rapid cleavage divisions without cell growth. Many organisms miss out one or other of the gap phases: <i>Dictyostelium discoideum</i> replicates its DNA immediately following mitosis (no G <sub>1</sub> ), whereas <i>S. cerevisiae</i> undergoes mitosis directly after DNA replication (no G <sub>2</sub> )
No S phase	Two rounds of division without intervening DNA synthesis occur during <i>meiosis</i> (q.v.)
No M phase	Multiple rounds of DNA synthesis without cell division occur in <i>Drosophila</i> secretory tissues to produce <i>polytene chromosomes</i> (q.v.)
<i>Stages extended</i>	
Indefinite arrest at G <sub>1</sub> , G <sub>2</sub> or M	Oocytes and eggs may be arrested at G <sub>1</sub> , G <sub>2</sub> or M depending on species. Fertilization releases the block and allows the cycle to resume
Withdrawal from the cell cycle at G <sub>1</sub> (G <sub>0</sub> )	Many animal cells can withdraw from the cell cycle at G <sub>1</sub> and enter a quiescent state, often termed G <sub>0</sub> , which may last months or years. This occurs if essential growth factors are withheld during early G <sub>1</sub> and involves the disassembly of the cell cycle control mechanism. Quiescent cells can be persuaded to re-enter the cycle if growth factors are made available, but there is a long delay before the initiation of the S phase while regulatory components are resynthesized. Normal somatic cells often enter G <sub>0</sub> after a characteristic number of divisions (the <b>Hayflick limit</b> ), a phenomenon termed <b>senescence</b> which may be related to <i>telomere</i> length (q.v.) in some animals, and can also be induced by certain plasmids in fungi. Some cells withdraw entirely from the cell cycle as part of their differentiation and become <b>postmitotic</b> , e.g. neurons and muscle cells

the context of a multicellular organism. The abnormal cell proliferation seen in cancer is caused by the failure of this regulatory mechanism (see *Oncogenes and Cancer*).

**Cell cycle checkpoints.** The primary function of the cell cycle is to duplicate the genome precisely and divide it equally between two daughter cells. For this reason, it is important that the events of the cell cycle proceed in the correct order, and that each stage of the cell cycle is complete before the next commences. DNA content remains constant only if DNA replication alternates with mitosis, if mitosis occurs after the *completion* of DNA replication, and replication commences after mitosis has precisely divided the DNA. The cell meets these criteria by organizing the cell cycle as a dependent series of events. Thus, if mitosis is blocked, the cell cycle arrests at the M phase until the block is removed — it does not go ahead and replicate the DNA anyway (i.e. DNA replication is dependent upon the completion of mitosis). Similarly, if DNA replication is prevented, the cell does not attempt to undergo mitosis, because mitosis is dependent upon the completion of DNA replication.

A further function of the cell cycle is to coordinate the chromosome cycle with cell growth, so there is no progressive loss or gain of cytoplasm, and no cell proliferation in an unsuitable environment. Progress through the cell cycle is thus also dependent upon cell size and is regulated by nutrient availability, the presence of mating pheromones (in yeast), and the presence of growth factors and hormones (in animals).

The cell possesses a number of regulatory systems which can sense the progress of the cell cycle and can inhibit subsequent stages in the event of failure. These regulatory mechanisms are termed **cell cycle checkpoints**, and represent intrinsic signaling systems of cell cycle control. The checkpoint mechanisms also respond to external signals so that arrest may occur in cases of nutrient deprivation or growth factor withdrawal. There are numerous checkpoints in the cell cycle, which are



**Figure 2.2:** Known checkpoints in the eukaryotic cell cycle. These represent the points at which specific protein kinases are activated/inactivated.

clustered in two major groups — those occurring at G<sub>1</sub> and regulating entry into the S phase, and those occurring at G<sub>2</sub> and regulating entry into the M phase (Figure 2.2). This clustering suggests that intrinsic and extrinsic signals may funnel into common components of cell cycle regulation. Additional checkpoints insure the orderly and dependent series of events which comprise mitosis.

Different organisms attach varying degrees of significance to the G<sub>1</sub> and G<sub>2</sub> checkpoints, reflecting the stage at which the cell receives input from the environment. The G<sub>1</sub> checkpoint is predominant in the budding yeast *Saccharomyces cerevisiae* (where it is called **START**) and in animal cells (where it is called the **restriction point** or **commitment point**). The yeast assesses nutrient availability and the presence of mating pheromones during G<sub>1</sub>, whereas animal cells respond to the presence of growth factors. Cells of both kingdoms will arrest at this checkpoint if the environment is unsuitable for growth, but once past it, they are committed to a round of DNA replication and mitosis regardless of their environment. Conversely, in the fission yeast *Schizosaccharomyces pombe*, the environment is monitored at the G<sub>2</sub> checkpoint, and under satisfactory conditions the cell will undergo mitosis, division and the next round of DNA replication before checking again. The advantage of pausing at G<sub>2</sub> rather than G<sub>1</sub> for the haploid yeast cells reflects the presence of two copies of the genome at G<sub>2</sub>, allowing any damage to DNA to be repaired by recombination.

**Studying cell cycle regulation.** Two complementary approaches have been used to characterize and isolate the regulatory components of the cell cycle. In the heterokaryon approach, nuclei at different stages of the chromosome cycle are joined in a common cytoplasm and their behavior observed. Cultured mammalian cells and amphibian eggs have been used for these experiments. The results of fusing cultured fibroblasts synchronized at different cell cycle stages are shown in Table 2.2. The ability of M-phase cells to induce mitosis in any interphase nucleus provided early evidence for the existence of an **M-phase promoting factor**. Similar results were obtained in *Xenopus* nuclear

**Table 2.2:** Heterokaryon experiments to investigate regulatory factors controlling the cell cycle

Fusion	Result	Conclusion
S × G <sub>1</sub>	Both nuclei replicate	S nucleus contains an S-phase promoting factor
S × G <sub>2</sub>	S-phase cell completes replication, G <sub>2</sub> -phase nucleus waits for S-phase nucleus to complete replication and then both cells enter the M phase	G <sub>2</sub> nucleus cannot respond to S-phase activator (a re-replication block), S-phase activator is also an inhibitor of mitosis
M × G <sub>1</sub> , S or G <sub>2</sub>	Interphase nucleus enters precocious mitosis (regardless of state of chromosome replication)	M nucleus contains an M-phase promoting factor
G <sub>1</sub> × G <sub>2</sub>	Neither nucleus undergoes replication or mitosis	Both S-phase and M-phase activators are present transiently



transplantation studies — interphase nuclei formed spindles when injected into eggs arrested at the metaphase of meiosis I, and cytoplasm from these eggs could induce meiosis in oocytes arrested at G<sub>2</sub>. The large size of the eggs was exploited to purify the substance responsible, which was called **maturation promoting factor**. Further studies showed that maturation promoting factor could also induce mitosis in somatic cells, and was in fact identical to M-phase promoting factor, which shared the same acronym (MPF).

The second approach has been to exploit the versatility of yeast genetics to isolate conditional mutants for cell cycle functions. Numerous *cdc* mutants (cell division cycle) have been identified which are blocked at various stages of the cell cycle, yet continue to grow. A second class of so-called *wee* mutants allows precocious transition of cell size checkpoints, and are smaller than wild-type cells. Many of the genes identified from *cdc* mutants are not specific cell cycle regulators, but control processes such as DNA replication, repair and mating, upon which the progress of the cell cycle depends. However, a number of *cdc* genes appear to play a direct role in the regulation of the cell cycle, as discussed in the following section. Satisfyingly, the biochemical analysis of MPF has shown that both approaches have converged on the same small group of molecules.

### 2.3 The molecular basis of cell cycle regulation

**Cyclins and cyclin-dependent kinases.** The sequential stages of the cell cycle reflect alternative states of phosphorylation for key proteins which mediate the different cell cycle events. The cell cycle transitions represent switches in those phosphorylation states. The G<sub>1</sub>–S transition involves the phosphorylation of proteins required for DNA replication, whilst the G<sub>2</sub>–M transition involves the phosphorylation of proteins required for mitosis. The basis of cell cycle regulation is a family of protein kinases which phosphorylate these target proteins and hence coordinate the different activities required for each transition.

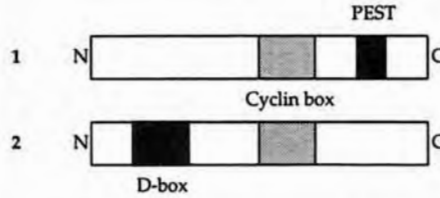
The involvement of protein kinases in cell cycle control was revealed when analysis of *S. cerevisiae* *cdc* mutants blocked at START identified the product of the *CDC28* gene, a 34 kD protein kinase, as the principal regulator of the G<sub>1</sub>–S transition. The *cdc2* gene, which played an equally important role in the G<sub>2</sub>–M transition in *S. pombe*, was found to encode a homologous protein kinase. Genes encoding similar kinases were subsequently isolated from vertebrates, and these could restore wild-type cell cycle function to yeast *cdc* mutants. Significantly, the *Xenopus* homolog of *CDC28/Cdc2* was found to be a component of MPF.

The kinases were found to be present constitutively in the nucleus, but to control cell cycle transitions their activity would have to oscillate. An explanation for their periodic activity came from the study of sea urchin eggs, wherein were discovered a family of molecules whose synthesis and activity oscillated with the cell cycle. These molecules were termed **cyclins** and they were subsequently found in many other eukaryotes including yeast and vertebrates. The second component of MPF was found to be a B-type cyclin.

The activity of MPF resides in the catalytic kinase subunit but it is dependent upon the cyclin subunit, which introduces a conformational change in its partner to stimulate kinase activity. The cell cycle kinases are thus described as **cyclin-dependent kinases (CDKs)** and function as CDK–cyclin holoenzymes. This strategy of cell cycle regulation appears to be conserved throughout the eukaryotes.

**CDK–cyclin diversity in the yeast and animal cell cycles.** A number of potential CDKs have been isolated from yeast, but only *CDC28* in *S. cerevisiae* and *Cdc2* in *S. pombe* appear to be directly involved in the cell cycle, and are required for the G<sub>1</sub>–S and the G<sub>2</sub>–M transitions in both species. In animal cells, there is a greater diversity of CDKs. The first to be discovered, the p34<sup>CDC28/Cdc2</sup> component of MPF, appears to function specifically at the G<sub>2</sub>–M transition. Ten or more further CDKs are present in animal cells; five of these are involved specifically in the early stages of the cell cycle.





**Figure 2.3:** Domain structure of cyclins. (1) Cyclins which possess a PEST motif are targeted for proteolytic degradation and are very unstable. This class includes the *S. cerevisiae* CLN cyclins and vertebrate cyclins of classes C, D, E and F. Most of these are G<sub>1</sub>/S cyclins. (2) Mitotic cyclins tend to be stable throughout interphase, but contain a destruction box required for their ubiquitin-dependent degradation during M phase. The first cyclins were isolated on the basis of their oscillating activity, but several are known to be synthesized constitutively and are defined on the basis of cyclin box homology rather than expression parameters.

The diversity of cyclins is greater than that of CDKs, as different cyclins are synthesized at different stages of the cell cycle in both animals and yeast. There are at least eight families of vertebrate cyclins (designated A–H). Since CDKs phosphorylate different targets at each cell cycle transition, cyclins are required not only for kinase activity, but also for substrate specificity. In animals, alternative cyclins may be differentially expressed in different cell types, which would facilitate the unique aspects of cell cycle control in distinct differentiated cells. There are generally three types of cyclin in all organisms: the **G<sub>1</sub> cyclins** which regulate the G<sub>1</sub>–S transition, the **S-phase cyclins** which are required for DNA replication, and the **M-phase cyclins** which are required for mitosis. M-phase cyclins include the *S. cerevisiae* CLB cyclins, the vertebrate A- and B-type cyclins and the *S. pombe* cyclin Cig13. They are stable proteins but share a conserved motif called a **destruction box**, which is required for targeted ubiquitination (q.v.) resulting in degradation during mitosis. Other cyclins are inherently unstable because they carry a **PEST domain** (q.v.), and their levels are determined primarily by the transcriptional activity of their genes. All cyclins carry a conserved motif, the **cyclin box**, which is required for CDK binding (Figure 2.3). The yeast and animal CDKs and cyclins involved in cell cycle regulation are summarized in Table 2.3.

**Regulation of CDK–cyclin activity.** Cell cycle transitions are characterized by bursts of CDK activity which cause sudden switches in the phosphorylation states of target proteins responsible for cell cycle events. Sudden spikes of kinase activity are not regulated by cyclin synthesis and degradation alone, as cyclins accumulate gradually in the cell and only the mitotic cyclins are degraded by rapid, targeted proteolysis.

**Table 2.3:** Principle CDKs and cyclins active at each stage of the yeast and mammalian cell cycles

Stage	<i>S. cerevisiae</i>	<i>S. pombe</i>	Mammals
G <sub>1</sub>	CDK: CDC28 Cyclins: CLN1–3	CDK: Cdc2 Cyclins: Cig2	CDK: CDK2, 4, 5, 6 Cyclins: D1–3
S-phase	CDK: CDC28 Cyclins: CLN5, CLN6	CDK: Cdc2 Cyclins: Cig2 ?	CDK: CDK2 Cyclins: E Class
G <sub>2</sub> /M-phase	CDK: CDC28 Cyclins: CLB1–4	CDK: Cdc2 Cyclins: Cdc13	CDK: CDK1 (CDC2) Cyclins: A and B

Two CDK–cyclin systems are active in G<sub>1</sub> of the mammalian cell cycle. The CDK2–cyclin E complex is required for the G<sub>1</sub>–S transition. The other CDKs and the D cyclins are responsible for interpreting growth factor signals for the environment, and act at the restriction point to channel the cell into either late G<sub>1</sub> or G<sub>0</sub>. The mammalian CDK1–cyclin B complex is MPF — the vertebrate homolog of yeast Cdc2/CDC28 may be termed Cdc2 or CDK1. CDK7/H–cyclin, which is CAK, the mammalian CDK Thr-161 kinase, is thought to be expressed constitutively.

CDKs phosphorylate target proteins, but are themselves also regulated by phosphorylation (see Signal Transduction). The yeast CDC28 and Cdc2 CDKs are phosphorylated on two key residues, Tyr-15 and Thr-161. Phosphorylated Thr-161 is required for kinase activity, whereas phosphorylated Tyr-15 is inhibitory and dominant to Thr-161 phosphorylation. The principle determinant of CDK-cyclin activity in yeast is thus the state of phosphorylation of Tyr-15, and some of the upstream regulatory components have been identified. In *S. pombe*, Wee1 is a tyrosine kinase which phosphorylates Cdc2 at Tyr-15 and thus inactivates it. Wee1 activity is antagonized by Cdc25 phosphatase, which removes phosphate groups from the same substrate. Both Wee1 and Cdc25 are themselves regulated by intrinsic and extrinsic signals, and this is believed to be the basis of the G<sub>2</sub>-M checkpoint in this species. The decision to proceed with mitosis or arrest in G<sub>2</sub> thus reflects the relative levels of these opposing activities, and the regulatory networks which feed into this checkpoint are considered below.

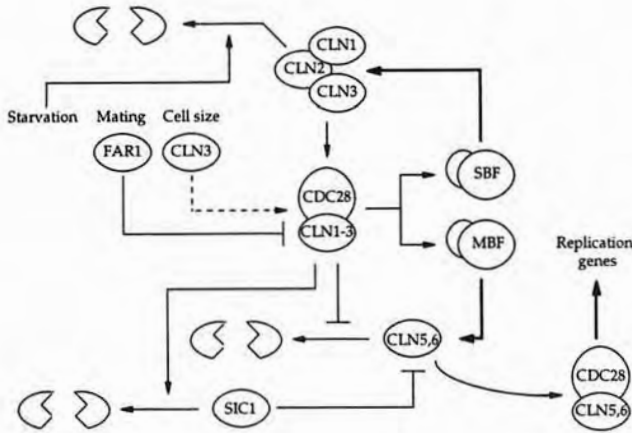
Homologs of *wee1* and *cdc25* have been identified in mammals, although the situation is more complex than in yeast because there are multiple isoforms which may demonstrate specificities for particular CDK-cyclin complexes. Additionally, there are three phosphorylation sites on mammalian CDC2 (CDK1), Thr-14 and Tyr-15, both of which are dominant inhibitors when phosphorylated, and Thr-161, whose phosphorylation is required for kinase activity. An enzyme has been identified in mammals which is responsible for Thr-161 phosphorylation. Remarkably, this turns out to be yet another CDK-cyclin complex comprising CDK7 and cyclin H; it is known as **CDK-activating kinase (CAK)** (q.v. *TFIIF*). The enzyme responsible for Thr-14 phosphorylation has not been identified.

CDK-cyclin complexes are also regulated by inhibitory proteins. This third level of control is used for both intrinsic and extrinsic regulation purposes. As discussed below, the *S. pombe* Rum1 protein is a specific inhibitor of the mitotic CDK-cyclin complex, and is synthesized throughout the G<sub>1</sub> and S phases, thus preventing the cycle skipping DNA replication and entering mitosis prematurely. The FAR1 protein is activated in response to signaling by mating-type pheromones in *S. cerevisiae* and inhibits the START CDK-cyclin complex, thus arresting the cell at G<sub>1</sub> in readiness for mating. Two families of **CDK-cyclin inhibitors (CKIs)** are found in animals. One blocks all CDK-cyclin activities and the other specifically inhibits D-cyclin complexes containing CDK4 and CDK6 (whose major substrate is the retinoblastoma protein). These regulatory mechanisms are discussed in more detail below.

## 2.4 Progress through the cell cycle

**Transition of the START checkpoint in yeast.** START is the predominant checkpoint in the *S. cerevisiae* cell cycle and is well characterized in this species (Figure 2.4). In G<sub>1</sub>, the CDC28 CDK can associate with any of three cyclins (CLN1, CLN2, CLN3) which are functionally redundant, at least under laboratory conditions (the genes can be deleted singly or in pairs with no effect on phenotype, but mutation in all three arrests the cell cycle at G<sub>1</sub>). The highly unstable CLN3 is expressed constitutively and acts as an indicator of cell growth. When the concentration of CDC28-CLN3 reaches a critical level, the kinase activates two transcription factors, SBF and MBF, which contain a common component, SWI6. SBF initiates transcription of the genes for the other G<sub>1</sub> cyclins, CLN1 and CLN2. These cyclins associate with CDC28 to form CDK-cyclin complexes with several functions related to the activation of the six CLB cyclins (the mitotic cyclins CLB1, CLB2, CLB3 and CLB4, and the S-phase cyclins CLB5 and CLB6), which share a conserved destruction box. Firstly, they reduce the rate of proteolysis of the CLB cyclins, and secondly, they increase the rate of hydrolysis of a CDK-CLB cyclin inhibitor, SIC1.

The second transcription factor, MBF, initiates transcription of the *CLB5* and *CLB6* genes. Thus, by stimulating the transcription of S-phase cyclins and increasing their stability, the G<sub>1</sub>-specific CDK-cyclin complexes induce the onset of the S phase when cell growth is sufficient. The S-phase-specific CDK-cyclin complexes stimulate the transcription or activity of replication proteins,



**Figure 2.4:** Summary of the molecular regulation of START in *S. cerevisiae*. Accumulation of CLN3 acts as an indicator of cell size (dashed line). At a critical concentration, the kinase activity of CDC28 is induced and activates two transcription factors. SBF promotes *CLN1–3* transcription, thus generating a positive feedback loop, although proteolytic degradation of CLN2 due to starvation can delay entry into the S phase. MBF activates the genes for CLN5 and CLN6 whilst activated CDC28 kinase inhibits destruction box-targeted cyclin degradation and promotes degradation of the inhibitor SIC1. CLN5 and CLN6 form complexes with CDC28 which activate the expression of replication genes.

perhaps including those found at the constitutive origin recognition complexes on yeast ARS elements (see Replication).

The cell cannot pass START until mitosis is complete and in *S. pombe* this is achieved by a simple negative feedback mechanism in which the mitotic CDK–cyclin complex inhibits a protein required for the initiation of DNA replication, Cdc18. In the absence of active Cdc18, DNA replication is blocked, and this repression is lifted when the mitotic kinase is destroyed (see below).

**Entry into the S phase in mammalian cells.** In mammals, four CDKs (CDK2, CDK3, CDK4 and CDK6) act in early  $G_1$ . All four associate with the D cyclins, which are synthesized early in  $G_1$  in a growth-factor-dependent manner (growth factor regulation of the cell cycle is discussed below). CDK4 and CDK6 play the principle roles in the regulation of downstream events. The only known target of these early CDK–cyclin complexes is the retinoblastoma protein RB-1, which is a negative regulator of the cell cycle. RB-1 in its unphosphorylated form binds transcription factors of the E2F family which normally activate genes required for entry into S-phase. RB-1 inhibits the expression of these genes in two ways: by sequestering the E2F activation domain, and direct repression by chromatin remodeling. Phosphorylation of RB-1 by the early  $G_1$  CDK–D cyclin complexes releases the repression. An important downstream effect of CDK–D cyclin activation is the synthesis of E cyclins, which are required for the  $G_1$ –S transition itself. The E cyclins form complexes with CDK2, and these activate CDC25A phosphatase, which may in turn activate the S-phase-specific complex CDK2/A cyclin. This complex is required for the initiation of replication, and has been localized at replication origins where it can be immunoprecipitated with PCNA (q.v.), a component of the major eukaryotic DNA polymerase. Its specific targets in the initiation complex remain to be characterized, but both cyclin A and cyclin E complexes are known to phosphorylate and regulate the activity of several transcription factors including the E2F family, p53, B-Myb and the inhibitory helix–loop–helix protein Id2.

**The restriction of replication to once per cell cycle.** The cell fusion studies described above show that the S-phase nucleus can induce the  $G_1$  nucleus to replicate but not the  $G_2$  or M nuclei, which need to complete mitosis before becoming competent for replication once again. In all cases,

replication occurs only once. A model proposed to explain these data involves a **replication licensing factor** displaying the following properties:

- it interacts with DNA origins in the latter stages of mitosis;
- it is essential for initiation of replication, but is inactivated or destroyed once initiation has taken place;
- it is unable to cross the nuclear envelope.

The model proposes that a cytoplasmic licensing factor interacts with DNA origins before reassembly of the nuclear envelope at the telophase. This factor is a target for regulation by S-phase-promoting CDK–cyclin complexes and initiates replication at each origin at the beginning of the S phase, becoming inactive in the process. During the remainder of the cell cycle, the new licensing factor is unable to enter the nucleus, and the precocious initiation of replication is prevented. The factor enters the nucleus at the start of mitosis, but is inactive or prevented from interacting with DNA because of the condensed state of the chromatin. The factor interacts with origins as the chromatin decondenses, but remains inactive, awaiting the S-phase signal. The nuclear envelope closes, preventing entry by further licensing factor, and any uncomplexed factor in the nucleus, and perhaps also in the cytoplasm, is targeted for degradation.

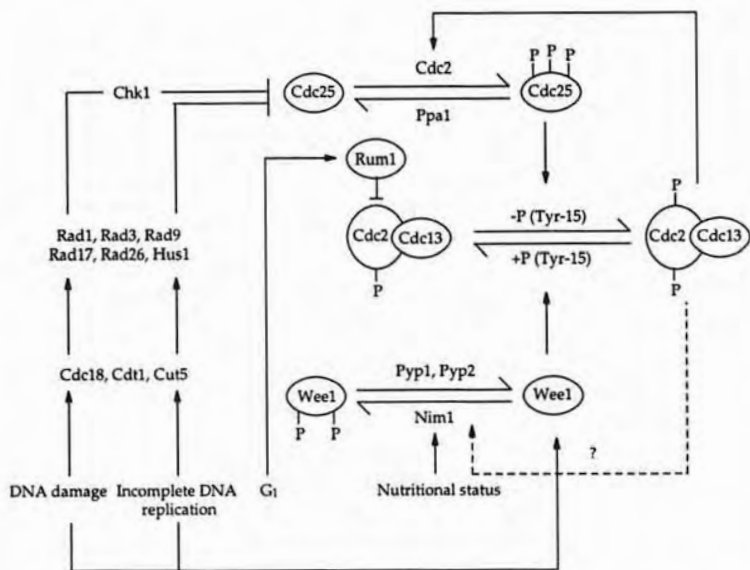
Proteins which display many of the properties of the licensing factor have been identified in *Xenopus*, mammals and yeast. In *S. cerevisiae*, the product of the *CDC46* gene is thought to be the licensing factor, or a component of it, although in yeast the nuclear envelope does not break down during mitosis and the licensing factor is transported into the nucleus by translocation.

**Control of entry into mitosis.** In yeast, entry into mitosis is controlled by the same cyclin dependent kinase that regulates START (*Cdc2*, *CDC28*), but in combination with specific mitotic cyclins. In *S. cerevisiae*, these are encoded by the *CLB1*, *CLB2*, *CLB3* and *CLB4* genes, and in *S. pombe* by the *cdc13* gene. In vertebrates, the CDK1(*CDC2*) kinase associates with A- and B-type cyclins which, like their yeast counterparts, contain mitotic destruction boxes. Much of the regulatory pathway of the G<sub>2</sub>–M checkpoint has been determined in *S. pombe*, and similar components are found in other eukaryotes. Essentially, mitosis is initiated by a burst of *Cdc2* activity generated by a posttranslational positive feedback mechanism (this differs from entry into the S phase in *S. cerevisiae*, which is governed by a transcriptional feedback loop). The mechanism is broadly conserved throughout the eukaryotes, but the checkpoints for entry into mitosis are regulated in different ways, as discussed below.

In *S. pombe*, *Cdc2* cyclin-dependent kinase appears to be regulated predominantly by phosphorylation of the Tyr-15 residue (Figure 2.5). It is activated by dephosphorylation, carried out by *Cdc25* and another phosphatase termed *Pyp3*, and inactivated by redundant kinases termed *Wee1* and *Mik1*. The G<sub>2</sub>–M transition coincides with a burst of *Cdc2* kinase activity which is controlled by a positive feedback loop in which the *Cdc2/Cdc13* kinase phosphorylates and activates *Cdc25*, which in turn dephosphorylates and activates *Cdc2* (*Cdc2* may also inactivate *Wee1*). This sudden exponential rise in kinase activity simultaneously changes the phosphorylation states of many target proteins and effectively throws a switch which initiates mitosis (Table 2.4). *Wee1* and *Cdc25* process intrinsic signals and integrate signals from the environment, and are probably the main checkpoint regulators. *Wee1* activation is regulated by a kinase *Nim1*, which is responsive to the nutritional status of the cell. Checks for completion of DNA replication or the presence of damaged DNA are mediated through a group of genes, including *cdc18*, *cdt1* and *cut5*, whose products are thought to form a recognition complex which is monitored by *Rad1*, *Rad3*, *Rad9*, *Rad17*, *Rad26* and *Hus1*. The downstream targets for these signals may be *Cdc25* or *Wee1* or both. *Rad3* has recently been shown to phosphorylate *Chk1*, which in turn phosphorylates *Cdc25* and causes it to be sequestered in a complex with 14–3–3 protein. In animals, there may also be a mechanism which regulates *CDC2* kinase activity independently of phosphorylation on Tyr-15.

The checkpoints for DNA damage and DNA replication are different in *S. cerevisiae*, which lacks a definite G<sub>2</sub> phase. In this species, the cell forms a spindle and arrests in metaphase in response to





**Figure 2.5:** Control of entry into mitosis in *S. pombe*. Entry into mitosis requires active Cdc2/Cdc13 kinase which phosphorylates and activates a number of proteins required for the appropriate structural changes in the cell. Cdc2/Cdc13 activity is controlled predominantly by phosphorylation of Cdc2 Tyr-15, which is regulated by the opposing activities of Cdc25 phosphatase and Wee1 kinase. Both these regulators are themselves regulated by intrinsic and extrinsic signals (e.g. incomplete DNA replication, nutritional status), as discussed in the text.

**Table 2.4:** Structural changes in the cell at the beginning of the M phase and the role of M-phase kinase

Substrates	Role in mitosis
Cdc25	Positive feedback loop to generate surge of M-phase kinase activity
Anaphase promoting complex (This is likely to be an indirect substrate)	Negative feedback loop to induce degradation of cyclin. Causes transient spike of M-phase kinase activity at onset of the M phase
Histone H1, HMG proteins	Condensation of chromatin
TFIID, TFIIB, poly(A) polymerase	Inhibition of basal transcription
SBF, SWI5 ( <i>S. cerevisiae</i> )	Regulation of transcription (phosphorylation of SWI5 inhibits nuclear import)
Lamins	Phosphorylation of nuclear lamins causes dissociation into subunits which breaks down nuclear lamina and may drive dissolution of the nuclear membrane
RMSA (regulator of mitotic spindle assembly)	Mitotic spindle assembly
Vimentin	Reorganization of cytoskeleton
Caldesmon, Myosin light chain	Caldesmon is an actin filament regulator and MLC interacts with actin to form contractile ring at the site of cytokinesis. Phosphorylation inhibits this activity
Caesin kinase II $\alpha$ and $\beta$ subunits, c-Src, c-Abl	Protein kinases with regulatory roles or specificity for other downstream targets?

Note that the nuclear membrane breaks up in higher eukaryotes but not in yeast, where the nucleus as well as the cell undergoes division (**karyokinesis**).

DNA damage or incomplete replication. The target for the checkpoint control is the anaphase-promoting complex (APC; see below), which is thought to promote the metaphase–anaphase transition by linking ubiquitin to the mitotic cyclin destruction box. The DNA damage checkpoint appears to be mediated by a protein called PDS1, which may control sister chromatid separation or may function indirectly by inhibiting APC. Interestingly, the DNA replication checkpoint of *Aspergillus nidulans* involves both control of Cdc2 (through Wee1 and Cdc25) and a downstream component, NimA kinase, through the product of the *bimE* gene. BimE is a component of the *A. nidulans* APC.

**Mitosis.** The M-phase CDK–cyclin complex phosphorylates target proteins mediating the structural changes occurring during mitosis (condensation of chromatin, assembly of the spindle apparatus, reorganization of the nucleus and assembly of the cytokinetic machinery), although few targets have been precisely defined (Table 2.4). One critical but probably indirect target of the M-phase kinase is the **anaphase-promoting complex (APC)**, a destruction box-dependent ubiquitin ligase which facilitates the sudden degradation of mitotic cyclins (in animals, cyclin A at metaphase and cyclin B at the metaphase–anaphase transition — the differential timing may reflect the efficiencies of the destruction boxes). Other APC targets include proteins required for maintaining the association of sister chromatids, possibly kinetochore components or perhaps a more generally distributed factor — a candidate is *S. cerevisiae* PDS1, as discussed above. Another important target of M-phase kinase is the myosin light chain, which in its phosphorylated state is prevented from binding to actin and thus from forming the cytoskeletal ring required for cytokinesis. Degradation of cyclin B is dependent upon chromatin alignment on the mitotic spindle, so it is possible that formation of the APC is dependent upon kinetochore attachment and inhibited by unattached kinetochores. The inactivation of M-phase kinase (triggered by degradation of cyclin B) allows M-phase kinase substrates such as myosin to be dephosphorylated, thus facilitating the cytoskeletal organization which precludes cytokinesis, the reconstruction of the nuclear lamina and decondensation of chromatin. The cell ensures the correct order of events in mitosis, as for the cell cycle as a whole, by organizing them into a dependent series (Table 2.5).

**Table 2.5:** The stages of mitosis with important cellular events indicated and molecular mechanisms (where known) in italics

Stage	Critical events
Prophase	<i>M-phase kinase activated.</i> Chromatin condenses and mitotic spindle begins to form. These are probably direct results of phosphorylation of target proteins by M-phase kinase
Prometaphase	In metazoans, breakdown of nuclear envelope allows mitotic spindle access to chromosomes. Some spindle microtubules attach to kinetochores. Nuclear reorganization is also directly regulated by M-phase kinase. When all chromosomes attached to spindle, <i>APC activated by CDC20/Slp1 following inhibition of MAD/BUB pathway</i>
Metaphase	<i>Cyclin A degraded.</i> Chromosomes guided to <b>metaphase plate</b> held by tension between opposing poles of spindle, which are attached to sister kinetochores
Anaphase	<i>Cyclin B degraded; M-phase kinase inactive.</i> Paired kinetochores separate towards poles as kinetochore microtubules shorten; poles also move apart as polar microtubules repel each other. <i>Proteolytic machinery responsible for cyclin B degradation probably also targets a protein which binds sister chromatids together (Scc1/Mcd1). This stage must be dependent on completed chromosome alignment. The nature of the early metaphase signal which prevents cyclin B destruction and precocious segregation is unknown, but constitutes a major cell cycle checkpoint</i>
Telophase	<i>Substrates of M-phase kinase dephosphorylated (including myosin light chain, nuclear lamins, histones).</i> Chromosomes reach poles, nuclear membrane reforms, chromatin decondenses, contractile ring assembles
Cytokinesis	Contractile ring completed. Cell constricts around remains of mitotic spindle and cleaves into two daughter cells



## 2.5 Special cell cycle systems in animals

**Exit from the cell cycle.** As discussed above, the withdrawal of growth factors from animal cells at a critical period during  $G_1$  causes them to cease growth and exit from the cell cycle, entering a quiescent state termed  $G_0$ . This may be a transient response to the withdrawal of growth factors, a permanent aspect of differentiation (e.g. in neurons and muscle cells), or regulated during development to control the final size of a growing structure. The absence of growth factors stimulates exit from the cell cycle because growth factors are required for the synthesis of the D cyclins; these are very unstable proteins replenished only by new transcription early in  $G_1$ . It is this transcription which is activated by growth factor signaling (see Signal Transduction). In summary, growth factors are ligands for receptor tyrosine kinases which initiate a signal transduction cascade through Ras, Raf and MAP kinases, eventually activating transcription factors in the nucleus. The targets for these transcription factors are a set of so-called **immediate early genes**, some of which also encode transcriptional regulators likely to activate D-cyclin transcription.  $G_0$  cells stimulated with growth factors enter the S phase several hours later, reflecting the time taken for D-cyclin mRNA to appear in the cytoplasm.

Withdrawal from the cell cycle can also be stimulated by growth inhibitors such as TGF- $\beta$ , contact inhibition or loss of substrate. These diverse signals act through a family of small cyclin-dependent kinase inhibitors (CKIs) targeting specific components of the cell cycle machinery. The p16 family of inhibitors specifically inhibit CDK-D cyclins by preventing both their assembly and their ability to phosphorylate the RB-1 protein (q.v. *tumor-suppressor genes*). The p21 and p27 families inhibit all CDK-cyclin activities and thus arrest the cell cycle by inhibiting the activity of CDK-cyclin D complexes by preventing their activation by CAK (which is itself a CDK-cyclin holoenzyme) and also by blocking the activity of cyclin E and cyclin A-containing complexes, and the PCNA/ $\delta$ -DNA polymerase complex itself. The CKIs also integrate signals from intrinsic pathways. A major pathway of DNA damage regulation works through the p53 regulator which activates p21 (see Oncogenes and Cancer).

As well as stimulating D-cyclin synthesis, growth factors facilitate cell growth by repressing many of the CKIs discussed above. Ironically, it has been found that p21 or p27 are constitutively present in the cell and comprise part of the normal, functional CDK-cyclin complex. Stoichiometric binding of the inhibitory proteins does not affect kinase activity, but inhibition occurs at higher concentrations. p21 is also capable of direct interaction with PCNA/DNA polymerase  $\delta$ , so that DNA damage blocks DNA replication directly rather than signaling through a CDK-dependent kinase cascade.

**Apoptosis as an alternative to cell cycle arrest.** Apoptosis is programmed cell death, a form of cell death initiated by the cell itself in response to various intrinsic and external signals (as opposed to necrosis which is caused by damage, infection or injury). Apoptosis has an important role in development and is also a defence against chemical insult and infection. The inhibition of apoptosis can lead to cancer.

The behavior of a cell undergoing apoptosis is distinct from that of a necrotic cell. It involves condensation and peripheralization of chromatin, loss of cytoplasm, fragmentation of the nucleus, compaction of organelles, the disestablishment of communication with neighboring cells, fusion of the endoplasmic reticulum with the outer cell membrane and finally fragmentation of the cell into numerous **apoptotic bodies** which are engulfed by surrounding cells. There is no inflammation as seen with necrosis. Prior to chromatin condensation, DNA is degraded into small fragments representing multiples of the nucleosome unit.

Central to the control of apoptosis is the BCL-2 family of cell death regulators. BCL-2 itself is a **survival factor** (i.e. it inhibits apoptosis) and is homologous to the *C. elegans* CED-9 protein expressed in all surviving cells during development. Other family members (e.g. BAX, BAD and BAK) are stimulators of apoptosis. The proteins form dimers, and competition between them,

reflecting their relative abundances, determines the fate of the cell. The effectors of BCL-2 signaling are a family of cysteine proteases related to the interleukin-1 $\beta$  converting enzyme (**ICE proteases**). A number of such enzymes are synthesized by mammalian cells as inactive zymogens and are activated in a proteolytic cascade, culminating in the activation of proteins such as poly(ADP)ribose polymerase and nuclear lamins, the direct effectors of apoptotic cell behavior.

The signal transduction pathways leading to apoptosis are not fully understood. One pathway involves p53 (*see* Oncogenes and Cancer), which initiates apoptosis in response to DNA damage in some cell types (e.g. skin cells) by inhibiting BCL-2 and activating BAX. In other cells, p53 arrests the cell cycle in G<sub>1</sub> by activating the CKI proteins which inhibit CDK-cyclin complexes and *stress-activated protein kinases* (q.v.). This pathway appears not to be involved in apoptosis occurring during development (e.g. in the interdigital regions of the mouse limb bud) as mice homozygous for a deletion of *TP53* develop normally.

Other signals, e.g. growth inhibitory factors such as TGF- $\beta$ , can also induce apoptosis, and do so through BCL-2 family proteins but independently of p53. There are also signaling pathways influencing the ICE family proteases directly. For example, tumor necrosis factor (TNF) initiates the apoptotic signaling cascade by revealing an 80 residue cytoplasmic **death domain** which is shared by the TNF receptor and several other receptors and intracellular apoptotic signaling molecules (e.g. TRADD, Fas, MORT1 and RIP), all of which activate the ICE proteases. CrmA is an ICE protease inhibitor encoded by cowpox virus. This is a survival factor which blocks apoptosis induced by p53, TGF- $\beta$  and TNF, allowing the virus to avoid the lethal consequences of this response to infection.

## References

- Hartwell, L. (1995) Introduction to cell cycle controls. In: *Cell Cycle Control: Frontiers in Molecular Biology* (eds C.M. Hutchinson and D.M. Glover), pp. 1–15. IRL Press, Oxford.
- Murray, A.W. and Hunt, T. (1993) *The Cell Cycle: An Introduction*. W.H. Freeman, New York.
- Osmani, S.A. and Xiang, S.Y. (1997) Targets of checkpoints controlling mitosis: Lessons from lower eukaryotes. *Trends Cell Biol.* 7: 283–288.
- Baylin, S.B. (1997) Tying it all together: Epigenetics, genetics, cell cycle and cancer. *Science* 277: 1948–1949.
- Donachie, W.D. (1993) The cell cycle of *E. coli*. *Annu. Rev. Microbiol.* 47: 199–230.
- Dynlacht, B.D. (1997) Regulation of transcription by proteins that control the cell cycle. *Nature* 389: 149–152.
- Gottesfeld, J.M. and Forbes, D.J. (1997) Mitotic repression of the transcriptional machinery. *Trends Biochem. Sci.* 22: 197–202.
- Lane, H.A. and Nigg, E.A. (1997) Cell cycle control: POLO-like kinases join the outer circle. *Trends Cell Biol.* 7: 63–68.
- Lutkenhaus, J. (1997) Bacterial cytokinesis: Let the light shine in. *Curr. Biol.* 7: R573–R575.
- McCall, K. and Steller, H. (1997) Facing death in the fly: Genetic analysis of apoptosis in *Drosophila*. *Trends Genet.* 13: 222–226.
- Nasmyth, K. (1996) At the heart of the budding yeast cell cycle. *Trends Genet.* 12: 405–412.
- Noble, M.E.M., Endicott, J.A., Brown, N.R. and Johnson, L.N. (1997) The cyclin box fold: protein recognition in the cell cycle and transcriptional control. *Trends Biochem. Sci.* 22: 482–487.
- Pines, J. and Hunter, T. (1995) Cyclin-dependent kinases: An embarrassment of riches? In: *Cell Cycle Control: Frontiers in Molecular Biology* (eds C.M. Hutchinson and D.M. Glover), pp. 144–176. IRL Press, Oxford.
- Rothfield, L.I. and Garcia-Lara, J. (1996) Cell division. In: *Regulation of Gene Expression in Escherichia coli* (eds E.C.C. Lin and A.S. Lynch), pp. 547–570. R. G. Landes/Chapman & Hall, New York.
- Sherr, C.J. and Roberts, J.M. (1995) Inhibitors of mammalian cyclin dependent kinases. *Genes Devel.* 9: 1144–1163.
- Vaux, D.L. and Strasser, A. (1996) The molecular biology of apoptosis. *Proc. Natl Acad. Sci. USA.* 93: 2239–2244.
- Wake, R.G. and Errington, J. (1995) Chromosome partitioning in bacteria. *Annu. Rev. Genet.* 29: 41–67.

## Chapter 3

# Chromatin

### Fundamental concepts and definitions

- The DNA in eukaryotic chromosomes exists in a complex with an approximately equal mass of protein, forming a highly ordered nucleoprotein substance termed **chromatin**.
- The protein component of chromatin consists primarily of **histones** plus a small amount of other proteins, collectively termed **nonhistones**. The histones are relatively homogeneous and are responsible for the fundamental aspects of chromatin structure, whereas the nonhistone proteins are heterogeneous and perform many different functions. The DNA in each chromosome is a single, very long molecule (the **unineme model**).
- A eukaryotic nucleus is no more than 1  $\mu\text{m}$  in diameter, yet it contains up to  $10^9$  bp of DNA, which would measure several meters if extended. The DNA must therefore be highly folded, a property expressed as its **packaging ratio** — the ratio of the length of DNA in the nucleus to its theoretical free length. The packaging ratio in the interphase nucleus is approximately  $10^3$ .
- The packaging of DNA into chromatin is highly organized and is achieved through several hierarchical orders of structure. Chromatin is not merely a structural phenomenon, but has a great impact on gene activity. The organization of chromatin is used to regulate gene expression and is a basis of epigenetic cell memory.

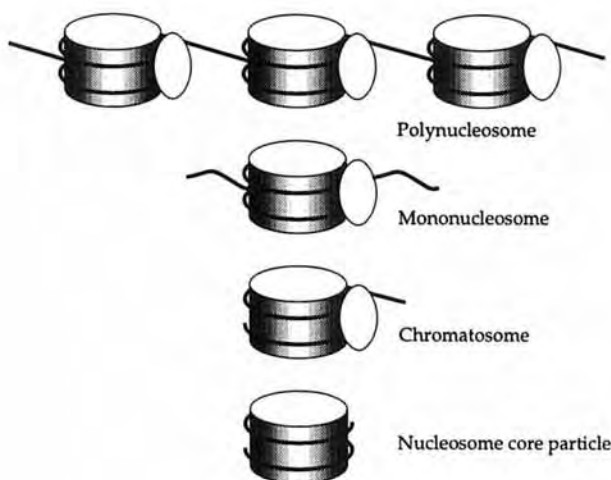
### 3.1 Nucleosomes

**Nucleosome structure.** Nucleosomes are the fundamental units of chromatin. They are found in all eukaryotes (with the exception of the dinoflagellates) and are conserved in structure. When nuclei are lysed in a low salt solution, the chromatin decondenses and nucleosomes can be observed resembling beads on a string.

The relationship between DNA and the nucleosome can be investigated using nucleases (*Figure 3.1*). Limiting digestion of decondensed chromatin separates the nucleosomes into short fragments consisting of single nucleosome units (**mononucleosomes**) and multimers thereof. The DNA associated with the mononucleosomes is of a constant length, usually in the order of 200 bp (the exact length is species- and cell-type-specific). Electrophoretic separation of DNA liberated from digested chromatin thus produces a characteristic ladder of fragments with a periodicity reflecting the length of the mononucleosomal DNA. This suggests that the nuclease cleaves at the same relative position between each nucleosome.

In most species, further nuclease digestion reduces the length of DNA isolated from the nucleosomes to 165 bp. This occurs in a single step, suggesting that the initial cleavage is followed by rapid trimming of the trailing DNA ends. The DNA that remains is protected by the proteins of the nucleosome particle, which at this stage is termed a **chromatosome**.

Further nuclease digestion reduces the DNA to 146 bp in length, and liberates some protein from the nucleosome. This length of DNA is conserved throughout the eukaryotes and is relatively resistant to further nuclease attack. It represents the DNA intimately associated with the proteins which comprise the nucleosome core. The particle thus formed, the **nucleosome core particle**, consists of a histone octamer containing two copies each of histones H2A, H2B, H3 and H4, and 146 bp of **core DNA** which is wrapped around this octamer twice. The core particles are joined together by **linker DNA**, whose length varies between species and between different cell types. In most species, as the



**Figure 3.1:** The nucleosome structure of chromatin. Liberated DNA appears as beads on a string, which are nucleosomes joined by linker DNA. Mild cleavage with micrococcal nuclease generates mononucleosomes; further digestion removes most linker DNA to generate single chromatosomes. Intense digestion removes DNA associated with histone H1, which dissociates leaving a nucleosome core particle containing 146 bp of DNA.

DNA enters and leaves the core particle it is associated with a single copy of a **linker histone** (histone H1) which is thought to seal the DNA in place.

The consequences of nuclease treatment can be summarized as follows: initial nuclease digestion cleaves the linker DNA at a specific point and then trims the ends until the remaining DNA is protected by histones. Further digestion removes the DNA associated with histone H1, whose loss suggests that it lies outside the core, but close to it. The remaining DNA is that protected by the nucleosome core itself. In yeast chromatin, there appears to be no linker histone. Initial cleavage generates mononucleosome DNA of just 164 bp, which is trimmed in a single step to the 146 bp of core DNA. The linker DNA between nucleosome core particles is thus 18 bp in length, the shortest linker DNA known. However, histone H1 genes have been identified in the yeast genome.

**Histones.** Histones are highly basic proteins (rich in positively charged lysine and arginine residues) which fold to form a compact core with a protruding N-terminal tail. The positive residues interact with DNA by forming salt bridges with the negatively charged DNA backbone, while the tails are targets for posttranslational modification and facilitate nucleosome–nucleosome interactions, and interactions with other chromatin proteins. Several hundred histone sequences have been determined and they show remarkable conservation across the eukaryote kingdom. Histones H3 and H4 are the most strongly conserved, whilst the linker histones show the most diversity. All the histones except H4 exist as multiple isoforms (**isohistones**) whose relative predominance in chromatin varies in a cell-type-specific manner. The linker histones show the most isotypic diversity and can be divided into several subclasses (e.g. H1, H5, H1<sup>0</sup>).

Histones possess an inherent ability to associate and form various complexes in the absence of DNA, notably the (H3–H4)<sub>2</sub> tetramer and the H2A–H2B dimer. Each histone contains a three- $\alpha$ -helix motif called a **histone fold** which facilitates these interactions. The tetramer can organize DNA into nucleosome-like particles *in vitro*, and if the other histones are added, chromatin will assemble spontaneously. This occurs slowly, however, and *molecular chaperones* (q.v.) are required to help assemble nucleosomes *in vivo* (see below). The structure of the histone octamer has been investigated using X-ray crystallography and cross-linking studies, and has recently been solved at a resolution of 2.8 Å.



H2A–H2B dimers fit above and below a central (H3–H4)<sub>2</sub> tetramer, and the cylindrical particle thus formed possesses an outer surface which describes the superhelical path taken by DNA. The path is not smooth, however, but distorted because of several bends. As DNA enters and leaves the octamer, it is bound by extensions of the H3 histone protein.

All histones appear to undergo some form of posttranslational modification, usually on the N-terminal tail, and in many cases, specific patterns of modification correlate to changes in chromatin function. The acetylation of histones H3 and H4 is a marker of accessible 'open' chromatin, although a number of recent experiments suggest the relationship between acetylation and genetic activity is not a simple one. Histones H3 and H1 are phosphorylated and, at least in the slime mould *Physarum*, the level of H1 phosphorylation varies in a cell cycle-dependent manner. Various histones are also methylated, or conjugated with ubiquitin or poly-ADP-ribose, although the significance of these modifications is unclear. Histone tails are long and unstructured, but ordered by nucleosome–nucleosome interactions. The modification of histones is therefore thought to disrupt interactions between nucleosomes and thus perturb higher order chromatin structure.

**Nonhistone proteins.** The histones represent the bulk of protein in chromatin and are relatively homogeneous in nature. The remaining chromatin proteins, collectively described as **nonhistone proteins**, represent a small but extremely heterogeneous fraction. The nonhistone proteins include enzymes involved in DNA and histone metabolism, *replication*, *recombination* and *transcriptional regulation* (q.v.). They also include the scaffold proteins which organize higher order chromatin structure (as discussed below), and the **high mobility group proteins (HMG proteins)**, which are highly charged proteins with various functions in gene regulation and structural organization. Of these, the HMG14/17 family of nucleosome-binding proteins are enriched in active chromatin and are thought to help decondense higher order chromatin structure. The HMGI/Y family proteins, whose precise role is not known, preferentially bind to repetitive AT-rich DNA, and like histone H1, they can be phosphorylated by cyclin-dependent kinases (*see The Cell Cycle*). The principal effect of HMG proteins in packaging and transcriptional activation is to introduce sharp bends into the DNA. In a packaging context, this may be required for DNA to adopt particular three-dimensional configurations, whereas in a transcriptional context it may bring regulatory factors at different sites into proximity (q.v. *SRY factor*, *enhanceosome*). A further class of nonhistone proteins, termed **protamines**, facilitate the packaging of DNA into the sperm head. These proteins align the major grooves of adjacent DNA duplexes and fold the DNA into a highly compact array of parallel fibers.

**Histone–DNA docking and the linking number paradox.** DNase I or free radical cleavage of nucleosome core particles generates DNA fragments with 10–11 bp periodicity, suggesting that the DNA wrapped round the histone octamer is in the B-conformation (*see Nucleic Acid Structure*). The pattern of cleavage varies across the surface of the nucleosome, with more frequent cleavage at the ends and less frequent cleavage in the middle, indicating that the structure of the DNA changes as it wraps around the octamer. The DNA is wrapped twice around the core, and the path it follows should create approximately 1.8 turns of negative supercoil. However, experiments designed to measure the degree of supercoiling generated by the sequestration of DNA into nucleosomes show that each nucleosome actually generates just one turn of negative supercoil. The discrepancy is termed the **linking number paradox** (q.v. *DNA topology*) and is explained by the decreased *pitch* (number of base pairs per turn) of nucleosomal DNA (10.2) compared with free DNA.

**Position of nucleosomes on DNA.** Nucleosomes do not form randomly on DNA, but occupy preferential sites. This property is termed **nucleosome phasing**, and can be demonstrated by treating chromatin with micrococcal nuclease and a restriction endonuclease. A discrete DNA fragment is obtained because the restriction site in the DNA is found at the same position relative to the nucleosome in each chromatin strand. Nucleosome phasing may occur in two ways, either by positioning the histone octamer at a particular sequence because its structure is favorable for winding, or by

positioning it relative to a particular boundary where nucleosomes are excluded. Both types of phasing have been observed. The tendency of particular DNA sequences to curve influences nucleosome positioning, as does the presence of other proteins in nucleosome-free regions, e.g. at promoters and enhancers. In yeast chromatin, nucleosome phasing is observed in transcriptionally repressed chromatin, whereas in active chromatin, the nucleosomes adopt random positions. This suggests that nucleosome phasing may be an initial requirement for higher order chromatin structure.

### 3.2 Higher order chromatin organization

**The 30 nm fiber.** The winding of DNA into nucleosomes represents only the first level of structural organization. When nuclei are lysed in a low salt solution, the characteristic beads on a string structure, termed the **10 nm fiber**, represents a packaging ratio of five and lacks histone H1. At a higher salt concentration, chromatin adopts a more compact structure which requires histone H1. This is probably a coiled fiber of 25–45 nm in diameter (it is termed the **30 nm fiber**), although alternative models suggest a zig-zagging organization, or that nucleosomes do not form a regular structure, but irregular clumps, or **superbeads**, which are arranged in linear fashion to form a fiber.

A number of different structures have been proposed for the coiled 30 nm fiber based on its density (6–8 nucleosomes for every nucleosome in the 10 nm fiber — a packing ratio of 40), and on X-ray diffraction evidence suggesting that the nucleosome discs are packed with their flat faces parallel to the helical axis, although with a variable degree of tilt. Most controversy surrounds the role of linker DNA and the position of histone H1, which may form polarized filaments perhaps by cooperative binding.

**Chromatin loops.** The gross organization of chromatin in the interphase nucleus is poorly understood, but it is thought that the 30 nm fiber is attached at various points to the nuclear matrix<sup>1</sup> to form a series of loops containing 30–100 kbp of DNA (the **folded fiber model**). The loops can be seen in scanning electron micrographs of protein-depleted chromatin, with their bases attached to scaffold proteins of the nuclear matrix (also q.v. *lampbrush chromosome*). The functional significance of the loops is not understood, but they may correspond to the *chromatin domains* discussed below, which have been identified by genetic and biochemical analysis.

**Euchromatin and heterochromatin.** Interphase chromatin exists in two distinct forms: diffuse **euchromatin**, which is believed to comprise looped 30 nm fibers mingling and tangling together in the nucleoplasm, as discussed in the preceding paragraph, and highly condensed **heterochromatin**, which is believed to adopt a higher order structure and tends to cluster around the nuclear periphery.

Heterochromatin is usually transcriptionally repressed, and is assumed to have adopted a higher order structure similar to mitotic chromatin which excludes transcription regulatory proteins. The nature of this structure is unknown although the DNA sequence is not critical, for as well as **constitutive heterochromatin** (which is condensed in all cells at all times and is often found near the centromere), eukaryotic cells contain **facultative heterochromatin**, which is maintained in a repressed form in some cells but not others (an example is the inactive X-chromosome). It is likely, therefore, that specific nuclear proteins are involved in the control of heterochromatin structure, and candidates have been identified in several organisms on the basis that they regulate the related

<sup>1</sup>The **nuclear matrix** (also called the **nuclear scaffold**) is a group of structures which remain in the nucleus when it has been extracted with detergents, nucleases and high concentrations of salt. It comprises a fibrous proteinaceous network which may well subdivide the nucleus into compartments and organize individual chromosomes. The composition of the matrix is poorly characterized, but it does contain topoisomerase II, which is also a component of the **central scaffold** of the metaphase chromosome, or **chromosome core**. There is tantalizing evidence for the involvement of chromatin-matrix interactions in the initiation of replication and transcription, but there is much to be learned about the role of the matrix in DNA function.



phenomenon of *position effect variegation* (see below). Heterochromatin also contains a high proportion of specifically modified histones and, in higher eukaryotes, correlates with increased levels of DNA methylation (q.v.).

**Metaphase chromosomes.** Chromatin is most condensed at mitotic metaphase when each chromosome is packaged as a tiny discrete structure to facilitate segregation. Again, little is known about the organization of chromatin in this highly compact form (packaging ratio  $10^5$ ). It is known that highly condensed fibers are arranged in loops upon a proteinaceous scaffold which forms a central helix from which the loops of chromatin radiate. It is not known whether these loops are equivalent to loops of chromatin in the interphase nucleus although, as discussed below, similar DNA motifs may be involved in their attachment. The scaffolds of each sister chromatid twist in opposite directions and appear to be joined together in the initial stages of mitosis.

A remarkable aspect of mitotic chromatin is that even in its highly condensed state it retains a memory of which genes were transcriptionally active and which repressed in the previous interphase. DNA is therefore not packaged uniformly in the mitotic chromosome, allowing the specific arrangement of open and closed chromatin domains to be reinstated when the chromatin decondenses at the following interphase. The differential packaging of mitotic chromatin is revealed by disruptive *chromosome banding* techniques (q.v.) which generate reproducible patterns of dark and light bands, corresponding to domains of genetic activity and inactivity determined through biochemical analysis.

### 3.3 Chromatin and chromosome function

**Chromatin and access to the information in DNA.** The organization of DNA into highly ordered structures (within which it is closely associated with proteins along its whole length) presents a problem for the proteins which mediate DNA function. In particular, both replication and transcription must occur in the context of nucleosomal organization, both involving large enzyme complexes which translocate processively along DNA and unwind it. Both DNA polymerase and RNA polymerase are substantially larger than a nucleosome. It is important to consider how these proteins have access to the information in DNA when it is organized into chromatin.

**Nucleosome structure during DNA replication.** The unique pattern of nuclease sensitivity within the replication fork indicates that the separation of parental strands during replication displaces histones from the DNA. The free histones are thought to reassociate immediately with the daughter duplexes, as histones displaced by replication *in vitro* rapidly assemble into nucleosomes on competitor naked DNA. This is confirmed by scanning electron micrographs of replicating SV40 DNA, which show nucleosome beading on the parental strand and both daughter strands immediately adjacent to the replication fork.

The replication of chromatin is dispersive rather than conservative. Histones displaced from the parental duplex demonstrate no preference for either daughter duplex, and appear to mix with newly synthesized histones which accumulate during the  $G_1$  stage of the *cell cycle* (q.v.). The precise assembly mechanism *in vivo* is not understood, but it is thought that a molecular chaperone called N1/N2 initiates nucleosome formation by loading the  $(H3-H4)_2$  tetramer onto DNA, whilst another chaperone, nucleoplasmin, facilitates the docking of H2A-H2B dimers. A challenging problem concerning chromatin function is how active and repressed chromatin domains are stably propagated through successive rounds of replication. This is likely to reflect the distribution of specifically modified parental histones onto both daughter duplexes immediately following the passage of the replication fork, preserving preexisting chromatin structure. However, replication would allow competition between nucleosomes and transcriptional complexes for occupation of strategic DNA sites, providing an opportunity for the state of commitment of the cell to be changed, especially if a new transcriptional regulator is synthesized prior to replication. The outcome of such competitions would

presumably reflect the affinity of regulatory complexes for DNA and their effective concentration in the cell (see below).

**Nucleosome structure during transcription.** Transcription, like replication, displaces nucleosomes from DNA, and reassembly appears to occur in the wake of the RNA polymerase. Most transcribed genes thus retain a nucleosome structure, although the pattern of nucleosome phasing characteristic of nontranscribed genes is lost, resulting in a smear of DNA fragments following digestion with micrococcal nuclease and a restriction enzyme, rather than a discrete band. Experiments which examine the progress of the polymerase complex through the nucleosome have shown that pausing occurs about half-way through the core DNA, which may reflect the build-up of torsional strain as the enzyme attempts to negotiate the first coil released from the nucleosome. The strain is released as the enzyme moves past this point, indicating that the octamer is expelled. The octamer then reassociates with DNA behind the enzyme, perhaps because it remains attached to the nontranscribed strand, or perhaps because it is transiently associated with the enzyme itself.

In heavily transcribed genes such as the rRNA genes of *lampbrush chromosomes* (q.v.), the extended conformation of chromatin indicates that it is nucleosome-free. This probably reflects failure of the displaced histones to reassemble on posttranscribed DNA because of a following transcriptional elongation complex. In very active genes, there would be a convoy of RNA polymerases which would maintain an indefinite nucleosome-free region of chromatin.

**Chromatin domains.** The structural properties of transcriptionally active or potentially active chromatin are distinct from those of inactive chromatin. Transcribed chromatin has a general increased sensitivity to DNase I digestion, which may reflect its less condensed packaging. This is described as **open chromatin**, and whilst it is true that transcription itself disrupts nucleosome structure, the extension of DNase I sensitivity for several kilobases either side of the actual transcription unit, and the maintenance of sensitivity in the absence of transcription suggests that this phenomenon involves organization at a higher level than the nucleosome.

The extent of DNase I sensitivity defines a conceptual **chromatin domain**, a region of chromatin whose activity is independent from that of other domains. Superimposed upon the **general DNase I sensitivity** are further **DNase I hypersensitive sites**. These are found flanking the transcription unit, and are preferentially cleaved at low DNase I concentrations. Such sites are usually about 200 bp in length and often correspond to *cis*-acting regulatory elements (q.v. *enhancer*, *locus control region*); they are believed to represent nucleosome-free regions where transcription regulatory complexes bind to DNA. This has been confirmed in the case of the SV40 genome, where molecular analysis of DNase I hypersensitive sites has shown correspondence to regulatory elements, and topographical analysis by scanning electron microscopy has identified nucleosome-free regions. The molecular basis of open chromatin structure is not fully understood, although there are some interesting correlations. Open chromatin contains a generally higher proportion of N-terminal acetylated core histones and HMG14/21 proteins than bulk chromatin, and is relatively depleted for linker histones. Since linker histones are required for chromatin to adopt the 30 nm fiber, and since histone N-terminal tails facilitate nucleosome–nucleosome interactions, it is possible that the transition from repressed chromatin to open chromatin involves the decondensation of the 30 nm fiber to a simple 10 nm fiber organization, stabilized by nonhistone proteins. Histone acetyltransferases and deacetylases have been shown to be recruited by some transcription factors, providing a mechanism for **chromatin remodeling** as an initial step in transcriptional activation and repression. In mammals and plants, **repressed chromatin** is often associated with high levels of DNA methylation, which may also play a major role in the epigenetic regulation of gene expression (see DNA Methylation). However, DNA methylation is absent from many lower eukaryotes, including, for example, yeast and *Drosophila*.

The division of eukaryotic genomes into functionally discrete domains is particularly conspicuous in mammals because the chromatin structure is reflected in the topography of the chromosomes at the M phase (see Chromosome Structure and Function). A variety of disruptive staining procedures (notably *G-banding* and *R-banding*, q.v.) reveal a reproducible pattern of dark and light transverse bands which are thought to correspond to areas distinguished generally by the density of chromatin structure. The G-bands correspond to chromosome bands revealed by other methods, such as transient replication banding (which can specifically label early and late replicating DNA) and D-banding, which identifies regions of DNase I sensitivity. Molecular analysis shows that the DNA within these regions differs with respect to sequence content and architecture, providing evidence for a high level *biphasic organization* (q.v.) of the mammalian genome (see Genomes and Mapping).

**Domains and boundary functions.** Chromatin is physically divided into discrete and topologically isolated regions. This is demonstrated by the loop and scaffold appearance of protein-depleted chromatin, the lateral loops of *lampbrush chromosomes* (q.v.) and the reproducible banding pattern of *Drosophila polytene chromosomes* (q.v.). Chromatin loops appear to be tethered to the nuclear matrix at their bases, and it is likely that this involves specific nucleoprotein complexes. The significance of the chromatin loops is unclear, but it is possible that they represent the functionally independent domains discussed above. This would allow adjacent domains to adopt different orders of chromatin packaging, and would provide a mechanism for **enhancer monogamy**, where the activity of a distant enhancer is confined to its target gene.

If loops are equivalent to domains, it can be predicted that specific DNA sequences would be associated with the nuclear matrix, that these sequences should map near the borders of biochemically defined chromatin domains, and that they should be able to isolate genes from the effects of adjacent domains, i.e. they should act as **insulators** or **boundary elements**.

Putative **matrix-associated regions (MARs)** (also known as **scaffold attachment regions (SARs)**) have been identified in two complementary approaches by their ability to bind to nuclear matrix proteins. Fragmented DNA exposed to matrix components can trap putative MARs in the insoluble fraction, and protein-depleted chromatin digested with nucleases should leave only MARs protected from nuclease activity. These procedures have identified a number of AT-rich elements with no strong sequence conservation except a recognition site for topoisomerase II, a component of both the nuclear matrix and the metaphase scaffold. It is possible that topoisomerase II could divide chromatin into topologically isolated domains.

Boundary elements have also been identified by their cytological and biochemical properties, i.e. they map at the boundaries of known chromatin domains and contain nuclease-hypersensitive sites. Such elements include the *Drosophila* special chromatin structures (scs and scs') and the chicken and human  $\beta$ -globin HS5 element from the *locus control region* (q.v.). A genuine boundary element would be able to establish an independent chromatin domain in an autonomous fashion, and this is thought to be the role of the  $\beta$ -globin locus control region (LCR) *in vivo*. This function can be tested in two ways: by assaying for protection of a transgene against endogenous position effects using flanking boundary elements, and by testing for the insulation of a gene from the effects of an adjacent enhancer by the interposition of a boundary element. This has been achieved in the case of the *Drosophila* scs-like elements, the chicken  $\beta$ -globin LCR 5HS site and the chicken lysozyme gene A element. There is a degree of functional interchangeability between the elements. The  $\beta$ -globin LCR 5HS site, for instance, functions in *Drosophila*, but the scs-like elements do not function in transgenic mice. The chicken lysozyme A elements suppress position effects both in transgenic mice and in transgenic plants (also q.v. *gypsy transposon*).

Although the results of experiments designed to test the function of the MARs isolated by physical methods are inconclusive, the presence of MAR-type elements in some boundary elements identified functionally suggests that physical division of chromatin into loops could be the basis of the functional division of chromatin domains.



**Heterochromatin structure and epigenetic gene regulation.** The translocation of euchromatic DNA into heterochromatin often results in the spread of transcriptional repression into the euchromatic region. The extent of heterochromatinization varies from cell to cell, but is clonally propagated so that genes near the breakpoint are expressed in a variegated manner. This phenomenon, which is termed **position effect variegation (PEV)**, suggests that a *cis*-acting silencing activity can spread across the chromosome breakpoint. The spreading effect is linear and uninterrupted; gene loci nearest the breakpoint are inactivated most often and the repression never skips over genes. This suggests that heterochromatin normally spreads by proteins adding on to preexisting heterochromatin and extending it until some boundary is reached. Translocation presumably removes the boundary, and the variable spreading which causes PEV probably reflects variable amounts of heterochromatin-sponsoring proteins in different cells.

It is likely that heterochromatin and euchromatin proteins are found in equilibrium between chromatin and the nucleoplasm, and that mutations encoding genes with chromatin functions would alter the balance of these proteins and result in a shift in the extent of PEV following a translocation event (these are termed **antipodal effects**). Genetically, such mutations would behave as either suppressors or enhancers of PEV, and would identify the genes involved in higher-order chromatin structure.

In *Drosophila*, a search for such PEV modifiers has identified several cellular components with a role in heterochromatin formation. Heterochromatin itself is one of the best PEV modifiers; by increasing the dosage of the Y-chromosome (which is mostly heterochromatin), PEV can be suppressed, probably resulting from the depletion of heterochromatin proteins in the nucleoplasm. Histones have also been identified as an important component: heterochromatin is generally poor in acetylated H4 but rich in a *particular* acetylated form. Other modifiers correspond to specific proteins which play a direct role in heterochromatin structure. These include the general protein HP1 and some regulators of gene expression (e.g. Polycomb; q.v. *homeotic genes, maintenance of differentiation*), which mediate their effects at the level of chromatin structure. Such proteins often possess a conserved structure, termed a **chromodomain**.

In yeast, heterochromatin proteins have been identified on the basis that mutants de-repress the silent mating-type loci, which are usually sequestered in repressed chromatin near the telomeres (q.v. *mating type switching*). As well as histones H3 and H4, the silent information regulators SIR3 and SIR4 and a further protein termed RAP1 have been identified, the spreading of heterochromatin apparently reflecting interactions between the histones and the SIR products, facilitated by RAP1 (SIR3, for instance, interacts with the N-terminal tail of histone H4).

### 3.4 Molecular structure of the bacterial nucleoid

**Organization of bacterial chromosomes.** Bacterial cells possess a single chromosome which, like eukaryotic chromosomes, exists as a nucleoprotein complex. This **nucleoid** does not condense prior to cell division and thus does not display the compact structural features of eukaryotic chromosomes. However, it is organized into a series of looped domains which reflect functional as well as structural order.

In *E. coli*, the looped domains are approximately 50 kbp in length and there are approximately 100 in the whole genome. Each domain appears to be topologically isolated, which may facilitate individual topological control over different promoters. The molecular basis of this organization is almost totally uncharacterized, i.e. it is unknown whether the loops define specific regions of the chromosome or form randomly, it is unknown what sequences of the chromosome are involved, and it is unknown how the loops are formed and maintained. Several protein components of the nucleoid have been determined (e.g. HU, H1), although these have been identified as mutations in genes affecting a variety of other systems, not as mutations affecting nucleoid function. This suggests that nucleoid proteins may be functionally redundant and may influence other systems

indirectly, e.g. by modulating the topological status of DNA, which is an important factor for transcriptional activity (see Transcription, Nucleic Acid Structure).

## References

- Elgin, S.C.R. (Ed.) (1995) *Chromatin Structure and Gene Expression*. IRL Press, Oxford.
- Van Holde, K.E. (1989) *Chromatin*. Springer, New York.
- Wolffe, A.P. (1992) *Chromatin: Structure and Function*. Academic Press, London.
- ## Further reading
- Geyer, P.K. (1997) The role of insulator elements in defining domains of gene expression. *Curr. Opin. Genet. Dev.* 7: 242–248.
- Hartzog, G.A. and Winston, F. (1997) Nucleosomes and transcription: Recent lessons from genetics. *Curr. Opin. Genet. Dev.* 7: 192–198.
- Loo, S. and Rine, J. (1995) Silencing and heritable domains of gene expression. *Annu. Rev. Cell Biol.* 11: 519–548.
- Nash, H.A. (1996) The HU and IHF proteins: Accessory factors for complex protein–DNA assemblies. In: *Regulation of Gene Expression in Escherichia coli* (eds E.C.C. Lin and A.S. Lynch), pp. 149–180. R.G. Landes/Chapman and Hall, New York.
- Pirrotta, V. (1997) Chromatin-silencing mechanisms in *Drosophila* maintain patterns of gene expression. *Trends Genet.* 13: 314–318.
- Sherman, J.M. and Pillus, L. (1997) An uncertain silence. *Trends Genet.* 13: 308–313.
- Tsukiyama, T. and Wu, C. (1997) Chromatin remodeling and transcription. *Curr. Opin. Genet. Dev.* 7: 182–191.
- Wade, P.A., Pruss, D. and Wolffe, A.P. (1997) Histone acetylation: Chromatin in action. *Trends Biochem. Sci.* 22: 128–132.
- Widom, J. (1997) Chromatin: The nucleosome unwrapped. *Curr. Biol.* 7: R653–655.
- Wolffe, A.P. and Pruss, D. (1996) Deviant nucleosomes — the functional specialization of chromatin. *Trends Genet.* 12: 58–62.

**This Page Intentionally Left Blank**



## Chapter 4

# Chromosome Mutation

### Fundamental concepts and definitions

- A **chromosome mutation** (or **chromosome aberration**) is a mutation involving a large segment of the genome. Such a mutation usually affects many genes and is often observable at the cytogenetic level, i.e. it can be seen with a light microscope.
- Chromosome mutations are **numerical** if they involve a deviation from normal chromosome number, and **structural** if they involve breakage and rearrangement of chromosome segments. Numerical mutations often result from chromosome segregation errors caused by structural mutations, but may also reflect aberrant replication or errors at fertilization. Structural mutations result from the faulty repair of broken chromosomes or from nonallelic recombination events. In mammals and in *Drosophila*, subtle structural mutations can be detected because they disrupt *chromosome banding patterns* (q.v.). Structural mutations are **balanced** if DNA is rearranged but there is no loss or gain of material, or **unbalanced** if DNA is lost or gained. All numerical mutations are unbalanced.
- There are four consequences of chromosome mutations: disruption, fusion, position and dosage effects. **Disruption and fusion effects** occur in structural mutations and reflect the nature of the chromosome breakpoints before and after mutation (e.g. a breakpoint can interrupt a gene or separate a gene from its promoter, resulting in loss of gene function, and fusion can join two genes, allowing a composite product to be synthesized, perhaps with novel functions). **Position effects** also occur in structural mutations and reflect global influences on gene expression conferred by chromatin structure (e.g. a translocation may bring a normally active gene adjacent to a region of heterochromatin causing transcriptional repression; q.v. *chromatin domain, position-effect variegation*). **Dosage effects** occur in both structural and numerical mutations and concern the number of copies of each gene in the cell. The dosage levels of many autosomal genes are flexible, but, others demonstrate *haploinsufficiency* (q.v.), which is one type of dosage effect. Unbalanced chromosome mutations, especially full aneuploidies, alter the dosage of many genes at the same time, and therefore generate multiple dosage effects. A change in dosage of many contiguous genes, especially an entire chromosome, is termed **chromosome imbalance**.
- **Constitutional chromosome mutations** arise in the germline or in the zygote and thus affect every cell in the body. In humans, these give rise to specific clinical syndromes (e.g. Down's syndrome, Klinefelter's syndrome). Constitutional numerical mutations often arise due to errors occurring during meiosis (e.g. homologous chromosomes failing to pair) or fertilization (e.g. dispermy — the fertilization of an egg by two sperm). **Somatic chromosome mutations** affect individual cells and clones derived from them; somatic numerical mutations often reflect errors that occur during mitosis. Somatic mutations arising early in development may generate a mosaic organism of different cell lines.
- The number and nature of chromosomes in the normal genome is expressed as the *karyotype* (q.v.). Specific aberrations are identified by adding information to the conventional karyotype designation. The nomenclature is summarized in *Table 4.1*.

### 4.1 Numerical chromosome mutations

**Numerical mutations involving full chromosome sets.** The number of full chromosome sets in a given eukaryotic cell is defined as its **ploidy** (*Table 4.2*). Many multicellular eukaryotes have a diploid vegetative state — which includes the somatic cells — as this represents the minimal ploidy

**Table 4.1:** Cytogenetic nomenclature for chromosome aberrations, with examples from human karyotypes

Karyotype symbol	Meaning	Example of nomenclature
<i>Symbols to identify chromosomes and chromosome arms</i>		
1–22	Autosomes (human genome)	
X, Y	Sex-chromosomes	
p, q	Short and long arms, respectively	
pter, qter	Terminal portions of short and long arms	
<i>Symbols to identify mutations</i>		
+	Extra chromosome or part of chromosome <sup>a</sup>	47, XY, +21
–	Missing chromosome or part of chromosome <sup>a</sup>	46, XX, 17q–
del	Deletion: Terminal Interstitial	46, XX, del (12) (p13.3→pter) 46, XX, del (12) (p12.1→12.3)
der	<i>Chromosome derivative</i> (q.v.)	der (12)
dup	Duplication	46, XY, dup (2) (q21.1→q23.1)
fra	Fragile site	46, XY, fra (Xq27)
h	Extra heterochromatin	46, XX, 16qh+
i	Isochromosome	46, X, i(Xq)
inv	Inversion: Pericentric Paracentric	46, XY, inv (2) (p12.1q23.2) 46, XY, inv (2) (p11.1p12.3)
marker	Unidentified chromosome	47, XX, + marker
r	Ring chromosome	46, XY, r(19) (p13q13)
s	Extra satellite material	46, XX, 21ps+
t	Translocation: Reciprocal Robertsonian	46, XX, t(7;19) (q34.2; p13.1) 45, XY, –14, –21, t(14q;21q)

<sup>a</sup>The + and – symbols indicate an extra/missing chromosome when placed in front of a chromosome designation (e.g. +21) or unspecified extra/missing part of a chromosome when placed after the chromosome designation (e.g. 17q–).

required for sexual reproduction. In mammals and birds, constitutional changes away from diploidy are poorly tolerated, leading to grossly abnormal offspring. This reflects the dosage balance of autosomes and active sex-chromosomes, and of imprinted genes (q.v. *dosage compensation, parental imprinting*).

Conversely, amongst plants both monoploidy and polyploidy are well tolerated, and polyploidy is common in nature. Significantly, polyploidy tends to increase the size of plants and is thus a desirable property exploited by plant breeders. This results from an increase in cell size rather than cell number, indicating that DNA content influences the cell size *checkpoint* (q.v.) of the cell cycle. There are also examples of naturally occurring haploidy or polyploidy in lower animals (e.g. some insects are haploid, some leeches are polyploid), and abnormal states of ploidy are better tolerated in these animals than in vertebrates. Polyploid animals such as leeches also have larger cells than diploids, but do not increase in size overall because cell number is reduced, reflecting a regulative mechanism during development (*see* Development: Molecular Aspects).

Although individuals with abnormal ploidy may be viable, they may also be sterile. Generally, any species which is **anisoploid** (has an odd number of chromosome sets) will be sterile because of meiotic failure. In triploids, for instance, three homologous chromosomes attempt to pair and segregate at meiosis, generating grossly unbalanced products. This can be exploited in commercially important fruit crops to produce seedless varieties. Monoploids cannot undergo meiosis and must reproduce asexually. This occurs in haploid insects which reproduce by parthenogenesis, and in plants whose life cycle involves alternation of haploid/diploid generations. Monoploid varieties of

**Table 4.2:** Some terms used to describe states of ploidy

Terms	Definitions
Ploidy	The number of full chromosome sets in a cell
Diploid	Two chromosome sets
Monoploid	One chromosome set
Haploid	Strictly, half the normal number of chromosome sets found in meiotic cells (e.g. applies to gametes), but often used as synonymous with monoploid because most gametes have one set of chromosomes (q.v. <i>haploid number</i> , <i>monoploid number</i> )
Polyploid	More than two chromosome sets: triploid = 3; tetraploid = 4; hexaploid = 6, etc. Over 10 copies of the genome are represented by numbers, e.g. 12-ploid
Polyploidy and polyteny	Polyploidy refers to the possession of more than two sets of homologous chromosomes. Polyteny refers to the possession of many (identical) chromatids per chromosome, as seen in <i>Drosophila</i> salivary glands
Isoploid and anisoploid	Even and odd numbers of chromosome sets, respectively. Anisoploid individuals are generally sterile because of unbalanced segregation at meiosis
Autopolyploid and allopolyploidy	Autopolyploidy is polyploidy arising from intrinsic genome duplication, whereas allopolyploidy results from genome duplication of a species hybrid
Homoploid and heteroploid	Possessing a chromosome number typical or atypical of the species, respectively

plants are particularly useful because, like bacteria and yeast, cultured cells can be subjected to mutagenesis screens and selected on the basis of mutant phenotypes (e.g. herbicide tolerance) which become manifest immediately. Cells with desirable traits can then be rendered fertile by **diploidization** (a doubling of the number of chromosomes induced by artificially blocking mitosis), and regenerated into whole plants by hormone treatment (*see* Mutation and Selection).

The forms of polyploidy discussed above are **autopolyploidy** because they involve the doubling of the *endogenous* genome. **Allopolyploidy** is a different form of polyploidy, where the individual possesses multiple copies of chromosomes from different species. This strategy is used to combine the desirable features of two plant species where simple crossing has failed. The hybrid sterility resulting from interspecific crosses reflects meiotic failure, i.e. the chromosomes from the different species do not form compatible pairs. If the hybrid genome is induced to double, however, each chromosome will have a homologous partner and the hybrid species may be fertile. Species-specific pairing between chromosomes in an allopolyploid is termed **autosynapsis**, whereas cross-species pairing is termed **allosynapsis**. Partially homologous chromosomes derived from different species are described as **homeologous chromosomes** (c.f. *homologous chromosomes*).

**Numerical aberrations involving individual chromosomes.** The cells of a normal, **euploid** individual contain full chromosome sets, whereas an individual with missing or extra chromosomes is described as **aneuploid** (Table 4.3). Aneuploidy is deleterious in both plants and animals, although the effects in animals are more severe. The loss or gain of chromosomes causes **chromosome imbalance**, where there is a net loss or gain of many genes with consequent multiple dosage effects. Such effects are usually lethal in humans, with the result that out of 22 possible autosomal trisomies, only three are normally found in live births, and only one is viable in the long term (Table 4.4). Sex-chromosome aneuploidy is better tolerated because of the *dosage compensation* mechanisms (q.v.) which act to redress natural variations in sex-chromosome number. Thus monosomy and polysomy for the X-chromosome generate relatively mild phenotypes due to *X-chromosome inactivation* (q.v.). The abnormalities which are seen in Turner's syndrome and XXX syndrome (Table 4.4) probably reflect

**Table 4.3:** Some terms used to describe the number of individual chromosomes in a karyotype

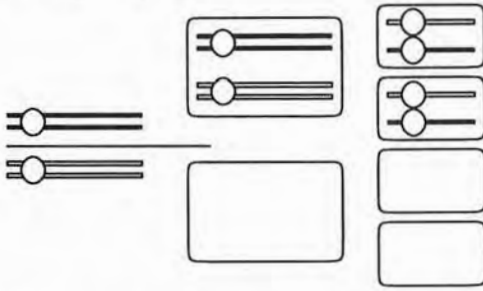
Terms	Definitions
Euploidy, aneuploidy	Possessing full chromosome sets, having extra or missing chromosomes
Chromosome imbalance	The loss or gain of chromosomes with consequent multiple gene dosage effects
Hyperploidy, hypoploidy	Possessing more chromosomes than usual, possessing fewer chromosomes than usual
Pseudodiploidy	Possessing the correct diploid chromosome number but an abnormal karyotype due to simultaneous monosomy and trisomy for different chromosomes
Eusomy, aneusomy	Possessing the correct number of copies of a given chromosome, possessing the incorrect number of copies of a given chromosome: nullisomy = no copies; monosomy = one copy; disomy = two copies; polysomy = more than two copies (trisomy = 3, tetrasomy = 4, etc.)

**Table 4.4:** Aneuploidies commonly seen in human live births

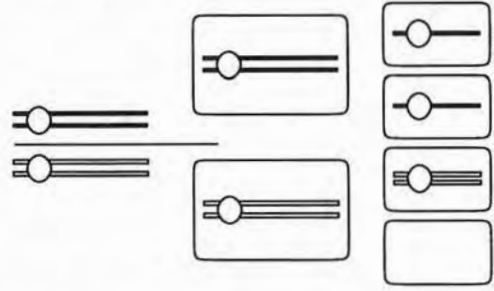
Aneuploidy	Occurrence and clinical phenotype
<i>Autosomal aneuploidy</i>	
Trisomy 21 (Down's syndrome)	Occurs 1:700 live births, clinical phenotype includes mental and growth retardation, characteristic facial features (upward slanting eyes, small nose, open mouth with protruding tongue, low-set ears), partial syndactyly and a simian crease. Recurrence risk is normally low although it increases with maternal age. Recurrence risk is high if trisomy caused by translocation (q.v. <i>translocation Down's syndrome</i> ). Trisomy 21 individuals have a lower life expectancy than normal, and many die in childhood
Trisomy 18 (Edward's syndrome)	Occurs 1:3–6000 live births, clinical phenotype includes elongated head with low-set, dysplastic ears, curved feet, clenched, overlapping fingers and cardiovascular and renal disorders. Perinatal death
Trisomy 13, occasionally 14 or 15 (Patau's syndrome)	Occurs 1:5–10000 live births, clinical phenotype includes gross head and brain malformations, reduced or missing eyes, polydactyly, low set and dysplastic ears, cleft lip and palate, fused nostrils and cardiovascular disorders. Perinatal death
<i>Sex-chromosome aneuploidy</i>	
Monosomy X (Turner's syndrome)	Occurs 1:10000 live female births, clinical phenotype includes short stature, webbed neck, wide carrying angle (cubitus valgus), widely spaced nipples, primary amenorrhoea and infertility
Trisomy X (XXX syndrome)	Occurs 1:1000 live female births, clinical phenotype may include failure of the development of primary and secondary sexual characteristics and mild mental retardation, but many individuals are normal and fertile
Polysomy X with Y (Klinefelter syndrome)	Primarily associated with karyotype 47, XXY, although more bizarre aneuploidies involving multiple X and Y chromosomes have been observed. Occurs 1:1000 live male births, clinical phenotype may include infertility, gynaecomastia (breast development in males), testicular atrophy, malformations of the penis, nonfunctional sperm, tall stature and mild mental retardation. The phenotype tends to increase in severity with the number of X-chromosomes
Disomy Y (‘Supermale’ syndrome)	Primarily associated with karyotype 47, XYY although rare cases of 48, XXYY, have been observed. Occurs 1:1000 live male births. Clinical phenotype includes tall stature. Karyotype tenuously associated with increased tendency to antisocial and violent behavior, although this remains to be shown conclusively. Individuals are usually fertile



1) Primary nondisjunction



2) Secondary nondisjunction



**Figure 4.1:** Aneuploidy in the gametes caused by nondisjunction. (1) Homologous chromosome pairs fail to segregate at first meiotic division (primary nondisjunction). (2) Chromatids fail to segregate at second meiotic division (secondary nondisjunction).

the dosage of genes which escape inactivation and effects occurring early in development before X-inactivation takes place. Disomy for the Y-chromosome has little phenotypic impact, and the behavioral abnormalities which have been associated with this karyotype are controversial.

**Causes of numerical chromosome aberrations.** Changes in ploidy can arise from cell-cycle failure, either by omission of DNA replication/unscheduled chromosome segregation (to halve the ploidy) or omission of chromosome segregation/unscheduled DNA replication (to double the ploidy). Changes in ploidy can also arise at fertilization, e.g. in mammals, triploidy can result from either diploidy of the sperm or ovum, or more commonly due to dispermy (fertilization by two sperm). Such aberrant fertilization events produce developmental abnormalities termed **hydatidiform** or **vesicular moles** (q.v. *parental imprinting*).

Aneuploidies reflect more localized disturbances to the cell cycle and arise in several ways, the most common being **nonconjunction** (where homologous chromosome pairs fail to find each other during meiotic prophase; q.v. *meiosis*) and **nondisjunction** (where chromosomes pair, but segregation fails). **Primary nondisjunction** occurs in the first meiotic division and results in both homologues of a given pair migrating to the same pole. **Secondary nondisjunction** occurs in the second meiotic division and results in both chromatids migrating to the same pole (Figure 4.1). This is often caused by failure of the centromere to divide. An analogous situation may occur during mitosis (**mitotic nondisjunction**), resulting in a somatic aneuploidy. Partial nondisjunction may also occur in the first meiotic division if the centromere divides prematurely, which releases one chromatid of a pair to segregate with the two chromatids of the homologous chromosome (to which it is attached by chiasmata). Several structural mutations increase the chances of nondisjunction (q.v. *ring chromosome*, *attached chromosome*, *Robertsonian translocation*).

Other ways in which aneuploidies arise include **chromosome gain** (the result of unscheduled replication of a single chromosome) and **chromosome loss**, which occurs if a chromosome segregates too slowly to be included in the daughter nuclei (**anaphase lag**), or fails to attach to the spindle in the first place.

## 4.2 Structural chromosome mutations

**Breakpoints.** Structural chromosome mutations are caused by the faulty repair of chromosome breaks or by recombination between homologous but nonallelic sites. The point at which fracture or recombination occurs is termed a **chromosome breakpoint** and structurally rearranged chromosomes are termed **chromosome derivatives**. If derivatives contain parts of different chromosomes, they are

named according to the origin of the centromere. Thus if human chromosomes 12 and 17 are broken and segments exchanged, the chromosome containing the centromere from chromosome 12 is termed the chromosome 12 derivative. Structural mutations may be balanced or unbalanced, as discussed below. Both balanced and unbalanced mutants may be characterized genetically by their altered linkage maps and, in heterozygotes, cytologically by the presence of unusual pairing configurations between homologous chromosomes at meiosis. Animals and plants differ in their tolerance of structural mutations. Animal gametes tend to be functional even if severely unbalanced and will take part in fertilization normally, generating an unbalanced zygote. While the same is true of plant ovules, pollen grains containing chromosome structural aberrations tend to abort, and fertilization does not take place.

**Unbalanced structural mutations.** Unbalanced mutations involve the loss or gain of genetic material and behave as partial monosomies or trisomies — i.e. they show the multiple dosage effects of losing or gaining a set of contiguous genes. Structural mutations involving DNA loss are termed **deletions** or **deficiencies**, whereas those involving DNA gain are termed **insertions** or **duplications**; they are considered as chromosome mutations if they involve visible amounts of genetic material. Large deletions can be recognized in heterozygotes by the presence of a **deletion loop**, a structure which extrudes from the paired chromosomes at meiosis because it has no homologous segment with which to pair. Insertions may behave in a similar fashion, but tandem duplications can generate more complex structures: the duplicated region may be extruded as a loop, or the two copies of the duplicated region may pair with each other within the chromosome causing the homologous region of the normal chromosome to be looped out (also q.v. *gene amplification*).

Even the smallest visible deletions and duplications in human chromosomes affect several megabases of DNA and therefore span many gene loci. They often cause **contiguous gene syndromes**, i.e. disease phenotypes resulting from the simultaneous loss or duplication of many linked genes (Table 4.5). The mutations are usually not precisely defined because the exact boundaries cannot be determined at the cytogenetic level. Contiguous gene syndromes may be characterized by a variable phenotype comprising a set of overlapping clinical symptoms. The correspondence of individual symptoms in different patients may allow a gene map to be constructed, termed a **morbid map**. Where similar phenotypes arise without visible mutation, deletion or duplication

**Table 4.5:** Human syndromes associated with unbalanced structural chromosome aberrations (deletions and duplications)

Syndrome	Karyotype and clinical symptoms
<i>Cri du chat</i> syndrome	del (5) (p14-pter). Small head with associated mental retardation; characteristic cat-like cry
Beckwith–Wiedemann	dup (11p15). Fetal overgrowth leading to gigantism, enlarged tongue and a predisposition to childhood tumors
WAGR syndrome	del (11p13). Name is acronym for principle symptoms: Wilm’s tumor, aniridia, ambiguous genitalia, and mental retardation
Cat eye syndrome	+ iso (22) (pter→q11). Eye defects, anal atresia, cardiovascular defects, mental retardation
Prader–Willi syndrome	del (15) (q11–q13) of paternal chromosome. Small stature and obesity, small genitalia, behavioral difficulties, mental retardation
Angelman syndrome	del (15) (q11–q13) of maternal chromosome. Mental retardation, ataxia, inappropriate laughter

Some, such as WAGR, are true contiguous gene syndromes where overlapping clinical symptoms reflect the deletion of groups of contiguous genes. Others, such as Beckwith–Wiedemann syndrome, may be associated primarily with a single gene defect (in this case, overactivity of the *IGF2* gene; q.v. *parental imprinting*).



below the resolution of light microscopy is suspected (**microdeletions** and **microduplications**), or the source of the phenotype may be a *point mutation* (q.v.).

A simultaneous whole-arm deletion and duplication occurs when division of the centromere in the wrong plane generates chromosomes with two long arms and no short arms and *vice versa*. These perfectly metacentric chromosomes are termed **isochromosomes** and cause monosomy for one arm and trisomy for the other, with the corresponding gene dosage effects.

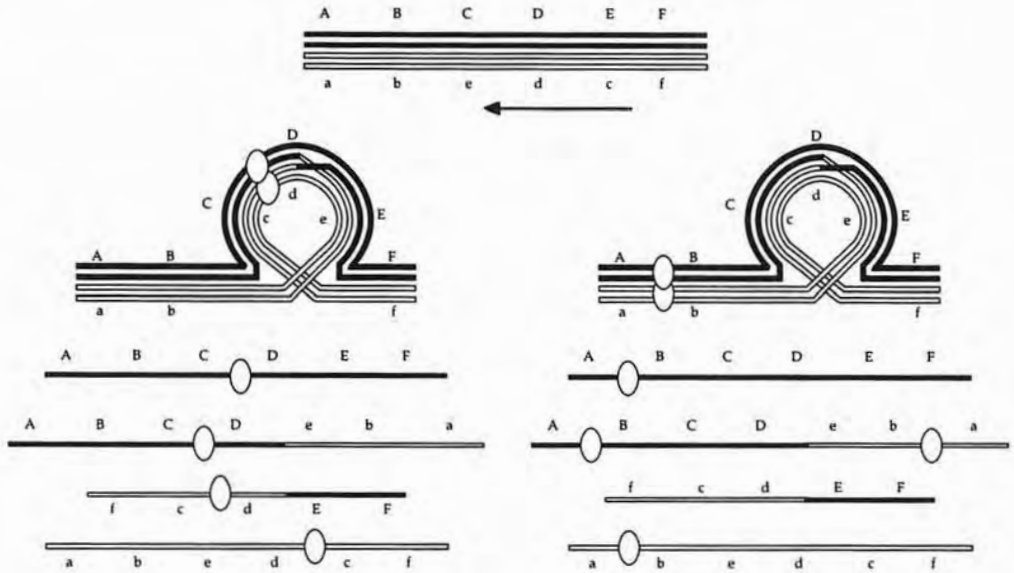
**Balanced structural mutations.** Balanced mutations are otherwise known as **chromosome rearrangements**, and involve the rearrangement of DNA but no net loss or gain of genetic material. Individuals carrying such rearrangements are often phenotypically normal, but heterozygotes may be semisterile because (a) the rearranged chromosomes behave anomalously at meiosis, often segregating to yield unbalanced gametes, and (b) crossing over within the rearrangement generates unbalanced recombinant products which are nonviable. Recombination is suppressed at the breakpoints of such rearrangements due to steric hindrance (q.v. *recombination coldspots*). These principles are exploited to suppress crossing over in *Drosophila*, allowing particular chromosomes to be transmitted without recombination, for the creation of specific strains of flies. Chromosomes containing complex multiple inversions are able to suppress crossing over so that each chromosome is inherited as an intact unit. Such recombination suppressors are termed **balancer chromosomes**.

**Balanced structural mutations of single chromosomes.** **Inversions** are chromosome rearrangements which involve the rotation of a segment of DNA within a single chromosome. They can be caused by the misrepair of chromosome breaks or by intramolecular recombination between inverted repeats. Inversions have no phenotype unless (a) the breakpoints interrupt a gene, or (b) the rearrangement results in unfavorable *position effects* (q.v.), but invertants can be identified genetically by their altered linkage map, or in heterozygotes for large inversions, by the formation of meiotic **inversion loops**, where one chromosome twists on its axis to incorporate the reversed sequence of loci on its partner.

Because inversions involve only a single chromosome, they do not affect meiotic segregation *per se* (c.f. *translocation*). They do suppress the recovery of recombinants, however, because of the generation of severely unbalanced cross-over products. The nature of cross-over products depends upon whether or not the inversion spans the centromere (*Figure 4.2*). If the centromere is included in the inversion (**pericentric inversion**), recombinant products containing deletions and duplications are generated. However, if the centromere lies outside the inversion (**paracentric inversion**) a potentially more serious problem arises where recombination joins two centromeres together to generate a dicentric chromosome and an acentric fragment. The latter is usually lost because it has no means of attaching to the spindle apparatus. The dicentric fragment is stretched between two nuclei and is either excluded due to anaphase lag, or fragmented randomly generating a broken end which may initiate a *breakage-fusion-bridge* cycle (q.v.).

A second type of intramolecular structural aberration occurs when a chromosome which has lost telomeres from both ends fuses the broken ends together. Such **ring chromosomes** concatenate following replication, and if they cannot be separated, nondisjunction is inevitable.

**Balanced structural mutations of multiple chromosomes.** **Translocations** are structural aberrations in which part of one chromosome is transferred to another. Translocations can be reciprocal or non-reciprocal. A **reciprocal translocation** occurs when two chromosomes are broken and the distal fragments are swapped and rejoined. **Nonreciprocal translocations** involve the one-way transfer of material. An **internal translocation** occurs when a chromosome segment is inserted into an interstitial site within another chromosome. Similarly, a **terminal translocation** involves material added onto the end of another chromosome. A **jumping translocation** is a special kind of terminal translocation where the same chromosome fragment jumps sequentially to different chromosomes and appears in different terminal positions in different cells. The extreme form of nonreciprocal translo-



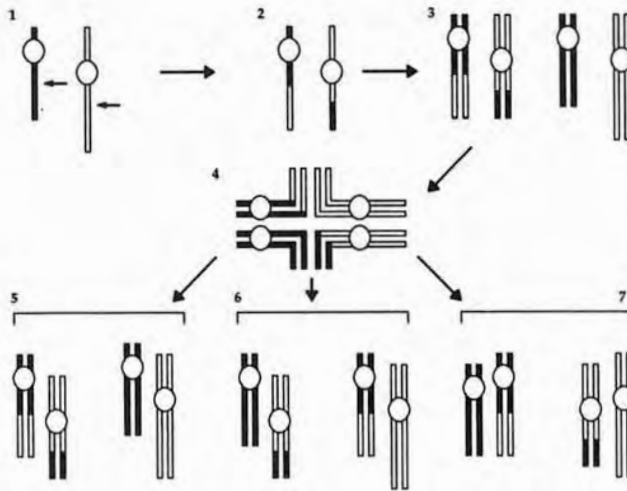
**Figure 4.2:** The effects of crossing over in inversion heterozygotes. An inversion reverses the orientation of loci C, D and E on one chromosome and an inversion loop forms at meiosis. In a pericentric inversion (left), the centromere lies within the inversion and crossing over generates unbalanced products containing deletions or duplications. In a paracentric inversion (right), the centromere lies outside the inversion and crossing over within the loop generates an acentric fragment (which is lost) and a dicentric bridge.

cation is a **Robertsonian** or **whole arm translocation**, where two *acrocentric* chromosomes (q.v.) are fused at or near the centromere to generate a **compound chromosome**<sup>1</sup> which segregates as a single unit.

Like inversions, translocations are balanced, but unlike inversions, the rearrangement involves more than one chromosome. Thus, in a heterozygote, pairing at meiosis may involve more than two partners. Translocation heterozygotes are therefore often sterile not simply due to the effects of crossing over, but also because segregation itself is aberrant and generates unbalanced gametes (Figure 4.3). Because Robertsonian translocations involve whole chromosomes, unbalanced segregation generates monosomic and trisomic gametes. This is reflected in the recurrence risk of Down's syndrome, which is very low when caused by spontaneous nondisjunction, but much higher when caused by a Robertsonian translocation involving chromosome 21. In the most extreme case, a Robertsonian translocation can fuse both homologous copies of chromosome 21 together. If this happens, every meiotic segregation will generate gametes with a chromosome 21 imbalance, and all viable offspring will have Down's syndrome (i.e. the recurrence risk is 100%).

**Fragile sites.** Unstained chromosome gaps identified as secondary constrictions induced by stressful growth conditions are thought to be hotspots for chromosome breakage, and are termed **fragile sites**. The sites are classified according to the growth conditions required to induce them. In human

<sup>1</sup>A **compound chromosome** is two whole chromosomes fused together. Robertsonian translocations involve the fusion of *acrocentric* chromosomes (q.v.) where loss of part of the short arm appears to have no effect on phenotype. The two short arms fuse to generate a *submetacentric* chromosome (q.v.) containing the long arms of both original acrocentric chromosomes and a common centromere (hence *whole arm translocation*). It is also possible for two submetacentric chromosomes, such as X-chromosomes, to become joined together. Such chromosomes are dicentric (two centromeres) and are termed **attached chromosomes**. Attached-X chromosomes were used to prove the chromosome theory of inheritance.



**Figure 4.3:** Behavior of reciprocal translocation heterozygotes at meiosis. (1) Two chromosomes suffer double-strand breaks (shown by arrows), and (2) repair by end-joining causes a reciprocal translocation (alternatively, the translocation could result from recombination between sites on each chromosome). (3) At meiosis, the translocants attempt to align with their normal homologs, often generating a four-chromosome structure termed a **tetraivalent** (4). The tetraivalent may segregate in a number of ways. In **alternate segregation** (5), the two normal chromosomes segregate to one pole and the two translocants to the other and gametes are balanced. In **adjacent-1 segregation** (6), each normal chromosome segregates with the nonhomologous translocant and gametes are unbalanced. In **adjacent-2 segregation** (7), each normal chromosome segregates with the homologous translocant and gametes are unbalanced (adjacent-2-segregation occurs very rarely). In Robertsonian translocation (not shown), the compound chromosome attempts to pair with its homologs forming a three-chromosome structure termed a **trivalent**. In alternate segregation the compound chromosome segregates to one pole and the two normal chromosomes to the other and gametes are balanced. In adjacent segregation, the compound chromosome segregates with one of the normal homologs and the remaining normal chromosome segregates to the other pole. The gametes are not only unbalanced but aneuploid.

chromosomes, over 50 **common fragile sites** have been identified which are present in most people. About 20 **rare fragile sites** have also been described, which are present in a minority of individuals. Generally, fragile sites have no phenotypic effect. An exception is the folate-sensitive rare fragile site at Xq27, which corresponds to the presence of mental retardation, large testes and characteristic facial features in males, and mild mental retardation in some female carriers (**Martin-Bell syndrome, fragile-X syndrome**). The cause of the syndrome is a polymorphic triplet repeat mutation in the gene *FMR1* which explains the complex inheritance patterns observed (q.v. *triplet repeat syndromes*). The properties of the locus which make it a fragile site probably have nothing to do with the mutation *per se*, although fragile sites are induced by conditions which influence DNA replication and repair.

**The causes of structural mutation.** Deletions, duplications, inversions and translocations can all result from incorrectly repaired DNA or aberrant forms of homologous and site-specific recombination (q.v. *illegitimate recombination*). Unequal crossing over between tandemly repeated DNA can generate deletions and duplications, whilst recombination between dispersed repetitive sequences such as transposable elements on different chromosomes can cause translocation. Translocations can also occur due to illegitimate *V-D-J recombination* (q.v.), and such aberrations often lie at the root of lymphoid disorders (see *Oncogenes and Cancer*).

Chromosome breaks occur naturally, but can be induced by chemical and physical agents, termed **clastogens**. The frequency of breaks also increases where there are underlying genetic weaknesses, such as mutations in genes responsible for DNA recombination and repair. The frequency of

chromosome breaks is related to the frequency of **sister chromatid exchange (SCE)** (exchange of segments of DNA between the two chromatids of a chromosome), because homologous recombination is stimulated by double stand breaks (*see* Recombination). The frequency of SCE can be quantified by differential staining of sister chromatids (q.v. *replication banding*). An increased frequency of SCE is seen in several **chromosome instability syndromes**: Bloom's syndrome, Fanconi's anemia and ataxia telangiectasia, which result from mutation in DNA repair genes or DNA damage response genes (q.v. *repair deficiency syndromes*). The distribution of breakpoints in chromosomal DNA is not random. Although different clastogens are known to have different breakage hot spots, these cluster in areas of chromatin defined by light G-bands (q.v. *chromosome banding*), suggesting that susceptibility is related to chromatin structure (*see* Chromatin).

**Karyotype mixing within individuals.** The above sections discuss **constitutional chromosome mutations**, i.e. those which affect every cell in the body. A second class of chromosome aberrations, termed **somatic** or **acquired chromosome mutations**, occur during development, resulting in a mixture of different karyotypes within the same individual (Table 4.6) (q.v. *somatic mutation, cancer*). Such an individual is termed a **mosaic**, but similar karyotype mixtures can arise where two independently derived clones are mixed together. An individual thus formed is termed a **chimera**. In both cases, genetically distinct cells contribute to the same organism, but only in the case of the mosaic are both types of cell derived from a common ancestor. Chimerism occurs naturally at a low frequency: **dispermic chimeras** result from fusion of dizygotic twins, **blood chimeras** as a result of colonization of one twin by cells from the co-twin (also q.v. *transgenic mice, ES cell*).

The critical determinants of the effects of mixed karyotype are the clonal extent of each cell line and the autonomy of the genes involved. A deleterious karyotype may have no overall effect if it is confined to a small sector of tissue or if the defect is cell-nonautonomous (i.e. it can be complemented by products from other cells). Conversely, a karyotype may be lethal if it is widespread, or if it affects a specific organ, or if the defect is cell-autonomous resulting in widespread cell death. In animals, mosaics or chimeras for male and female cells (**gynandromorphs**) can develop normally as either sex, as **intersexes** (individuals with an ambiguous or intermediate sexual phenotype) or as

**Table 4.6:** Terms associated with the appearance of different phenotypes in otherwise equivalent cells, and their underlying causes

Term	Definition
Variegation	Within a tissue, the occurrence of sectors displaying different phenotypes (e.g. white/green variegation in some plants). This may or may not reflect an underlying difference in genotype
Mixoploidy (aneusomaty)	A term which describes an individual carrying somatic cells with different karyotypes. A mixoploid can be a mosaic or a chimaera (see below). <b>Eusomaty</b> describes an individuals whose cells have the same karyotype
Mosaic	An individual carrying somatic cells with different genotypes (or karyotypes) which have arisen from a common ancestor, e.g. by mutation or somatic cell recombination
Chimera	An individual carrying somatic cells with different genotypes (or karyotypes) from different origins. Chimerism usually arises through embryo fusion or colonization (q.v. <i>transgenic mice, ES cells</i> )
Allopheny	Variegation in one tissue arising from mosaicism in another, e.g. the differences in the variegated tissue arise not from intrinsic differences but from inductive interactions with other cells which have different genotypes. The creation of allophenic animals allows the cell autonomy of a mutation to be tested. Allophenic <i>Drosophila</i> can be created by X-ray-induced somatic recombination, e.g. to study inductive interactions in eye development (q.v. <i>Sevenless signaling pathway</i> )



**hermaphrodites** (individuals capable of producing both male and female gametes), depending upon the clonal extent of each cell line.

The term mosaic is often used loosely to describe any situation where clones of otherwise equivalent cells display different phenotypes. Mosaicism, however, strictly refers to a mixture of genotypes and thus does not include different phenotypes arising because of variable penetrance or expressivity, e.g. resulting from environmental effects or epigenetic regulation (e.g. *position effect variegation* (q.v.) or the inheritance of different active and inactive X-chromosomes). Where an underlying difference in genotype is not detected, the term **variegation** should be used.

## References

- Borgankar, D.S. (1989) *Chromosome Variation in Man: A Catalogue of Chromosome Variants and Anomalies*. Liss, New York.
- Therman, E. and Susman, M. (1993) *Human Chromosomes. Structure, Behaviour and Effects*. 3rd edn. Springer, Berlin.
- Antonarakis, S.E. (1993) Human chromosome 21 — genome mapping and exploration, circa 1993. *Trends Genet.* 9: 142–148.
- Epstein, C.J. (1988). Mechanisms of the effects of aneuploidy in mammals. *Annu. Rev. Genet.* 22: 51–75.
- Hassold, T.J. and Jacobs, P.A. (1984) Trisomy in man. *Annu. Rev. Genet.* 18: 69–98.
- Moore, G., Foote, T., Helentjaris, T., Devos, K., Kurata, N. and Gale, M. (1996) Was there an ancestral cereal chromosome? *Trends Genet.* 11: 82–83.
- Zinn, A.R., Page, D.C. and Fisher, E.M.C. (1993) Turner syndrome — the case of the missing sex-chromosome. *Trends Genet.* 9: 90–93.

## Further reading



**This Page Intentionally Left Blank**

## Chapter 5

# Chromosome Structure and Function

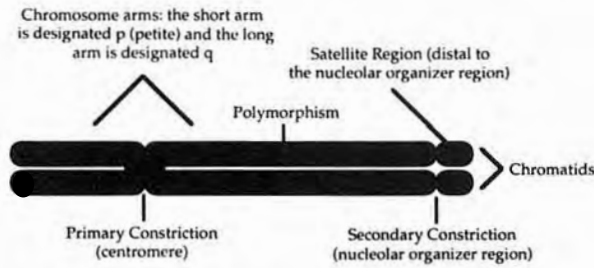
### Fundamental concepts and definitions

- A **chromosome** is a discrete DNA molecule which carries essential genetic information, together with any associated proteins which define its structure. The scope of the term can include the bacterial nucleoid, organelle genomes and virus genomes, as well as eukaryotic nuclear chromosomes, but only the last is considered in this chapter.
- The original suggestion that genes were carried on chromosomes was based on the parallel behavior of chromosomes and Mendelian genes (occurring in pairs, equal segregation, independent assortment). This is the **chromosome theory of inheritance**.
- There are three levels of duality in diploid eukaryotes. Chromosomes occur in homologous pairs, and when individual chromosomes become visible at mitosis, they can be seen to comprise a pair of **chromatids**, joined at a common centromere, each chromatid containing two DNA strands in a duplex. Despite the presence of eight DNA strands, the cells are still diploid. Only one strand of each DNA duplex actually carries the (transcribed) information, and the presence of two chromatids represents the doubling of information in preparation for segregation into daughter cells — cells are born with only one copy (a chromatid becomes a chromosome when it segregates). Furthermore, the paired chromatids of each mitotic chromosome are identical, whereas homologous chromosomes are not. Throughout the cell cycle, therefore, each locus is represented by just two alleles.
- Two types of chromosome can be distinguished at meiosis: **autosomes**, which have **homomorphic** (structurally identical) partners and thus form **homologous chromosome pairs**, and **heterosomes** (or **allosomes**) which have **heteromorphic** (structurally dissimilar) partners and form pairs which are homologous over only part of their lengths. The region of homology between two heteromorphic chromosomes is termed the **pseudoautosomal region** because the genes found within this region show the same pattern of inheritance as those situated on autosomes. Heterosomes are usually **sex-chromosomes** (q.v. *pedigree analysis*, *sex-linked inheritance*, *sex determination*).
- The role of the chromosome is to provide a framework which allows each linear segment of the genome to replicate and segregate efficiently. Failure in either of those processes causes chromosome imbalance in daughter cells. Three specific *cis*-acting sites are required for stable chromosome maintenance, an *origin of replication* (q.v.), a centromere and telomeres.

### 5.1 Normal chromosomes — gross morphology

**Cytogenetic features of normal chromosomes.** For most of the cell cycle, eukaryotic chromosomes exist as loosely packed chromatin and cannot be distinguished in the nucleus. They become visible at the onset of mitosis (or meiosis) when the chromatin condenses, forming discrete structures which stain densely when nuclei are treated with appropriate dyes (Box 5.1). The morphological features of the metaphase chromosome are shown in Figure 5.1, and these allow individual chromosomes to be recognized and aberrations to be identified (see Chromosome Mutation).

**Chromosome banding.** The features of chromosomes described in Figure 5.1 are based on **homogeneous staining** with DNA-binding dyes such as Feulgen. Mammalian chromosomes can also be stained in a heterogeneous manner using a variety of disruptive techniques which reveal a highly reproducible and specific pattern of alternate light and dark transverse bands (Table 5.1). Such **chromosome banding** methods allow chromosomes with similar gross morphology to be discriminated,



**Figure 5.1:** Morphological features of normal metaphase chromosomes. The **primary constriction**, which indicates the position of the centromere, stains densely and joins all four arms. **Secondary constrictions** are pale staining and usually represent **nucleolar organizer regions (NORs)**, the positions of tandemly repeated rRNA genes. Interstitial secondary constrictions are found on human chromosomes 1, 9 and 16, whilst distal constrictions appear on chromosomes 13, 14, 15, 21 and 22. Telomeric chromosome segments found distal to NORs are termed **satellite regions** because they may appear detached from the main body of the chromosome. The **satellite regions** of different chromosomes are often found grouped together (**satellite association**), because the nucleolar organizers contribute to a common **nucleolus**. Other secondary constrictions can be induced by growing cells under abnormal conditions and are termed *fragile sites* (q.v.). **Chromosome polymorphisms (heteromorphisms)** are heritable morphological features of chromosomes which vary within a population but have no phenotype. Examples include areas of variable heterochromatin (as shown), *inversions* (q.v.) and rare fragile sites. These can be used as cytogenetic markers.

and provide a cytogenetic basis for gene mapping. They are most useful for characterizing chromosome aberrations. A low-resolution banding pattern is observed in metaphase chromosomes, but each metaphase band may be resolved into many sub-bands in the less condensed prometaphase chromosomes. The **International System for Human Cytogenetic Nomenclature (ISCN)** is based on the bands and sub-bands in early and late prometaphase chromosomes and metaphase chromosomes (these have a resolution of 400, 550 and 850 bands, respectively). Each band is identified by chromosome and arm (e.g. 1p, 2q) and is numbered from the centromere (*cen*). Increasing resolution is designated by the use of more numbers (e.g. 1p3→1p34→1p34.1). Chromosome banding patterns reflect the structural and functional organization of the mammalian genome (q.v. *isochores*), with, for example, light Giemsa bands corresponding to regions of general transcriptional activity, early replication, low repetitive DNA content and DNase I sensitivity. This is only a very general organization, however, as each band corresponds to up to 10 Mb of DNA (see Chromatin, Genomes and Mapping). Lower eukaryotic chromosomes do not stain in response to the banding techniques applied to mammalian chromosomes, but the polytene chromosomes of dipteran insects display a natural banding pattern of high resolution (see below).

## 5.2 Special chromosome structures

**A and B chromosomes.** In most eukaryotes, each cell of a normal individual carries a defined and invariant set of chromosomes which are diagnostic of the species (q.v. *karyotype*; c.f. *double minute chromosomes*, *gene amplification*). Some species, however, carry extra chromosomes, often appearing to be composed entirely of heterochromatin, which have no effect on phenotype and often vary between populations, within a given population or even between cells within an individual. These structures, which are variously referred to as **accessory chromosomes**, **satellite chromosomes**, **supernumerary chromosomes**, or, in plants, **B-chromosomes** (to distinguish them from the invariant and essential **A-chromosomes**) often do not take part in mitosis and thus segregate randomly. They may be regarded as giant linear plasmids.

**Polytene chromosomes.** In *Drosophila* and other dipterans, the cells of certain larval secretory tissues (e.g. salivary glands) contain **giant chromosomes** comprising up to 1000 chromatids. These

**Table 5.1:** A selection of chromosome banding techniques

Banding technique	Method and applications
C-banding	A Giemsa-based technique which includes incubation with barium hydroxide. Identifies regions of <i>heterochromatin</i> (q.v.) and is therefore useful for determining the position of centromeres
C <sub>d</sub> -banding	A technique for identifying <i>kinetochores</i> (q.v.)
D-banding	A technique for identifying regions of DNase I sensitivity, generally corresponding to regions of open chromatin which are potentially transcriptionally active
G-banding	The most widely applied chromosome banding technique, the reproducible pattern of bands being the basis of the international standard human and mouse cytogenetic maps. Chromosomes are partially digested with trypsin and incubated with <b>Giemsa's stain</b> (a mixture of methylene blue, eosin and other dyes dissolved in methanol)
G12-banding	G-banding carried out at high pH, which stains mouse and human chromosomes differently, allowing them to be discriminated, e.g. in somatic cell hybrids
N-banding	A technique for specifically staining nucleolar organizer regions, e.g. silver nitrate staining
Q-banding	A banding technique using the fluorescent dye quinacrine. This produces a similar pattern to G-banding, but also causes areas of heterochromatin to fluoresce brightly and is useful, e.g. for identifying the Y-chromosome
R-banding	A Giemsa-based technique including a heat-denaturation step which results in a reversed banding pattern from that obtained with conventional G-banding. This technique is useful in, e.g. identifying terminal deletions
Replication banding	This technique involves the incorporation of bromodeoxyuridine into replicating DNA. A brief pulse during the S phase allows replication timing to be investigated and if prolonged, generates a banding pattern similar to Giemsa's stain, suggesting G-bands correlate to <i>replication time zones</i> (q.v.). Incorporation over two rounds of replication allows sister chromatids to be discriminated on the basis that one has bromodeoxyuridine incorporated in both strands and the other in only one strand. The chromatids then stain differently in the presence of Giemsa's stain and the fluorescent dye Hoechst 33258. This technique, known as <b>harlequin staining</b> , is useful for the detection of <i>sister chromatid exchanges</i> (q.v.)
T-banding	A variation of R-banding in which only telomeric DNA is stained. T-bands are particularly gene-rich (q.v. <i>isochore</i> , <i>transcriptional mapping</i> )

In many cases, these methods are empirical, i.e. it is not understood how they work. It is remarkable that these diverse techniques generate similar banding patterns, strongly indicating the structural organization of the genome into discrete isochores.

may also be termed **polytene chromosomes** or **Balbani chromosomes** after their discoverer, and are generated by multiple rounds of replication in the absence of mitosis as a strategy for gene amplification. Polytene chromosomes are not only thicker than normal chromosomes, but are also longer because the association of many chromatids prevents the adoption of normal chromatin structure.

A remarkable feature of polytene chromosomes is the highly reproducible pattern of transverse dark **bands** (or **chromomeres**) and light **interbands**. These are thought to reflect regional differences in chromatin density and to correspond functionally to individual *chromatin domains* (q.v.). The resolution of the banding pattern is much finer than even prometaphase bands on mammalian chromosomes (each polytene band ranges in size from 1 to 100 kbp of DNA), but can be used in the same way, to construct detailed cytogenetic maps and characterize chromosome mutations.

A second feature of polytene chromosomes is the presence of **chromosome puffs** and **Balbani rings**. These are small and large, respectively, distensions of the chromosome which correspond to regions of transcriptional activity, presumably reflecting local decondensation of chromatin struc-

ture. Reproducible patterns of puffs and rings can be induced by certain environmental stimuli (e.g. treatment with the moulting hormone ecdysone), allowing the positions of the activated genes to be determined by direct observation.

**Lampbrush chromosomes.** In amphibian oocytes, where the cell cycle is arrested at *diplonema* (q.v.) of the first meiotic division, a single nucleus serves the needs of an extremely large cell. The four chromatids of the meiotic bivalent are held together by chiasmata, and can be seen to extrude long, uncoiled **lateral loops** of transcriptionally active chromatin from an axis of densely packed **chromomeres** (the same term is used to describe the densely packed, presumed inactive, chromatin in both lampbrush and polytene chromosomes, and in decondensing mammalian chromosomes). The loops occur in pairs, one originating from each chromatid, and they are surrounded by a halo of ribonucleoprotein. This, and the extensively decondensed structure of the loops, is thought to reflect the continuous and prodigious transcriptional activity of the chromosome. There are some 10–15 000 loops active in the cell as a whole, but most of the DNA remains condensed as chromomeres. Like polytene bands and interbands, each loop is thought to correspond to a *chromatin domain* (q.v.).

### 5.3 Molecular aspects of chromosome structure

**Molecular structures required for chromosome function.** The gross morphology of the chromosome reveals little about how it functions in the cell. Primarily, the role of the chromosome is to provide a framework which allows each linear segment of the genome to replicate and segregate efficiently, failure in either of those processes resulting in chromosome imbalance in daughter cells (see Chromosome Mutation). Since each eukaryotic chromosome carries a different array of genes, the particular information carried on the chromosome does not influence its function. Three specific *cis*-acting sites are required for stable chromosome maintenance: an *origin of replication* (q.v.), a centromere and telomeres. The origin of replication is essential for the replication of the chromosome because it provides a site for the assembly of replication initiation proteins (see Replication). The **centromere** is essential for segregation because it provides a site for kinetochore assembly and facilitates microtubule attachment. The **telomeres** are essential for chromosome stability because they allow the completion of chromosome ends during DNA replication and prevent illegitimate end-joining to other chromosomes. *Artificial chromosomes* (q.v.) carrying arbitrary DNA can be stably co-maintained with the endogenous genome as long as these three elements are present.

**Molecular nature of the centromere.** In *Saccharomyces cerevisiae*, centromeres have been defined genetically by their ability to confer mitotic stability upon a plasmid (this is known as a **CEN function**). Several centromeres have been cloned by *chromosome walking* (q.v.) and appear to be functionally interchangeable. Sequence comparison has identified three conserved elements, termed CDI, CDII and CDIII. CDI has a short consensus sequence that appears to function primarily in meiotic segregation. CDII is an extremely AT-rich sequence of about 100 bp whose function is unclear. Mutations in both these elements influence mitotic segregation but do not abolish it. CDIII is a highly conserved 26 bp element displaying dyad symmetry, which appears to be essential for centromere activity, as point mutations at the centre of symmetry abolish centromere function, resulting in unstable segregation. A multiprotein complex termed Cbf-III binds to this element and displays microtubule motor activity. Cbf-III may thus represent the site of microtubule attachment and the engine for segregation at anaphase.

In *Schizosaccharomyces pombe* and higher eukaryotes, centromeres span several tens of kilobase pairs, compared with the minimal *Saccharomyces cerevisiae* centromere of 125 bp, and this may reflect the nature of spindle attachment. In *S. cerevisiae*, a single spindle fiber attaches to each chromosome, whereas in *Schizosaccharomyces pombe* and in mammals, numerous fibers are involved, and the centromere contains repetitive DNA. In *S. pombe* this displays dyad symmetry, whilst in mammals, the centromere contains a large proportion of *satellite DNA* (q.v.), which in primates consists of 170 bp



tandem repeats with local perturbations. Proteins specifically associate with the mammalian centromere, one of which binds to satellite DNA *in vitro* and may represent the site of kinetochore formation.

**Molecular nature of the telomere.** The possession of a linear chromosome presents two problems to the eukaryotic cell. Firstly, because of the properties of DNA polymerases (q.v.), the 5' ends of each strand cannot be completed during replication (see Replication). Secondly, because of the abundance of end-joining enzymes for DNA repair, the chromosome ends could be ligated together generating polycentric compound chromosomes or ring chromosomes which would fail to segregate properly (see Chromosome Mutation). Both these potentially lethal processes are prevented by **telomeres**, which are specialized structures which are added to the chromosome ends in a replication-independent manner. Telomeres also appear to act as initiators of synapsis. They are associated with the nuclear envelope and are the first chromosome regions to pair up. In yeast and trypanosomes, subtelomeric DNA plays an important role in the regulation of gene expression by housing silent copies of information which is transferred to expression-competent sites by nonreciprocal recombination (q.v. *mating type switching*, *antigen switching*).

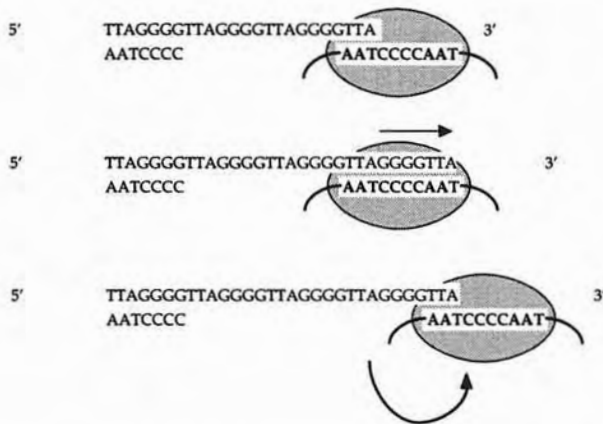
Telomeric DNA consists of short, tandemly repeated sequences. These have been characterized from a number of eukaryotes and are generally GC-rich, with guanidine residues clustered on one strand and cytidine residues on the other (Table 5.2). They may form unusual quadruplex structures by unorthodox interactions between guanosine residues, and which may play a role in protecting the telomere from end-joining reactions (see Nucleic Acid Structure). The addition of telomeric repeats to the termini of unstable linear plasmids confers stability as long as a centromere is also present.

Telomeres are added to the ends of DNA by a specialized ribonucleoprotein complex termed **telomerase**. This comprises several polypeptides and a single RNA molecule which contains two copies of the cytidine-rich strand sequence found in telomeric DNA. The protein component of telomerase possesses reverse transcriptase activity, but the activity appears to be limited to using the telomerase-specific RNA as a template. Based on this information, a current model suggests that telomere repeats are added to the 3' ends of existing telomeres by a primer extension/template translocation strategy, as shown in Figure 5.2. It is thought that the most distal telomere repeats can form a structure which blocks the telomere ends, and thus prevents illegitimate end-joining. This may involve looping of the DNA and/or the association with telomere-binding proteins. Looping of the terminal DNA could prime synthesis of the G-rich DNA strand.

Components of telomerase and other proteins associated with telomere activity can be identified from the analysis of mutations which affect telomere function. Mutations which affect telomere length, a strain-specific characteristic in yeast which is associated with a senescent phenotype, have identified the *TLC1* gene (which encodes telomerase RNA) and several *EST* loci (even shorter telomeres) which code for telomerase polypeptides or telomere binding proteins like mammalian TRF1. The senescent phenotype suggests that telomeres may play a critical role in the life-span of a given cell. This is supported by the observation that the length of telomeres decreases with age in certain human somatic tissues, whilst it is maintained in germ cells (also q.v. *growth transformation*; see Oncogenes and Cancer). In mice, however, which have a much shorter life-span but much longer telomeres than humans, no age-dependent shortening has been observed. It has been suggested that

**Table 5.2:** Telomeric repeat sequences in different eukaryotes

Organism	Telomere repeat
<i>Tetrahymena</i>	TTGGGG
<i>Trypanosoma</i>	TAGGG
<i>Saccharomyces cerevisiae</i>	(TG) <sub>1-3</sub> TTGGG
<i>Arabidopsis thaliana</i>	TTAGGG
Mammals	TTAGGG



**Figure 5.2:** A model for telomerase activity. (1) Telomerase RNA (bold) pairs with DNA at chromosome terminus. (2) Telomerase RNA acts as a template for reverse transcription; DNA primer is extended to the end of the template. (3) Telomerase complex translocates to new chromosome terminus.

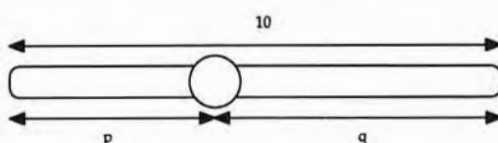
reduction of telomere length to a critical limit induces a DNA repair mechanism which arrests the cell cycle. If so, telomerase activity would be required for cell proliferation, and this hypothesis is supported by the presence of telomerase activity in the majority of human tumors; telomerase may thus be a suitable target for anti-cancer therapy. However, the recently generated telomere-knock-out mouse has been shown to be as susceptible to cancer as its wild-type littermates (*see Replication*).

In *Drosophila*, the distal tips of the telomeres are similar in structure to those of other eukaryotes and presumably function in the same way — by forming a cap which blocks end-joining. Telomere length, however, is maintained in a novel manner which involves the repeated transposition of LINE1-like retroelements specifically to the chromosome ends. The manner in which these elements are targeted is unclear.

**Box 5.1: Chromosome classification and nomenclature**

**Chromosome parameters.** All known eukaryotes possess more than two chromosome per haploid set, so it is necessary to define a system of descriptive nomenclature which allows individual chromosomes to be distinguished from others in the same nucleus. The nomenclature which is used is summarized below and specifies the type and position of the centromere and the relative lengths of the chromosome arms. Most chromosomes can be defined in terms of three parameters, the *d*-value, the *r*-value and the centromeric index (*i*), which delimit regions of the chromosome in which the centromere can be placed irrespective of chromosome size. The *d*-value is the difference in length between the long and short arms of the chromosome, i.e.  $d = q - p$ , where *q* and *p* are the lengths of the long and short arms, respectively, and the total length of the chromosome has been

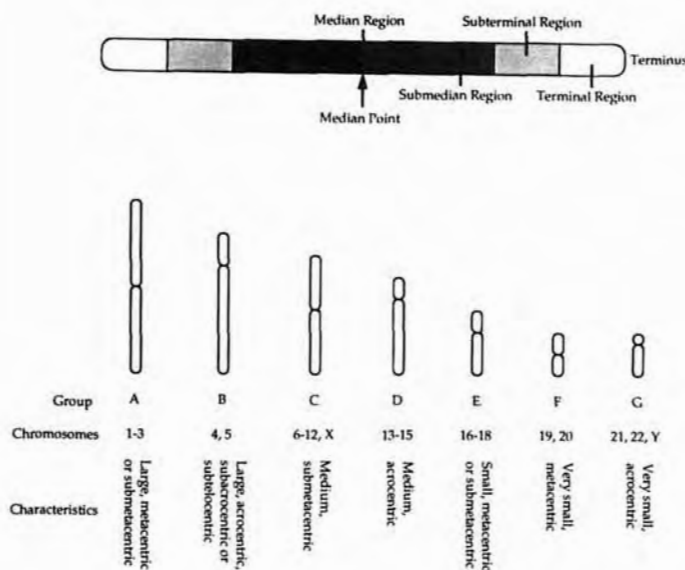
divided into 10 arbitrary units. The *r*-value is the ratio of the lengths of the two chromosome arms, i.e.  $r = q/p$ . The **centromeric index** is the distance from the centromere to the tip of the short arm, expressed as a percentage of the total chromosome length, i.e.  $i = 100p/(p + q)$ . Human chromosomes are placed into seven categories based on these parameters and their size as shown.



Measurements used to define the morphological parameters *d*, *r* and *i*.

Term	Definition
<b>Monocentric</b>	A chromosome with a single defined centromere. Most normal chromosomes are monocentric
<b>Holocentric</b>	A chromosome with a diffuse centromere, so that spindle attachment occurs over the entire chromosome ( <b>lateral attachment</b> ), as found in the <i>Lepidoptera</i>
<b>Telocentric (telomictic)</b>	A monocentric chromosome with a terminal centromere ( $p = 0$ , $q = 10$ , $d = 10$ , $r = \infty$ and $i = 0$ ), also known as a <b>T-chromosome</b> or <b>monobrachial chromosome</b> . Such chromosomes are very rare in nature, most supposedly telocentric chromosomes having a short but definite <i>p</i> arm. The term is therefore widely misused. It can be used to describe aberrant chromosomes broken through the centromere
<b>Atelocentric (atelomictic)</b>	A monocentric chromosome with a nonterminal centromere (a <b>dirachial chromosome</b> ), further classified as shown below:
<b>Metacentric</b>	A monocentric chromosome with a central or near central centromere. If the centromere is exactly at the median point, it is classified as an <b>M-chromosome</b> ( $d = 0$ , $r = 1.0$ and $i = 50$ ), whilst if the centromere is not central but lies within the <b>median region</b> , it is classified as an <b>m-chromosome</b> ( $d = >0 - 2.5$ , $r = >1.0 - 1.7$ and $i = 37.5 - <50$ ). True M-chromosomes are rare in nature (q.v. <i>isochromosome</i> )
<b>Submetacentric</b>	A monocentric chromosome with a centromere in the <b>submedian region</b> ( $d = 2.5 - 5.0$ , $r = 1.7 - 3.0$ and $i = 25 - 37.5$ )
<b>Subacrocentric</b>	A monocentric chromosome with a centromere in the <b>subterminal region</b> ( $d = 5.0 - 7.5$ , $r = 3.0 - 7.0$ and $i = 12.5 - 25$ )
<b>Acrocentric</b>	A monocentric chromosome with its centromere in the <b>terminal region</b> ( $d = 7.5 - <10$ , $r = 7.0 - \infty$ , $i = >0 - 12.5$ ), also known as a <b>t-chromosome</b> . Most so-called telocentric chromosomes are actually acrocentric

Continued



Classification of human chromosomes using the nomenclature discussed above. Human autosomes are numbered by size: 1 (largest; - 22 (smallest) although chromosome 22 is actually bigger than 21.

## References

- ISCN (1995) *An International System for Human Genetic Nomenclature* (ed F. Mittelman). Karger, Basel.
- Bickmore, W.A. and Sumner, A.T. (1989) Mammalian chromosome banding. *Trends Genet.* 5: 144-178.
- Clarke, L. (1990) Centromeres of budding and fission yeasts. *Trends Genet.* 6: 150-154.
- Greider, C.W. (1996) Telomere length regulation. *Annu. Rev. Biochem.* 65: 337-365.
- Mason, J.M. and Biessmann, H. (1995) The unusual telomeres of *Drosophila*. *Trends Genet.* 11: 58-62.
- Therman, E. and Susman, M. (1993) *Human Chromosomes. Structure, Behaviour and Effects*. 3rd Edn. Springer, Berlin.
- Smith, S. and de Lange, T. (1997) TRF1, a mammalian telomeric protein. *Trends Genet.* 13: 21-26.
- Zakian, V.A. (1996) Structure, function and replication of *Saccharomyces cerevisiae* telomeres. *Annu. Rev. Genet.* 30: 141-172.

## Chapter 6

# Development, Molecular Aspects

### Fundamental concepts and definitions

- **Development** is the course by which a complex, multicellular organism arises. It usually begins with a single cell, a fertilized egg, but in certain species it may be initiated asexually, e.g. by parthenogenesis or by budding from an adult.
- Development occurs by **epigenesis**, the progressive diversification of parts, rather than by **pre-formation**, the growth of a prestructured organism in miniature. Thus a simple and crude pattern is formed in the early embryo and filled with detail as the embryo grows.
- Development can be described by three overlapping and interdependent processes. **Differentiation** is the process by which cells become specialized to their different functions and reflects the synthesis of different sets of proteins. **Regional specification** is the process by which cells acquire **positional information**, i.e. instructions which govern their behavior and allow them to form specific structures in appropriate positions, the basis of **pattern formation**. **Morphogenesis** is the creation of form and reflects different aspects of cell behavior: growth and division, movement relative to other cells, changes in shape and size, interactions with other cells and the extracellular matrix, and cell death.
- With a few exceptions, the genome remains the same in all cells of a developing organism, i.e. development is predominantly controlled by the regulation of gene expression.
- The developmental program is regulated by intrinsic and extrinsic factors. In most organisms, the first aspects of differentiation and patterning occur through preexisting (intrinsic) asymmetry in the egg. Further diversification and pattern formation is achieved by extrinsic processes — cell-to-cell communication and the interaction between cells and their environment.
- Development begins at fertilization, but it is more difficult to state when it stops. Many cells become terminally differentiated, which can be thought of as their final fate, whereas others — stem cells — do not; many developmental processes are reiterative. Some regard aging as a developmental process, in which case development eventually ceases when the organism dies.

### 6.1 Differentiation

**Genomic equivalence.** During differentiation, cells become specialized to perform their particular functions, and this reflects the synthesis of characteristic sets of proteins. With few exceptions (*Box 6.1*), the genome of all cells in a developing organism remains the same, which indicates that the basis of differentiation is selective gene expression. Gene expression can be regulated at many levels (*see Gene Expression and Regulation*), and some examples of developmental gene regulation are shown in *Table 6.1*.

Evidence for genomic equivalence comes from both molecular and functional studies. Molecular analysis of genomic DNA (e.g. by Southern hybridization or PCR) shows that genome structure is typically the same in all cells regardless of which genes are expressed. Functional experiments show that differentiated cells are, under certain circumstances, able to **dedifferentiate** and follow alternative pathways of development, a phenomenon termed **transdifferentiation** or **metaplasia**. This shows that genes which are not utilized in a particular cell type are latent and can be reactivated. The ultimate functional test is to regenerate an entire organism from a single differentiated cell. This is quite possible in plants, where differentiated cells of several species can be routinely dedifferentiated in culture and will then proliferate to form a clone of disorganized undifferentiated cells



**Table 6.1:** The regulation of differentiation at different levels of gene expression

Level of gene expression	Developmental system
Transcription	Differentiation of muscle cells is controlled by the synthesis of transcription factors such as MyoD1 which activate muscle-specific genes
RNA processing	Differentiation of the sexes in <i>Drosophila</i> is controlled by alternative splicing of the primary transcripts of the <i>sex lethal</i> , <i>transformer</i> and <i>doublesex</i> genes. Precocious translation of a number of mRNAs in oocytes of the moth <i>Manduca sexta</i> is prevented by delayed capping
RNA turnover	Limb bud development in the chick is controlled in part by the expression of FGF-2. Antisense RNA complementary to FGF-2 mRNA is expressed in the chick limb bud and targets the message for degradation
Protein synthesis	Establishment of anterior and posterior structures in early <i>Drosophila</i> development is controlled in part by repression of translation. Bicoid protein represses translation of <i>caudal</i> mRNA in the anterior, and Nanos protein represses translation of <i>hunchback</i> mRNA in the posterior. Synthesis of the <i>Xenopus</i> fibroblast growth factor receptor is blocked in oocytes by proteins bound to the <i>Xfgfr1</i> mRNA; the repression is lifted in mature eggs
Protein function	Establishment of dorso-ventral polarity in the <i>Drosophila</i> embryo depends on the transfer of the Dorsal protein to the nucleus, which is brought about by phosphorylating (and hence inactivating) the inhibitory protein Cactus

termed a **callus**. Calluses can be exposed to plant hormones and will regenerate to form a complete plant, a process which has been exploited in the genetic manipulation of commercially important plant species (see Recombinant DNA). It has been impossible to obtain the same results using differentiated animal cells, but it has been possible to produce viable embryos by transferring the *nucleus* of a differentiated amphibian cell into an enucleated egg. Although the embryos do not develop into adults, these experiments indicate that whereas differentiated animal cells, unlike plant cells, are unable to fully reverse their developmental commitment (even when isolated and placed in an unusual environment), the nucleus maintains some potency. Performing the same technique using nuclei from cells of an amphibian *embryo*, however, allows development to the adult stage, and nuclei from preimplantation embryos of certain mammals will also support full development. These results suggest that the animal nucleus progressively loses its potency as development proceeds, and some reasons for this are considered below. Despite this restriction, a sheep has been cloned recently by transferring the nucleus of a fully differentiated mammary gland cell into an enucleated egg, demonstrating for the first time that the differentiated animal cell nucleus retains all the information required to produce a fully functional organism.

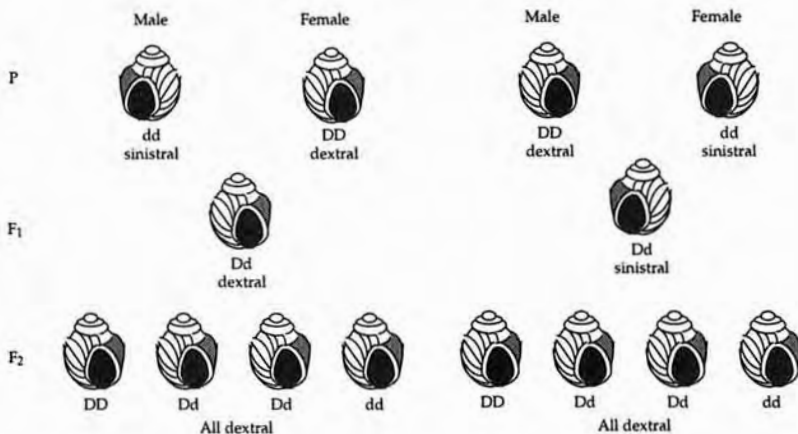
**Simple models of differentiation.** The development of a multicellular organism is a complex process because it involves not only the diversification of cell types, but also their precise spatial organization. Unicellular organisms may also differentiate in response to their environment, and they provide simple model systems involving the same principles seen in multicellular development — differential gene expression controlled by intrinsic and extrinsic factors — but in the absence of pattern formation. The simplest developmental model is probably sporulation in *Bacillus subtilis*, where two cell types, a spore and a mother cell, differentiate upon nutrient depletion (Box 6.2). Yeast cells demonstrate a simple form of transdifferentiation in their ability to spontaneously switch from one mating type to another, a process which involves programmed nonreciprocal recombination (q.v. *mating type switching*). The slime mould *Dictyostelium discoideum* alternates between a unicellular

existence and that of a simple multicellular organism, and is perhaps the most informative model for development in complex organisms. The aggregate of cells demonstrates many of the properties of animal embryos including morphogenesis, simple mechanisms of regional specification, and the ability to regulate for missing parts (Box 6.3).

**Patterning and differentiation in early development.** In multicellular organisms, the processes of differentiation and pattern formation are inextricably linked. The early stages of development involve the segregation of different lineages (an example being the three germ layers ectoderm, mesoderm and endoderm in vertebrate embryos) and the establishment of the first crude positional cues, in the form of the major body axes.

As discussed above, the developmental program is controlled by both intrinsic and extrinsic information. In most animals and plants, the course of development is initiated by intrinsic biochemical asymmetry in the egg. Particular molecules are segregated into different cells as the egg cleaves, and cells arising at different positions in the embryo are therefore nonequivalent. Such molecules are termed **cytoplasmic determinants** and are products of the *maternal* genome deliberately placed in the egg at certain positions. Hence it is the maternal genotype which determines the early developmental phenotype of the embryo, a phenomenon referred to as the **maternal effect** (Figure 6.1).

Further developmental cues may arise through external stimuli. In all vertebrate embryos, for instance, the dorso-ventral axis is specified by an external physical stimulus in the environment. Gastrulation then begins on the dorsal surface of the embryo, and in doing so establishes the antero-posterior axis. In frogs, the dorso-ventral axis of the embryo is specified by the act of fertilization — the region opposite the site of sperm entry becomes the dorsal side via a mechanism which involves cortical rotation. Gravity and rotation are also important in the polarity of avian and (presumably) mammalian embryos. In each case, homologous molecules are expressed in the dorsal organizing centers (see Box 6.4).

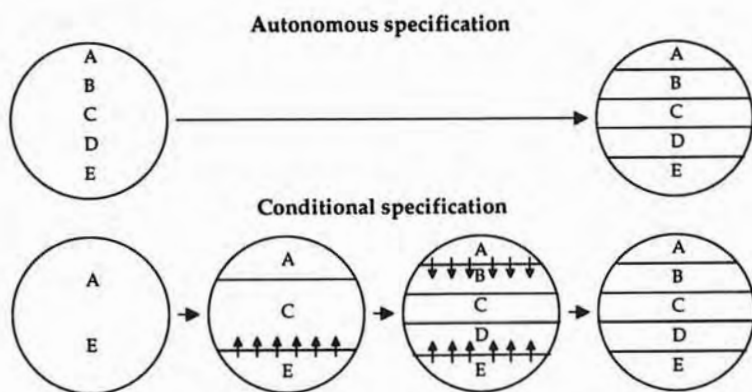


**Figure 6.1:** Maternal effect in snails of the genus *Limnaea*. The direction of coiling is determined at a single locus *D*, with the dominant allele *D* specifying dextral coiling and the recessive allele *d* specifying sinistral coiling. The *D* allele encodes a cytoplasmic protein synthesized in the oocyte, which determines the orientation of the mitotic spindle at the second cleavage division; this factor is missing from sinistral oocytes, suggesting that left-handed coiling is the default pathway. Reciprocal crosses generate F<sub>1</sub> snails with identical heterozygous genotypes but coiling phenotypes that depend upon the mother's ability to synthesize the *D* protein in her oocytes, i.e. it is dependent on her genotype, and is different in each cross. If the F<sub>1</sub> snails are self-crossed, the F<sub>2</sub> generation is uniformly dextral, because the heterozygous F<sub>1</sub> snails can synthesize the *D* protein in their oocytes and specify dextral coiling regardless of their own phenotype. Note that the segregation of alleles follows normal Mendelian principles (q.v. *Mendelian inheritance*).

There is a limit to how much detail can be mapped out by cytoplasmic determinants in the egg. The use of intrinsic cues as a sole mechanism of diversification is therefore limited to early development, whereas later specification is controlled by extrinsic cues, i.e. those arising from the interactions between cells. In this way, a crude pattern established by the segregation of cytoplasmic determinants and the interpretation of physical stimuli can be elaborated and filled with finer layers of detail by progressive interactions between cells.

**Mosaic and regulative development.** The relative predominance of cytoplasmic determinants and cell-cell interactions in early development differs across the animal kingdom. Generally, cytoplasmic determinants appear to be most important in invertebrate embryos and cell-cell interactions in vertebrate development.

These two sources of information define two extreme mechanisms of development, termed mosaic and regulative (Figure 6.2). A **fate map** of an egg or an embryo shows the **fate** (normal developmental outcome) of each region. In **mosaic development**, the fate of an individual cell is determined entirely by its intrinsic characteristics (i.e. its cytoplasmic determinants), whilst in **regulative development**, the fate of each cell is determined by its interactions with other cells. In mosaic development, therefore, each cell undergoes **autonomous specification**, i.e. in principle, if removed from the embryo it will develop according to its intrinsic instructions. An isolated cell will thus differentiate into the appropriate part of the embryo and the remainder of the embryo will lack that part (c.f. *community effect*). The fate map of the embryo can be reconstructed from the fate of individual cells differentiating in isolation. In regulative development, each cell undergoes **conditional specification**, i.e. if it is removed from the embryo it will fail to fulfil its fate because it lacks the appropriate interactions with other cells. The remainder of the embryo, however, can **regulate** itself and replace the missing parts. This is because the appropriate interactions have yet to occur, and therefore the fates of the remaining cells are not determined. The region of the embryo driving such interactions is termed an **organizer**; it has the special property of being able to induce the formation of new structures when transplanted to another region of the embryo because it is the source of the primary inductive signals (Figure 6.2).



**Figure 6.2:** Autonomous and conditional specification. In autonomous specification (upper figure), the five regions of the embryo (A–E) are already specified by cytoplasmic determinants in the egg and become segregated by cleavage. In conditional specification (lower figure), the five regions of the embryo are specified by progressive cell–cell interactions. Both embryos have the same fate map, but would respond differently to isolation experiments. If the region corresponding to tissue B were to be removed from the early embryo in each case, the upper embryo would show mosaic development (the isolated cells would differentiate into tissue B and the remainder of the embryo would develop lacking B), whilst the lower embryo would show regulative development (the isolated cells would differentiate into tissue A because they have yet to be specified as B, and the remainder of the embryo would regulate for the missing parts, i.e. E would act as an organizer and induce A to generate C, then A would induce C to generate B).

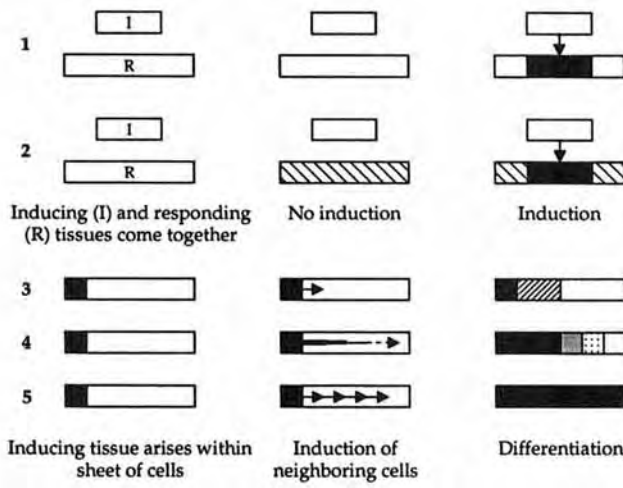
The most extreme example of mosaic development occurs in tunicate embryos where the majority of embryonic cell types are already specified by cytoplasmic determinants in the egg. In some species, e.g. *Halocynthia roretzi*, the distribution of the determinants is obvious because they are different colors! Even so, some inductive interactions are required for proper development, e.g. the nervous system is specified by interactions between blastomeres just prior to gastrulation. Mammals provide an extreme example of regulative development. There appears to be no intrinsic asymmetry in the egg, and in the first few divisions, any cell in the blastocyst can be isolated and will develop into a complete embryo. The earliest differentiation segregates the trophoctoderm from the inner cells mass (roughly speaking, these will develop into the extraembryonic membranes and the embryo proper, respectively). This decision appears to be entirely stochastic, with the fate of each cell apparently dependent upon its location, either on the surface or within the blastocyst. Most organisms fall somewhere between these extremes and show examples of both autonomous and conditional specification in early development. The frog *Xenopus laevis* is used as an example (Box 6.4). Many insects, e.g. *Drosophila*, follow a unique style of early development where specification occurs in the context of a syncytium, and cell fates are determined prior to cellularization by the interaction of diffusible regulators with individual nuclei. After cellularization, further specification occurs conditionally by cell-cell interactions. The development of *Drosophila melanogaster* has been particularly well characterized in this respect, and is discussed in Boxes 6.5 and 6.6.

**Cell-cell interactions in development.** Cells respond to signals from other cells and from the environment by altering patterns of gene expression and the activities of protein molecules already in the cell (see Signal Transduction). Both these responses are potential routes to differentiation, and where one cell (or more properly, a substance produced by it) influences the developmental fate of another, the process is termed **induction** (e.g. see Box 6.4).

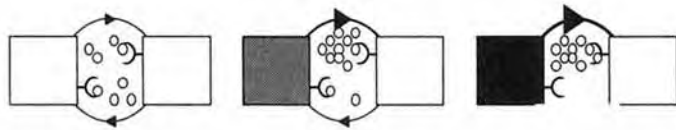
Inductive signaling takes many forms. Firstly, one can distinguish between interactions involving different cell types and those involving equivalent cells. In the former, signaling between cells may be mediated by locally secreted substances active over short distances (paracrine signaling) or by signals released by distant cells and transported to their target through the vascular system (endocrine signaling). Cell-cell contact may be required for signal transduction (sometimes termed juxtacrine signaling), or the cell may respond to molecules secreted into the extracellular matrix. There are two sorts of inductive interactions. In **instructive induction** the responding cells adopt different fates in the presence and absence of the inducer, i.e. the inducing cells *instruct* the responding cells to follow a particular developmental pathway. In **permissive induction** the responding cells are already committed to a certain fate but need a particular signal to permit them to continue, i.e. there is no choice involved. Several responses to induction can also be discriminated. The classical response is uniform, i.e. the inductive signal causes the responding cells to adopt a single, characteristic fate. This often occurs when two layers of tissue are brought together (appositional induction) or where induction is considered at a single cell level. A **morphogen**, however, is a diffusible signal which establishes a concentration gradient over a given population of cells (a morphogenetic field) and elicits different responses at different concentration thresholds. **Homeogenetic induction** is a propagated response where the responding tissue is induced to differentiate into the inducing tissue, and then induces its neighboring cells to follow the same fate. These different types of induction are summarized in Figure 6.3.

The induction mechanisms considered above all concern interactions between nonequivalent cells. Differences can also arise within populations of equivalent cells by a process termed **lateral inhibition**. In this model (Figure 6.4), equivalent, undifferentiated cells secrete a signal which suppresses differentiation in surrounding cells and inhibits further signaling. An equilibrium is thus established where all cells signal to their neighbors in low amounts. A slight imbalance arising by chance can disrupt this signaling process and become amplified by feedback. A random burst of signal released by a central cell will repress signaling in the surrounding cells and enable the





**Figure 6.3:** Mechanisms of induction. The upper panel shows the difference between permissive and instructive induction. In permissive induction (1) the fate of the responding tissue is fixed and requires the inducing signal to fulfill that fate. There is no choice involved and in the absence of induction, the tissue fails to differentiate. In instructive induction (2) the responding tissue will differentiate one way in the presence of the inducer and another way in its absence. The lower panel shows the consequences of different forms of inductive signal. In appositional induction (3), the inducing cells evoke a single response from the responding cells. Conversely, a morphogen (4) is a diffusible signal which forms a gradient across a field of cells and causes different cells to adopt different fates according to their position. In homeogenetic induction (5), the inducing cell causes the responding cell to adopt the same fate as the inducing cell. The responding cell thus becomes an inducer also capable of inducing neighboring cells, and the signal is propagated throughout the population of cells until they all adopt the same fate.



**Figure 6.4:** Lateral inhibition. Differentiation is repressed in a group of equivalent cells by a signaling molecule which downregulates signal production in neighboring cells and may also upregulate synthesis of the receptor. Initial equivalence is maintained by mutual repression, but a chance event, in which a particular cell momentarily synthesizes more of the signal than its neighbors, inhibits the surrounding cells' ability to produce the signal and increases their receptivity to the signal. A positive feedback loop is initiated whereby the neighboring cell eventually ceases signal production and can only receive it. This is seen, for example, in the Notch/Delta signaling pathway during *Drosophila* neurogenesis.

central cell to differentiate. However, the surrounding cells not only cease to inhibit differentiation, they also cease to inhibit signaling from the central cell. The central cell thus establishes dominance by constitutive high-level signaling, whilst the surrounding cells are able to neither signal nor differentiate. This mechanism is thought to underlie the differentiation of neural precursor cells in both the *Drosophila* ectoderm and the vertebrate brain, and the differentiation of the anchor cell in the *Caenorhabditis elegans* uterus (see Box 6.6).

The converse of lateral inhibition is the **community effect**, where cells behave differently in isolation compared with a population of equivalent cells. This effect is brought about by autocrine signaling (the ability of a cell to respond to a signal secreted by itself). Here, equivalent cells secrete a signal which is necessary for their own differentiation and that of their neighbors. The concerted



effort of a population of cells can produce enough of the signal to alter their collective fate, but a single cell cannot achieve the same response.

There are many examples of induction in development, some of which are involved in early fundamental processes (e.g. mesoderm induction and neural induction, discussed in Box 6.4) and some of which occur later in the genesis of particular organs. Well-characterized examples of the latter include vulval specification in *Caenorhabditis elegans* (Box 6.6) and specification of the R7 photoreceptor cell in *Drosophila* eye development (Box 6.7).

**Commitment and maintenance of the differentiated state.** The diversification of the embryo by progressive cell-cell interactions allows an initially small number of cell types to be elaborated into the many distinct cell types in the mature animal or plant. This occurs as a hierarchy: as more cell types arise, more types of inductive interactions are possible. A prominent feature of animal cells, discussed briefly at the beginning of this chapter, is their tendency to become committed to particular developmental pathways. Once the early vertebrate embryo has diversified into ectoderm, mesoderm and endoderm, the cells of those lineages are restricted in their fate and can only differentiate into the spectrum of cell types characteristic of their own lineage, not those of the other lineages. For example, once a cell has differentiated as ectoderm, its fate is broadly restricted to epidermis or neural tissue and it can no longer follow the pathway to mesodermal derivatives such as muscle. Animal development thus proceeds through a series of irreversible decisions, where cells become progressively restricted to narrower and narrower fates until they are terminally differentiated. At this point, the cells either cease to proliferate (e.g. neurons, muscle) or produce only the same cell type as themselves (e.g. keratinocytes). The major exceptions to this rule are stem cells and germ cells (see Table 6.2).

Commitment occurs in several stages. A cell may be **specified** to adopt a certain fate by its intrinsic properties, and will differentiate this way in isolation. However, when exposed to a range of different environments, its fate may be altered by interactions with other cells. A cell becomes **determined** when its commitment is irreversible, i.e. it will follow the same developmental pathway regardless of its environment. The timing of determination can be demonstrated by transplant experiments. Ectoderm in *Xenopus* embryos is specified to become epidermis but can be induced by underlying dorsal mesoderm to form neural plate. Before determination, either presumptive epidermis or presumptive neural plate will differentiate into epidermis in isolation, but once the presumptive neural plate becomes determined it will become neural plate regardless of its environment — *in situ*, in isolation or when transplanted to an ectopic site. Specification and determination usually precede overt differentiation because they involve the synthesis of regulatory molecules which activate the downstream genes controlling differentiation. The difference between the two states of commitment reflects how those regulatory molecules are themselves controlled. Determination involves permanent maintenance of the regulatory circuit, which confers upon the cell a memory of its committed state. Maintenance can occur either through a cytoplasmic feedback loop or by structural reorganization of chromatin, and these mechanisms help explain the nuclear transplantation data presented earlier. In the case of the muscle cell lineage, determination is controlled by feedback. Muscle differentiation is controlled by a group of bHLH regulatory proteins known as the MyoD family. Once these regulators have been activated by inductive interactions in the somites, they can activate the transcription of muscle specific genes (e.g. skeletal actin and myosin) and thus initiate muscle differentiation. However, MyoD also activates *its own gene* as well as those for other members of this family, thus maintaining their own expression through positive feedback. Once a cell begins to express the MyoD regulators, it therefore becomes irreversibly committed to the muscle lineage. However, in principle, the cytoplasmic loop could be broken by removing the nucleus of that cell and placing it in a different cytoplasm. Conversely, maintenance of states of commitment in other cells occur at the level of the chromosome itself. In *Drosophila*, expression of the homeotic genes is maintained by reorganization of chromatin structure, mediated by the *Polycomb* and *trithorax* gene

**Table 6.2:** Terms relating to the commitment of cells to particular pathways of development

Term	Definition
Commitment	The property of a cell which causes it to follow a particular developmental pathway
Dedifferentiation	The ability to reverse the developmental pathway and move from a differentiated state to a more labile state
Determination	The irreversible commitment of a cell to a specific fate, a condition which drives the cell to the same fate irrespective of its environment
Differentiation	The adoption of a new phenotype by the synthesis and/or activation of a new set of proteins. <b>Terminal differentiation</b> is the last stage in any particular cell lineage, where a cell either becomes quiescent or produces a single type of progeny (c.f. <i>stem cell</i> , <i>germ cell</i> )
Fate	The type of differentiated cell a given cell will become in the future. A <b>fate map</b> is a map of an egg or embryo which shows the fates of all cells
Germ cell	A cell which will form gametes. Germ cells usually segregate from somatic cells early in animal development and are the only cells to retain totipotency in the mature animal. Conversely, plant germ cells arise from somatic cells in the mature plant. This fundamental difference in biology may explain the different potencies of animal and plant somatic cells in isolation
Potency	The sum of all possible fates a cell has (in any environment, not just in normal development). Also <b>totipotent</b> — able to adopt any fate, i.e. can in principle form an entire organism; <b>pluripotent</b> — able to adopt several, but not all fates
Specification	The commitment of a cell to a particular fate in the isolated context of a neutral environment. Defines the default developmental program, but may be altered by external influences
Stem cell	A cell which produces two types of descendent at each division: a copy of itself and a cell that will differentiate (c.f. <i>embryonic stem cell</i> )
Transdetermination	A change in commitment occasionally seen in <i>Drosophila</i> imaginal disc cells maintained in a proliferative state by serial transplant into the abdomens of adult flies, rather than being allowed to differentiate. The changes are not random, but occur in a preferred sequence, probably reflecting small changes in the activities of regulatory proteins
Transdifferentiation	The ability to dedifferentiate and follow an alternative developmental pathway. Often seen to occur in regenerating tissues following amputation

products (*Box 6.8*); in female mammals, chromatin structure and DNA methylation maintain one of the X-chromosomes in an inactive state (see DNA Methylation and Epigenetic Regulation). Commitment states maintained in this way cannot be reset by transferring the nucleus to a different cytoplasmic environment, and the prevalence of epigenetic commitment in animals may explain the failure of somatic cell nuclei to initiate development in enucleated eggs. The success of somatic cloning in plants suggests that states of commitment are maintained by extrinsic processes, i.e. cell–cell signaling rather than intrinsic regulatory systems, so that transferring the cell to a different environment is sufficient to reset the developmental program. Similar cases may be seen in animals cells which can dedifferentiate in response to injury and facilitate regeneration. In such cases, differentiation occurs in the absence of determination, allowing the differentiated state to be controlled by the local environment.

## 6.2 Pattern formation and positional information

**Pattern formation.** Pattern formation is the process causing cells to adopt a precise spatial organization, a phenomenon which ensures that all members of a species are morphologically similar. The components of pattern formation are regional specification and morphogenesis. **Regional specification** is the process by which cells acquire **positional values**, i.e. molecular information causing

them to behave in a manner appropriate for their position in the organism as a whole. **Morphogenesis** describes the processes by which cells form structures, by interpreting the positional information and behaving accordingly.

**Regional specification.** Two forms of regional specification can be distinguished, one which controls differentiation and causes particular differentiated cell types to arise in particular positions, and one which controls morphogenesis in cells which are equivalent with respect to their state of differentiation, i.e. it causes them to behave differently so that they give rise to regionally specific structures.

Examples of both processes can be found in the vertebrate limb. The limb comprises a number of different cell types which form the bone, muscle, skin, etc., and these become organized into a roughly concentric pattern so that the bone is on the inside, surrounded by muscle, with skin on the outside. Once this ground plan has been laid, structural differences arise between identical cell types in different locations, e.g. between the forelimb and the hindlimb, and between cells at various points along the three principal limb axes. Such differences reflect the individual behavior of the cells in each position, e.g. differential growth or changes in shape.

**Positional information.** The manner in which cells acquire their positional values is through **positional information** conferred by patterns of gene expression. The source of positional information was first determined in *Drosophila* through the analysis of mutants whose body pattern was fundamentally disrupted (suggesting that certain cells had received the *wrong* positional information). Two sets of genes control positional values along the anteroposterior axis, the **segment polarity genes** whose role is to specify the positions of specific cell types in each segment, and the **homeotic selector genes** whose role is to instruct those cells to behave in such a manner as to generate the appropriate regional structures. The combination of genes expressed in a particular cell thus gives it an 'address' and assigns it to a particular region of the embryo. Mutations in the segment polarity genes cause the loss of particular regions of each segment and their replacement by a mirror image of the remaining cells (i.e. cell types are specified incorrectly) whereas mutations in the homeotic selector genes generate spectacular and sometimes bizarre phenotypes where one body part is replaced by the likeness of another (i.e. the correct cell types are present but they behave aberrantly). The regulation and function of these genes in *Drosophila* is discussed in Box 6.8.

Remarkably, both the segment polarity genes and the homeotic genes have been conserved throughout animal evolution and appear to play an important role in the patterning of all animals, including vertebrates. The role of the vertebrate homeotic genes, and their similarity to the homologous *Drosophila* genes, is considered in Box 6.9. The developing limb (Box 6.10) provides an example of how both sets of genes function in vertebrate development, and how not only the genes themselves, but also their regulatory interrelationships are conserved.

The animal homeotic genes are not found in the plant kingdom: the fundamentally different body plans and developmental mechanisms suggest entirely different regulatory strategies. However, a family of transcription factors related by their possession of a MADS box (a motif also found in the mammalian serum response factor) have been identified in *Arabidopsis thaliana*, and their loss of function causes homeotic transformations in the reiterated structures of the flower (petals, sepals, stamens and carpels). It is likely that similar principles control the expression and activity of these factors and the animal homeotic genes, and thus the MADS system may be the basis of positional information in plants. However, much remains to be learned about the regulators and effectors of these molecules, and of cell-cell signaling in plants in general.

**The role of morphogenesis in development.** Morphogenesis is the creation of form and structure, and can be regarded as the 'third' component of development, taking cells which have been differentiated and endowed with positional values and translating the instructions they have been given into particular movements and behaviors which create precise arrangements in space. However, while morphogenesis may indeed act as the effector of the developmental program, it also plays a



**Table 6.3:** Morphogenetic processes in development

Mechanism of morphogenesis	Examples in development
Cell growth (increase in size)	Neurons in response to nerve growth factor
Proliferation	Organizer cells during amphibian gastrulation
Cell division	Unequal cleavage of <i>Xenopus</i> egg to generate large vegetal blastomeres and small animal blastomeres
Change of cell shape	Narrowing of apical pole of neural cells in the hinge region of the neural plate during neural tube closure
Cell fusion	Differentiation of myotubes from myoblasts
Cell death	Many neurons
	Death of mesenchyme cells in interdigital necrotic zones
Loss of cell-cell adhesion	Mammals and birds: delamination of cells from epiblast as they move through primitive streak
Gain of cell-cell adhesion	Condensation of cartilage mesenchyme in developing limbs
Cell-matrix interactions	Migration of neural crest cells
Loss of cell-matrix interactions	Delamination of cells from basal layer of skin stimulates differentiation into keratinocytes

very active role in its control. From the very earliest stages of development, the manner in which cells behave — how they grow and divide, how they adhere to each other, how they change shape — controls cellular interactions in the embryo and thus establishes the framework within which differentiation and regional specification take place. A primary example is gastrulation, where the structure of the animal embryo is fundamentally reorganized early in development: this involves many morphogenetic processes including changes in cell shape, cell proliferation, and differential cell affinity caused by changing patterns of cell adhesion molecule expression. The result of gastrulation in *Xenopus* embryos is shown in Box 6.4 — the dorsal mesoderm is endowed with positional information and is placed so that it can induce the overlying ectoderm to become neural plate. Hence morphogenesis has acted as the driving force behind both differentiation and regional specification.

**Mechanisms of morphogenesis.** The basis of morphogenesis is cell behavior. This may be considered as an isolated process or in terms of the cell interacting with its environment. Cell intrinsic processes which drive morphogenesis include the rate of cell growth and proliferation, the nature of cell division, changes in cell shape and cell death. Examples of these processes occurring during development are shown in Table 6.3. For some species, these processes are tightly regulated. In the nematode worm *Caenorhabditis elegans*, for instance, every cell division and cell death is written into an invariant developmental program. The predictable movements mean that all cell-cell inductive interactions are also invariant, and each individual contains precisely the same number of cells in precisely the same positions. Several genes have been identified in *C. elegans* which regulate this cell behavior. An example is *lin-14*, which encodes a transcription factor required in the first-stage larva. Modulations in the level of LIN-14 protein disrupt several cell lineages by causing cells to behave in a **heterochronic** fashion — i.e. appropriate for a previous or subsequent stage of development. Overexpression of *lin-14* causes many cells to behave immaturely and repeat divisions characteristic of an earlier stage of development. Loss-of-function mutations, conversely, cause cells to skip certain divisions and behave in a manner more appropriate for later development. In larger organisms cell growth and cell division tend to be controlled *en masse* by growth factors and hormones rather than by an intrinsic program. There is therefore a greater degree of variability in cell number and a statistical likelihood rather than a programmed certainty that a given cell will divide. Such stochastic influences probably play a major role in the phenomenon of *developmental noise* (q.v.), a source of variation in isogenic populations which are maintained in a constant environment.

Cell extrinsic processes which drive morphogenesis include cell-cell adhesion and cell-matrix

interactions. Cells can change the repertoire of cell adhesion molecules they express, and thus alter their relationship with the cells around them. Two families of cell adhesion molecules are recognized: the  $\text{Ca}^{2+}$ -dependent cadherins and the  $\text{Ca}^{2+}$ -independent cell-adhesion molecules of the immunoglobulin superfamily (e.g. N-CAM). By expressing different CAMs and cadherins, cells can choose to stick together or not, and can move in relation to each other, allowing them to form sheets or clumps, or to disperse. Such interactions are important for the complex morphogenetic movements observed during gastrulation, and as cells make and break contacts, the patterns of cell adhesion molecules expressed in the embryo can be seen to change. Cells also interact with molecules such as laminin and fibronectin in the extracellular matrix through cell-surface receptors termed integrins. Such cell-matrix interactions are important for migrating cells such as the neural crest cells and primordial germ cells, and for the differentiation of keratinocytes.

6.3 The environment in development

**Genes and the environment in development.** In the preceding sections, development has been presented as a genetic process programmed into the genome and followed in the correct cytoplasmic setting to generate a whole organism. The use of physical cues in early vertebrate development has been alluded to, but it is also necessary to consider the wider role of the environment. Much of the phenotypic variation observed in populations arises through the influence of the environment during development as well as the stochastic influences behind *developmental noise* (q.v.). Like many developmental processes, sex determination involves both genetic and environmental factors, but this process is unusual in the extent to which these factors can be separated, and their relative predominance in different organisms (Table 6.4). The development of primary sex characteristics (the type of gamete produced) and secondary sex characteristics (the appearance of the individual) are regulated by different pathways with some common components. An important feature of sex-determination in many animals is the presence of *sex-chromosomes* (q.v.). These are asymmetrically distributed between the sexes, and **dosage compensation mechanisms** are required to redress the effects of having double the dose of certain gene products in one sex compared with the other. In *Drosophila* and mammals, females possess two X-chromosomes and males only one. In *Drosophila*, X-dosage compensation is mediated by reducing the rate of transcription for X-linked loci to 50% in females (Table 6.4). In mammals, the same result is achieved by inactivating one of the X-chromosomes (see DNA Methylation Epigenetic Regulation).

**Table 6.4:** Some examples of predominantly genetic and predominantly environmental sex-determination mechanisms

Mechanism	Description
<i>Predominantly genetic mechanisms</i>	
Mammals — chromosomal determination	In mammals, primary sex characteristics (gonad physiology) are determined by the presence of the <i>SRY</i> gene on the Y chromosome. Individuals with a Y chromosome are usually male regardless of the number of X-chromosomes, and XO individuals are female. <i>SRY</i> encodes an HMG box transcription factor which induces the expression of other regulators (including SOX9 and SF1) and may repress Wnt-4a signaling in the gonad. SF1 is required for early gonad development in both sexes, but levels decline in the female in the absence of <i>SRY</i> . Where SF1 expression is maintained in the male, it induces expression of Mullerian inhibiting substance in Sertoli cells and testosterone in Leydig cells of the testis. These two hormones promote survival of the mesonephric ducts (which become the vas deferentia) and degeneration of the paramesonephric ducts (future female genital tubes). The

Continued



Mechanism	Description
	<p>effect of SRY in abolishing Wnt-4a may be to counteract the effects of an X-linked gene <i>DAX</i>. When present in a single copy, <i>DAX</i> expression is insufficient to reverse the effects of SRY, but duplications of the <i>DAX</i> locus cause XY individuals to develop as females</p> <p>The secondary sex characteristics of mammals are determined by hormones. In this case, the female is the default state. In the absence of testosterone receptors, XY individuals develop (internal) male gonads but female secondary sexual characteristics. By the same principle, female cows can be masculinized <i>in utero</i> by testosterone production in a male twin, producing an animal known as a <b>freemartin</b></p>
<i>Drosophila</i> — the balance mechanism	<p>In <i>Drosophila</i>, primary sex is determined by the balance between autosomes and X-chromosomes, and while the Y-chromosome is required for male fertility, it is not the primary determinant of male sex characteristics. Hence XY individuals are males and XO individuals are sterile males. Individuals with a 1:1 X:autosome ratio are females and those with a 1:2 X:autosome ratio are males, those with intermediate ratios have intermediate phenotypes and are described as <b>intersexes</b>, and those with more extreme ratios show more extreme 'maleness' and 'femaleness' and are described as <b>metamales</b> and <b>metafemales</b></p> <p>The balance mechanism reflects the relative levels of various helix-loop-helix transcription factors encoded by the X-chromosome and autosomes. The X-linked genes <i>sisterless-a</i>, <i>sisterless-b</i> and <i>sisterless-c</i> encode factors which activate the gene <i>Sex-lethal</i>, whilst the autosomal gene <i>deadpan</i> encodes an inhibitor which probably acts by sequestering the above into inactive heterodimers. <i>Sex-lethal</i> is activated in females where the Sisterless factors are predominant, whereas in males, the gene is repressed due to the relatively high levels of Deadpan. The Sex-lethal protein performs two functions. Firstly it initiates an alternative RNA splicing cascade which produces a female-specific transcript of the <i>doublesex</i> gene, and secondly it represses a battery of genes which regulate the male-specific transcription rate of X-linked genes (q.v. <i>dosage compensation</i>). The RNA splicing cascade in <i>Drosophila</i> sex determination is shown in Figure 27.4 (see RNA Processing)</p>
<i>Predominantly environmental mechanisms</i>	
Turtles — temperature dependency	<p>Most reptiles use a chromosomal sex-determination mechanism similar to mammals, but in turtles and crocodiles, sex is temperature-dependent. This reflects the temperature-sensitive synthesis or activity of an enzyme which converts testosterone into estrogen. The activity of the enzyme itself may be affected by temperature, or the dependency may lie with an upstream factor which regulates enzyme synthesis</p>
Snails — substrate dependency	<p>Snails of the genus <i>Crepidula</i> form mounds containing 5–10 individuals which are initially all males. In older snails, the reproductive system breaks down and can be regenerated either as male or female depending on the substrate. In a population of young males, females arise if they are attached to dead snails, and snails attached to females become males. The default state is female — snails raised in isolation are always female</p>

**Box 6.1: Genomic nonequivalence in development**

**Random and programmed changes.** Whilst most cells of a developing organism contain the same genetic information, differences can arise in two ways. The first is by mutation, which is random with respect to both the type of change and the site at which it occurs (see Mutation and Selection). The second is a programmed change where both the type of alteration and its place in the developmental program are tightly regulated. Both processes may involve DNA gain, DNA loss or DNA rearrangement.

**Gene amplification.** Gene amplification is the process by which the copy number of a given gene or other DNA sequence is increased. Two modes of programmed amplification can be distinguished *in vivo*: whole genome amplification and selective amplification. **Whole genome amplification** increases the copy number of a given gene by increasing the number of copies of the genome in the cell. In eukaryotes, genome duplication can occur if the cell fails to undergo mitosis after DNA replication, and this is seen, for example, in mammalian liver cells, which are tetraploid, but most obviously in the secretory tissues of dipterans (e.g. *Drosophila*), where many rounds of DNA replication occur to generate the giant *polytene chromosomes* (q.v.) containing up to 1000 chromatids. **Selective amplification** is a regulated developmental process occasionally used to increase the output of high-demand gene products. Three well-characterized amplification systems have been described. Many ciliated protozoa (e.g. *Tetrahymena* spp.) and some multicellular animals (e.g. frogs) selectively amplify their rRNA genes. In both cases, this occurs in the context of a very large cell (the maturing oocyte in frogs) where there is an unusually high demand for protein synthesis, and reflects the fact that there is no other way to increase output of the rRNA product — it is not translated. In both cases, amplification involves excision of the rDNA from the genome. The protozoans possess a single rRNA gene, whereas frogs already possess up to 1000 genes in the germline. In *Tetrahymena*, the excised rDNA is duplicated by hairpin priming, generating a linear inverted repeat element which is then amplified. In frogs, the excised rDNA is circularized prior to amplification. The integrity of the genome is main-

tained during amplification, so it is likely that rounds of extra DNA replication precede excision, with recombination causing individual rDNA units to be excised. Protein-encoding genes are rarely amplified because output can be boosted at the level of protein synthesis. The chorion protein genes of *Drosophila* are an exception — the chorion proteins are required in large amounts over a very short period of time, and this is facilitated by selective gene amplification in the follicle cells. Unlike the rRNA genes, the chorion protein genes are not excised but are amplified within the genome by nested rounds of replication, to generate a series of concentric amplicons (the **onion skin model**). In other insects, multiple copies of the chorion protein genes are found in the germline, and it is unknown why this amplification should have evolved as an alternative in *Drosophila*.

**Chromatin diminution and other instances of DNA loss.** As with gene amplification, DNA loss can be selective or can involve whole copies of genome. Whole genome reduction occurs in all cells undergoing meiosis: two successive rounds of division occur without intervening DNA replication. A special incidence of DNA loss is the extrusion of the nucleus from mammalian reticulocytes, which differentiate into anucleate erythrocytes (**nulliploidy**). Selective DNA loss is seen in many lower eukaryotes and is associated with the differentiation of the somatic cells and germ cells. Generally, the germ cells retain the entire complement of DNA, whilst there is selective DNA loss or **chromatin diminution** in the somatic cell lineage. This process involves site-specific DNA cleavage and segregation. Its significance is unclear, but it appears to be an essential aspect of somatic cell development and is controlled by the regional distribution of cytoplasmic components.

**Programmed recombination.** The differentiation of a few cell types is dependent upon recombination. Programmed nonreciprocal recombination controls mating type switching in yeast and antigen switching in trypanosomes, whilst programmed site-specific recombination controls many aspects of vertebrate B-cell and T-cell differentiation (see Recombination).

**Box 6.2:** Sporulation in *Bacillus subtilis*

**Overview of sporulation pathway.** The starvation of vegetative *B. subtilis* cells induces sporulation — one copy of the genome is segregated as a forespore and encased in a spore coat whilst the other remains in the mother cell, which is eventually destroyed. This is a simple form of differentiation — a cell gives rise to two cells with distinct phenotypes and different functions. The process of differentiation is controlled primarily at the level of transcription by the synthesis of alternative  $\sigma$ -factors (components of the bacterial RNA polymerase; see Transcription) and other transcriptional regulators. The pathway to sporulation involves cascades of transcriptional regulators in each cell coordinated by criss-cross posttranscriptional regulation.

**Initiation of sporulation.** Sporulation begins when starvation induces a protein kinase cascade culminating in the phosphorylation of the transcriptional regulator SpoOA. Activated SpoOA induces the transcription of a number of genes which control entry into the sporulation pathway by interacting with the vegetative  $\sigma$ -factors  $\sigma^A$  and  $\sigma^H$ . Two novel  $\sigma$ -factors,  $\sigma^F$  and  $\sigma^E$ , are synthesized as part of this initial response. There is also a switch from medial to polar septation, with rings of the protein FtzZ forming at polar positions at both ends of the cell (q.v. *bacterial cell cycle*). A septum forms at only one end of the cell and the small compartment distal to the septum becomes the forespore. It is unknown how the cell chooses which end is to become the spore, but the formation of a *single* spore is  $\sigma^E$ -dependent. Chromosome segregation occurs after septation, so the chromosome must be translocated across the septum into the forespore and this is carried out by the SpoIIIE protein.

**A cascade of  $\sigma$ -factors.** The process of sporulation can be divided into a number of morphologically distinct stages which are dependent upon the expression of specific sets of genes. Gene expression is coordinated by a hierarchical cascade of transcriptional regulatory proteins. The advent of spore differentiation involves activation of  $\sigma^F$ , specifically in the forespore.  $\sigma^F$  is initially distributed uniformly as an inactive complex with the protein SpoIIAB. SpoIIAB is a kinase which phosphorylates the protein SpoIIAA; its activity is antagonized by the membrane-bound phosphatase SpoIIIE. In its unphosphorylated form, SpoIIAA can interact with the  $\sigma^F$ /SpoIIAB complex and displace the  $\sigma$ -factor; thus the concentration of dephosphorylated SpoIIAA is the principle determinant of  $\sigma^F$  activity. It is thought that SpoIIIE is displayed on both sides of

the septal membrane, but because of the small volume of the forespore, its relative concentration is higher in the smaller compartment and this shifts the equilibrium in favor of SpoIIAA dephosphorylation, resulting in displacement and activation of  $\sigma^F$  specifically in the forespore.

The next stage concerns the activation of  $\sigma^E$  specifically in the mother cell. Like  $\sigma^F$ ,  $\sigma^E$  is initially distributed uniformly in an inactive form, but its inactivity stems from the fact that it is synthesized as a preprotein, and it must be cleaved by a protease. The protease is a transmembrane receptor encoded by the *spoIIIGA* gene and its protease activity is stimulated by the ligand SpoIIIR, which is synthesized specifically in the forespore under the regulation of  $\sigma^F$ . Thus  $\sigma^F$  in the forespore is required for the activation of  $\sigma^E$  in the mother cell. SpoIIIR is secreted from the forespore into the intercompartmental space and activates the protease which cleaves  $\sigma^E$ .  $\sigma^E$  is not activated in the forespore itself because it is completely degraded, apparently under the control of the SpoIIIE protein.

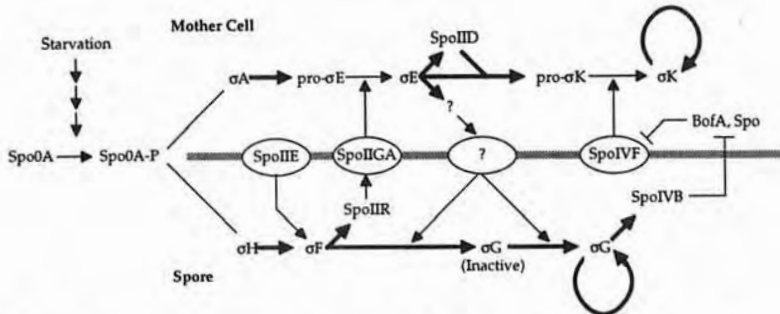
$\sigma^E$ -dependent genes are transcribed in two phases: the primary phase genes include *spoIIID*, which encodes a transcriptional regulator required in addition to  $\sigma^E$  for the transcription of the secondary phase genes. Among the secondary  $\sigma^E$ -dependent genes is *sigK*, which encodes another  $\sigma$ -factor,  $\sigma^K$ . The regulation of  $\sigma^K$  expression is complex, for the gene is interrupted by a cryptic prophage called *skin* which must be excised. This process is carried out by a site-specific recombinase encoded within *skin*, SpoIVCA, whose expression is dependent upon  $\sigma^E$  and SpoIIID. Once the two halves of the *sigK* locus have been rejoined, the gene is transcribed under the control of  $\sigma^E$  and SpoIIID and translated to yield an inactive preprotein in a manner similar to pro- $\sigma^E$ . SpoIIID can act as both an activator and repressor of  $\sigma^E$ - and  $\sigma^K$ -dependent genes.

In the forespore, further genes regulated by  $\sigma^F$  are induced about 1 h after  $\sigma^F$  itself becomes active. These include *spoIIIG*, which encodes another  $\sigma$ -factor,  $\sigma^G$ , and the delay appears to reflect some dependence upon  $\sigma^E$  function in the mother cell for transcription (although the putative signaling molecules involved have not been identified).  $\sigma^E$  is also required for  $\sigma^G$  activation, although again the mechanism is unknown.

$\sigma^G$  directs expression of the late sporulation genes in the forespore, one of which is *spoIVB*. This encodes a signaling protein which is involved in the activation of  $\sigma^K$  in the mother cell. Like pro- $\sigma^E$ , pro- $\sigma^K$  is activated by a membrane-bound protease (in

this case SpoIVFB) which is initially inactive and becomes activated by signaling from the forespore. However, unlike SpoIIIGA, which is inactive in its default state, SpoIVFB is constitutively active, but maintained in an inactive complex by inhibitory proteins encoded by the *spoIVFA* and *bofA* genes. The role of SpoIVB is thus to inhibit the inhibitors rather than to activate the inert protease directly.

$\sigma^G$  and  $\sigma^K$  control the expression of late genes in the forespore and mother cell respectively, and both are able to positively regulate their own genes. In addition,  $\sigma^K$ -dependent genes are transcribed in two phases. The primary phase includes *gerE*, which encodes a transcriptional regulator which activates some and represses other  $\sigma^K$ -dependent genes.



Summary of the regulatory network controlling sporulation in *B. subtilis*. Transcriptional regulation is shown by thick arrows and posttranslational regulation by thin arrows. Integral membrane proteins are shown as circles.

### Box 6.3: The life cycle of *Dictyostelium discoideum*

**Overview of the life cycle.** In its vegetative state, *D. discoideum* exists as unicellular, haploid cells termed **myxamoebae** which reproduce asexually. Remarkably, when the single cells deplete their nutrient supply, they congregate into streams which migrate to a central point where they form a multicellular aggregate. The cells in the aggregate differentiate, forming **prespore cells** and **prestalk cells**. The prestalk cells form at the tip of the aggregate and the prespore cells at its base. The aggregate may tip onto its side and migrate *en masse* as a **slug** (also called a **pseudoplasmodium** or **grex**). When the slug moves from darkness into light, it differentiates into a **fruiting body** comprising a body of spore cells upon an elevated stalk. The spore cells disperse to form new myxamoebae.

**Molecular basis of aggregation.** The decision to aggregate is based on nutrient availability and cell density. The myxamoebae can measure cell density by monitoring levels of a secreted protein termed **prestarvation factor (PSF)**. Starved cells release cAMP into their surroundings, and if the levels of both cAMP and PSF are high enough, surrounding cells will begin to aggregate. Initially, cAMP is

released by a small number of individual cells which have depleted their food supply. Neighboring cells respond by moving towards the source of the signal, and releasing cAMP themselves, thus relaying the signal to more peripheral cells. Once an individual cell has responded in this manner, there is a short resolution period before it can respond again. Hence, the signal is propagated as discontinuous pulses emanating from the original source. As they migrate, the cells begin to synthesize new cell adhesion molecules, initially enabling them to form streams and eventually the larger aggregate (which may contain up to  $10^5$  cells).

**Cellular differentiation in the aggregate.** The cells of the aggregate differentiate into prestalk and prespore cells which sort themselves so that the prespore cells are located in the body (which becomes the posterior of the slug) and most prestalk cells are located in the tip (which becomes the anterior of the slug). The posterior of the slug also contains a scattering of prestalk cells, which are known as **anterior-like cells (ALCs)**. The process of differentiation is regulated by cAMP and a family of related lipophilic molecules collectively termed **differentia-**

*Continued*



**tion inducing factor (DIF).** The cAMP is required for prespore cell formation, as cAMP induces prespore genes and stabilizes their mRNAs. DIF is required for prestalk cell differentiation and induces the expression of prestalk genes. Cell sorting occurs by several mechanisms: prestalk cells are thought to sort to the anterior of the slug because they continue to migrate towards the initial source of cAMP, whilst the prespore cells do not. The differential expression of extracellular matrix proteins controls the distribution of different subclasses of prestalk cells. Most prestalk cells express the *ecmA* gene and sort to the anterior of the slug — those which express the gene strongly are termed PstA cells and occupy the anterior tip; those which express it weakly are termed PstO cells and occupy a more posterior domain. Within the PstA cells, a cone-shaped group of cells also express the *ecmB* gene. These PstAB cells will eventually initiate stalk formation. Finally, a group of cells expressing *ecmB* alone are found scattered through the spore cells. The initial choice as to which cells become prespore and which prestalk cells may depend upon the stage of the cell cycle stage at which growth arrest occurred and differentiation commenced: cells arrested in G<sub>1</sub> synthesize fewer cell adhesion molecules than those arrested in the late cell cycle and consequently migrate further before sticking to other cells.

**Culmination.** The construction of the fruiting body is termed **culmination** and involves the reorganization of the different cell populations in the slug and their terminal differentiation into stalk and spore cells. Slug migration ceases upon illumination or in low humidity, both of which cause ammonia (which

is produced in great amounts by the migrating slug) to diffuse away more rapidly. Ammonia depletion releases the repression of many genes involved in differentiation by allowing the production of cAMP. This de-represses differentiation genes through the activity of protein kinase A, which is thought to phosphorylate (and inactivate) one or more repressor proteins. The act of culmination involves the formation of an elevated stalk, which is accomplished by the migration of PstA cells towards the centre of the early culminant and their conversion into PstAB cells by induction of the *ecmB* gene. The PstAB cells push down through the prespore cells into the base and elevate the spores so they can be dispersed. In doing so, they become vacuolized and synthesize cellulose; they die and form a rigid stalk. The PstB and PstO cells also migrate and differentiate: the former migrate downwards and form the base of the fruiting body, and the latter, together with remaining PstA cells, migrate outwards and form the case of the spore body and the upper and lower cups.

**Developmental mutants.** Many genes involved in *D. discoideum* development have been identified by mutagenesis. Such mutants may be unable to aggregate because they lack the ability to send, receive or respond to the cAMP signaling. Other mutants are blocked at later stages because they are unable to migrate, differentiate or respond to signaling by DIF or ammonia. Still others differentiate abnormally, e.g. by defaulting to a particular developmental fate, by disrupting the normal prespore:prestalk ratio, or by entering the final differentiation pathway directly. There may be up to 400 genes specifically involved in this developmental process.

#### Box 6.4: Early events in *Xenopus laevis* development

**Overview of axis formation.** The frog *Xenopus laevis* and other amphibians show both autonomous and conditional specification in early development. The egg is asymmetrical in both molecular and gross physiological terms, and thus an **animal-vegetal axis** can be discerned prior to fertilization. The future dorso-ventral axis of the embryo is determined by a physical cue — the point of sperm entry. Fertilization causes cortical rotation, and the mixing of cytoplasm opposite the point of sperm entry induces dorsalizing signals in the vegetal cytoplasm, in a region known as the **Nieuwkoop center**. During cleavage, the vegetal cells induce overlying animal cells to form mesoderm. The cells in the Nieuwkoop center

induce dorsal mesoderm, which has organizer activity, whilst the ventral vegetal cells induce ventral mesoderm. The **organizer** is so called because it initiates gastrulation and specifies the anteroposterior axis of the embryo, and can do so when transferred to an ectopic site. The organizer also has a number of further specialized properties. Firstly, it can differentiate into dorsal mesoderm structures (notochord, somites, head mesenchyme); secondly, prior to gastrulation, it can signal the adjacent ventral mesoderm and induce lateral and intermediate mesoderm structures; thirdly, during gastrulation, it can dorsalize the overlying ectoderm, i.e. induce it to form neural plate, and in so doing impart its own antero-

*Continued*



posterior positional information to the neural tissue to define the anteroposterior neuraxis. The molecular bases of these interactions are discussed below and summarized in the accompanying figure.

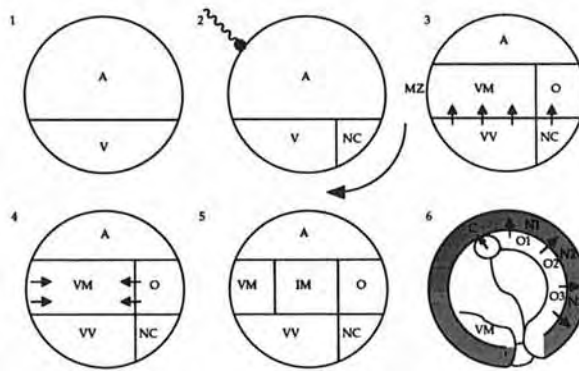
**Establishment of the Nieuwkoop center.** The Nieuwkoop center is established by cortical rotation, and this appears to involve the accumulation of the transcription factor  $\beta$ -catenin in the dorsal part of the egg. The protein is initially synthesized throughout the egg, and it is thought that fertilization locally induces the activity of the  $\beta$ -catenin inhibitor glycogen synthase kinase-3, although the precise mechanism is unclear. The activity of  $\beta$ -catenin is enhanced by the TGF- $\beta$  signal transduction pathway, and it is thought that the vegetally localized maternal protein Vg-1 may play this role *in vivo*. Hence, active  $\beta$ -catenin would be restricted to the future dorsal side of the egg by the act of fertilization, and would be stimulated by a vegetal-specific cytoplasmic determinant. This would be sufficient to localize the Nieuwkoop center to the dorsiventral quadrant of the egg.  $\beta$ -catenin induces the expression of a transcription factor termed *Siamois* which presumably regulates the transcription of genes central to the function of the Nieuwkoop center.

**Mesoderm induction.** The vegetal cells induce overlying animal cells in the marginal layer to become mesoderm, and molecules of both the fibroblast growth factor (FGF) and transforming growth factor- $\beta$  (TGF- $\beta$ ) families have been implicated in this signaling pathway. Both types of molecule are present in the *Xenopus* embryo, and inactivation of each type of receptor by the overexpression of dominant negative receptor molecules severely disrupts mesoderm development. Inactivation of the receptor for the TGF- $\beta$ -like factor activin completely abolishes mesoderm formation, whilst similar experiments involving FGF receptors result in the loss of ventral mesoderm derivatives. These results suggest that TGF- $\beta$ -like factors are instrumental in the formation of mesoderm, but that FGFs are required for the specification of ventral mesoderm. Vg-1 is a member of the TGF- $\beta$  family, and has been shown to be required for the function of the Nieuwkoop center. *In vitro*, increasing doses of Vg-1 induce the expression of progressively more dorsal mesoderm markers in isolated animal tissue, although *in vivo*,  $\beta$ -catenin is also required, suggesting that Vg-1 may act in concert with an unknown product activated by the *Siamois* transcription factor. FGF signaling may thus modulate the response to Vg-1 in the ventral side of the embryo to specify ventral mesoderm-specific cell fates. However, the ubiquitous signaling protein BMP-4 appears to play a direct role in

ventral mesoderm specificity by inducing ventral mesoderm-specific transcription factors (see below). The induction of ventral mesoderm may therefore involve the combined effects of Vg-1, FGFs and BMP-4. These molecular interactions are specific to *Xenopus*. Many of the genes involved are conserved in the mouse but appear to play different roles, as shown by the effects of *gene knockout* experiments (q.v.).

**Molecular control of the organizer.** Signals from the Nieuwkoop center induce the synthesis of a number of transcription factors specifically in the organizer, and these activate genes concerned with dorsal mesoderm development, signaling into the adjacent ventral mesoderm, the movements of gastrulation and subsequent signaling to the overlying ectoderm. A number of transcription factors have been localized specifically to the organizer, including XANF1, Pintallavis, Lim1 and Goosecoid, and their overlapping expression patterns appear to subdivide the organizer into functionally specific domains. Dorsalizing signals from the organizer induce adjacent ventral mesoderm to express intermediate mesoderm markers. These signals appear to work by antagonizing the activity of BMP-4 (which is expressed throughout the embryo and is a potent ventralizing signal, inducing the activity of several transcription factors — Xvent-1, Xom and Vox — which not only activate downstream genes concerned with ventral mesoderm differentiation, but also repress the activity of dorsalizing factors such as Goosecoid). The organizer synthesizes at least three secreted proteins — Noggin, Chordin and Follistatin — whose role appears to be the repression of BMP-4 signaling. In the case of Noggin and Chordin, this is achieved by directly binding to BMP-4, whilst Follistatin binds to BMP-7, which is upstream of BMP-4. Opposing gradients of BMP-4 and its inhibitors across the embryo thus specify a range of intermediate mesoderm phenotypes adjacent to the organizer.

**Neural induction and specification of the anteroposterior neuraxis.** The molecular basis of neural induction is not understood, but some of the molecules which impart positional information to the neural plate have been determined. Noggin, Chordin and Follistatin cause the overlying ectoderm to form anterior neural structures (forebrain). It is thought that a graded posterior signal causes caudalization, and that this may be FGF. The very first cells to enter the blastopore synthesize a secreted protein called Cerberus which specifies the anterior-most structures of the head.



1. Preexisting animal-vegetal asymmetry in the egg (A, V). 2. Fertilization causes cortical rotation and establishes Nieuwkoop center (NC) by regional activation of  $\beta$ -catenin. 3. Vegetal cells induce overlying animal cells in the marginal zone (MZ) to become mesoderm; Nieuwkoop center induces dorsal mesoderm (organizer, O), probably through synergy of Vg-1 and products activated by the transcription factor Siamois; ventral vegetal cells (VV) induce ventral mesoderm (VM) in a pathway that involves FGF factors and BMP-4. 4. Lateral signals from the organizer (Chordin, Follistatin, Noggin) interact with BMP-4 in the ventral mesoderm. 5. Graded repression of BMP-4 establishes a range of intermediate mesoderm types (IM). 6. The organizer initiates gastrulation. Intermediate mesoderm is displaced to the side of the embryo. The most anterior organizer cells secrete the protein Cerberus which induces anterior head structures such as the cement gland (C). The remaining organizer tissue induces neural plate (N); the cells which migrate through the blastopore first (O1) become anterior mesoderm and induce anterior neural plate (N1). The nature of the neuralizing signal is not known, but it may activate protein kinase C and adenylate cyclase in the ectoderm. Chordin, Follistatin and Noggin induce anterior neural fates, whilst a second signal, possibly FGF, caudalizes the neurectoderm to generate posterior neural structures (N2, N3).

#### Box 6.5: Syncytial specification of the body axes in *Drosophila*

**Overview.** As discussed in the main text, early insect development occurs in the context of a syncytium which becomes cellularized prior to gastrulation. In *Drosophila*, the anteroposterior body axis is specified by three sets of **maternal genes** (genes expressed in maternal cells and whose products are transported into the egg) which establish anterior, posterior and terminal organizing centers in the syncytial embryo. These genes encode regulatory factors which set up opposing morphogen gradients and activate downstream **zygotic genes** (genes expressed in the embryo during development, rather than in the surrounding maternal cells) in a concentration-dependent manner. Progressive hierarchical activation of different classes of zygotic

genes divides the syncytium into stripes defined by domains of gene expression. These form the basis of the segments which become patterned by later-acting genes.

**Maternal genes of the anterior and posterior groups.** There are four primary genes which control the anterior and posterior organization of the fly; these are *bicoid* and *hunchback*, which specify anterior structures, and *nanos* and *caudal*, which specify posterior structures. In each case, the mRNA is synthesized in the nurse cells and is placed in the future anterior of the egg.

The 3' untranslated region (UTR) of the *bicoid* mRNA is attached to the anterior cytoskeleton by

*Continued*

the products of the *exuperantia* and *swallow* genes. Translation of *bicoid* mRNA is initially prevented by its very short polyadenylate tail, but after fertilization, the products of the *cortex*, *groucho* and *staufer* genes facilitate further polyadenylation and allow synthesis of Bicoid protein. The *nanos* mRNA is transported to the posterior of the egg by the products of the *oskar*, *staufer*, *tudor*, *vasa* and *valois* genes and its 3' UTR is attached to the posterior cytoskeleton. The mRNAs for *hunchback* and *caudal* are not localized and disperse uniformly throughout the egg.

Bicoid is a homeodomain-containing protein which binds to both DNA and RNA. It activates transcription of the *hunchback* gene in the zygotic nuclei and represses translation of the maternal *caudal* mRNA in the anterior of the egg. Conversely, Nanos protein represses translation of the maternal *hunchback* mRNA in the posterior of the egg (it is not a transcription factor). The Pumilio protein binds to the *hunchback* mRNA and provides a docking site for Nanos. The interactions between the products of the four genes generate an anterior-posterior gradient of Bicoid and Hunchback and a posterior-anterior gradient of Nanos and Caudal.

**Maternal genes of the terminal group.** The genes of the terminal group specify the **acron** and **telson**, the structures which form the extremities of the anteroposterior axis. The key players in this group are the *torso* and *torso-like* genes. Torso is an integral membrane receptor of the receptor tyrosine kinase class. The mRNA for *torso* is ubiquitous in the egg, and the receptor itself is evenly distributed throughout plasma membrane. Terminal specificity is controlled by the ligand, *Torso-like*, which is secreted from the anterior and posterior follicle cells only. Activation of Torso by its ligand induces a signaling cascade involving Ras, Raf and MAP kinase (see Signal Transduction), culminating in the activation of an unknown transcription factor which induces the transcription of the gap genes *huckebein* and *tailless*. These encode transcription factors which activate genes specifying terminal structures. Acron-specific genes also require Bicoid protein. Constitutively active Torso mutants are *epistatic* (q.v.) to the anterior and posterior group genes — the phenotype is 'hyperterminal', with large acron and telson but no body in between.

**Anteroposterior zygotic genes.** The maternal genes *bicoid*, *hunchback* and *caudal* encode transcription factors which control the expression of zygotic genes involved in the formation of the segmental body plan. Five classes of zygotic gene can be recognized.

**Gap genes** divide the embryo into broad regions corresponding to several parasegments, and loss-of-function mutations cause the loss of these contiguous parasegments in the larva (a **parasegment** comprises the anterior portion of one segment and the posterior portion of the adjacent segment; the embryonic *Drosophila* body plan is initially divided into parasegments by domains of gene expression even though the recognizable morphological structures formed are segments).

**Primary pair rule genes** are regulated by the gap proteins and are expressed in alternative parasegments. These structures are deleted in loss-of-function mutants.

**Secondary pair rule genes** are the same as above but act later than the primary pair rule genes and require the primary pair rule proteins as well as the gap proteins to establish their expression domains. Together, the pair rule genes define the parasegments of the embryo.

**Segment polarity genes** are expressed in serial parasegments and in the same relative position in each one. They are responsible for the specification of cell types within the parasegments, and loss-of-function mutations cause the loss of parts of each parasegment and replacement by the mirror image of the remaining portion.

**Homeotic selector genes** are expressed in specific domains encompassing one or more parasegments and specify segmental identity, i.e. what structures will develop within each segment.

The gap genes and pair rule genes thus interpret the axis-defining maternal cues and convert them into a repeating segmental structure, whereas the segment polarity genes and homeotic genes confer positional values upon the segments. The gap and pair rule genes are discussed in more detail below, the segment polarity and homeotic genes in Box 6.8.

**The gap genes.** The gap genes are initially expressed in wide domains, but become localized as the gradient of Hunchback protein is stabilized. Bicoid and Hunchback activate *giant* and *Krüppel* in the anterior of the embryo and repress *knirps*. Caudal protein is responsible for the transcription of *giant* and *knirps* in the posterior of the embryo. The three head-specific gap genes *orthodenticle*, *empty spiracles* and *buttonhead* are expressed only where there is a very high concentration of Bicoid protein. The *tailless* and *huckebein* genes are regulated by transcription factors activated by Torso signalling at the termini of the embryo. The resulting pattern of gap gene expression is shown below. The products of the gap genes are all themselves transcriptional regulators and they have two functions

— to feed back and regulate each other, and to regulate the expression of the downstream pair rule genes. This regulation appears to be primarily negative in both cases. Once the mRNAs are translated, the proteins diffuse laterally and repress the expression of neighboring gap genes so that sharp boundaries of transcription are established.

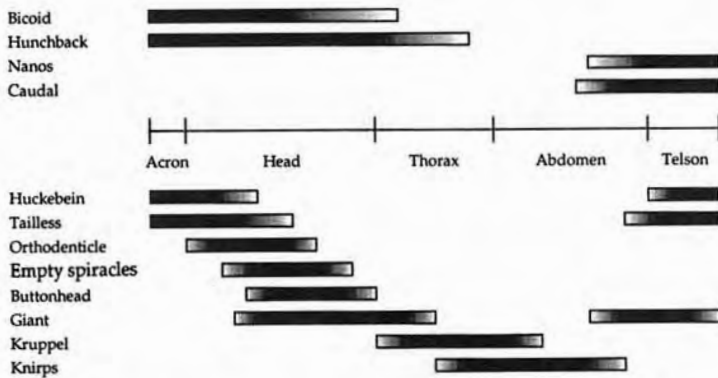
**The pair rule genes.** The pair rule genes are expressed during cellularization of the blastoderm and the mRNA appears in stripes along the antero-posterior axis. Each stripe is regulated individually by the local gap gene products. The different pair rule genes respond to different gap protein concentrations, so that adjacent stripes of cells express different combinations of pair-rule gene products in a repeating pattern. There are three primary pair rule genes, *even-skipped*, *hairy* and *runt*, whose expression is essential for setting up the metameric pattern. The expression patterns of the primary pair rule genes are stabilized by cross-regulation and help to specify the patterns generated by the secondary pair rule genes, which include *fushi tarazu* and *odd-skipped*. The control of the primary pair rule gene *even-skipped* (*eve*) has been particularly well characterized and appears to be predominantly negative. The *eve* mRNA is expressed where the levels of most of the gap proteins are low, i.e. at the boundaries of the gap gene domains. As shown below, each stripe of *eve* expression is under the control of several gap genes, and this is reflected by the modular structure of the *eve* promoter, with separate regulatory elements controlling the expression of stripes 4–6, stripe 1, stripe 3 and stripes 2 and 7. The control elements for stripe 2 have been mapped in detail, and this region of the promoter contains binding sites for Bicoid, Hunchback, Giant and Kruppel. The binding sites for the positive regulators (Bicoid and Hunchback) often overlap those for the negative regulators (Giant and Kruppel), indicating that competition between the various regulators is an important mechanism in stripe positioning.

**Dorso-ventral polarity.** The dorso-ventral polarity of the fly is established after cellularization by a gradient of nuclear Dorsal protein. Dorsal is synthe-

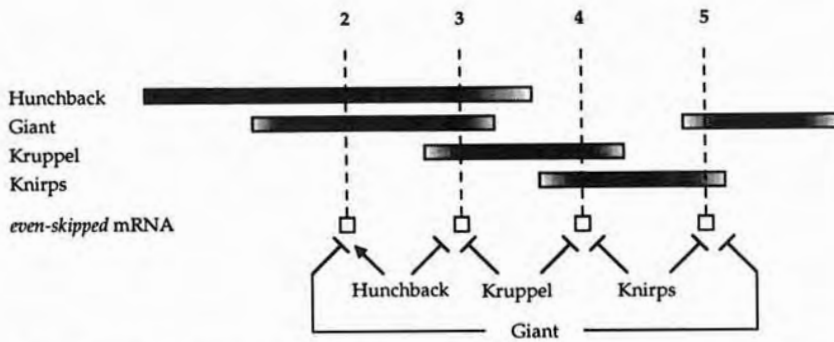
sized throughout the fly, but is translocated to the nucleus specifically on the ventral side of the embryo. On the dorsal side of the embryo, Dorsal is associated with a second protein, Cactus, which masks the Dorsal nuclear localization signal. The degradation of Cactus is induced by phosphorylation, which occurs through a signal transduction cascade emanating from the ventral side of the embryo. Dorsal is homologous to the vertebrate transcription factor NF- $\kappa$ B and cactus to the inhibitor protein I- $\kappa$ B; the signaling pathway to activation is also conserved. The pathway to dorso-ventral polarity begins with the oocyte nucleus itself, which dorsalizes the overlying follicle cells through the products of the genes *gurken* and *cornichon*. Gurken is homologous to mammalian epidermal growth factor (EGF) and activates the receptor Torpedo (which is homologous to the mammalian EGF-receptor) in the follicle cell membrane. Activation of the Torpedo receptor induces a signaling cascade which represses transcription of the genes *nudel*, *windbeutel* and *pipe*. In the ventral side of the embryo, where these three genes are active, their products form a membrane complex whose function is to activate three serine proteases secreted by the embryo and encoded by the *easter*, *snake* and *gastrulation defective* genes. The activation of Gastrulation defective initiates a protease cascade: Gastrulation defective cleaves Snake, which then cleaves Easter, which in turn cleaves the product of another dorsal group gene *spatzle*, the ligand for the Toll receptor. Toll is distributed throughout the egg membrane, but because the protease cascade activates Spatzle only on the ventral side, Toll signaling is also ventral-specific. Toll activates the protein tyrosine kinase Pelle in a signal transduction pathway which involves the product of the *tube* gene. Pelle is the enzyme which phosphorylates Cactus and induces its degradation, allowing Dorsal protein to be taken into the ventral nuclei. In the nucleus, Dorsal acts as both an activator and a repressor of transcription. It activates the *rhomboid*, *snail* and *twist* genes whilst repressing the dorsalizing genes *zerknüllt*, *tolloid* and *decapentaplegic*.

*Continued*





Distribution of the products of the key maternal egg polarity genes (upper) and zygotic gap genes (lower) in the syncytial *Drosophila* embryo after 12 nuclear divisions.



Regulation of *even-skipped* transcription by zygotic gap gene products. The *eve* mRNA stripes are positioned where gap protein levels are lowest, except for stripe 2 which is positively regulated by Bicoid and Hunchback.

### Box 6.6: Vulval specification in *Caenorhabditis elegans*

**The vulval lineage.** The vulva of the nematode worm *Caenorhabditis elegans* is an opening in the ventral hypodermis (the external layer of the animal, equivalent to the epidermis of mammals) through which eggs are delivered from the overlying uterus. It comprises 22 cells which are derived from three hypodermal cells over three generations of cell divisions, whilst other hypodermal cells divide only once and produce only hypodermal cells. The hypodermal cells are induced to form the vulva by a single cell in the overlying gonad which is known as the **anchor cell**. If the anchor cell alone is destroyed, all the ventral hypodermal cells divide once and form hypodermis. If all gonadal cells except the anchor cell are destroyed, the vulva still

forms. If the anchor cell is moved mechanically, neighboring hypodermal cells can respond to it and form a vulva, but only six hypodermal cells in total are competent to do so.

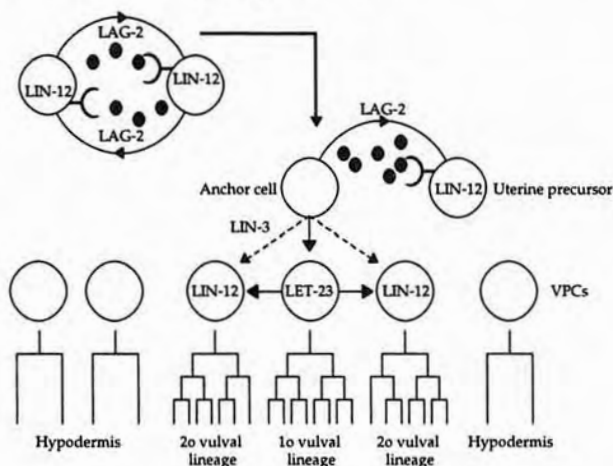
**Specification of vulval precursor cell phenotype.** Genes involved in vulval specification have been identified from mutagenesis screens for vulvaless and multiple vulva phenotypes, demonstrating the power of genetics in simple systems to define developmental pathways. The vulval model provides examples of several types of cell-cell interactions. Initially, two equivalent cells are competent to become the anchor cell, and reciprocal signaling between them, involving the signaling molecule LAG-2 and the receptor LIN-12, results in lateral



inhibition of one of the cells. If this signaling is abolished by mutating either *lag-2* or *lin-12*, both cells become anchor cells. In constitutive signaling mutants with dominant gain of function *lin-12* alleles, both cells become uterine tissue precursors.

Once the anchor cell is specified it induces the underlying hypoblast cells to form the vulva. The central cell becomes the central vulval cells and the two flanking cells the lateral vulval cells. A particularly interesting subgroup of genes encode components of a signaling pathway homologous to the RTK-Ras-Raf pathway in mammals (see Signal Transduction). The ligand, encoded by the *lin-3* gene, is a member of the EGF family of growth

factors and is required in the anchor cell. Other components, encoded by *let-23*, *sem-5*, *let-60* and *lin-45*, are required in the responding hypoblast cells and correspond to the receptor, adaptor protein, Ras and Raf. Homologs of MEK and MAP kinase have also been identified. It is possible that LIN-3 is a morphogen which evokes different responses from the immediately underlying vulval precursor cell and the lateral cells. The central cell also secretes an additional signal which induces expression of the *lin-12* gene in the flanking cells which prevents them differentiating into central vulval cells.



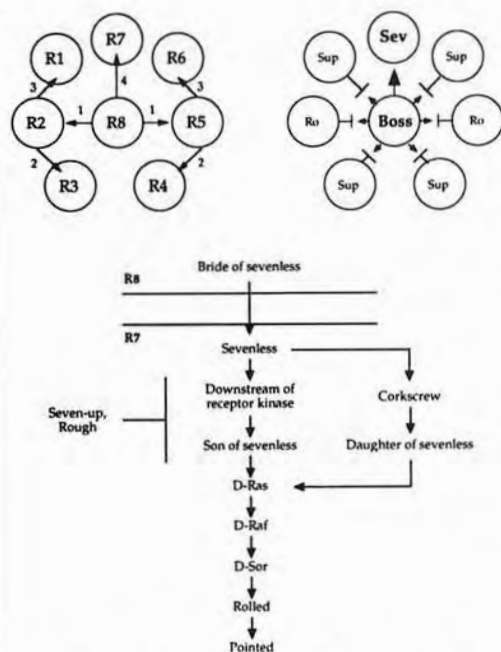
Summary of vulval specification in *C. elegans*. Lateral inhibition establishes the anchor cell which then induces the underlying hypodermis to differentiate into vulval cells. The primary and secondary cell lineages may be specified by a gradient of LIN-3, or by lateral signaling from the central vulval cell.

#### Box 6.7: Eye development in *Drosophila melanogaster*

**Photoreceptor cells.** The *Drosophila* compound eye comprises about 800 functional units termed **ommatidia**, each composed of 20 cells, eight of which are photoreceptors. The differentiation of the eye cells involves a cascade of instructive inductions at the single cell level, with lens representing the default (uninduced) state. The eye originates as a sheet of cells and the differentiation of the ommatidia occurs in a predictable sequence through progressive cell-cell interactions. The first cell to become determined is the central R8 cell. The mechanism is unclear, but may involve signaling molecules such

as Hedgehog and Decapentaplegic in the morphogenetic furrow which sweeps across the eye primordium. Signals emanating from R8 induce adjacent cells along the anteroposterior axis to differentiate into functionally equivalent R2 and R5 cells. These cells then signal the adjacent undifferentiated cells on each side to become the R1, R3, R4 and R6 cells, which again are functionally equivalent. The final event is the specification of the R7 photoreceptor cell, a process whose mechanism has been determined in detail. All other cells become lens.

*Continued*



*Drosophila* eye development. (left) Order of inductive reactions which specify the fates of the photoreceptor cells. (right) R7 photoreceptor specificity is controlled by regulators in the other cells which block Boss-Sev signal transduction. (bottom) The Boss-Sev signal transduction pathway to R7 differentiation (see Signal Transduction).

**Specification of the R7 photoreceptor.** A number of genes have been identified by their abnormal R7 cell developmental phenotype. The *sevenless* (*sev*) and *bride of sevenless* (*boss*) genes appear to func-

tion at the start of the pathway, and mosaic analysis has shown that *sev* is required in the future R7 cell whilst *boss* is required in the inducing R8 cell. Boss is a signaling molecule which activates the Sevenless receptor tyrosine kinase initiating a signaling cascade in the R7 cell. Many of the components of this cascade have been identified — Drk (Downstream of receptor kinase) is an adaptor protein with SH2 and SH3 domains, Sos (Son of sevenless) is a guanosine nucleotide exchange factor which acts on Ras. Ras, Raf and MAP kinase homologs have also been identified (although mutations in these genes are pleiotropic and lethal because the same proteins mediate many other signaling cascades in the fly), and a target transcription factor, Pointed. There also appear to be alternative branches to the pathway, involving the protein tyrosine phosphatase Corkscrew and its substrate Dos (Daughter of sevenless) which also activates Ras.

The decision to become R7 thus reflects the presence of a receptor and signaling pathways which can respond to the Boss signal displayed by the R8 cell. However, all the photoreceptor cells initially synthesize Sevenless, which means that there must be a mechanism for inhibiting the pathway in the other cells. A transcription factor called Seven-up probably mediates this function in R1, R3, R4 and R6 cells because loss-of-function mutations cause these four cells to differentiate into R7-like cells. Expression of the Rough transcription factor in R2 and R5 cells is necessary for *sup* expression in R3 and R4, and it is likely that similar mechanisms induce *sup* expression in R1 and R6, and block the Boss-Sev signal transduction pathway in the R2 and R5 cells themselves.

#### Box 6.8: Segmentation and segment identity in *Drosophila*

**The segment polarity genes.** The maternal genes, and the zygotic gap and pair rule genes of the anteroposterior axis all encode regulators of gene expression (either transcriptional or translational) because development up to this point has occurred in the context of the syncytium and the regulators are free to diffuse and interact with the nuclei (see Box 6.5). Cellularization occurs at the time pair rule genes are expressed, and downstream processes thus occur in a multicellular environment. The segment polarity genes therefore encode signaling proteins as well as transcription factors.

The segment polarity genes have two functions — they maintain and reinforce the metameric pattern of

parasegments by reciprocal signaling and they establish the fate of the individual cells within each parasegment. The first process is understood in detail in the case of the cells flanking the parasegment borders. Initially, *engrailed* expressing cells arise at the anterior border of each parasegment under the control of either *even-skipped* or *fushi tarazu*. Cells expressing *wingless* arise at the posterior border of each parasegment, where *eve* and *ftz* are not expressed, thus *wingless* may be regulated by Odd-paired or another pair rule protein. Once established, the expression of *wingless* and *engrailed* is maintained by the reciprocal signaling pathway illustrated below. Secreted Wingless

*Continued*

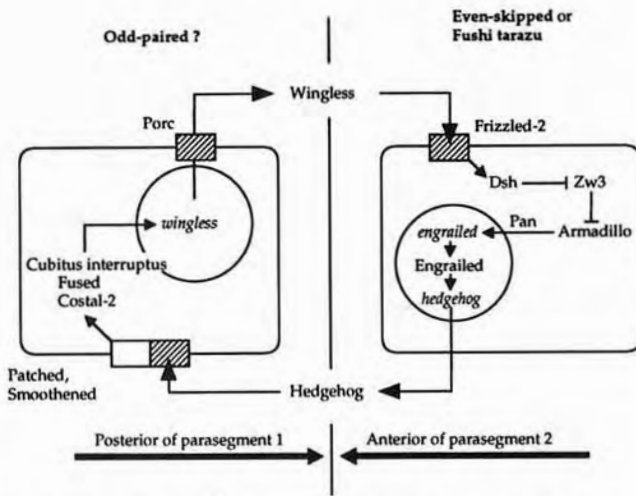
activates the Frizzled receptor and intracellular signaling results in the activation of a transcription factor, Armadillo, which activates transcription of the *engrailed* gene. Engrailed induces the transcription of *hedgehog*. This encodes a signaling molecule which interacts with the Patched receptor on the neighboring cell, releasing the signaling protein Smoothened from inhibition. Signaling by Smoothened through Cubitus interruptus induces the transcription of the *wingless* gene, resulting in the secretion of Wingless protein from the cell.

The second process is less well understood. Under the control of different pair rule genes, the segment polarity genes are expressed in specific cells in the context of each parasegment. There they may regulate the genes which cause the differentiation of each cell into a regionally appropriate cell type. However, the Wingless and Engrailed cells described above appear to play a predominant role in this process, suggesting that Wingless and Engrailed, as well as establishing the parasegmental boundaries, propagate graded signals across the parasegment which specify individual cell fates.

**Homeotic selector genes.** The homeotic selector genes provide the segments with positional information, i.e. they govern the individual development of segments to produce regionally specific structures from a common group of cell types. Consequently, mutations in these genes have the effect of converting one body part into the likeness of another, a **homeotic transformation**. Most of the *Drosophila* homeotic genes are found in two clusters on chromosome 3, termed the **Antennapedia complex** (ANT-C, which contains the genes *labial*, *Antennapedia*, *Sex combs reduced*, *Deformed* and *proboscipedia* and specifies head and thoracic segments) and the **Bithorax complex** (BX-C, which contains the genes *Ultrabithorax*, *abdominal A* and *Abdominal B* and specifies abdominal segments). Together these comprise the **homeotic complex, HOM-C**. The homeotic genes are expressed in specific regions of the embryo, in some cases corresponding to a particular segment or parasegment, in others spanning several parasegments (these

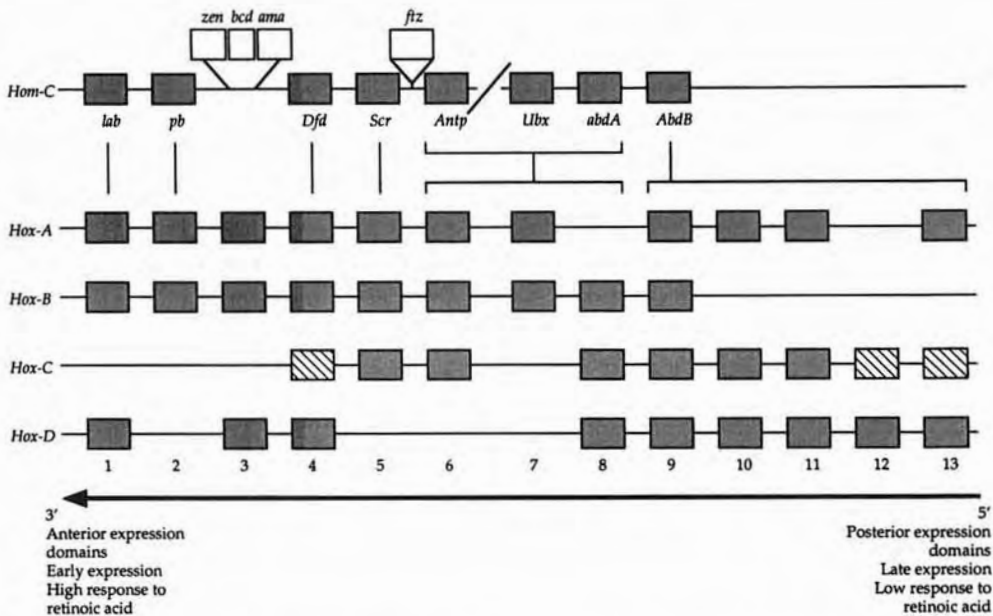
expression parameters are shown in Box 6.9 compared with those of the mammalian *Hox-B* cluster genes). Remarkably, the order of genes along the chromosome is recapitulated by the spatial domains of the expression patterns in the embryo, with the most 3' genes expressed in the most anterior domains (see Box 6.9). Homeotic gene expression is initiated by the combined activities of gap gene and pair rule gene products. Like the segment polarity genes, however, the homeotic genes maintain their expression patterns by cross-regulation. The effects of homeotic mutations thus reflect not only the primary response to the mutation, but also its effect on the domains of other homeotic genes. Again, the cross-regulation is primarily negative in nature, with each homeotic gene repressed by the genes expressed in more posterior domains. The effect of a loss-of-function homeotic mutation is thus an expansion in the domain of the anterior gene, with the resultant conversion of posterior segments into anterior ones. The patterns of homeotic gene transcriptional activity are stabilized by the modulation of chromatin structure. Inactive homeotic genes are sequestered into repressed chromatin, which is stabilized by proteins of the Polycomb family. Conversely, proteins of the Trithorax family appear to maintain active chromatin domains (see Chromatin). The homeotic genes encode transcriptional regulators containing a conserved DNA binding domain called a homeodomain (see Nucleic Acid-Binding Proteins). They appear to have overlapping DNA-binding specificities *in vitro* and are thought to work with cofactors *in vivo* which determine the exact DNA binding site. Several candidate coactivators have been found, including the products of the *Extradenticle* and *teashirt* genes. A number of downstream target genes for homeodomain protein regulation have been identified. The *salmon* and *distal-less* genes are required for eye and leg development, respectively, and themselves encode transcription factors, whilst *decapentaplegic* encodes a signaling molecule involved in the specification of leg fate.

*Continued*



Summary of the signaling pathway which maintains the parasegmental boundaries in *Drosophila*. The *engrailed* gene is initially activated by Even-skipped or Fushi tarazu, whereas *wingless* is activated in the adjacent cell by another pair rule protein, probably Odd-skipped. Pair-rule gene expression is only transient, but once activated, Wingless and Engrailed maintain each other through reciprocal signaling.

### Box 6.9: The vertebrate *Hox* genes

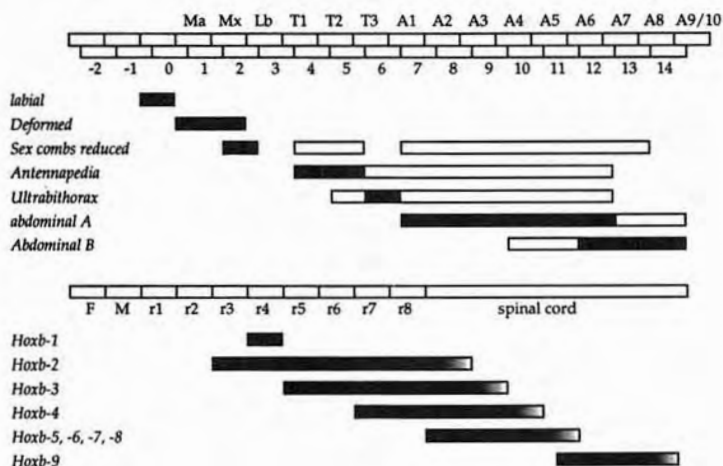


Architecture of the *Drosophila* and vertebrate *Hox* clusters. The fly cluster is interrupted by several non-*Hox* genes including *bicoid*. The ancestral vertebrate cluster underwent a 5' end expansion before duplication and several genes have been lost.

Continued

Vertebrate genomes contain four copies of the *Drosophila* homeotic complex, designated *Hox-A*, *Hox-B*, *Hox-C* and *Hox-D*. Although there are significant differences between the vertebrate and fly complexes, the similarities are remarkable. The same types of homeobox gene are present, allowing classification into 13 cognate groups (paralogous subgroups) based on homeobox structure. Furthermore, the genes are arranged in more or less the same order along the chromosome and are expressed in a similar manner, with the most 3' genes expressed in the most anterior domains and the most 5' genes in the most posterior domains. Since the divergence of *Drosophila* and vertebrates, the fly HOM-C has been split into two subcomplexes, whilst the vertebrate cluster has undergone a 5' end expansion and has been duplicated in its entirety to generate four complexes. Each of the four complexes has suffered individual losses, which may be different between species. For instance, the hatched boxes in the above figure represent *Hox-C* genes present in humans but missing in mice.

The similarity between the *Drosophila* and vertebrate homeobox-containing genes, in terms of both structure (see figure above) and expression patterns (see figure below), is strong evidence for a conserved function. This has been confirmed by the use of cloned human *HOX* genes to rescue *Drosophila* homeotic mutants, and targeted disruptions of mouse *Hox* genes do indeed generate partial homeotic transformations (q.v. *gene targeting*). Deletion of the *Hoxc-8* gene, for instance, results in the partial transformation of lumbar vertebrae into vertebrae with a more anterior characteristic (in this case, a thoracic vertebra, complete with a rib). The paralogous *Hox* genes appear to have overlapping but nonidentical functions and may cooperate with each other in certain cases. Individual gene knock-outs of genes in paralogous subgroup 3, for instance, cause different types of disturbances to the structures of the neck, but when combined in the same mouse, severe defects are observed including missing vertebrae.



Expression domains of the *Drosophila* HOM-C complex genes and the mouse *Hox-B* cluster genes in the nervous systems of each organism. *Drosophila* expression patterns are shown relative to segments and parasegments. Dark boxes indicate strong expression where abolition causes a homeotic transformation, and unfilled boxes indicate weak expression where abolition has no effect. Mouse expression patterns are shown relative to the rhombomeres of the hindbrain, and the spinal cord. The spinal cord is not to scale with the brain. Fly: Ma, mandible; Mx, maxillae; Lb, labium; T, thoracic segment; A, abdominal segment. Mouse: F, forebrain; M, midbrain; r, rhombomere of hindbrain.



**Box 6.10: Vertebrate limb development****Origin and development of the vertebrate limb.**

Vertebrate limbs develop from **limb fields** which arise in specific positions along the body axis determined by the anteroposterior *Hox* code (see Box 6.9). Proliferation of the lateral plate mesoderm (which contributes to the skeletal elements) and the somites (which contribute to the muscular elements) forms a protuberance termed a **limb bud**. The proliferation may be induced by FGF-8 secreted from the mesonephros.

A significant event in limb development is the formation of the **apical ectodermal ridge (AER)**. This is a raised crest of ectoderm which maintains the proliferation of the underlying mesoderm and plays an important regulatory role in axis specification. Only ectoderm lying at the dorsal/ventral boundary of the embryo is competent to form the AER, and this reflects the synthesis of a protein named **Radical fringe** in the dorsal ectoderm. Radical fringe may be involved in a reciprocal signaling event with the adjacent ventral cells that express *en-2*, a vertebrate homolog of the *Drosophila engrailed* gene. The boundary cells are induced to elevate and synthesize FGF-8, which enables them to sustain the proliferation of the underlying mesoderm. As the limb bud extends, a discrete **progress zone** of about 200  $\mu\text{m}$  can be identified under the AER where cells continue to proliferate. Behind the progress zone, cells differentiate into structures appropriate for their position along the proximodistal (shoulder to fingers), anteroposterior (thumb to fingers) and dorso-ventral (knuckle to palm) axes. The size of the progress zone reflects the range of the FGF8 signal.

**Forelimb or hindlimb?** The decision to differentiate into structures appropriate for the forelimb or hindlimb reflects the positional information in the mesodermal components of the limb imparted by the anteroposterior *Hox* code. Different combinations of *Hox* genes are expressed in the fore- and hindlimbs, as are other transcription factors (e.g. *Tbx4* and *Tbx5* which are expressed in the mouse forelimb and hindlimb, respectively). These factors control responses to morphogenetic instructions such as growth and differentiation signals, allowing cells to behave differently in each limb and form limb-specific structures.

**The proximodistal axis.** Regional-specific differentiation along the proximodistal axis (e.g. shoulder, humerus, radius/ulna, metacarpals or digits) reflects the length of time the cells have remained in the progress zone. Heterochronic grafts (where the

progress zone of one limb bud is grafted under the AER of another from a different developmental stage) show that the limb develops according to the instructions carried by the donor progress zone. Older cells leaving the progress zone express progressively more distal genes of the *Hox-A* and *Hox-D* clusters, and compound knockouts of paralogous *Hox* genes cause deletions of particular proximodistal structures. Therefore it seems that the age of the cell as it begins to differentiate, perhaps indicated by the number of divisions it has undergone, may in some way regulate *Hox* gene expression and impart proximodistal positional information to differentiating cells.

**The dorso-ventral axis.** The dorsal side of the limb is specified by the signaling molecule *Wnt-7a*, which is expressed in the dorsal ectoderm. Mice lacking the *wnt-7a* gene have ventralized limbs and also lack posterior structures because *Wnt-7a* is required for the specification of the anteroposterior axis (see below). A transcription factor, *Lmx1*, has been identified which is required for dorsal specification and its gene is induced by the *Wnt-7a* signaling pathway.

**The anteroposterior axis.** The anteroposterior axis is specified by a posterior **zone of polarizing activity (ZPA)** which arises in the limb field at about the time the AER is formed. The ZPA acts as an organizer, and when grafted onto a different position in the limb, induces a secondary axis. Signals emanating from the ZPA induce concentric nested expression patterns of the distal *Hox-D* genes, and targeted disruption and ectopic expression experiments have shown that the particular combination of *Hox-D* genes determines anteroposterior positional value, e.g. which type of digit is formed. The *Hox-D* expression pattern is set up in a complex manner that involves the ZPA protein *Sonic hedgehog* (*Shh*), which cooperates with other signaling molecules, FGF-4 and BMP-2, which are synthesized in the AER.

The expression of *shh* is initiated by FGF-8 secreted from the early AER and *Wnt-7a* which is expressed in the dorsal ectoderm. *shh* expression may be restricted to the posterior of the limb bud by the anterior boundary of *hoxb-8* expression. *Shh* induces *fgf-4* expression in the posterior AER, and the two proteins maintain each other through an autoregulatory loop. Retinoic acid, which can induce *shh* expression and mimic the ZPA, may also be involved in the induction of *fgf-4* expression. Thus the expression and function of *Shh* is depen-

dent on molecules which are involved in specification of all three limb axes.

**Positional information in *Drosophila* and vertebrates.** The limb model shows how the *Hox* genes perform a similar role to the *Drosophila* homeotic genes in the specification of positional values. In both organisms, loss of *Hox* gene expression results in homeotic transformations, although in vertebrates the situation is complicated by the redundancy of some of the components. The segment polarity genes in *Drosophila* are required to maintain cell boundaries and establish cell fates and function, not only in segmental specification but also in other structures, including the legs and wings. Particularly important are interactions between the Hedgehog,

Wingless and Engrailed proteins in axis specification. In the vertebrate limb and in other regions, the same three molecules are seen to be involved. Again there has been duplication since the divergence of flies and vertebrates, and in mice there are three Hedgehog-related molecules, more than ten Wingless homologs and two Engrailed-related transcription factors. In the limb bud, Sonic hedgehog, Wnt-7a (a Wingless-related protein) and Engrailed-2 play an important role in axis specification and act upstream of the *Hox* genes. It is remarkable that so many of the components of the *Drosophila* regional specification network should be reiterated in vertebrates when their early development is so fundamentally different.

## References

- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1994) In: *Molecular Biology of the Cell*. 3rd edn, pp. 1037–1119. Garland Publishing, New York.
- Bard, J.B.L. (ed.) (1994) *Embryos. Colour Atlas of Development*. Wolfe, London.
- Gilbert, S.F. (1997) *Developmental Biology*. 5th edn. Sinauer, Sunderland MA.
- Slack, J.M.W. (1991) *From Egg to Embryo: Regional Specification in Early Development*. 2nd edn. Cambridge University Press, Cambridge.
- Cohn, M.J. and Tickle, C. (1996) Limbs — a model for pattern formation within the vertebrate body plan. *Trends Genet.* 12: 253–257.
- Coyne, R.S., Chalker, D.L. and Yao, M.C. (1996) Genome downsizing during ciliate development: Nuclear division of labour through chromosome restructuring. *Annu. Rev. Genet.* 30: 557–578.
- Dickson, B. (1995) Nuclear factors in sevenless signaling. *Trends Genet.* 11: 106–111.
- Hammerschmidt, M., Brooke, A. and McMahon, A.P. (1997) The world according to Hedgehog. *Trends Genet.* 13: 14–21.
- Joyner, A.L. (1996) *Engrailed*, *Wnt* and *Pax* genes regulate midbrain–hindbrain development. *Trends Genet.* 12: 15–20.
- Kornfield, K. (1997) Vulval development in *Caenorhabditis elegans*. *Trends Genet.* 13: 55–61.
- Leamire, P. and Kodjabachian, L. (1997) The vertebrate organizer: Structure and molecules. *Trends Genet.* 12: 525–531.
- MacOnochie, M., Nonchev, S., Morrison, A. and Krumlauf, R. (1996) Paralogous *Hox* genes: Function and regulation. *Annu. Rev. Genet.* 30: 529–556.
- Moon, R.T., Brown, J.D. and Torres, M. (1997) WNTs modulate cell fate and behaviour during vertebrate development. *Trends Genet.* 13: 157–162.
- Morisato, D. and Anderson, K.V. (1995) Signaling pathways that establish the dorsal–ventral patterns of the *Drosophila* embryo. *Annu. Rev. Genet.* 29: 371–399.
- Parent, C.A. and Devreotes, P.N. (1996) Molecular genetics of signal transduction in *Dictyostelium*. *Annu. Rev. Biochem.* 65: 411–440.
- Riverapomar, R. and Jackle, H. (1996) From gradients to stripes in *Drosophila* embryogenesis — filling in the gaps. *Trends Genet.* 12: 478–483.
- Stragier, P. and Losick, R. (1996) Molecular genetics of sporulation in *Bacillus subtilis*. *Annu. Rev. Genet.* 30: 297–341.
- Werner, M.H., Huth, J.R., Gronenborn, A.M. and Clore, G.M. (1996) Molecular determinants of mammalian sex. *Trends Biochem. Sci.* 21: 302–308.

## Further reading

## Chapter 7

# DNA Methylation and Epigenetic Regulation

### Fundamental concepts and definitions

- Nucleic acids and proteins may be modified during or after synthesis by the addition of specific chemical groups. RNAs and proteins are frequently modified (*see* RNA Processing, Proteins), but DNA modification is limited. Some viruses alter specific residues in their genome, a process which may protect the DNA from nucleases or may facilitate packaging into the capsid (*Table 7.1*). The major exception is **DNA methylation**, which is frequently observed in both prokaryotes and higher eukaryotes, and has many roles concerning the recognition and function of DNA.
- Both DNA methylation and viral genome modification involve the covalent modification of DNA bases, but neither process alters base-pairing specificity, and the information carried in the DNA is preserved (c.f. *mutagen*). Methylation can influence DNA function, however; for example by changing the way it interacts with transcription factors and other proteins. The state of DNA methylation may thus regulate gene expression and have a direct bearing on the phenotype of the organism.
- Enzymes which add methyl groups to DNA, **DNA methyltransferases** (or **DNA methylases**), use S-adenosylmethionine as the methyl donor. Two types of enzyme can be distinguished: **de novo methylases** add methyl groups to unmethylated DNA at specific sites and can therefore initiate a pattern of methylation, whereas **maintenance methylases** add methyl groups to DNA which is already methylated on one strand (**hemimethylated DNA**) and thus perpetuate patterns of methylation through successive rounds of replication. The target sites for maintenance methylases often show dyad symmetry, so the same enzyme can methylate the nascent strand of both daughter duplexes.
- DNA carries two forms of information: **genetic information** in its nucleotide sequence and **epigenetic information** in its structure. Epigenetic information is any heritable property of DNA which influences its activity (i.e. ultimately contributes to the phenotype of the organism) but which lies outside the nucleotide sequence itself. DNA methylation is one source of epigenetic information: it is heritable (due to maintenance methylation) and able to control gene expression. A methylation state is thus termed an **epigenotype** and a change in methylation state an **epimutation**. Other forms of epigenetic information include chromatin structure (*see* Chromatin) and the topological and conformational properties of the DNA molecule (*see* Nucleic Acid Structure).

### 7.1 DNA methylation in prokaryotes

**Methylation in restriction-modification systems.** Many bacteria encode endonucleases which cleave DNA in a sequence-specific manner. These enzymes represent defense systems which cut up invasive DNA (such as phage genomes) as they enter the cell. A susceptible phage thus demonstrates a low efficiency of plating on a host strain synthesizing the endonuclease, but a much higher efficiency of plating on a strain which lacks it. The first host is said to be **restricting** the propagation of the phage, and the enzymes are thus termed **restriction endonucleases**.

The host protects its own DNA from endonucleolytic cleavage by modifying specific bases within the endonuclease recognition sequence, a process facilitated by DNA methylation. Each endonuclease thus has a **cognate methylase**, which may be a distinct enzyme or part of a common holoenzyme. Such **restriction-modification systems** are widespread in bacteria: hundreds of

**Table 7.1:** Modified DNA nucleosides found in viral genomes

Virus	Modified nucleoside
FV3	~20% 5-Methylcytidine for cytidine
φW-14	α-Putrescinythymidine for thymidine
PBS1	Deoxyuridine for thymidine
SPO1	5-Hydroxyuridine for thymidine
T-even family	Hydroxymethylcytidine for cytidine

FV3 is a eukaryotic virus which encodes its own DNA cytosine methyltransferase. The other viruses are bacteriophage.

restriction enzymes and methylases have been isolated. They are widely exploited for the manipulation of DNA *in vitro* (see Recombinant DNA). DNA phage have evolved a range of anti-restriction strategies such as the modification of their own genomic DNA or the synthesis of restriction endonuclease inhibitors. Many viruses also lack the restriction sites targeted by their hosts.

**Dam methylation.** In *E. coli*, adenine residues in the sequence GATC are methylated at the N<sup>6</sup> position by the enzyme **DNA adenine methylase (Dam)**. The GATC site displays dyad symmetry, and adenine residues on both strands are methylated. The primary role of this modification is to allow the cell to discriminate between the parent and daughter strands following replication, when the newly synthesized strand is transiently unmethylated. This facilitates *post-replicative mismatch repair* (q.v.), a DNA repair system which corrects mismatches arising though replication errors. The lack of methylation directs the repair enzymes to excise the incorrect nucleotide from the daughter strand, rather than the correct nucleotide from the parental strand.

Dam methylation also plays a direct role in DNA replication. The *E. coli* origin of replication, *oriC*, contains 14 Dam methylation sites, and an additional site is found in the promoter of the *dnaA* gene, which encodes an essential replication initiator protein (q.v. *initiation of replication*). Hemimethylated origins, which arise directly following the initiation of replication, are unable to undergo reinitiation, possibly because they interact with components of the cell membrane. The daughter strand *oriC* and *dnaA* Dam methylation sites remain unmethylated for much longer than general Dam sites and may be under direct regulation by factors which control the *bacterial cell cycle* (q.v.). The GATC site in the *dnaA* promoter may act as a transcriptional silencer to prevent reinitiation. Similarly, GATC sites found in various bacterial transposons may prevent transposition by blocking transcription of the transposase gene. Alternatively, they may act posttranscriptionally by preventing the transposase protein interacting with DNA. In either case, transposition is restricted to a short period following replication, ensuring that two genomes are present in the cell to facilitate recombination-mediated repair of excision damage. Interestingly, DNA methylation plays a major role in the control of transposition in eukaryotic cells, as discussed below (see Mobile Genetic Elements).

**Dcm methylation.** In *E. coli*, internal cytosine residues in the sequence CCWGG are converted to 5-methylcytosine by **DNA cytosine methylase (Dcm)**. The function of this methylation system is unknown, although it might protect the genome from the restriction enzyme *EcoRII*. A DNA repair system encoded by the *usr* gene corrects the G:T mismatches which frequently occur within this target site caused by the deamination of 5-methylcytosine to thymine (q.v. *very short patch mismatch repair*). Remarkably, *usr* and *dcm* are part of a common operon and the open reading frames overlap by six codons (q.v. *overlapping genes*).

**7.2 DNA methylation in eukaryotes**

**Patterns of cytosine methylation in eukaryotes.** In eukaryotes, the only modified base commonly found in DNA is 5-methylcytosine (5-meC), and this probably represents the sole programmed modification. The abundance of 5me-C varies between taxa, being low in fungi and invertebrates



(e.g. 5me-C is undetectable in both *S. cerevisiae* and *Drosophila* spp.), moderate in vertebrates (up to 10% of cytosine residues may be methylated) and high in many plants (up to 30% of cytosine residues).

In vertebrates, most cytosine methylation occurs within the dinucleotide motif 5'-CG-3', whereas in plants, both 5'-CG-3' and 5'-CNG-3' motifs are methylated. Both sites display dyad symmetry, and all the eukaryotic DNA methyltransferases isolated so far are maintenance methylases. Much DNA methylation observed in eukaryotic genomes is constitutive, and maintenance methylation is sufficient for its propagation. However, as discussed below, global changes in methylation states observed during both animal and plant development predict the existence of *de novo* methylases and demethylases which have yet to be characterized. Furthermore, additional methylated cytosines occur at asymmetric sites and are implicated in the control of DNA replication (q.v. *densely methylated island*). The modification of these sites also requires *de novo* methylase activity.

**Cytosine methylation, mutation and genome evolution.** The CG motif occurs at only 20% of its expected frequency in vertebrate DNA because of the hypermutability of 5-meC. *Deamination* (q.v.) is a common form of spontaneous DNA damage, and the deamination product of unmodified cytosine is uracil, which is efficiently removed from DNA by *base excision repair* (q.v.). The deamination product of 5-methylcytosine, however, is thymine, a legitimate DNA base. As in bacteria, a specific repair system exists to correct the resulting T:G mismatches, but this is leaky, resulting in net cytosine depletion over an evolutionary timescale. Notably, C→T transitions are predominant in human diseases caused by point mutations.

**CpG islands.** Analysis of the distribution of 5me-C in eukaryotic genomes has revealed the existence of **CpG islands** (or **HTF islands**<sup>1</sup>). These are nonmethylated regions, 1–2 kbp in size and predominantly associated with the 5' ends of some genes. They are generally GC-rich, and show an abundance of CpG dinucleotide motifs. This indicates that the absence of methylation has prevented the depletion of cytosine residues. These characteristics are highly diagnostic and can be used to identify potential genes in large genomic clones (q.v. *positional cloning*).

About 50 000 CpG islands exist in the human genome. About half are associated with house-keeping genes, and half with cell-type-specific genes. The islands of both types of genes are constitutively unmethylated whether or not the genes are expressed, with the exception of genes on the inactive X-chromosome and those subject to parental imprinting (discussed below). The position and nature of CpG islands suggests that an undermethylated promoter may be required for, but not always sufficient for, transcriptional activity. In agreement with this, the artificial methylation of CpG islands causes transcriptional repression. It has been proposed that the islands are normally protected from the activity of DNA methylases by the constitutive binding of transcription factors to the DNA. This would explain the lack of methylation, the association with transcriptional activity and the effects of ectopic methylation. For cell-type-specific genes it can be assumed that some regulatory complexes bind to the DNA constitutively (creating the island), whereas additional cell-type-specific factors are also required for active transcription. Many genes are not associated with CpG islands.

**Methylation and gene regulation in mammals.** Many lines of evidence suggest a link between DNA methylation and transcription, specifically between **undermethylation** or **hypomethylation** (the lack of methylation at certain sites) and transcriptional activity. Evidence arises from the manipulation of CpG islands, as discussed above, the abolition of transgene expression if the transgene is methylated before introduction into the cell, the induction of gene expression by methylation-blocking drugs

<sup>1</sup>CpG islands are termed HTF islands because they usually contain a cluster of restriction sites for the 5-meC sensitive restriction endonuclease *Hpa* I. In normal (methylated) DNA, *Hpa* I cuts rarely because its target site is both depleted and frequently modified. In CpG islands there is an abundance of unmodified targets and these sites diagnostically generate *Hpa* I tiny fragments (q.v. *restriction enzymes*).



such as 5-azacytidine, the observation of cell-type-specific methylation patterns, and **hypermethylation** (increased methylation at certain sites) in CpG islands of the inactive X-chromosome and at loci subject to parental imprinting.

In principle, methylated DNA could influence transcription in several ways. It could directly interfere with the binding of transcription factors with cytosine in their recognition site. Alternatively, it could recruit proteins which specifically recognize methylated DNA and hence block or displace transcription factors — two **methyl-CpG-binding proteins** of this nature have been isolated. Finally, it could regulate chromatin structure by influencing either nucleosome positioning or interactions with other chromatin proteins.

Although there is no doubt that DNA methylation can influence gene expression, there is no compelling evidence that it is used in a regulatory capacity *in vivo* other than in the specialized cases of parental imprinting and X-chromosome inactivation (see below). Most of the evidence for methylation as a general regulatory mechanism comes from artificial systems in cell culture, where patterns of methylation often differ from those of endogenous tissues. While cell-type-specific methylation patterns are also observed *in vivo*, it has yet to be demonstrated that these differences are *responsible* for differential gene expression. It is probable that such differences arise secondarily to differential gene activity, e.g. if the binding of a regulatory protein blocks the access of maintenance methylase to the DNA, and may reinforce earlier decisions and maintain the differentiated state. Homozygous mice deficient for the DNA methyltransferase gene die shortly after gastrulation, indicating that DNA methylation is essential in somatic cells for later stages of development. However, these effects may result from the ensuing inactivation of all copies of the X-chromosome, and to a lesser extent from the deregulation of imprinted genes (see below).

If DNA methylation ultimately regulates only a small number of genes, what is the role of the widespread methylation observed in higher eukaryotic genomes? One theory proposes that it reduces the effects of aberrant gene expression in the large genomes of higher eukaryotes, i.e. it reduces 'genetic noise'. An interesting recent observation is that most methylated cytidine residues in mammalian genomes lie within *transposable elements* (q.v.). Two classes of *retroelement* (q.v.), LINES and SINES (for long and short interspersed element, respectively), are particularly abundant in mammalian DNA and comprise much of the middle repetitive DNA of the mammalian genome (see Mobile Genetic Elements, Genomes and Mapping). It is well established that DNA methylation represses the transposition of plant and bacterial transposable elements, so it is likely to be used in the same way in vertebrates. In agreement with this, recombinant retroviruses introduced into the mouse genome are often silenced by hypermethylation. Interestingly, the incidence of mutations and chromosome rearrangements caused by mobile elements is low in mammals but high in *Drosophila*, where DNA is unmethylated.

**Methylation and gene regulation in plants.** The integration of extra copies of an endogenous gene or multiple copies of a transgene into a plant genome often results in epigenetic silencing of both endogenous and exogenous genes. The silencing mechanism can occur at two levels, either transcriptionally (**homology-dependent transcriptional gene silencing** which may involve *cis* or *trans* DNA pairing), or posttranscriptionally (**homology-dependent posttranscriptional gene silencing** or **cosuppression**, which may involve *antisense RNA* (q.v. and **degradosomes**), but there is no change in DNA sequence. In many cases, epigenetic silencing correlates to hypermethylation of the repetitive sequences.

The genetic effects of transcriptional silencing are similar to the natural process of **paramutation**, an allelic interaction where one **paramutagenic** allele suppresses another **paramutable** allele in the heterozygote. The result of paramutation in most cases is that the paramutable allele becomes paramutagenic, and the change is heritable through meiosis but ultimately reversible. There is no change in DNA sequence, and the effect appears to be related to gene copy number and chromosome position. In some, but not all, cases, paramutation is associated with a change in the amount of DNA methylation.

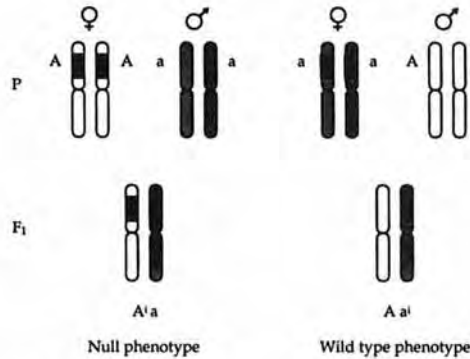
The hypermethylation of repetitive DNA in plants and some fungi may represent a defence mechanism to prevent the spread of transposable elements (as it is in vertebrates). The involvement of repetitive sequences in both cosuppression and paramutation would seem to support this theory. Like animals, plants undergo global changes of methylation states in development which give rise to transposition bursts where many families of transposable elements become activated at once. *Arabidopsis* mutants lacking DNA methyltransferase are viable but show frequent developmental defects, which may correspond to the activation of transposon families. However, a general regulatory role for DNA methylation in plants has yet to be ruled out, and the loss of methylation may thus reflect a deregulation of epigenetic programming rather than increased mutagenesis.

### 7.3 Epigenetic gene regulation by DNA methylation in mammals

**Global regulation of methylation patterns.** In mammals, the pattern of methylation characteristic of the adult is established just prior to gastrulation. The level of methylation in the zygote is high, but global demethylation occurs during cleavage, and subsequent *de novo* methylation occurs only in the epiblast lineage which gives rise to the somatic tissues of the adult. The extraembryonic lineage, which includes the extraembryonic membranes and germ cell primordia, remains undermethylated until gametogenesis, when the germ cells are also *de novo* methylated. As discussed above, most of the methylation and demethylation is constitutive and may fulfill a protective role, repressing the activity of transposable elements and suppressing 'genetic noise'. There are two exceptions, however: imprinted genes and the *Xist* locus on the X-chromosome. These are methylated in a sex-specific manner in the germ cells and escape both the global demethylation and the global *de novo* methylation which occurs early in development.

**Parental imprinting.** An imprint is a form of epigenetic information, and **parental imprinting** is thus an epigenetic mechanism by which a chromosome region can retain an epigenetic memory of its parental origin. In somatic cells, most genes are expressed on both maternal and paternal chromosomes, but imprinted genes are expressed in a parent chromosome-specific manner. Conventionally, an allele is said to be imprinted if it is repressed (e.g. at maternally imprinted loci, only the paternal allele is expressed), but this is an arbitrary definition and does not necessarily reflect the intrinsic character of the imprinting mechanism. At the molecular level, the imprint is DNA methylation, but in some loci methylation corresponds to gene activity and in others it corresponds to inactivity. As discussed below, most imprinted genes are found in clusters, and a single methylation imprint at a *cis*-acting control site can direct maternal-specific expression of some genes and paternal-specific expression of others. The consequences of parental imprinting in a genetic cross are shown in Figure 7.1. Although both parents contribute equally in terms of chromosomes, the locus is functionally hemizygous and the results of reciprocal crosses are therefore nonequivalent.

The first evidence for parental imprinting was the failure of parthenogenetic development (development from unfertilized eggs) in mammals. Nuclear transplantation experiments then showed that both parental genomes are required for normal development. **Uniparental** mouse embryos, i.e. those derived from zygotes containing either two male nuclei (**androgenotes**) or two female nuclei (**gynogenotes**), are abnormal and die before birth. Gynogenous embryos undergo normal early development but the extraembryonic membranes are undeveloped. Androgenous embryos are growth-retarded although the membranes develop normally — such aberrant concepti are termed a **complete hydatidiform moles** (**partial moles** contain a triploid foetus, with two paternal and one maternal genome). Experiments using mouse strains with **isodisomic** chromosome fragments (both copies derived from the same parent) identified the chromosome regions responsible for imprinting effects. The *syntenic* (q.v.) regions in the human genome are associated with classic imprinting disorders such as the Prader-Willi and Angelman syndromes and Beckwith-Wiedemann syndrome. In the animal kingdom, imprinting effects appear to be restricted



**Figure 7.1:** The effects of parental imprinting in a reciprocal cross. A homozygous wild-type individual (AA) is crossed to a homozygous null mutant (aa) to generate a heterozygous F<sub>1</sub> generation (Aa). The maternal allele carries an imprint (dark bar) and is not expressed, but the paternal allele is expressed. Where the mother is the null parent, the F<sub>1</sub> generation shows the dominant phenotype associated with the paternal allele, A. Where the mother is the wild-type parent, the F<sub>1</sub> generation shows the mutant phenotype associated with the paternal null allele, a.

to mammals, but imprinting is also found in plants. Plants are more tolerant of dosage effects, so the phenotypes are morphological rather than disease-related.

**Parental imprinting and DNA methylation.** At least 20 imprinted genes have been identified in mice and humans, many encoding products concerned with growth regulation. Consistent with the identification of imprinted chromosome regions in the mouse, the majority of human imprinted genes map to three clusters, and there is evidence that specific **imprinting boxes** in each cluster are required in *cis* to regulate parent-specific expression. These sites show parent-specific methylation patterns established and reset independently of the global *de novo* methylation and demethylation occurring during early mammalian development. The imprinting boxes are differentially methylated in the male and female gametes and direct the parent-specific gene expression patterns. The maternal and paternal chromosomes also differ with respect to recombination frequency, chromatin structure and replication timing.

The imprinted cluster on human chromosome 11p15.5 is associated with the fetal overgrowth disorder Beckwith–Wiedemann syndrome (BWS). The cluster contains at least seven genes (Table 7.2), with paternal-specific *IGF2* surrounded by maternal-specific *H19*, *KCNA9*, *CDKN1C* and *NAP2*. *IGF2* encodes a growth factor whose deregulated expression plays the predominant role in BWS. In the homologous mouse cluster, three further imprinted genes have been characterized. *Ins2* is paternally expressed: this encodes another growth factor and is homologous to the *INS* gene in the human cluster, whose expression parameters are not established. *Mash-2* and *Igf2r* are maternally expressed: *Igf2r* encodes a receptor for the paternal-specific *Igf2* (however, human *IGF2R* shows biallelic expression). In the mouse, the *Igf2* and *Ins2* genes show biallelic expression in certain tissues, indicating that the imprinting effect is lineage-specific, i.e. the genes are **conditionally imprinted**.

Central to the regulation of the human BWS cluster is the mutually exclusive expression of *H19* and *IGF2*. Most cases of BWS involve biallelic expression of *IGF2* (normally expressed only on the paternal chromosome), and many of these reflect paternal uniparental disomy or paternal trisomy. However, in cases where one chromosome is inherited from each parent as normal, most show either deletion or methylation of the maternal *H19* locus. This contains a paternal-specific methylation site, and methylation causes loss of *H19* gene expression. The epistatic effect of *H19* mutations and epimutations on *IGF2* transcription has led to an **enhancer competition model** where *H19* and *IGF2* are proposed to depend on the same transcriptional enhancer, but *H19* transcription sequesters the enhancer, blocking *IGF2* expression. The *IGF2* gene is therefore expressed only when *H19* transcrip-



**Table 7.2:** Human imprinted genes of the BWS cluster and at the X-inactivation center (XIC)

Gene	Product	Position	Expression
Beckwith–Wiedemann syndrome cluster			
<i>IGF2</i>	Growth factor	11p15.5	Paternal
<i>INS</i>	Growth factor		Unknown — paternal in mouse
<i>H19</i>	Untranslated RNA		Maternal
<i>IGF2R</i>	Growth factor receptor		Biallelic — maternal in mouse
<i>KCNA9</i>	K <sup>+</sup> channel		Maternal
<i>CDKN1C</i>	CDK inhibitor		Maternal
X-inactivation center			
<i>XIST</i>	RNA	Xq13	Paternal prior to gastrulation

Note the predominance of growth regulators in the BWS cluster. *IGF2* and *INS* encode growth factors, *IGF2R* is a growth factor receptor, *H19* regulates *IGF2* transcription, and *CDKN1C* is a putative cell cycle regulator. The mouse BWS cluster also contains the maternally expressed gene *Mash-2*, which encodes a transcription factor.

tion is prevented by methylation. The *H19* product is an untranslated RNA, and remarkably appears to have no intrinsic function except that its transcription represses transcription of *IGF2*.

Recently, experiments using mice transgenic for YACs containing the entire mouse BWS cluster have suggested that antisense RNA plays a major role in the imprinted expression of *Igf2r* (normally expressed only on the maternal chromosome). There is a paternal-specific methylation site in the *Igf2r* promoter, and a maternal-specific methylation site within an intron of the gene. These sites are present in both humans and mice, even though the human gene shows biallelic expression. The intronic methylation site marks the position of a promoter of an antisense transcript, which is normally expressed on the unmethylated paternal chromosome, leading to inhibition of paternal *Igf2r* expression. The maternal antisense promoter is methylated and *Igf2r* transcription is uninhibited. It thus appears that in at least these two examples parental imprinting is initiated by differential methylation and regulated by competition mechanisms. The nature of the competition is unknown for *Igf2r* — antisense transcription could cause countertranscription against the *Igf2r* gene; it could coat the chromosome and induce heterochromatinization like *Xist* RNA (see below), or there could be competition for regulatory elements like *Igf2/H19*.

**The role of imprinting: development or parental conflict?** Various theories have been put forward to explain the significance and evolution of imprinting in mammals, many suggesting that it plays an intrinsic role in mammalian development. Although parthenogenetic mice fail to develop, chimeras of parthenogenetic and normal cells are viable. The distribution of parthenogenetic cells in chimeric mouse embryos suggests that differentially imprinted cells play an important role in the development of the brain, e.g. parthenogenetic cells are excluded from the hypothalamus but abundant in forebrain. The alternative **parental conflict model** suggests that imprinting arose as a result of competition between maternal and paternal genomes in polygamous and promiscuous mammals. In promiscuous species, females gestate offspring from multiple fathers who will try to maximize the ability of their offspring to outcompete their half-siblings for maternal resources, whereas the mother will try to maximize the number of surviving offspring. Thus, paternal alleles would be selected for increased growth, whereas maternal alleles would be selected for reduced growth. In the mouse BWS cluster, paternal *Igf2* and *Ins2* encode growth factors, whereas maternal *H19* is an inhibitor of *Igf2* transcription and maternal *Igf2r* is a scavenger receptor for *Igf2*, and so reduces the level of available growth factor in the serum. Knockout mice with targeted deletions in the *H19* or *Igf2r* genes increase the amount of circulating *Igf2* by increasing the amount of mRNA and free protein, respectively, and show an increased birth weight (conversely *Igf2* knockouts have 60% normal growth rate if the targeted allele is inherited on the paternal chromosome). Interestingly, when a promiscuous species of mouse is

crossed to a species which is usually monogamous, the birth weight of the F<sub>1</sub> offspring shows parental imprinting. Promiscuous males crossed to monogamous females produce larger offspring than monogamous males crossed to promiscuous females. This fits the parental conflict model as monogamous males would be under less pressure to boost the growth their offspring and the females would be under less pressure to compensate. In the interspecific crosses, the birth weight of the F<sub>1</sub> would thus be driven by the competition strategies of the promiscuous species. Overall, this model predicts that rather than reflecting an intrinsic aspect of mammalian development, imprinting is ultimately dispensable. This is supported by the phenotype of double knockout mice with targeted deletions of *Igf2* with *Igf2r*, or *Igf2* with *H19*: both types of double knockout are normal.

**X-chromosome inactivation.** In the somatic cells of female mammals, *dosage compensation* (q.v.) for the X-chromosome is facilitated by **X-chromosome inactivation**. One of the X-chromosomes becomes condensed into heterochromatin and so transcriptionally inert. The **inactive X (Xi)** remains condensed throughout the cell cycle as a densely staining **Barr body**, while the **active X (Xa)** behaves as normal.

In the female zygote, both X-chromosomes are active. The inactivation process initiates at the late blastocyst stage in the extraembryonic lineage, and slightly later in the epiblast lineage which gives rise to the somatic cells of the embryo. In embryonic tissues of placental mammals, the choice of active versus inactive X is *random*. Once chosen, the process is irreversible: the active and inactive chromosomes are clonally propagated, so that an individual heterozygous for a visible X-linked marker demonstrates patchy variegation for the marker phenotype (e.g. tortoiseshell coat colours in cats). In extraembryonic tissues of many mammals (but not humans), the paternal X is preferentially inactivated (an example of *parental imprinting*, see above), and paternal preference occurs in both embryonic and extraembryonic tissue of marsupials. In both cases, paternal imprinting appears not to be an essential component of the inactivation process, because normal inactivation occurs in embryos with two maternally derived X-chromosomes.

**The molecular basis of X-inactivation.** X-inactivation can be divided into four stages: counting, initiation, propagation (spreading) and maintenance. An X-linked *cis*-acting site, apparently involved in all four stages, is the **X-inactivation center (Xic)**, which has been mapped by studying deletions and translocations of the X-chromosome resulting in the disruption of inactivation. If the region containing *Xic* is deleted, the mutant X remains active; if the X-chromosome is involved in a translocation, the fragment containing *Xic* is inactivated (as is part of the autosome to which it is attached) while the other fragment remains active.

Fine mapping of *Xic* originally revealed a single locus, *Xist*, which is expressed specifically on Xi (i.e. Xi-specific transcript). *Xist* encodes a large untranslated RNA which associates with Xi in the nucleus and is thought to induce heterochromatinization by coating the chromosome. Several ingenious experiments have provided clear evidence for the pivotal role of *Xist* RNA in the inactivation process, e.g. integration of transfected *Xist* DNA into an autosome in a male *ES cell* (q.v.) causes random inactivation of either the recombinant autosome or the endogenous X-chromosome upon differentiation (i.e. the autosome containing *Xist* is seen as a second X-chromosome by the cell). In mice, a second locus, *Xce* (X-controlling element), is linked to *Xist* and influences the random nature of X-inactivation — the chromosome containing the weaker *Xce* allele is preferentially inactivated.

**Counting and initiation.** The overall regulation of X-inactivation reflects the balance between X-chromosomes and autosomes. *Polysomy X* (q.v.) is tolerated because supernumerary X-chromosomes are inactivated, generating multiple Barr bodies, whereas *polyploidy* (q.v.) causes more X-chromosomes to remain active. The nature of the counting mechanism is unknown, as transgenic mouse studies have provided contradictory evidence for the role of the *Xce*. It is unclear whether multiple *Xce* transgenes in *cis* are counted individually — X-inactivation does not occur in human males with *Xce* duplications.

During gametogenesis the *Xist* promoter is subject to sex-specific methylation, generating a



maternal imprint in the zygote (see *parental imprinting* above). This results in preferential expression of the paternal *Xist* allele and hence preferential inactivation of the paternal X-chromosome. In the mouse extraembryonic lineage, this pattern of methylation is maintained (and the paternal X remains inactive), while in the epiblast lineage the methylation pattern is reset by a global demethylase, and *de novo* methylation of the *Xist* promoter occurs at random. The choice of active versus inactive X thus lies ultimately with this *de novo* methylase, and it is still not clear how the choice is made. In humans, random X-inactivation occurs in both extraembryonic and epiblast lineages, and in some extraembryonic cells both X-chromosomes remain active.

It has been suggested that cells show preferential X-inactivation in response to a chromosome mutation. Thus an individual heterozygous for a large X-linked deletion would preferentially inactivate the deficient chromosome to avoid the lethal consequences of maintaining the deficiency at the expense of the normal chromosome. In fact, there is no bias in chromosome selection — the observed homogeneity reflects **cell selection**, the differential survival of cells containing alternative Xi and Xa early in development. Cells with the mutant Xa will tend not to survive, leaving cells with the normal Xa to populate the entire embryo.

**Spreading and maintenance of inactivation.** The inactivity of Xi reflects changes in chromatin structure and DNA methylation patterns. The chromatin is highly condensed, corresponding to depletion for acetylated histone H4 (see *Chromatin*), and there is hypermethylation of CpG islands upstream of many X-linked genes. The signals causing these changes are unknown, and chromatin proteins specifically involved in X-silencing have not been identified (cf. *position effect variegation*, *SIR proteins*). It is likely that *Xist* RNA itself may play a major role in chromatin remodelling, but it is not sufficient for full inactivity: in Xi-autosome translocations, inactivity may spread across the translocation breakpoint into the autosome but never to the ends of the autosome, even though *Xist* RNA coats the entire chromosome. This suggests the existence of boundary elements in the autosomes which block the spread of inactivation, or *cis*-acting inactivation stations, **booster elements**, in the X-chromosome.

In normal X-chromosomes, inactivation spreads out from *Xic* over a number of days and inactivation occurs before the CpG islands become hypermethylated. Spreading may represent a cooperative effect where specific proteins bind to pre-existing inactive chromatin, similar to the spread of heterochromatin in autosomal translocations (q.v. *position effect variegation*). In humans, several X-linked genes 'escape' inactivation (e.g. the gene encoding steroid sulfatase). Such genes map in clusters, indicating that there are regions of local activity within the Xi. The corresponding genes in mice are often inactivated so it is not possible to extrapolate such observations between different mammals. DNA methylation thus appears to play a dual role in X-chromosome inactivation — first in the establishment of *Xist* expression, and secondly in the maintenance of transcriptional repression, i.e. it allows clonal propagation of the inactive X through sequential rounds of DNA replication by the activity of maintenance methylase.

### Further reading

- Bird, A. (1995) Gene number, noise reduction and biological complexity. *Trends Genet.* 11: 94–100.
- Hollick, J.B., Dorweiler, J.E. and Chandler, V.L. (1997) Paramutation and related allelic interactions. *Trends Genet.* 13: 302–308.
- Jaenisch, R. (1997) DNA methylation and imprinting: Why bother? *Trends Genet.* 13: 323–329.
- Lalande, M. (1996) Parental imprinting and human disease. *Annu. Rev. Genet.* 30: 173–195.
- Lee, J.T. and Jaenisch, R. (1997) The (epi)genetic control of mammalian X-chromosome inactivation. *Curr. Opin. Genet. Dev.* 7: 274–280.
- Reik, W. and Maher, E.R. (1997) Imprinting in clusters: Lessons from Beckwith–Wiedemann syndrome. *Trends Genet.* 13: 330–334.
- Richards, E.J. (1997) DNA methylation and plant development. *Trends Genet.* 13: 319–323.
- Siegfried, Z. and Cedar, H. (1997) DNA methylation: A molecular lock. *Curr. Biol.* 7: R305–R307.
- Yoder, J.A., Walsh, C.P. and Bestor, T.H. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* 13: 335–340.

**This Page Intentionally Left Blank**

## Chapter 8

# The Gene

### Fundamental concepts and definitions

- A **gene** is a physical and functional unit of genetic information with the potential to be *expressed*, i.e. to be used as a template to generate one or more *gene products* of RNA or protein. The *genome* (q.v.) is the total genetic information in the cell, and that information is stored as the nucleotide sequence. While all genetic information can be stored and transmitted from generation to generation as part of a chromosome, only the genes are expressed. Other loci function solely at the DNA level and include regulatory elements, origins of replication, centromeres and telomeres.
- A **cistron** is a unit of genetic function defined by a complementation test. The bacterial gene is the same as a cistron — a unit of genetic function which corresponds to an open reading frame. The gene-cistron relationship in eukaryotes is more complex because the information in a eukaryotic gene can be used selectively, altered posttranscriptionally or combined with that from other genes to generate a single gene product.

### 8.1 The concept of the gene

**Evolution of the gene concept.** The term *gene* was coined by Wilhelm Johansen in 1909 to describe a heritable factor responsible for the transmission and expression of a given biological character, but without reference to any particular theory of inheritance. Therefore, in its original context, the gene had no specific material basis and could be treated purely in abstract terms, as it is in the study of transmission genetics (q.v. *Mendelian inheritance*).

A more precise idea of the physical and functional basis of the gene arose from several sources in the first half of the twentieth century. In 1902, Archibald Garrod showed that the metabolic disorder alkaptonurea resulted from the failure of a specific enzyme and could be transmitted in an autosomal recessive fashion; this he called an **inborn error of metabolism**. Garrod was unfamiliar with Mendelian inheritance and the significance of his discovery was not realised for 30 years, when George Beadle and Edward Tatum found that X-ray-induced mutations in fungi often caused specific biochemical defects. This led to the **one gene one enzyme** model of gene function. In 1911, Thomas Hunt Morgan showed that genes were located on chromosomes and were physically linked together (q.v. *chromosome theory of inheritance*), and in 1944, Oswald Avery and colleagues showed by elimination that DNA was the genetic material. Thus evolved a simple picture of the gene — a length of DNA in a chromosome which encoded the information for an enzyme.

This definition has to be expanded to encompass what is known today about gene function: not all genes encode enzymes (some encode polypeptides with different functions, and some encode functional RNA molecules such as rRNA and tRNA); also, in viruses, genes may be RNA not DNA. Furthermore, the information in the gene may be used selectively to generate more than one product. The Mendelian concept of a gene controlling a single character has also expanded to take into account genes which affect several characters simultaneously (**pleiotropy**) and genes which cooperate in groups to control individual characters (q.v. *quantitative inheritance*). With the advent of molecular biology, it is now possible to define the gene exactly in terms of its structure and function, although this definition differs between bacteria and eukaryotes due to their very different strategies for gene expression and genome organization (also q.v. *prions*).

## 8.2 Units of genetic structure and genetic function

**Indivisible units of genetic structure.** After it became established that genes were carried on chromosomes and could be mapped on them by recombination (q.v. *genetic mapping*), individual genes came to be regarded as fundamental and indivisible units of genetic information, in terms of both structure and function, linked together in a linear fashion (the **beads on a string theory**). This view was challenged by recombination studies in bacteriophage, which showed that the gene could be subdivided into smaller units. Seymour Benzer introduced the terms **muton**, **recon** and **cistron** to define indivisible units of mutation, recombination and function, respectively. By crossing independently derived mutants of the same gene in a phage infection, it was shown that wild-type phage could be produced. This could only happen by **intragenic recombination** (recombination within a gene) if distinct subelements of the gene were mutated. This showed that the gene was divided into smaller units which could undergo recombination and mutation. The muton and recon are equivalent to a single nucleotide pair.

**The gene as an indivisible unit of genetic function.** Benzer's term cistron means an indivisible unit of genetic function. This can be determined by **complementation analysis** where the gene (or more specifically, its product) is tested for its ability to compensate for a mutation in a homologous gene in the same cell. Successful complementation restores the wild-type phenotype.

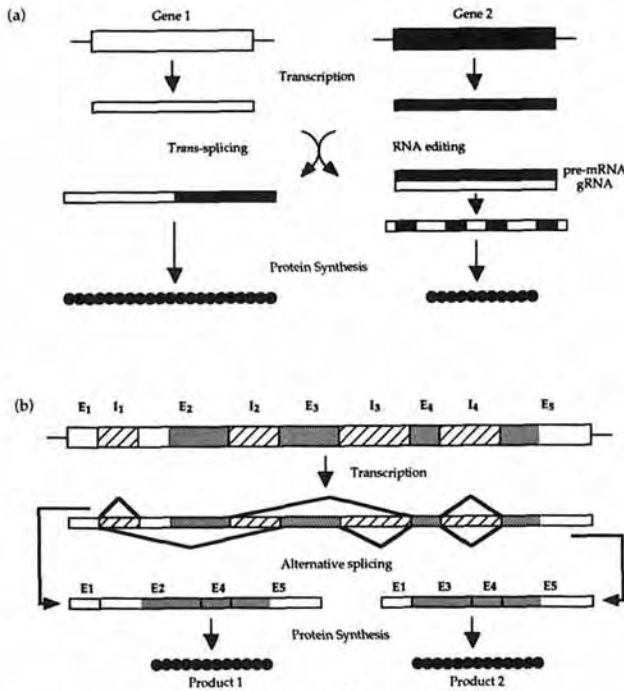
The basis of complementation analysis is the **cis-trans test** (from which the term cistron arises; Box 8.1), where pairs of independently derived mutations are considered in the *cis* and *trans* configurations. The *cis* test is a control, because if both mutations are present in one genome, the other should be wild-type at both loci and produce the normal gene product(s), thus conferring the wild-type phenotype upon the cell. The *trans*-test is the complementation test and defines the boundaries of the unit of function. If both mutations are within the same gene, when they are present in the *trans*-configuration each genome carries a mutant copy of the gene and no functional product would be generated in the cell — there would be no complementation. If the mutations lie in different genes, when they are present in the *trans*-configuration, each genome can supply the gene product the other lacks. With all necessary gene products, the cell is wild-type — there is **positive complementation**.

Complementation analysis in bacteria and yeast established that the gene is a cistron, i.e. the gene could be defined as a unit of function. The test was also useful for determining the number of genes acting in a given pathway and, by cross-feeding between mutants, determining the order in which they acted. Complementation can also be exploited to confirm gene function and to clone genes whose function is known (q.v. *marker rescue*, *functional cloning*, *expression cloning*).

## 8.3 Gene-cistron relationship in prokaryotes and eukaryotes

**Gene-cistron equivalence in simple genomes.** In prokaryotes and lower eukaryotes, there is usually a simple relationship between the gene and its product. In most cases, there is a one gene-one product correspondence and the gene is usually colinear with its product. In these organisms, therefore, the gene and cistron are equivalent: the gene is a unit of genetic function; a unit of expressed genetic information. In bacteria, the gene is synonymous with the coding region (open reading frame), whereas in eukaryotes it is synonymous with the transcription unit. This is because bacterial genes are often arranged as an *operon* (q.v.), so that several products are translated from a common **polycistronic mRNA**. In eukaryotes, conversely, most genes are transcribed as **monocistronic mRNA** (c.f. *rRNA genes*, *trans-splicing*, *internal ribosome entry site*).

**Gene-cistron nonequivalence in complex genomes.** In higher eukaryotic genomes, there is often a complex relationship between the gene and its product (Figure 8.1). Most higher eukaryotic genes contain *introns*, which are intervening sequences which are not represented in the final product, and are therefore not part of its function (see Genomes and Mapping). While the gene represents the



**Figure 8.1:** Examples of gene-cistron nonequivalence in eukaryotes (thick boxes are DNA, thin boxes RNA and chains are protein). (a) The expression of more than one gene is required to make a single polypeptide in cases of *trans*-splicing and RNA editing, but both genes are part of the same unit of function and comprise a single cistron. The open boxes represent information from one gene and the filled boxes represent information from the other. Note that in all known cases of *trans*-splicing, the 5' spliced transcript is untranslated, although in principle there is no reason why such a mechanism should not also be used to generate proteins. (b) One gene can generate several products by alternative splicing and other selective information usage (see text). The gene thus contains overlapping cistrons. Introns are represented by hatched boxes and are spliced out during RNA processing. Exons are shown as boxes, unfilled if they are untranslated, and filled to represent the coding region. Note that introns may interrupt both coding and noncoding exons, and that exons may contain both translated and untranslated information (i.e. exons 2 and 5).

entire transcription unit, the cistron would be interrupted by the introns. The cistron is therefore equivalent to the *exons* of the eukaryotic gene.

A further complication arises in those eukaryotic genes where the information is used selectively to generate several products. This is often achieved by *alternative splicing* (q.v.), which may reflect regulation at the level of RNA processing or alternative promoter or polyadenylation site usage during transcription (see RNA Processing). The structurally related products often have different functions, and the gene thus comprises a series of overlapping cistrons.

The opposite situation occurs where two genes are required to generate a single product, e.g. *trans*-splicing, where two separately encoded mRNA fragments are spliced together and translated, and *RNA editing* (q.v.) in Trypanosomes, where the mRNA and gRNA are required to produce a mature template for polypeptide synthesis. Both genes are required to produce a single function; they are part of the same cistron.

There are also cases where several proteins are derived from a single open reading frame. Translation initially produces a **polyprotein**, which is cleaved to generate the individual functional products. This strategy is used by some RNA viruses to meet the challenge of the monocistronic environment of eukaryotic cells, but also occurs in some endogenous genes, e.g. the mammalian preprodynorphin gene generates seven different peptides with different functions in the brain. The segment of the open reading frame encoding each peptide can be regarded as a cistron.



**Overlapping and nested genes.** Most genes are discrete nonoverlapping units, i.e. they do not share information with other genes. **Overlapping genes** are independently regulated but utilize some of the same sequence. In principle, genes can overlap at two levels. In bacterial systems, and in other situations where space conservation is necessary (e.g. in RNA virus genomes and in animal mitochondrial DNA), genes may overlap at the level of the reading frame, so that the same information is used to generate two or more unrelated proteins. The open reading frames may be transcribed from opposite strands, may be translated in different directions, and may be out of frame with respect to each other, i.e. the genes have nothing in common except that they share some of the same space. An example is the lysis protein gene of the leviviruses (which include bacteriophage MS2); this overlaps with the replicase and coat protein genes but is translated in the opposite direction and in a different reading frame. In some species, the lysis protein gene is completely inset within the replicase gene. In eukaryotic systems, genes can overlap at the level of the transcription unit, but exons can remain discrete. Thus the same information is never expressed in the protein products of both genes, because DNA regarded as exon material in one gene is part of the intron of the overlapping gene (e.g. human *TCRA* and *TCRD* T-cell receptor genes overlap at the exon level). Occasionally, complete genes may be embedded within the intron of a larger gene: small open reading frames encoding proteins concerned with intron metabolism are often found in self-splicing introns (see RNA Processing). Three small genes are also found in intron 26 of the large human gene *NF-1*. Overlapping genes may also reflect a mechanism of regulation. In plasmids, genes encoding antisense RNA tend to overlap with the genes they regulate (also q.v. *countertranscription*).

A **nested gene** is a single gene which produces two or more nested products by modulating the end point of protein synthesis. This can occur by leaky readthrough of a termination codon (e.g. in the case of the Q $\beta$  virus coat protein gene) or by cotranslational frameshifting (e.g. in the case of the *E. coli* *dnaX* gene and the F plasmid *traX* gene). Similar strategies are used by eukaryotic RNA viruses (e.g. retroviruses), and nested products can also be produced from eukaryotic genes by alternative splicing or use of alternative polyadenylation sites.

## 8.4 Gene structure and architecture

**Structural components of the gene.** The gene can be divided into discrete regions with specific functions (Table 8.1).

At any given locus, the DNA which is transcribed can be termed a **transcription unit**. In prokaryotes, a transcription unit may consist of several genes (constituting an *operon*), whereas in eukaryotes, transcription units are almost always equivalent to a single gene (the polycistronic rRNA gene transcribed by RNA polymerase I, and genes of RNA viruses and organellar genomes are exceptions; q.v. *internal ribosome entry site*, *trans-splicing*).

For genes which encode proteins, a distinction can be made between information translated into polypeptide sequence and untranslated information. In bacteria, the translated region (**open reading frame**, **coding region**) is equivalent to the gene, and genes are usually separated by short **internal noncoding regions**. The extreme genes of the operon are also flanked by a noncoding region which may be termed the 5' **untranslated region (UTR)** or **leader sequence** and the 3' **UTR** or **trailer sequence**. These sequences are often regulatory in nature; the 5' UTR controls ribosome binding and may facilitate *attenuator control* (q.v.); the 3' region often plays an important role in mRNA stability. In eukaryotes, the coding region is also flanked by regulatory UTRs, and both UTRs and the open reading frame may be interrupted by noncoding sequences, introns, which are spliced out before RNA export from the nucleus, i.e. they are not represented in the mature transcript.

RNA genes may be transcribed singly or as part of an operon in both prokaryotes and eukaryotes. The equivalent to protein coding regions are the regions which eventually form the mature RNA. Some RNA genes are transcribed as mature transcripts, others must undergo cleavage, processing and intron splicing. Any sequences which are eventually discarded are termed **tran-**

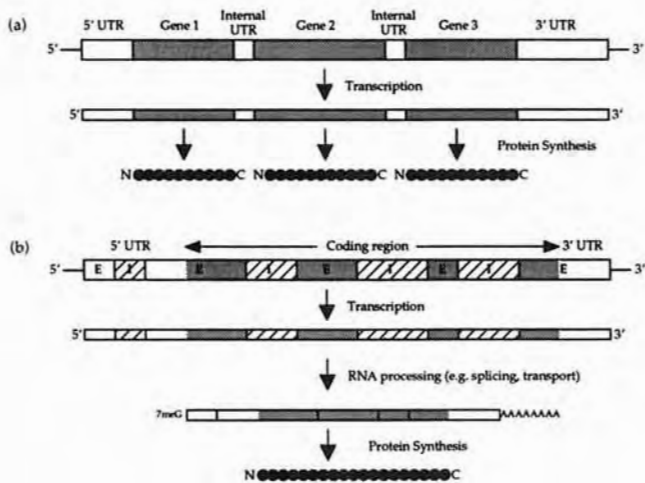
**Table 8.1:** Terms used to define functional parts of genes

Term	Definition
Allele	A sequence variant of a gene (or other genetic marker, e.g. RFLP, VNTR sequence)
Cistron	A unit of genetic function, a region of DNA which encodes a specific product
Coding region, open reading frame (ORF)	A region of DNA which is translated into protein. In bacteria, this is a gene. In eukaryotes, the coding region may be interrupted by introns
Divided gene	A split gene with exons at distinct loci which must be transcribed separately and joined by <i>trans</i> -splicing. Actually a misnomer as each locus should be considered as a discrete gene
Gene	In bacteria, a unit of genetic function encoding either a discrete polypeptide or RNA molecule. In eukaryotes, a transcription unit which may encode one or more products or may contribute to a product
Gene locus	The position of a gene on a chromosome, including flanking regulatory elements. The term locus used on its own refers to the position of any marker — gene, regulatory element, origin of replication, cytogenetic landmark, etc.
Operon	A bacterial locus containing several genes (which are transcribed as a single polycistronic transcript) and their common regulatory elements
Pseudogene	A nonfunctional sequence which resembles a gene (see Mutation and Selection)
Split gene, interrupted gene	A gene containing introns
Transcribed spacer	Any part of the transcription unit of an RNA gene or RNA gene operon which does not feature in the mature RNA molecule(s)
Transcription unit, transcribed region	A region of DNA which is transcribed into RNA. In eukaryotes, this is a gene. In bacteria it may encompass multiple genes
Untranslated region (UTR), noncoding region (NCR)	Any part of the transcription unit which is not translated into protein. UTRs flanking a coding region or operon are termed 5' and 3' UTRs (or leaders and trailers)

**scribed spacer sequences.** The typical structural organization of bacterial and eukaryotic genes is shown in *Figure 8.2*.

**Gene nomenclature.** Genes are named according to species convention (*Table 8.2*). It is normal to present the names of genes, alleles, genotypes and, where necessary, the mRNA transcribed from a gene in *italic*, and the protein products and phenotypes in *roman*. Confusion often arises when research in different organisms converges on a single genetic mechanism, as can be seen in the nomenclature of the cell cycle genes of *S. cerevisiae* and *S. pombe* (see *The Cell Cycle*). In addition, many genes are isolated several times from the same organism in different experiments, and given different names: the prominent *Drosophila* developmental gene *torpedo* is an example — it has been identified three times in screens for different phenotypes and has been given three different names. *Drosophila* provides the most colorful examples of genetic nomenclature and, especially in developmental biology (see *Development: Molecular Aspects*), this trend is spreading to vertebrates.

In most species, genes are designated by a symbol comprising several letters and numbers. Some species conventions (e.g. *Drosophila*, *E. coli*) dictate the use of lower case letters for genes identified as recessive mutations and initial capital letters for those identified as dominant mutations. In other species, including humans, genes are designated by all capital letters. Many genes are now identified by large scale sequencing approaches, and a unified approach to nomenclature is required.



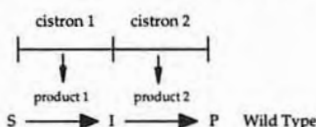
**Figure 8.2:** Typical structure of protein coding genes in prokaryotes and eukaryotes (thick boxes are DNA, thin boxes are RNA and chains are protein). (a) In bacteria, several genes are present in the same transcription unit. They are transcribed as a common mRNA which can be divided into coding regions (genes), shown in gray, and noncoding regions, shown in white. Each gene encodes a separate polypeptide. (b) In eukaryotes, transcription units comprise a single gene. The gene comprises a central coding region shown in gray, and flanking noncoding regions shown in white. The gene is often interrupted by one or more introns (hatched boxes), which are spliced out of the primary transcript. In this example, the mature transcript encodes a single polypeptide, but alternative splicing can be used to generate a series of different products.

**Table 8.3:** A brief summary of conventional genetic nomenclature

Species	Convention
<i>E. coli</i> and other bacteria	Three lower case letters to designate an operon followed by uppercase letters to denote different loci, e.g. <i>lac</i> operon; loci: <i>lacZ</i> , <i>lacY</i> , <i>lacA</i> . Proteins LacZ, LacY, LacA. A special convention adopted for sporulation genes of <i>B. subtilis</i> (see Box 6.2). Genes designated <i>spo</i> followed by roman numerals to designate morphological stage of sporulation and capital letters to designate operon then gene, e.g. <i>spoIIIGA</i> is expressed in stage II and refers to the first locus of operon G
Yeast	Three letters identify gene function followed by a number to specify different loci <i>S. cerevisiae</i> genes <i>GAL4</i> , <i>CDC28</i> ; proteins GAL4, CDC28 <i>S. pombe</i> genes <i>gal4</i> , <i>cdc2</i> ; proteins Gal4, Cdc2
<i>C. elegans</i>	Three lower case letters indicative of the mutant phenotype hyphenated to a number if more than one locus involved, e.g. genes <i>unc-86</i> , <i>ced-9</i> ; proteins UNC-86; CED-9
<i>Drosophila</i>	Name descriptive of mutant phenotype, may be represented by symbol of 1–4 letters, e.g. genes <i>white</i> ( <i>w</i> ), <i>tailless</i> ( <i>tlf</i> ), <i>hedgehog</i> ( <i>hh</i> ); proteins White, Tailless, Hedgehog
Plants	Although no convention covers all plants, most named with 1–3 letter lower case symbols. <i>Arabidopsis</i> genes named in style of <i>Drosophila</i> but using capital letters, e.g. gene <i>AGAMOUS</i> ( <i>AG</i> ), protein AGAMOUS
Vertebrates	Generally a 1–4 letter lower case symbol with letters and numbers, representative of gene function, e.g. genes <i>sey</i> , <i>myc</i> , proteins Sey, Myc
Humans	As above but in capitals, e.g. genes <i>MYC</i> , <i>ENO1</i> , proteins MYC, ENO1

**Box 8.1: The *cis-trans* test**

**The basis of the test.** The *cis-trans* test considers pairs of mutations in the same cell and asks whether mutual complementation can occur to restore the wild-type phenotype (i.e. does each genome compensate for the deficiency in the other). Consider an example where two gene products are required to convert a substrate (S) through an intermediate (I) into a pigment (P) which characterizes the wild-type phenotype. Failure to produce the pigment results in the mutant phenotype, which is identified by its color. The normal pathway is shown below.

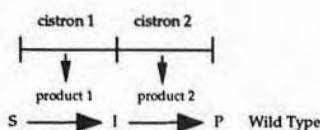
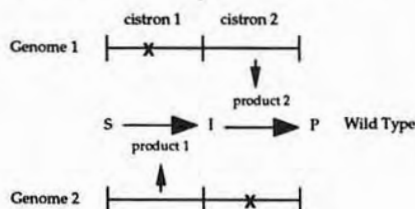


**The *trans*-test.** The *trans* part of the test is the test for complementation. Two mutations are considered in the *trans* configuration. **Positive complementation**, as shown by the restoration of the wild-type phenotype, usually only occurs if the two mutations are in different cistrons, i.e. the diploid cell carries at least one wild-type copy of each gene. This is known as **intercistronic** (or **intergenic**) **complementation**. If the mutations are in the same cistron, complementation usually does not occur because neither genome has a wild-type copy of the gene and no functional product is made by either genome.

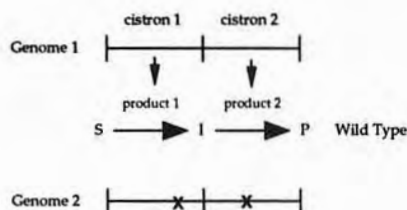
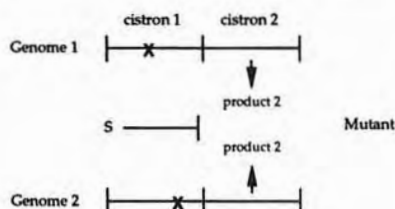
**Allelic complementation.** Occasionally, different mutations in the same cistron will complement each

other, and this is known as **allelic**, **intracistronic** or **intragenic complementation**. For this to occur, the gene product must function as a multimer, and one mutant gene product compensates for the deficiency in the other in the multimeric complex. Allelic complementation can usually be distinguished from normal intergenic complementation because the mutant heteromultimer is less active than a wild-type homomultimer. A practical example of allelic complementation is  **$\alpha$ -complementation**: many cloning vectors carry a 5' portion of the *E. coli lacZ* gene which is unable to produce a functional enzyme. These vectors can be used in host cells with 5' *lacZ* deletions. Neither truncated allele of *lacZ* is functional, but the two partial proteins can associate to form a functional enzyme.  $\alpha$ -complementation is used for recombinant selection in cloning experiments (q.v. *blue and white selection*).

**The *cis*-test.** The *cis* part of the test is a control, because when two mutations are found in the same genome, the second genome is wild-type, and hence the cell should display the wild-type phenotype. When this does not occur, it suggests that the mutant gene product produced by one genome is dominant to the wild-type product produced by the other. This is termed **negative complementation** and often occurs, like allelic complementation, when the gene product is a multimer. In this case, the inactive mutant product interacts with the wild-type to produce an inactive heteromultimer, i.e. it sequesters wild-type polypeptides into an inactive complex.



Positive (intercistronic) complementation



No complementation



Such mutations are described as **trans-dominant** or **dominant negative**.

**Cis-dominance.** Complementation occurs only between genes because they produce diffusible gene products. Some mutations cannot be compensated by the supply of a wild-type allele in the same cell, and these are described as **cis-dominant**. They reflect mutations of *cis*-acting elements (e.g. a promoter) rather than genes.

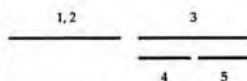
**Polar mutations.** A complication to complementation analysis in bacteria is the effects of **polar mutations**. Because many bacterial genes are arranged in operons which are transcribed as polycistronic mRNAs, a mutation in one gene can interfere with the expression of downstream genes due to the dependence of ribosome binding on completion of the translation of the upstream gene (q.v. *lac operon*).

**Complementation, reversion and recombination.** These three processes can all produce a wild-type phenotype. Reversion is a backwards mutation, from a loss of function mutant to the wild-type allele, i.e. it involves a change of genotype. Recombination between two mutant alleles may also produce a wild-type genome, as well as a double mutant, and also involves a change of genotype. Complementation involves no change of genotype and reflects the interaction of gene products. Reversion and recombination can generate aberrant results in complementation analysis, but both revertants and recombinants can be distinguished because they maintain their wild-type phenotype after segregation. Complications due to recombination can be avoided by using recombination-deficient strains.

**Complementation mapping.** A collection of mutants mutually unable to complement each other is termed a **complementation group** and corresponds to a cistron. By performing pairwise complementation analysis on many mutants, a **complementation map** can be constructed which represents complementation groups as discrete, nonoverlapping lines. Such a strategy can determine the number of genes in a given biochemical pathway, and by cross-feeding (exposing one mutant to the products of another), the order of gene activity can be established. A complementation map is not a map in the strict sense, because it gives no indication of the physical relationship between genes.

A simple example of complementation mapping is shown below. Five strains of yeast which are unable to synthesize a specific metabolite are tested for complementation and the results are shown on the left: + for positive complementation (restoration of wild-type phenotype) and - for no complementation. The map shows that mutants 1 and 2 form one complementation group and 3, 4 and 5 form another, i.e. there are two genes. Within the second gene, mutants 4 and 5 complement each other. This is allelic complementation, and the product of the second gene is thus likely to act as a multimer.

	1	2	3	4	5
1	-	-	+	+	+
2	-	-	+	+	+
3	+	+	-	-	-
4	+	+	-	-	+
5	+	+	-	+	-



## References

- Lewin, B. (1997). *Genes VI*. Oxford University Press, Oxford.
- Singer, M. and Berg, P. (1991) *Genes and Genomes: A Changing Perspective*. University Science Books, CA.
- Blumenthal, T. (1995) *Trans-splicing and polycistronic transcription in Caenorhabditis elegans*. *Trends Genet.* 11: 132-136.
- Kable, M.L., Heidmann, S. and Stuart, K.D. (1997) RNA editing: Getting U into RNA. *Trends Biochem. Sci.* 22: 162-166.
- Stewart, A. (ed) (1995) *The Trends in Genetics. Genetic Nomenclature Guide*. *Trends Genet.* 11 (Suppl).

## Further reading



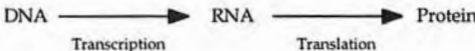
## Chapter 9

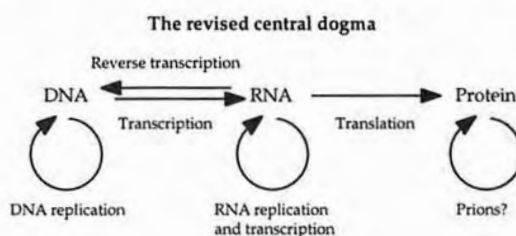
# Gene Expression and Regulation

### Fundamental concepts and definitions

- *Genes* (q.v.) differ from other forms of genetic information because they are *expressed*. **Gene expression** involves taking the information from the gene and using it to generate a **gene product**, which may be RNA or protein.
- In its simplest form, gene expression can be summarized by the **central dogma** of molecular biology, which states that genetic information flows unidirectionally from DNA through RNA to protein (*Table 9.1*). This holds true for most cellular genes, but information flow between nucleic acids in viruses is comparatively promiscuous. However, there is never information transfer directly from DNA to protein, nor from protein back to nucleic acid.
- Gene expression proceeds through a series of different levels before the information of the gene is released as a final product. The major levels are transcription and protein synthesis, but there are many intermediate fine tuning stages which may filter and modify the information. In principle any of these stages can be regulated. The relative predominance of regulation at each level of gene expression differs between prokaryotes and eukaryotes.
- Gene regulation is used to control the amount of each gene product produced by the cell. Regulation involves *cis*-acting elements and *trans*-acting factors, and can be positive or negative, inducible or repressible, and global or specific. These terms are discussed below.

**Table 9.1:** The original central dogma and modifications to take into account alternative routes for the flow of genetic information

The central dogma	
	
Deviation from central dogma	Example
DNA → RNA (no protein)	RNA genes
RNA → protein (no DNA)	RNA genomes
RNA → DNA (reverse transcription)	Retroid virus replication
RNA → RNA (RNA replication, RNA transcription)	RNA virus replication and transcription



### 9.1 Gene expression

**The multiple levels of gene expression.** A level of gene expression (Table 9.2) can be thought of as a discrete stage in the pathway of information transfer from gene to product where regulation is possible. Transcription is usually regarded as the initial stage of gene expression and is the predominant level for gene regulation, but before transcription is possible, the gene must be made accessible in a form suitable for transcription. This preparatory stage is particularly important in eukaryotes, where DNA methylation and the packaging of DNA in chromatin has a profound effect on its ability to be transcribed. Both transcription and protein synthesis can be divided into initiation, elongation and termination phases, each of which may be independently regulated (*see* Transcription, Protein Synthesis). In bacteria, where transcription, protein synthesis and RNA degradation occur simultaneously in the same cellular compartment, there may be cross-regulation between different levels of gene expression (q.v. *attenuation*, *retroregulation*). In eukaryotes, where pre-mRNA is extensively processed in the nucleus and then exported into the cytoplasm for translation, both processing and export may be subject to regulation (*see* RNA Processing). After synthesis, proteins may be subject to both covalent modification and noncovalent interactions with other molecules which regulate their activity (*see* Proteins: Structure, Function and Evolution). An

**Table 9.2:** Principle levels of gene expression and mechanisms of regulation

Level of gene expression	Regulatable processes involved
Gene preparation	Uncoating (viruses) Synthesis of template genome strand (viruses) Modification of chromatin structure (eukaryotes) Changes in topological and conformational properties of DNA Changes in DNA methylation states Genomic rearrangements Gene amplification
Transcription	Promoter usage Formation of active transcription complexes Promoter escape Elongation vs. pausing Termination vs. antitermination (prokaryotes) Attenuation (prokaryotes)
RNA processing (eukaryotes)	Capping Polyadenylation Splicing of introns RNA editing
RNA export	Export of mRNA from nucleus (eukaryotes) RNA targeting
RNA turnover	RNA degradation Retroregulation (prokaryotes)
Protein synthesis	Ribosome binding/initiation of protein synthesis Codon usage Frameshifting Readthrough/selenocysteine incorporation
Protein modification	Chemical modification of residues Cleavage of polypeptides Splicing of inteins Adoption of quaternary structure
Protein targeting	Interaction with regulatory proteins Targeting to cellular compartments Secretion
Protein turnover	Protein loss and degradation

Each level is covered in more detail in other chapters.

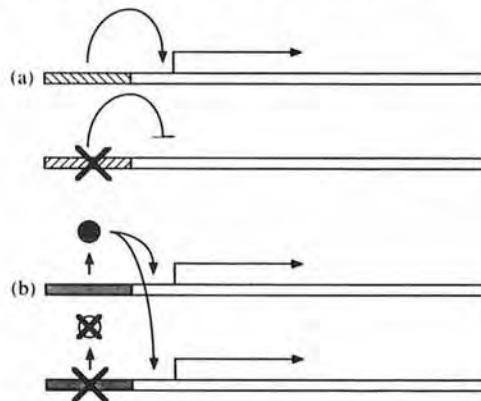
important but rarely emphasized level of gene regulation is turnover, i.e. the rate of RNA and protein degradation. The destruction of proteins can be regarded as the final level of gene expression, and is particularly important in the *cell cycle* (q.v.).

## 9.2 Gene regulation

**Principles of gene regulation.** The term **gene regulation** refers to the many mechanisms used by cells to control the amount of each gene product they produce. Even the simplest cells have hundreds of genes, the products of which are not all required at the same time nor in the same amounts. However, gene regulation has evolved not only as a mechanism to conserve resources by expressing individual genes only as needed, but also to prevent unproductive interactions between gene products. Furthermore, without gene regulation there could be no differentiation in **metazoans** (multicellular organisms), i.e. there could be no different cell types. Regulation can be *constitutive*, i.e. the rate of expression is fixed based on the general requirements of the cell, or *modulated* so that the rate of expression can vary to suit the needs of the cell under different circumstances, or in metazoans, to reflect the specialized properties of different cell types. In principle, gene regulation can occur at any of the levels of gene expression shown in Table 9.2, but in practice, occurs predominantly at the level of transcriptional initiation. This is the first level of expression, so resources are not wasted on unnecessary transcription.

**Cis-acting elements and trans-acting factors.** Regulatory processes of nucleic acids work either in *cis* or in *trans* (Figure 9.1). In DNA, **cis-acting elements** are nucleotide sequences which regulate genes to which they are attached (i.e. genes on the same chromosome; genes in the *cis*-configuration) while having no effect upon genes on other chromosomes (those in the *trans*-configuration). *Cis*-acting elements present in DNA regulate transcription, and those transcribed into mRNA can regulate RNA processing, turnover and protein synthesis. Mutations affecting such elements are **cis-dominant** because they affect only the gene to which they are attached and cannot be complemented by a wild-type copy of the same element supplied in *trans* in the same cell (q.v. *complementation*).

**Trans-acting factors** are diffusible molecules (usually proteins, but sometimes RNAs) regulating the expression of the genes with which they come into contact. *Trans*-acting factors regulating transcription are termed **transcription factors** (q.v.); others control RNA processing and protein synthesis. Mutations affecting such factors (i.e. mutations in the genes which encode them) are generally **recessive** (q.v.) because a wild-type copy of the gene in the same cell can supply more of the diffusible



**Figure 9.1:** *Cis*-acting elements and *trans*-acting factors. Mutations are shown as crosses. (a) Mutations in *cis*-acting elements are dominant in the *cis*-configuration because they cannot be complemented by a wild-type element in the same cell. (b) Mutations affecting (diffusible) *trans*-acting factors are usually recessive because a second source of wild-type factor will complement the mutation. Occasionally, mutant factors (white circle) may disrupt the function of the wild-type factors (black circle) resulting in *trans*-dominance.

factor. A special class of **trans-dominant** mutation causes the mutant factor to interfere with wild-type function, either by binding irreversibly to target sequences or by forming inactive multimers (q.v. *dominant negative mutation*).

Occasionally, *cis*-acting factors and *trans*-acting elements are used. *Cis*-acting proteins associate immediately with the DNA strand that encoded them and become irreversibly bound. Such factors exist only in prokaryotes, where transcription and protein synthesis are tightly coupled; examples include the transposase encoded by *Tn5* (see Mobile Genetic Elements) and the A protein encoded by phage  $\phi$ X174 (see Viruses). *Trans*-acting elements are encountered for a group of phenomena termed **trans-sensing**, where regulatory elements in one chromosome can influence gene expression on the homologous partner in *trans*. Prevalent amongst these processes is **transvection**, where a promoter in one homologous chromosome can control gene expression in the paired homolog, possibly because the transcriptional activator spans both duplexes. Such **synapsis-dependent processes** occur in *Drosophila*, where homologous chromosomes are paired in mitotic cells; they may occasionally be seen in other eukaryotes, including humans, when homologous chromosomes associate fortuitously (also q.v. *homology dependent silencing*).

**Modes of gene regulation.** Gene regulation may be described as positive or negative and inducible or repressible.

If a gene is under **positive regulation**, the presence of a particular regulatory factor increases (**upregulates**) gene expression, whereas its absence results in a decrease (**downregulation**); the factor is an **activator**. In **negative regulation**, the presence of a particular regulatory factor reduces gene expression and its absence results in an increase; the factor is a **repressor**.

The default state of gene expression usually dictates how the gene is regulated. **Inducible** genes are expressed at low levels, either because they are repressed (negative regulation), or because an activator is absent (positive regulation). Derepression (removal of the repressor) or activation (supply of the activator) results in the induction of gene expression. **Repressible** genes are usually active, either because a repressor is absent (negative regulation), or because an activator is present (positive regulation). Repression by removing the activator or supplying the repressor results in loss of gene expression.

These four modes of gene expression are often found in transcriptional control, but the concepts apply to any level of gene expression. In bacteria, all four modes of regulation are probably equally common at the level of transcription, whereas in eukaryotes, positive inducible control is predominant because the basal transcriptional apparatus is inherently unstable, and genes are transcriptionally repressed unless positively acting factors are supplied (see Transcription).

**The scope of gene regulation.** Global regulation, wide domain regulation, coregulation and narrow domain regulation are terms used to describe the scope of regulatory mechanisms. Global regulatory mechanisms control many genes at once and reflect fundamental changes in the cell. Global transcriptional regulation in eukaryotes may be controlled by chromatin structure (e.g. transcriptional shutdown during mitosis). Similarly, protein synthesis may be globally regulated by either preventing pre-mRNA processing or regulating the synthesis of the translation elongation factors. Wide- and narrow-domain regulatory mechanisms refer to the control of specific genes on a group basis or an individual basis, respectively, e.g. by transcriptional activation at individual promoters and enhancers. Examples of wide-domain regulation include the coregulation of yeast amino acid biosynthetic genes by starvation, the *dosage compensation* (q.v.) for genes on the female *Drosophila* X-chromosomes, and the SOS response in *E. coli*. Narrow-domain control includes examples of individual gene regulation and coordinated regulation of small groups of genes, such as the regulation of the three genes of the *E. coli lac* operon by lactose.



### 9.3 Gene expression in prokaryotes and eukaryotes

**Comparison of gene regulation strategies.** Although both prokaryotes and eukaryotes adhere to the central dogma, there are several differences in fundamental biology which reflect upon their mechanisms and strategies for gene expression and regulation.

(1) *The eukaryote nucleus.* Bacteria lack a nucleus, and therefore all stages of gene expression occur in the same compartment. Transcription, protein synthesis and RNA turnover are therefore closely linked, permitting regulatory mechanisms such as *attenuation* and *retroregulation* (q.v.) which do not exist in eukaryotes. The presence of a nucleus in eukaryotes separates transcription and protein synthesis both spatially and temporally. An additional level of control is thus imposed by export of the mRNA from the nucleus.

(2) *Coordinated cis-regulation.* In bacteria, genes with common functions are often arranged in a cluster termed an *operon* (q.v.) under common transcriptional control. Multiple genes are expressed as a polycistronic mRNA, which permits gene regulation based on the position of the gene within the operon (e.g. antitermination of transcription and sequential dependence of ribosome binding for protein synthesis). The close spacing of genes means that strategies such as *countertranscription* (q.v.) from adjacent promoters can be used to regulate gene expression. Operon structure is almost entirely absent from eukaryotic systems because ribosomes can bind only to the modified 5' end of mRNA, and not internally (c.f. *internal ribosome entry site*). Eukaryotic genes are also further apart than bacterial genes. Both these factors contribute to the absence of coordinated *cis*-transcription in eukaryotes, although transcriptional *cis*-coregulation is possible due to the sharing of distant regulatory elements.

(3) *Introns.* Higher eukaryotic genes are generally interrupted by one or more *introns* (q.v.) whose information is not represented in the final gene product. Introns are rare in some lower eukaryotes (e.g. the yeast *S. cerevisiae*) and are almost entirely absent from bacteria. Introns are removed from eukaryotic RNA by *splicing* (q.v.), allowing (a) regulation at a level between transcription and protein synthesis, and (b) selective use of the information in the gene. Bacterial genes therefore normally encode a single product whereas many eukaryotic genes can encode multiple overlapping products.

(4) *Chromatin.* Regulation of gene expression by DNA structure is a common global mechanism in both prokaryotes and eukaryotes because structure influences the binding of transcriptional regulators. In eukaryotes, DNA is organized into a highly ordered nucleoprotein complex termed *chromatin* (q.v.), whose structure can alternate between an accessible and potentially transcriptionally active form and a condensed form which is transcriptionally repressed. The regulation of gene activity by chromatin structure allows large sections of DNA to be repressed or activated *en masse*, which may be an advantage to organisms with large genomes.

(5) *RNA polymerase activity.* In bacteria, the RNA polymerase holoenzyme has an integral subunit, the  $\sigma$ -factor, which facilitates promoter recognition and binding. If a consensus promoter is present, efficient transcription is possible in the absence of regulation. In eukaryotes, RNA polymerase is unable to recognize the promoter and requires a collection of basal transcription factors to facilitate loading and template processing. This initiation complex is unstable and supports only minimal transcription. Efficient transcription therefore requires constitutive positive regulators to stabilize the initiation complex.

(6) *Transcriptional regulation.* In bacteria, the regulation of transcription is mediated predominantly by factors directly interacting with RNA polymerase, and such factors are often under allosteric control. This, together with the high turnover of mRNA, facilitates a rapid response to dynamic environmental conditions. In eukaryotes, the regulation of transcription often depends upon factors which bind at a distance from the basal promoter, and interaction is mediated by looping out of intervening DNA. This allows a form of coordinated *cis*-regulation where genes compete



for distant enhancers and other regulatory elements (q.v. *enhancer competition*, *locus control region*). In bacteria, the allosteric modification of preexisting transcription factors is used to mediate a rapid response to inductive events outside the cell. In eukaryotes, posttranslational modification of transcription factors is also a common regulatory strategy, but the *de novo* synthesis of new transcription factors is used for relatively slow and lasting responses such as differentiation during development (see Development: Molecular Aspects).

### Further reading

- See the other chapters concerning gene regulation in this book: Transcription, RNA Processing, Protein Synthesis, Chromatin, DNA Methylation and Epigenetic Regulation, Nucleic Acid Structure, Recombination, Signal Transduction, The Cell Cycle, Oncogenes and Cancer, Development: Molecular Aspects.
- Latchman, D. (1995) *Gene Regulation. A Eukaryotic Perspective*. Chapman & Hall, London.
- Lewin, B. (1997) *Genes VI*. Oxford University Press, Oxford.
- Lin, E.C.C. and Lynch, A.S. (Eds) (1996) *Regulation of Gene Expression in Escherichia coli*. Chapman & Hall, London.

## Chapter 10

# Gene Transfer in Bacteria

### Fundamental concepts and definitions

- **Gene transfer** describes the introduction of genetic information into a cell from an exogenous source (ultimately, another cell). This process occurs naturally in both bacteria and eukaryotes, and may be termed **horizontal** or **lateral genetic transmission** to distinguish it from the transmission of genetic information from parent to offspring, which is **vertical genetic transmission**.
- Intraspecific gene transfer facilitates genetic mixing in asexual species and thus mimics the effects of sexual reproduction. Such **parasexual exchange mechanisms** have been exploited to map prokaryote genomes analogously to *meiotic mapping* (q.v.) in eukaryotes. Interspecific gene transfer can also occur, and is a natural mechanism of *transgenesis* (q.v.). Interspecific gene transfer is an important evolutionary process and has been responsible for some of the most fundamental evolutionary events (e.g. the endosymbiont origin of eukaryotic organelles) as well as facilitating specific interactions between bacteria and eukaryotes (e.g. tumor-induction by *Agrobacterium tumefaciens*; q.v. *Ti plasmid*).
- The source of the transferred information is the **donor** and the genetic information transferred is the **exogenote** (exogenous genome, usually only a fragment of the donor genome). The target of the gene transfer, the **recipient**, possesses the **endogenote** (endogenous genome). If the exogenote is homologous to part of the endogenote, gene transfer will make the recipient cell partially diploid (a **merozygote**), in which case recombination can occur, which may involve allele replacement (**marker exchange**).
- There are four major mechanisms of gene transfer in bacteria: cell fusion, conjugation, transformation and transduction (Table 10.1). All cases of gene transfer must involve introduction of DNA into the recipient and a resolution phase, where the fate of the exogenote is determined. There are four consequences of gene transfer: rejection (destruction), maintenance (i.e. in the cytoplasm), replacement (recombination with the endogenote) and addition (integration into the endogenote). Replacement can only occur if the exogenote is homologous to part of the endogenote.
- Artificial gene transfer has been exploited for cloning and the expression of cloned genes, strain construction, targeted mutagenesis, the generation of transgenic organisms and the analysis of gene expression. In these cases, the exogenote is a recombinant DNA molecule and for a discussion of such approaches see Recombinant DNA.

### 10.1 Conjugation

**Requirements for conjugation.** Bacterial conjugation involves the transfer of genetic information from one cell to another while the cells are in physical contact. The donor cell is arbitrarily defined as the **male** and the recipient cell the **female**, and following gene transfer, the recipient may be termed a **transconjugant**. The ability to transfer DNA by conjugation is conferred by a **conjugative plasmid** (or more rarely a **conjugative transposon**), a **self-transmissible** element which encodes all the functions required to transfer a copy of itself to another cell by conjugation. If the conjugative plasmid can also facilitate the transfer of chromosomal genes, it is termed a **sex factor**.

The **F plasmid** (**F factor**, **fertility factor**) is an *E. coli* conjugative plasmid which confers no phenotype upon its host except that of **fertility** (the ability to instigate conjugal transfer). The F factor was the first conjugation system to be discovered and is the best characterized (see Box 10.1). In its autonomous state, F promotes self-transfer from donor (male, F<sup>+</sup>) cells to recipient (female, F<sup>-</sup>) cells with great efficiency because, unlike other conjugative plasmids, its conjugation genes are

**Table 10.1:** The four principle mechanisms of gene transfer in bacteria

Gene transfer mechanism	Defining features
Conjugation	DNA is transferred directly from cell to cell, often through a specialized conduit Conjugation evolved as a mechanism of plasmid transmission and requires a number of specialized functions, usually plasmid-encoded
Transformation	Naked DNA is taken into cells from their surrounding medium. In bacteria, transformation refers to the uptake of naked <i>chromosomal</i> or <i>plasmid</i> DNA from surrounding medium. The uptake of naked <i>viral</i> genomic DNA/RNA is termed <b>transfection</b> . Few species are naturally competent for transformation, but an artificial state of competence can be induced in some refractory species by various chemical treatments (see Recombinant DNA) In animal cells, the uptake of any naked DNA from the surrounding medium is termed <b>transfection</b> , although the term transformation may be used to specify a change in genotype thus brought about. The introduction of naked DNA into yeast and plant cells may be described as transfection or transformation
Transduction	The transfer of chromosomal or plasmid DNA into a cell mediated by a virus <i>particle</i> . The DNA may be packaged into the capsid instead of the virus genome (generalized transduction), or may be covalently attached to the virus genome itself (specialized transduction)
Cell fusion	Gene transfer occurring in some bacterial species (e.g. <i>Streptomyces</i> spp.) mediated by the fusion of plasma membranes

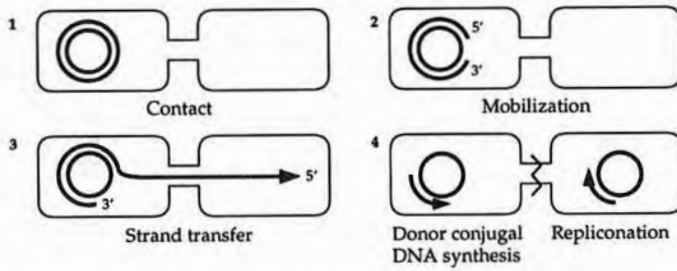
Note how usage of the terms transformation and transfection differs between bacteria and eukaryotes.

constitutively derepressed. Thus in a mixed population of male and female cells, most female cells receive a copy of F and become male, although at high male:female ratios the females may be killed by conjugal transfer, a phenomenon termed **lethal zygotis**. **R plasmids** (**R factors**, **resistance factors**) are conjugative plasmids which also carry antibiotic resistance markers. R plasmids do not promote conjugation with the same efficiency as F because the transfer genes are repressed; they also demonstrate **fertility inhibition** over F by repression of F transfer genes in *trans*. This can occur in the F-system because F-type plasmids are spread over two *incompatibility groups* (q.v.).

**Transfer and fate of plasmid DNA.** Conjugation requires contact between cells so that DNA can be passed from donor to recipient. In Gram-negative bacteria (e.g. *E. coli*), such contact involves a plasmid-encoded proteinaceous tube (**pilus**, **sex pilus**, **conjugation tube**; plural **pili**) which extends from the donor and binds to the surface of the recipient before retracting and pulling the conjugating cells close together. The pilus is obligatory for successful conjugation and can dictate the conditions under which conjugation occurs. Once successful contact has been made, a signal is released which promotes the mobilization of DNA. In Gram-positive bacteria (e.g. *Streptococcus* spp.), conjugation occurs in the absence of pili and requires direct contact between cells.

Once contact has been established, the DNA of the conjugative plasmid is **mobilized** (prepared for transfer). In the case of the F plasmid, this involves generating single-stranded DNA by introducing a nick at the **origin of transfer**, *oriT*, and unwinding single-stranded DNA to pass into the pilus. The DNA is transferred 5' end first, and thus loci 5' to the nick at *oriT* are the first to enter the recipient. In the recipient cell, the single-stranded DNA is used as a template to generate a double-stranded molecule which recircularizes, a process known as **replication**. In the donor cell, the remaining single strand is also used as a template to replace the transferred strand (**donor conjugal DNA synthesis**). Conjugation is thus a semiconservative process (Figure 10.1).

**Transfer and fate of chromosomal DNA.** A conjugative plasmid is not only self-transmissible, but may also facilitate the conjugal transfer of chromosomal or other plasmid DNA which is not



**Figure 10.1:** Steps in the conjugal transfer of plasmid DNA.

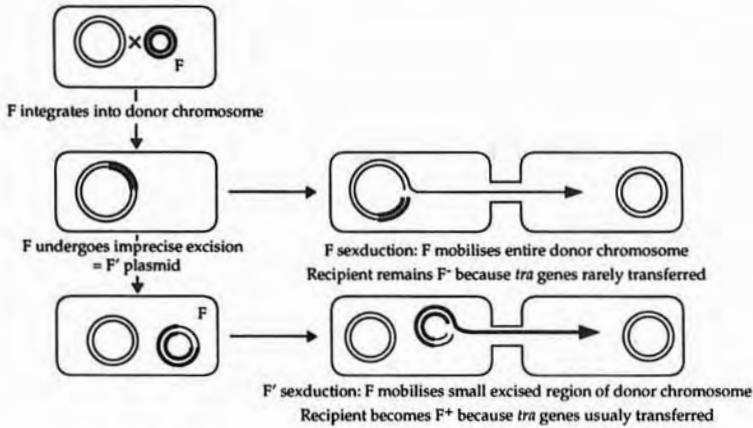
self-transmissible. This may occur in *cis* if the foreign DNA becomes covalently joined to the conjugative plasmid (**cis-mobilization** or **conduction**) or in *trans* if another plasmid is mobilizable (i.e. it can respond to *trans*-acting mobilization factors such as the endonuclease TraY but does not encode such factors itself; **trans-mobilization** or **donation**).

Where conduction involves the mobilization and transfer of chromosomal genes, the process may facilitate parasexual exchange and is termed **sexduction** (or **F-duction** may be used in the case of the F-plasmid) and can be exploited to map prokaryote genes (*Box 10.2*). The F-plasmid can integrate into the host chromosome following homologous recombination between *IS* elements (q.v.). Once integrated, it can *cis*-mobilize the chromosomal DNA to which it is covalently joined, allowing the conduction of chromosomal DNA from one individual to another. The recipient cell thus becomes diploid for the conducted genes, and if the donor and recipient cells are of different genotypes, marker exchange can occur following homologous recombination between the exogenote and endogenote. Strains of bacteria with integrated F plasmids are thus known as **Hfr strains** because they facilitate a *high frequency of recombination* between chromosomal markers of donor and recipient origin. A low frequency of marker exchange is observed when donor strains with autonomous F plasmids are used, and this is thought to result from the small number of episomal F-carriers within it. Whereas conjugal transfer with autonomous F plasmids always results in the F<sup>-</sup> (female) recipient cells becoming F<sup>+</sup> (male), this never occurs when using Hfr strains. The reason for this reflects the position of *oriT* within the transfer region. Since *oriT* is located on the 5' edge of the transfer region (see *Box 10.1*) and loci immediately 5' to *oriT* enter the recipient cell first, the transfer genes are always the last to enter the recipient. With autonomous F this presents no problem because the whole plasmid is usually transferred; however, if the plasmid is integrated into the bacterial chromosome, conjugal transfer is almost always interrupted before the entire genome can be transferred and the recipient cells remain F<sup>-</sup>.

The episomal F plasmid occasionally undergoes imprecise excision from the host genome and takes with it part of the chromosomal DNA (q.v. *prime plasmid*). The augmented plasmid, now known as an **F' plasmid**, can be transferred to F<sup>-</sup> bacteria, making them F<sup>+</sup> and diploid for the extra genes. A library of F' plasmids (**fosmid library**, c.f. *cosmid*) can be isolated carrying regions comprising the entire *E. coli* chromosome, and such plasmids can be used for a variety of purposes: in recombination-deficient hosts, the partial diploids can be used to study the effects of mutations in *trans* by *complementation* (q.v.), and it is this approach which led to detailed characterization of the *lactose operon* (q.v.); in recombination-proficient recipients, marker exchange can occur between the chromosome and plasmid-borne genes, a sexduction process with mechanistic similarity to *specialized transduction* (q.v.) which can be exploited for strain construction. Sexduction by F plasmids and F' plasmids is compared in *Figure 10.2*.

## 10.2 Transformation

**Requirements for transformation.** In bacteria, **transformation** describes the uptake of naked DNA from the surrounding medium, and the change in genotype thus conferred upon the recipient cell,



**Figure 10.2:** Sexduction: the transfer of chromosomal DNA during conjugation. The F plasmid can mediate sexduction either by conducting the chromosome into which it has integrated, or by excising imprecisely and conducting the chromosomal genes it has captured.

the **transformant**. Transformation occurs naturally in many bacteria (e.g. *Bacillus*, *Streptomyces* and *Haemophilus* spp.) although **competence** (ability to take up exogenous DNA) is usually transient, being associated with a particular physiological state and requiring the expression of specific **competence factors**. Other species of bacteria, including *E. coli*, are refractory to natural transformation, but a state of competence can be induced artificially which allows DNA uptake; this has facilitated the use of *E. coli* for molecular cloning (see Recombinant DNA).

**Transfer and fate of DNA.** In naturally competent cells, donor DNA is first bound reversibly to surface receptors. In some bacteria (e.g. *B. subtilis*), the DNA is processed by cleavage and degradation, with only one strand eventually being internalized. In others (e.g. *H. influenzae*), both strands enter the cell. Artificial transformation of *E. coli* also results in the internalization of intact DNA, possibly because the treatments involved work by increasing the permeability of the cell membrane to DNA. Generally, the binding and internalization of DNA is nonspecific, although, for example, *H. influenzae* takes up only DNA which contains a specific **DNA uptake site**, a sequence which is found with great frequency in the *H. influenzae* genome, thus ensuring that cells are only transformed with DNA from the same species.

If the transforming DNA is a plasmid, it may be maintained in the recipient cell as an autonomous replicon. Linear chromosomal DNA may undergo recombination with the host chromosome, resulting in marker exchange. Both these events result in **stable** or **permanent transformation**. Otherwise, transforming DNA may be degraded, in which case any characteristics it confers are short lived (**transient transformation**). Occasionally, transforming DNA may integrate into the host chromosome by illegitimate end joining, although this is a more common occurrence when linear DNA is introduced in eukaryotic cells because of the abundance of end-joining repair enzymes (q.v. *transfection*, *transgenesis*, *illegitimate recombination*).

### 10.3 Transduction

**Transfer of DNA by generalized and specialized transduction.** **Transduction** is the process by which cellular genes can be transferred from a donor to a recipient cell by a virus particle, the recipient being known as a **transductant** following transfer (c.f. *signal transduction*). Natural transduction can occur in two ways (Table 10.2). In **generalized transduction**, chromosomal or plasmid DNA accidentally becomes packaged into phage heads instead of the phage genome. Since infection is a property conferred by the phage particle and not the nucleic acid it carries, this can be an efficient



**Table 10.2:** Properties of generalized and specialized transduction

	<b>Generalized transduction</b>	<b>Specialized transduction</b>
Contents of transducing particle	Host DNA only, theoretically any sequence	Host DNA covalently linked to phage DNA. Phage often defective. Host DNA limited to that flanking prophage insertion site
Mechanism	Particles formed during lytic cycle by mistaken packaging of host DNA into capsid	Particles formed following aberrant prophage excision
Resolution	<b>Complete transduction:</b> homologous recombination with host genome — transductant is haploid <b>Abortive transduction:</b> donor DNA remains in cytoplasm and is not replicated — transductant is partial diploid but only one cell in population contains transduced genes	<b>Replacement transduction:</b> homologous recombination with host genome — transductant is haploid <b>Addition transduction:</b> lysogeny by transducing phage — transductant is partial diploid
Requirements	Low virulence (phage must not destroy host DNA before packaging), and must package DNA nonspecifically, i.e. by the <i>headfull mechanism</i> (q.v.)	Must be a temperate phage, i.e. must integrate into host genome
Host recombination system	Complete transduction requires bacterial <i>rec</i> system. Ratio of complete:abortive transductants increased by double-strand breaks in donor DNA	Replacement transduction requires bacterial <i>rec</i> system. Addition transduction may or may not require <i>rec</i> , depending on properties of defective phage
Examples	Bacteriophage P1 of <i>E. coli</i> ; Bacteriophage P22 of <i>S. typhimurium</i>	Bacteriophage $\lambda$ of <i>E. coli</i> ; Bacteriophage SP $\beta$ of <i>B. subtilis</i>

mechanism of gene transfer between cells, and any region of the chromosome can in theory be transduced. In **specialized** (or **restricted**) **transduction**, imprecise excision of a prophage results in the removal and packaging of some host DNA flanking the site of prophage insertion. The transduced DNA is thus covalently linked to the phage genome, and the genes which can be transduced are limited to those which flank the prophage integration site. Specialized, but not generalized, transduction is also observed in eukaryotes (q.v. *acute transforming retrovirus*).

Virus genomes can be exploited as *cloning vectors* (q.v.), and the transfer of cloned genes to the cloning host by first packaging the recombinant virus into its capsid can be regarded as artificial transduction; the use of recombinant bacteriophage  $\lambda$  vectors is artificial specialized transduction, because the cloned DNA is covalently joined to the  $\lambda$  genome. Conversely, cloning using cosmid vectors is more like generalized transduction because the  $\lambda$  genome is not used at all (see Recombinant DNA).

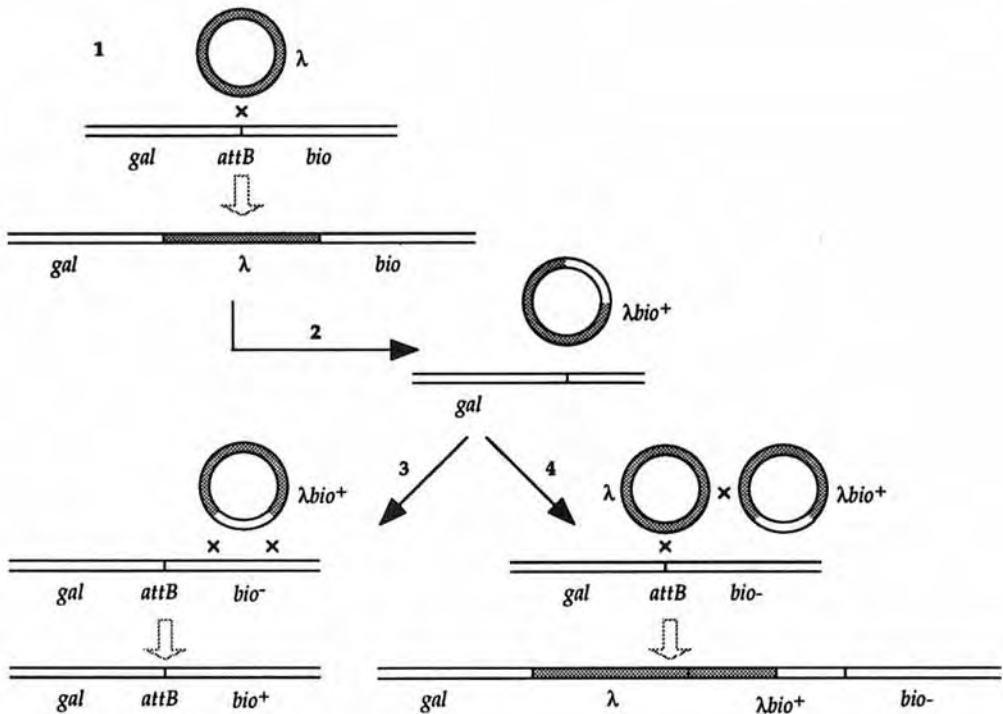
Most temperate bacteriophage can act as **specialized transducing phages**, but few phage are naturally competent for generalized transduction — this is because a **generalized transducing phage** must not destroy the host DNA, and must package DNA using a mechanism which does not require specific phage sequences, i.e. by the *headful mechanism* (see Viruses). A number of phages which cannot perform generalized transduction as wild-types can be rendered competent by specific mutations (e.g.  $\lambda$  and T7 phages).

**Generalized transducing phages.** The principal generalized transducing phages are P1 of *E. coli* and P22 of *S. typhimurium*. The genome of each is *circularly permuted* and *terminally redundant* (q.v.), which results from the packaging of concatemeric genomic DNA: the circular permutation results from the cutting of subsequent genomes from the same concatemer, and the terminal redundancy results from the filling of each head with a fixed length of genomic DNA which is greater than one genome in size, the *headfull mechanism* (q.v.). **Transducing particles** are formed when host DNA is incorporated into the head instead of the phage genome. Theoretically, any part of the host chromosome could be packaged and transduced with equal efficiency. In practice, however, different markers are transduced at frequencies, which vary by three orders of magnitude. This effect is especially apparent in P22 transductants, and reflects preferential packaging of specific chromosomal sites which resemble the chromosome **packaging sites** (*pac*) in the P22 genome. Mutant strains of phage which are deficient in packaging specificity have been generated, and these transduce all markers with similar frequency.

**Fate of generally transduced DNA.** Following the introduction of exogenous DNA into the recipient cell by the virus particle, several outcomes are possible. If the transduced DNA is a plasmid, it may be maintained in the cytoplasm as an autonomous replicon; large plasmids often become smaller following transduction (**transductional shortening**), a phenomenon reflecting the more efficient packaging of spontaneous deletion mutants. A linear fragment of transduced chromosomal DNA can synapse with a homologous region of the recipient genome and undergo *homologous recombination* (q.v.) and marker exchange; this is termed **complete transduction**. Linear DNA which remains in the cytoplasm may be degraded, or alternatively it may become stabilized in the cytoplasm as a deoxyribonucleoprotein particle; this is termed **abortive transduction**. The transduced genes may then be expressed, but lacking an origin they cannot be replicated, and proteins synthesized will be rapidly diluted from a growing population. Thus, if cells carrying an *auxotrophic* mutation (q.v.) are transduced with the corresponding wild-type allele, complete transductants will grow normally on selective media, whereas abortive transductants will grow very slowly and produce tiny colonies — this is because the transduced DNA may produce the enzyme required for cell growth but it will be inherited by only one of the daughter cells at each cell division. Thus, although enough enzyme may remain in each of the daughter cells to allow growth for several generations, it will eventually be diluted and degraded, so that growth ceases in all cells except those carrying the fragment itself.

**Specialized transduction by bacteriophage  $\lambda$ .** Specialized transduction arises from aberrant prophage excision events (see Viruses) and the host genes transduced are those flanking the site of prophage insertion. Many temperate phages have specific insertion sites. Bacteriophage Mu is an exception because it replicates by repeated transposition with little target-site preference. Mu can act as a general transducing phage, and deleted derivatives of Mu can act as specialized transducing phage, a process termed **mini-Mu**duction. In the case of bacteriophage  $\lambda$ , **specialized transducing particles** carry either the *gal* or *bio* loci (which flank the  $\lambda$  insertion site *attB*). Transduction occurs at the expense of phage genes from the other end of the genome, resulting in a **defective phage** (a phage lacking essential genes which can only infect a host if missing functions are supplied in *trans* by a wild-type phage, known as a **helper phage**). Such aberrant excision events occur at low frequency, and therefore infection of a *bio*<sup>-</sup> recipient culture with a lysate derived from a *bio*<sup>+</sup> host results in few cells being infected and subsequently transduced by a *bio*<sup>+</sup> specialized transducing particle ( $\lambda$ *bio*<sup>+</sup>). Such lysates are thus termed **low-frequency transducing (LFT) lysates**. Rarely, other genes can be transduced by  $\lambda$  if it integrates at a site other than *attB*.

**Fate of specially transduced DNA.** If  $\lambda$ *bio*<sup>+</sup> infects a second host, which is *bio*<sup>-</sup>, two outcomes are possible (Figure 10.3). The transduced gene can recombine with the chromosomal locus facilitating marker exchange. This is **replacement transduction**, and the transductant becomes *bio*<sup>+</sup> but remains haploid for the *bio* locus. Alternatively, the DNA from the specialized transducing particle may

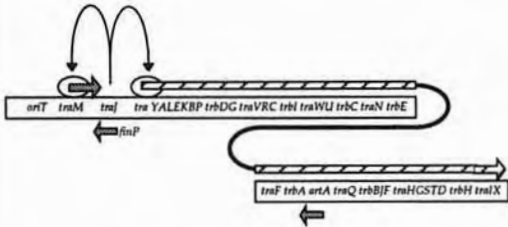


**Figure 10.3:** Specialized transduction in bacteriophage  $\lambda$ . (1)  $\lambda$  integrates at *attB* between the *gal* and *bio* loci of *E. coli*. (2) Aberrant excision generates a specialized transducing particle  $\lambda_{bio^+}$  which carries the *bio* gene. Subsequent infection of *bio*<sup>-</sup> host can lead to (3) replacement transduction by recombination, or (4) addition transduction by integration, the latter generating a  $\lambda:\lambda_{bio^+}$  double lysogen which generates high-frequency transduction lysates. Similar events can occur which involve the *gal* locus.

integrate into the host genome (**addition transduction**). This only occurs by homologous recombination with a helper prophage, as the specialized transducing particle DNA lacks efficient donor sites for site-specific recombination. Because there is a great excess wild-type phage in a LFT lysate, recipients infected with a transducing phage are also infected with wild-type particles. Integration of both phage genomes thus generates a **double lysogen** and the host is diploid for the *bio* locus. The transductant can be described as a **lysogenic merozygote** to show that the extra copy of *bio* arose through integration of a prophage. Subsequent induction of the double lysogen produces a lysate containing equal numbers of wild-type phage and specialized transducing particles which can transduce another population of *bio*<sup>-</sup> recipient cells at high frequency (**high-frequency transducing (HFT) lysate**).

**Box 10.1:** Transfer genes on the F plasmid

**The *tra* region and its regulation.** About one-third of the F plasmid comprises a single **transfer operon** whose  $\approx 35$  genes are positively regulated by the product of the *traJ* gene. At the 5' side of the operon are three loci, *oriT*, the origin of transfer, *traM*, a solitary transfer gene also under the control of TraJ, and *traJ* itself. These three loci and the downstream operon comprise the **transfer region**, carrying all the genes required for self-transmission. In other conjugative plasmids, *traJ* expression is repressed by the action of two genes *finO* and *finP*, the latter of which encodes a small RNA molecule which, in concert with FinO, binds to the *traJ* leader sequence and blocks its expression (the use of antisense RNA is a common strategy in *plasmid* (q.v.) gene regulation). In the F plasmid, the *finO* gene is disrupted by insertion of an *IS element* (q.v.) and the transfer genes are constitutively expressed. This is the *IS element* that recombines with the *E. coli* chromosome.



The *tra* region of the F plasmid, comprising (from 5'→3') *oriT*, *traM*, *traJ* and the *tra* operon. The *tra* operon is continuous; there is no physical break between *trbE* and *traF*, as shown above.

**Genes of the *tra* operon.** The functions of about 20 of the transfer genes are known. For others, the size

of the product and, in many cases, its subcellular localization and likely partners for interaction are known, but a precise function has yet to be determined. The *traI* locus encodes two products TraI and TraI\* (formerly TraZ) by nested translation. The function of TraI\* is unknown.

Gene	Function of gene product
<i>traM</i>	Envelope protein, possibly involved in cell-cell recognition
<i>traJ</i>	Positive transcriptional regulator of <i>traM</i> and <i>tra</i> operon
<i>traY</i>	Endonuclease, nicks DNA at <i>oriT</i>
<i>traA</i>	Encodes pilin, the major subunit of the pilus
<i>traL</i> , <i>traE</i> , <i>traK</i> , <i>traB</i> , <i>traV</i> , <i>traC</i> , <i>traW</i> , <i>traU</i> , <i>trbC</i> , <i>traF</i> , <i>traH</i> , <i>traG</i>	Pilus assembly and mating aggregate stabilization
<i>traN</i>	Mating aggregate stabilization
<i>traS</i> , <i>traT</i>	Surface exclusion <sup>1</sup>
<i>traD</i>	DNA transport during conjugation
<i>traI</i>	Helicase, required for unwinding DNA for transfer. Also has endonuclease activity

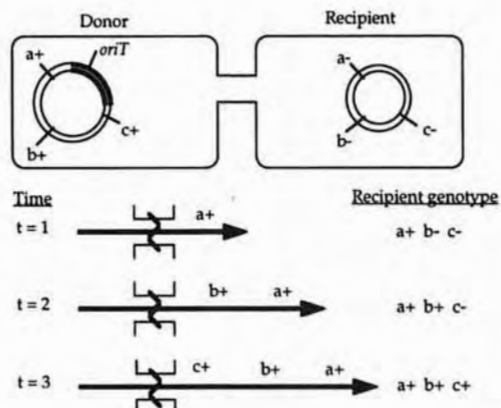
<sup>1</sup>**Surface exclusion** is the phenomenon which prevents donor bacteria conjugating with other donors carrying the same conjugal system.

**Box 10.2: Bacterial linkage mapping**

**Genetic mapping of bacteria.** Parasexual exchange in bacteria can be exploited to map genes in the same way that meiotic exchange is used in eukaryotes (see Recombination, Genomes and Mapping). Bacterial linkage mapping strategies are elegant and have provided detailed maps of several prokaryote genomes, but they have now been superseded by brute force physical approaches made possible by rapid and accurate methods for determining the order of genomic DNA clones, and rapid, high-throughput sequencing strategies (see Recombinant DNA, Genomes and Mapping). The mapping methods described below demonstrate the power of genetic analysis but are now only of historical interest.

**Mapping by sexduction.** If a wild-type Hfr strain of *E. coli* is mixed with F<sup>-</sup> cells carrying a number of mutations, wild-type alleles are passed from the donor to the recipient and marker exchange may occur, so that the transconjugant becomes wild-type at one or more of the mutant loci. Because the chromosome is transferred to the recipient in a linear fashion, starting with the DNA immediately 5' to *oriT*, markers in the donor chromosome which lie close to *oriT* on the 5' side will enter the cell first. In any population of cells, individual conjugating pairs may separate randomly; thus markers lying further from *oriT* are less likely to enter the recipient cell. This establishes a **gradient of transfer**, where markers proximal to the 5' border of *oriT* are transferred more frequently than distal markers, and undergo marker exchange at a greater frequency. The gradient of transfer can be exploited to order gene loci. If conjugation is initiated by mixing Hfr and F<sup>-</sup> cells and then samples removed and vortexed to break up conjugating pairs at various times, more of the chromosome will have been transferred in later samples, and markers further away from *oriT* will have undergone recombination. Such **interrupted mating** experiments can be used to order genes on a chromosome and estimate the map distance between them, which in this case would be measured in minutes (see figure below).

**Mapping by cotransduction and cotransformation.** Whilst interrupted mating gives the order of gene loci and an idea of the distance between them, it is not useful for fine mapping because only a few seconds may separate the transfer of closely linked markers. For detail, **cotransduction** or **cotransformation mapping** may be used, depending upon the species of bacteria. Generalized transduction may



The principle of mapping by interrupted mating. An Hfr donor conducts chromosomal DNA to the recipient in a linear manner. After time *t*, conjugation is interrupted by vortexing (zigzagged line). Loci nearest *oriT* on the 5' side will be transferred first, so at time *t=1* only marker *a* has been transferred and exchange generates transconjugants with genotype *a+b-c-*. At time *t=2*, marker *b* is transferred and exchange generates *a+b+c-* transconjugants. At time *t=3*, marker *c* is transferred and *a+b+c+* transconjugants are obtained. This establishes gene order and allows a rough map to be constructed with distances reflecting the time taken to transfer each marker.

occur frequently (1–5% of wild-type P1 and P22 virions are transducing particles), but the successful transduction of any particular locus occurs at low frequency (approximately  $10^{-6}$ ) and **cotransduction** (simultaneous transduction at two loci) is taken as evidence that the two genes are linked on the same DNA fragment. Fine-scale gene mapping can therefore be carried out by measuring the **cotransduction frequency** of pairs of markers on the donor chromosome, with higher cotransduction frequencies corresponding to tighter linkage. In *B. subtilis*, mapping by **cotransformation** (simultaneous transformation at two loci) works on the same principles. At limiting concentrations of DNA (so that each cell is likely to take up only one DNA fragment) **cotransformation** is evidence of linkage, and the **cotransformation frequency** reflects the distance between the loci. If a number of linked genes is cotransferred to a recipient cell by any of the three mechanisms described above, the distance between loci can be estimated by **recombination mapping**, i.e. determining the recombination frequency between pairs of markers and using this as an estimate of physical distance.



### Further reading

- Frost, L.S., Ippen-Ihler, K.A. and Skurray, R.A. (1995) Analysis of the sequence and gene products of the transfer region of the F sex factor. *Microbiol. Rev.* **58**: 162–210.
- Lanka, E. and Wilkins, B.M. (1996) DNA processing reactions in bacterial conjugation. *Annu. Rev. Biochem.* **64**: 141–169.
- Masters, M. (1985). Generalised transduction. In: *Genetics of Bacteria*, (eds Scaife, J., Leach, D. and Galizzi, A.) pp. 197–215. Academic Press, London.
- Solomon, J.M. and Grossman, A.D. (1996) Who's competent and when — regulation of natural genetic competence in bacteria. *Trends Genet.* **12**: 150–155.
- Summers, D.K. (1996) *Plasmid Biology*. Blackwell Science, Oxford.

# Chapter 11

## The Genetic Code

### Fundamental concepts and definitions

- **Genetic information** is the sequence of bases in DNA. Information may also be carried in the physical structure of the chromosome, which may be indirectly specified by the base sequence; this is termed **epigenetic information** (see Chromatin, DNA Methylation and Epigenetic Regulation, Nucleic Acid Structure).
- Genetic information that is expressed as a polypeptide sequence must be **translated** from the four-letter language of nucleotides into the 21-letter language of proteins. The system which allows base sequences in nucleic acids to specify amino acid sequences in polypeptides, and thus allows genes to encode the structure of proteins, is the **genetic code**. The genetic code is a near universal triplet code which is nonambiguous but highly degenerate.
- Translation is facilitated by transfer RNAs: small RNA molecules which act as adaptors possessing both an amino acid binding site and an anticodon which recognises codons in mRNA. The tRNAs therefore bring the correct amino acid to the ribosome for protein synthesis.
- Proteins contain many different types of amino acids, but only 21 are specified by the genetic code, the remainder arising through posttranslational modification (see Proteins).

### 11.1 An overview of the genetic code

**The nature of the code.** The genetic code uses combinations of three bases to specify a particular amino acid. A group of three bases is termed a **codon** or a **triplet**, the former applying to the mRNA, the latter to the gene. The linear order of **sense codons** in the mRNA dictates the linear order of amino acids in the polypeptide. Special codons also direct initiation and termination of the polypeptide chain. **Initiator codons** always encode the amino acid methionine (*N*-formylmethionine in prokaryotes and organelles), whereas **stop codons** (or **termination codons**) do not encode an amino acid but act as a signal for **release factors** which disassemble the ribosome. Stop codons are also termed **nonsense codons** because they have no translatable sense. The nonsense codons are identified by the names **amber** (UAG), **ochre** (UAA) and **opal** (or rarely **umber**, UAG). The properties of the genetic code are listed in *Table 11.1*. The universal genetic code is shown in *Figure 11.1*.



Codons are 'read' by the protein synthesis machinery from initiation to termination in a nonoverlapping and unpunctuated manner, i.e. each triplet of nucleotides is independent of any other and there are no gaps between codons. This allows any nucleotide sequence to be divided up into codons in three ways, i.e. there are three **reading frames** (*Figure 11.2*). The reading frame used for protein synthesis is determined by the position of the initiation codon, although the frame may occasionally be changed during translation (q.v. *frameshift mutation*, *cotranslational frameshift*). In genome sequence data, potential genes are often identified as **open reading frames (ORFs)**, i.e. long sequences of uninterrupted sense codons, because incorrect reading frames usually contain one or more stop codons (this, incidentally, is why many frameshift mutations cause truncation of the gene product). Where an open reading frame has been identified but the product has not been characterized, it may be termed an **unassigned reading frame (URF)**.

### 11.2 Translation

**Deciphering the code during protein synthesis.** The stage of protein synthesis described as **translation** refers to the decoding of genetic information, i.e. when a codon in mRNA is matched to its

**Table 11.1:** Properties of the genetic code and known exceptions

Property	Meaning	Exceptions
Universal	The code is the same in all organisms (the <b>universal genetic code</b> )	Minor exceptions are found in organelles and lower eukaryote genomes (see <i>Table 11.2</i> )
Nonoverlapping	Each codon is an independent unit, hence the code is read in the sequence 1-2-3, 4-5-6, 7-8-9 rather than 1-2-3, 2-3-4, 3-4-5, etc.	Cotranslational frameshifts and bypassing (q.v. <i>regulation of translation</i> )
Unpunctuated	There are no gaps between codons	Uncommon initiator codons
Unambiguous	Each codon specifies one amino acid only, or specifies termination	Selenocysteine incorporation at UGA
Degenerate	There are 61 sense codons and 21 amino acids. Each amino acid (or stop) may thus be specified by more than one codon	The code as a whole is degenerate, but tryptophan and (internal) methionine are each specified by a single codon in the universal code

		Second Position					
		U	C	A	G		
First position	U	Phe	Ser	Tyr	Cys	Third Position	U
		Leu					C
	C	Leu	Pro	His	Arg		A
				Gln			G
	A	Ile	Thr	Asn	Ser		U
		Met		Lys	Arg		C
	G	Val	Ala	Asp	Gly		U
				Glu			C

**Figure 11.1:** The universal genetic code. Amino acids are identified by their three-letter designation (see *Box 22.1*). Nonsense codons are identified by name. Note the dual functions of codons AUG and UGA.

Reading frame 1	.... <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> .....	....Ser-Ser-Ser-Ser....
Reading frame 2	....A <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> .....	....Ala-Ala-Ala-Ala....
Reading frame 3	....A <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> .....	....Gln-Gln-Gln-Gln....
Reading frame 1 fixed by position of initiation codon	.... <u>A</u> <u>T</u> <u>G</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> <u>C</u> <u>A</u> <u>G</u> .....	....Met-Ser-Ser-Ser....

**Figure 11.2:** Reading frames, and how the correct reading frame is chosen during protein synthesis.

amino acid. This process is facilitated by *transfer RNA (tRNA)* (q.v.), whose structure includes an **anticodon** (a sequence of three bases complementary to the codon) and an acceptor stem to which an amino acid is covalently joined. The tRNA is **charged** when attached to an amino acid and is referred to as an **aminoacyl-tRNA**. The aminoacyl-tRNA enters the *ribosomal A-site* and the

anticodon pairs with the codon. This places the amino acid adjacent to the *ribosomal peptidyltransferase site* which catalyses the transfer of the amino acid from the tRNA to the existing polypeptide chain (see Protein Synthesis for details).

**Fidelity of translation.** Accuracy during translation is insured by a diverse family of enzymes called **aminoacyl-tRNA synthetases** whose function is to charge tRNAs with their cognate amino acids. These enzymes are specific for their substrates, but do not conform to any particular structure and use idiosyncratic mechanisms of recognition (q.v. *transfer RNA* for discussion and see Nucleic Acid-Binding Proteins). There are as many aminoacyl-tRNA synthetases as there are amino acids, so each enzyme can recognize all *isoaccepting tRNAs* (q.v.).

The charging of tRNA occurs in two stages, termed activation and transfer, both of which may be proofread. During **activation**, the amino acid becomes linked to its aminoacyl-tRNA synthetase, generating an **activated complex**; this process is dependent upon ATP binding and hydrolysis. Many aminoacyl-tRNA synthetases bind the amino acid in a single recognition step, rejecting it if it does not fit the active site (tryptophan is recognized in this way). Those which choose between similar substrates (e.g. between valine and isoleucine) may bind either substrate during the activation step but reject the incorrect amino acid at a later stage. During **transfer**, the activated complex binds tRNA and the aminoacyl group is transferred from the enzyme to the 3' terminal adenosine of the tRNA, with the release of AMP. Two classes of aminoacyl-tRNA synthetases are discriminated by the nature of this reaction: class I enzymes transfer the amino acid to the 3' hydroxyl group whereas class II enzymes transfer the amino acid to the 2' hydroxyl group. The initial interaction with tRNA causes a conformational change in the enzyme which, as discussed above, rejects the amino acid if it is incorrect (**pretransfer proofreading**). After the amino acid has been transferred to the tRNA, the enzyme recognizes the shape of the product and hydrolyses the peptide bond if the tRNA has been charged incorrectly (**posttransfer proofreading**). Proofreading both before and after transfer is called the **two sieve proofreading mechanism**.

### 11.3 Special properties of the code

**Isoaccepting tRNAs and wobble base pairing.** The genetic code displays two types of degeneracy: **first and second base degeneracy** (where codons with different bases in the first two positions may encode the same amino acid) and **third base degeneracy** (where codons with different bases in the third position may encode the same amino acid). A collection of codons which specify the same amino acid is termed a **codon family** and the members are known as **synonymous codons**. The maximum size of a codon family is six, the minimum size is one (Figure 11.1).

First and second base degeneracy reflects the existence of **isoaccepting tRNAs**. Both prokaryotic and eukaryotic cells encode about 30 distinct species of tRNA, but because only 21 amino acids are specified by the genetic code, some of the tRNA species must bind the same amino acid. Such duplicate tRNAs may have the same anticodon sequence, in which case they are functionally interchangeable. Others carry different anticodon sequences and thus recognize different codons; they are termed isoaccepting tRNAs and their relative abundance may influence *codon usage* (q.v.).

Third base degeneracy is explained by the **wobble hypothesis** of Francis Crick, which predicts a relaxation in the normal base pairing rules between the third base of the codon and the first base of the anticodon (the **wobble position**), allowing a single tRNA species to recognize several different codons. The hypothesis suggests that normal bases become less discriminating in the wobble position, and in some cases can only recognize the type of base (purine or pyrimidine) rather than the specific base in the opposite strand. This explains the existence of degenerate codon families of two, in which the third position can be either purine or pyrimidine (e.g. AAR encodes lysine, AGY encodes serine). In other cases, the third base is irrelevant, explaining the existence of degenerate codon families of four (e.g. CCN encodes proline). Furthermore, the existence of rare bases in tRNA

permits promiscuous interactions (e.g. inosine can pair with adenine, cytosine or uracil). The exact wobble rules may differ between species, reflecting differences in tRNA modification.

**Codon usage.** Codon usage, choice, bias or preference all refer to the phenomenon where particular codons in a family are used preferentially in a particular organism. This differs from species to species, so that degenerate amino acids are encoded by only a proportion of their representative codons, but different codons are predominant in different organisms. First and second base preference usually reflects the relative proportions of different isoaccepting tRNA species. Third base preference may reflect complex wobble rules where particular pairing conformations are more stable. In either case, codon usage can be used as a mechanism of gene regulation (e.g. a rare codon can delay protein synthesis). Codon bias may also arise due to global effects, e.g. thermophiles favor codons containing guanine and cytosine to maintain high GC-content.

**Ambiguity in the genetic code.** The genetic code is for the most part unambiguous, and this property is essential for the faithful translation of genetic information. Two special circumstances exist where ambiguity is tolerated, but because of the uniqueness of each situation, there is no loss of fidelity.

Ambiguity is observed at initiation. The universal initiation codon AUG encodes methionine both at internal sites and at the initiator position (initiator methionine is modified to form *N*-formyl-methionine in prokaryotes and organelles). In bacteria, alternative initiation codons may be used: GUG is common in, e.g. *Micrococcus luteus* and GUG and UUG are used occasionally in *E. coli* (these codons specify valine and leucine at internal sites but always *N*-formylmethionine at the initiator position). Alternative initiator codons are very rare in eukaryote nuclear genomes, although CUG is used very occasionally. This type of ambiguity reflects the distinct molecular environments of the initiation and elongation stages of protein synthesis — methionine residues are not inserted at internal GUG and UUG sites in *E. coli* (see Protein Synthesis for further details).

A second situation where ambiguity arises involves insertion of the rare amino acid selenocysteine. This amino acid is similar to cysteine but contains selenium instead of sulfur, and is required for the efficient function of several gene products termed **selenoproteins** or **selenoenzymes**. Most of the unusual amino acids found in proteins are posttranslational modification products, but selenocysteine is generated by modification of serine *before* incorporation, and the amino acid therefore has its own cognate tRNA. Selenocysteine is specified by the codon UGA, which is usually recognized as a termination codon. In selenoprotein-encoding mRNAs, however, secondary structures known as **selenocysteine insertion sequences (SECIS)** cause the protein synthesis machinery to

Table 11.2: Variations in codon assignment

Codon(s)	Universal translation	System	Translation in exceptional system
AAA	Lysine	Some animal mitochondria	Aspartame
		<i>Drosophila</i> mitochondria	Serine
AGR	Argenine	Most animal mitochondria	Serine
		Vertebrate mitochondria	STOP
AUA	Isoleucine	Some animal and yeast mitochondria	Methionine
CUG	Leucine	<i>Candida cylindracea</i> nuclear genome	Serine
CUN	Leucine	Yeast mitochondria	Threonine
UAR	STOP	Some ciliate nuclear genomes (e.g. <i>Tetrahymena</i> )	Glutamine
UGA	STOP	Animal and yeast mitochondrial genomes <i>Mycoplasma capricolum</i> genome	Tryptophan
UGA	STOP	Some ciliate nuclear genomes (e.g. <i>Euplotes</i> )	Cysteine



translate the codon; the exact mechanism is unclear. In bacteria, SECIS are found in the coding region of the mRNA, whereas in eukaryotes, they are found in the 3' UTR. Proteins which interact with *E. coli* SECIS have been identified.

**Deviation from the standard genetic code.** The universality of the genetic code is remarkable, but as more organisms have been studied, subtle variations in **codon assignment** have been discovered (Table 11.2). Most deviations occur in mitochondrial genomes, probably reflecting the small number of proteins synthesized. However, in plant organelles, *RNA editing* (q.v.) is prevalent, and it is not clear whether all instances of deviation from the genetic code in plants are true variations or consequences of RNA editing prior to translation. Occasional changes also occur in bacterial genomes and eukaryote nuclear genomes, but usually involve the termination codons. The phylogenetic distribution of these changes indicates that the code is still evolving. Apart from these constitutive changes, there are also site-specific variations in codon assignment, i.e. effects where particular codons are interpreted in an unusual manner because of their position. Such effects include the insertion of selenocysteine at UGA codons, as discussed above, readthrough of stop codons, translational frameshifting and bypassing (q.v. *regulation of translation*). *RNA editing* (q.v.) can also be thought of as a deviation from the normal genetic code.

**Secondary genetic codes.** The recognition of nucleotide sequence information by tRNA is the cornerstone of the genetic code. Other biological systems also rely on the recognition of information in nucleic acids, primarily the DNA binding proteins which control transcription and other DNA functions. Efforts to identify an 'amino acid code' governing sequence-specific protein-DNA interactions have shown that no universal sequence-to-sequence correlation exists. However, for certain protein families, recognition codes are beginning to be characterized (see Nucleic Acid-Binding Proteins).

### Further reading

- Crick, F.C. (1990) *What Mad Pursuit: A Personal View of Scientific Discovery*. Penguin, London.
- Fox, T.D. (1987) Natural variation in the genetic code. *Annu. Rev. Genet.* 21: 67–91.
- Hatfield, D. and Diamond, A. (1993) UGA: A split personality in the universal genetic code. *Trends Genet.* 9: 69–70.
- Low, S.C. and Berry, J.M. (1996) Knowing when not to stop — selenocysteine incorporation in eukaryotes. *Trends Biochem. Sci.* 21: 203–208.
- Moras, D. (1992) Structural and functional relationships between aminoacyl-tRNA synthetases. *Trends Biochem. Sci.* 17: 159–164.

**This Page Intentionally Left Blank**

## Chapter 12

# Genomes and Mapping

### Fundamental concepts and definitions

- The **genome** is the full complement of genetic information in a cell, and contains the 'program' required for that cell to function. It can be thought of as either the total genetic material or, where there is more than one copy of the same information, the genetic material comprising a single copy of that information (the latter is sometimes termed the **haploid genome**). The number of redundant copies of the genome in the cell is its **ploidy**. In eukaryotes, over 99% of cellular DNA is found in the **nuclear genome**, but DNA is also found in organelles (see *Organelle Genomes*). Bacterial and organelle genomes are small and are usually single, circular chromosomes, although some linear bacterial genomes have been reported. Eukaryote nuclear genomes are comparatively very large and are split into multiple, linear chromosomes. Viruses show great diversity in genome structure (for discussion, see *Viruses*).
- Genomes are not simply random collections of genes. They have a functional higher-order structure and can be characterized in terms of their physico-chemical properties and sequence organization. The DNA of most organisms can be divided into several sequence components: **unique sequence DNA**, represented only once, and various classes of **repetitive DNA**. Most genes are found in unique sequence DNA, but some in moderately repetitive DNA correspond to highly conserved multigene families. Other repetitive DNA is not transcribed and consists of interspersed repeats (usually corresponding to active or mutated transposable elements), or in eukaryotes, tandem repeats of simple sequences, some of which may play a role in chromosome function.
- The structure of genes and their organization within the genome differs strikingly between bacteria and eukaryotes, and between higher and lower eukaryotes. Bacteria and many lower eukaryotes have small genomes of high complexity and high gene density, i.e. most of the genome is unique sequence DNA which is expressed. The genes are small and they usually lack introns. Conversely, higher eukaryotic genomes are large but contain predominantly noncoding DNA, both unique and repetitive. Genes vary considerably in size and usually contain multiple introns, which are generally larger than the exons. There are large intergenic distances. In bacteria, genes are often clustered in operons according to related function, but only rarely does this occur in eukaryotes. Vertebrate genomes show regional differences in gene density, corresponding to chromosome banding patterns. In some bacteria, gene orientation reflects position relative to the origin of replication.
- There is currently a considerable international effort to map and sequence the human genome, and the genomes of selected **model organisms**. These include vertebrates, such as the mouse and the puffer fish, whose genome maps can be exploited to advance the Human Genome Project as well as being useful in their own rights (q.v. *comparative mapping*), and species such as *E. coli*, *S. cerevisiae*, *C. elegans* and *D. melanogaster*, which have been extensively used as laboratory models to study a variety of biological systems, and represent the foundations of molecular biology research. There are essentially three types of genome map: cytogenetic, genetic and physical, in order of increasing resolution. The ultimate physical map is a genome sequence (i.e. a resolution of 1 nt). A complete genome sequence is invaluable, as it provides information concerning gene structure, regulation, function and expression, the evolutionary relationship between different organisms, the nature of higher-order genome organization, and genome evolution. Genome sequences also have many commercial applications, such as the development of drugs, vaccines and enzymes for industrial processes. The gene map is a prerequisite for *positional cloning* (q.v.). In the genomes of higher eukaryotes, which have a generally low gene density, it is useful to concentrate on the analysis of expressed DNA (q.v. *transcriptional mapping*).

## 12.1 Genomes, ploidy and chromosome number

**Ploidy.** The number of copies of a particular gene in a cell is defined as its **dosage**, whereas the number of copies of the entire genome is defined as the cell's **ploidy**. In eukaryotes, this is the number of chromosome sets. Eukaryote cells are **haploid** if they contain one chromosome set and **diploid** if they contain two<sup>1</sup>. However, the ploidy of a eukaryotic cell changes during the cell cycle. Following DNA replication, ploidy is effectively doubled, and then halved again during mitosis. The nominal ploidy of a proliferative cell can thus be defined as the number of chromosome sets it is *born* with. The effective ploidy of a bacterial cell changes with the growth rate because of nested replication (q.v. *Helmstetter–Cooper model*), and dosage is greater for genes nearest the origin of replication.

**Chromosome number.** In eukaryotes, the **monoploid number** ( $x$ ) is the number of chromosomes representing one copy of the genome, i.e. the number in one chromosome set. The **haploid number** ( $n$ ) is the number of chromosomes found in the gametes. In most eukaryotes, the gametes contain one set of chromosomes and  $n = x$ , but for plants which are normally polyploid,  $n$  would be a multiple of  $x$ . The **diploid number** ( $2n$ ) is a convenient way to describe the total number of chromosomes in the somatic cells of most animals, and is the basis of the *karyotype* (see below). The **C-value** is the amount of DNA in the haploid genome, and this may be expressed in base pairs, relative molecular mass or actual mass. Occasionally, ploidy may be expressed in terms of the C-value, e.g. diploid cells are  $2C$ .

The **karyotype** is a shorthand way to describe the total chromosome number and sex-chromosome configuration. For example, the karyotype of somatic cells in the human male is 46, XY, and in the human female 46, XX. In abnormal cells, the karyotype may be augmented with further information to indicate specific chromosome aberrations (see Table 4.1). A **karyogram**, on the other hand, is a picture or ideogram of stained chromosomes arranged in homologous pairs, used to identify chromosome aberrations (q.v. *chromosome banding*).

## 12.2 Physico-chemical properties of the genome

**Base composition.** Genomes may be characterized in terms of their physico-chemical properties. Since all cellular genomes are DNA, any physical or chemical differences between genomes must reflect one of two properties: bulk differences in **base composition** (the relative amounts of adenine/thymine and cytosine/guanine bases) or different amounts of *DNA methylation* (q.v.). A:T base pairs contain two hydrogen bonds and G:C base pairs three; thus a higher proportion of G:C pairs increases the physical stability of the genome because it takes more energy to separate the strands. High G:C content thus correlates to a high thermal melting temperature (q.v. *nucleic acid hybridization*). Also, G:C base pairs have a greater relative molecular mass than A:T base pairs, and GC-rich DNA has a greater buoyant density than AT-rich DNA. Methylation also increases the buoyant density of DNA (q.v. *buoyant density gradient centrifugation, satellite DNA*).

Base composition can be expressed in two ways. The **base ratio** (also known as the **dissymmetry ratio** or **Chargaff ratio**) is applied to microbial DNA. It is defined as  $(A+T)/(G+C)$  and is shown as a number. Species with base ratios greater than one are **AT types** and those with base ratios less than one are **GC types**. The **%GC content** is applicable to all genomes. It is defined as  $(G+C)/(A+T+C+G)$  and is shown as a percentage. Thermophiles are usually GC types (high %GC

<sup>1</sup>The term **haploid** strictly means 'half the ploidy' and was coined to describe the state of gametes (whose ploidy is half that of the meiotic cell). Since most meiotic cells are diploid, most haploid cells have one set of chromosomes and the term has been adopted with this meaning. However, the gametes of a plant with six chromosome sets should properly be described as haploid, even though they possess three chromosome sets and are also triploid. The term **monoploid** specifically indicates that a cell has one set of chromosomes (see Chromosome Mutation).

**Table 12.1:** Genome data for selected organisms

Species	Genome size	Complexity (%)	GC content (%)	No. of genes
Bacteriophage $\lambda$	45 kbp	>99	48	100
<i>Escherichia coli</i>	4.7 Mbp	99	51	4100
<i>Saccharomyces cerevisiae</i>	13.5 Mbp	90	41	6300
<i>Schizosaccharomyces pombe</i>	20 Mbp	90		6000
<i>Dictyostelium discoideum</i>	47 Mbp	70	23	7000
<i>Caenorhabditis elegans</i>	100 Mbp	83		14000
<i>Drosophila melanogaster</i>	165 Mbp	70	39	12000
<i>Fugu rubripes</i>	400 Mbp	>90	44	70000
<i>Danio rerio</i>	1.9 Gbp			70000
<i>Xenopus laevis</i>	2.9 Gbp	54	50	70000
<i>Mus musculus</i>	3.3 Gbp	58	41	70000
<i>Homo sapiens</i>	3.3 Gbp	64	40	70000
<i>Arabidopsis thaliana</i>	70 Mbp	80		25000

Complexity is given as percent unique sequence DNA. The number of genes is taken from genome sequencing projects where appropriate, but where a complete genome sequence is unavailable, it is estimated by extrapolation from existing sequence data.

content) as the G:C rich genome helps maintain DNA in a duplex at high temperatures. The %GC contents of various organisms are shown in Table 12.1.

The base ratio and %GC content are averages across the entire genome. However, the base composition varies within most genomes, giving rise to areas which are relatively AT-rich and others which are GC-rich. For small genomes, regional differences in base composition have been identified by **denaturation mapping**: when the genome is partially denatured and observed by electron microscopy, AT-rich areas are revealed as bubbles. The differential chemical behavior of AT-rich and GC-rich DNA contributes to the banding patterns observed in mammalian chromosomes (q.v. *isochore model*), and this can be exploited to separate individual chromosomes on the basis of quantitative differences in their ability to bind two different fluorescent dyes (q.v. *flow sorting*). The presence of a chromosome region with an uncharacteristic base composition is often indicative of horizontal transfer from a different species (q.v. *codon usage*); in bacteria, a number of **pathogenicity islands** — horizontally transferred virulence genes — have been identified in this manner.

### 12.3 Genome size and sequence components

**Genome size and complexity.** The total amount of DNA in the haploid genome (the C-value) might be expected to increase with the biological complexity of the species because of the requirement for more gene products. This is broadly true: vertebrates *generally* have more DNA than invertebrates, which have more than fungi, which have more than bacteria, which in turn have more than viruses.

The minimum genome size within each phylum appears to increase in proportion to biological complexity, but there are extraordinary differences in the C-value between similar species, generating a spread of C-values within each phylum. In the extreme case of amphibians, the smallest and largest genomes differ in size by two orders of magnitude. Furthermore, the largest insect genomes are bigger than the largest mammalian genomes, and the largest genomes of all ( $>10^{11}$  bp of DNA) belong to flowering plants. Such phenomena cannot be explained by the need for gene products alone, and collectively represent the **C-value paradox**.

The paradox is explained by the predominance of **noncoding DNA** in many eukaryotic genomes. This occurs both as repetitive DNA and as unique sequence DNA. The **complexity** of a genome is defined as the total amount of unique sequence DNA and may be expressed in physical units (i.e. base pairs, picograms) or more usually as a percentage of total genome size (Table 12.1). The presence of repetitive DNA was first shown by **reassociation kinetics** (Box 12.1) and accounts



for much of the C-value paradox. Differences in C-value within phyla appear predominantly to reflect differences in repetitive DNA content, which does not contribute to genome complexity. When repetitive DNA has been taken into account, however, there still appear to be disproportionate differences in genome size between species of similar biological complexity, especially when comparing certain groups of unicellular organisms. For example, the C-value of *Saccharomyces cerevisiae* is approximately 13.5 Mbp, whereas that of another yeast, *Schizosaccharomyces pombe*, is nearer 20 Mbp. Both organisms have similar structural complexity and little repetitive DNA. The discrepancy reflects differences in the amount of noncoding unique sequence DNA, i.e. intergenic DNA segments and introns: 40% of *S. pombe* genes contain introns, compared to 4% of genes in *Saccharomyces cerevisiae*. Both intergenic regions and introns are larger, and introns are more numerous, in higher eukaryotes, leading to an increase in the average size of the gene and the distance between genes.

**Distribution and function of DNA sequence components.** In bacteria, most of the genome is unique sequence DNA, representing genes and regulatory elements. Some genes and other sequences are repetitious, but the **copy number** (or **repetition frequency**) is generally low, usually <10. Examples of such repetitious sequences include the rRNA genes (there are seven in *E. coli*) and transposable elements such as *IS elements* (q.v.). Occasionally, certain sequence motifs may be moderately repetitive. In the 1.8 Mbp *Haemophilus influenzae* genome, there are about 1500 copies of the 30 bp DNA uptake site (q.v. *transformation*). Similarly, the *E. coli* genome contains many copies of two repetitive elements ERIC (enterobacterial repeated internal consensus) and REP (repeated extragenic palindromes). Together, however, repetitive DNA accounts for <1% of bacterial genomes and genome size is a direct reflection of complexity.

In eukaryotes, repetitive DNA accounts for a much greater proportion of the genome. This varies between different species, being as low as 5% in some microbial eukaryotes, 40–50% in mammals, and as high as 80% in some flowering plants. Reassociation experiments (*Box 12.1*) originally showed that eukaryotic DNA could be divided into three components, **unique sequence DNA**, **moderately repetitive DNA** and **highly repetitive DNA**, which are discussed in more detail below. Genomic DNA can also be partitioned on the basis of its function in the cell (*Table 12.2*).

## 12.4 Gene structure and higher-order genome organization

**Gene distribution in DNA sequence components.** In higher eukaryotes, most genes would be expected to lie in unique sequence DNA because their existence can be defined by a single mutation. If genes were repetitive, several mutations would have to occur simultaneously for a mutant phenotype to be observed and there would be a lack of selective pressure to maintain multiple copies (q.v. *functional redundancy*). Additionally, the gene-dense genomes of bacteria and lower eukaryotes are predominantly unique sequence DNA. Tracer reassociation experiments initially showed that most higher eukaryote genes were indeed found in unique DNA, and this has been confirmed in individual cloned genes by genomic Southern hybridization.

By carrying out hybridization analysis at progressively lower stringencies, however, more and more genes are found to occupy the moderately repetitive sequence component, indicating that *genes belong to multigene families with varying degrees of sequence conservation*. At high stringency, the only genes identified in repetitive DNA are highly conserved or identical multigene families (e.g. rRNA genes and histone genes). The proportion of genes belonging to multigene families differs from species to species. In *E. coli*, data from the recently published genome sequence suggest that up to half the genes of this bacterium may belong to multigene families, whereas the proportion from other bacteria is much lower (e.g. 25–30% in *H. influenzae*). This may reflect the diverse lifestyle of *E. coli*, i.e. in terms of its nutritional requirements. In mammals, the proportion of truly unique genes is much lower and the majority of human genes are thought to belong to multigene families,

**Table 12.2:** Eukaryotic genome components classified by abundance and by function

DNA class	Definition
<i>By abundance</i>	
Unique sequence (single copy, low copy, nonrepetitive DNA)	Sequences present as one or a very few copies per genome. Contains most genes and includes introns, regulatory sequences and other DNA of unknown function
Moderately repetitive DNA	Sequences present 10–10000 copies per genome. Generally dispersed repeats corresponding to highly conserved multigene families (functional genes and pseudogenes) and transposable elements. Occasionally clustered
Highly repetitive DNA	Sequences present 100000–1000000 copies per genome. Generally found as tandem repeats although some superabundant (dispersed) transposable elements also fall into this class — e.g. <i>Alu</i> elements
<i>By function</i>	
Genic DNA	Genes, i.e. DNA which is expressed. Genic DNA may be further classified as <b>mDNA</b> (protein encoding), <b>rDNA</b> , <b>tDNA</b> , <b>snDNA</b> , etc. representing the different classes of gene product
Regulatory DNA	DNA whose role is the regulation of gene expression (e.g. promoters, enhancers) or the regulation of DNA function (e.g. origins of replication, matrix-associated regions)
Intergenic DNA, spacer DNA	Introns and the DNA which separates genes from each other
Satellite DNA	Highly repetitive DNA found near centromeres, telomeres and at other sites. Some satellite DNA may play a role in chromosome function
Selfish DNA	DNA whose role appears to be to mediate its own replication and survival within the genome, e.g. some satellite DNA, and transposable elements
Junk DNA	DNA with no assigned function

some of which contain thousands of members. This probably reflects the whole genome duplications and family expansions which are thought to have occurred in the vertebrate lineage (see Proteins: Structure, Function and Evolution). The structure and organization of multigene families in the genome are discussed below.

**Gene size and intron–exon architecture.** Bacterial genes are characteristically small (average 1 kbp) and show little diversity in size, whereas those of higher eukaryotes are large (average 16 kbp) and show great size diversity. The smallest mammalian genes are comparable to bacterial genes (e.g. the human  $\alpha$ -interferon gene is <1 kbp in length) but many span more than 100 kbp of DNA, and the largest gene identified to date, the human dystrophin gene, is 2500 kbp in length.

Although higher eukaryotic genes are generally much longer than bacterial genes, the same is not true of the mRNAs derived from them. This discrepancy is caused by **introns**: intervening sequences which interrupt the transcription unit and must be spliced out at the RNA level (see RNA Processing). The remaining parts of the transcription unit, which become spliced together and expressed, are termed **exons**. Gene size is inversely proportional to the percentage of exon material in the gene. Bacterial genes generally lack introns and are therefore 100% exon material. Introns are also rare in many microbial eukaryotes (e.g. *Saccharomyces cerevisiae*), whose average gene size, 1–2 kbp, is similar to that of bacteria<sup>1</sup>. In humans, the smallest genes have the fewest and smallest sized introns (e.g. the 500 bp histone H4 gene has no introns). Conversely, the largest genes are >95% intron material. The dystrophin gene mentioned above has 78 introns, with an average size of 30 kbp; only 0.5% of the gene is exon material.

<sup>1</sup>Note that bacterial genes are usually defined as the open reading frame, whereas eukaryotic genes are usually defined as the transcription unit (see The Gene). The average size of a yeast gene is thus greater than that of bacterial genes partly because the untranslated regions of the yeast mRNA are included.

**Table 12.3:** Intron–exon organization in different eukaryotes. *S. cerevisiae* has few interrupted genes and gene length correlates to mRNA length. Higher eukaryotes show progressively larger average gene sizes, but the average mRNA sizes remain constant. Generally, gene length is proportional to intron number and inversely proportional to percent exon material

Species	Average gene length (kbp)	Average introns/gene	Average mRNA length (kbp)	%Exon material
<i>S. cerevisiae</i>	1.5	>95% uninterrupted	1.5	100
<i>C. elegans</i>	4	3–4	3	77
<i>D. melanogaster</i>	11	3–4	3	25
<i>H. sapiens</i>	16	6–7	2.5	13

Note that *C. elegans* has a similar intron number to *Drosophila* but the introns are smaller, producing a smaller average gene size.

Whereas the size and number of introns in higher eukaryotic genes is highly variable, exon size falls within a narrow range. For human interrupted genes, exon length is on average 170 bp and varies between 50 bp and 300 bp. There are some notable exceptions, e.g. exon 26 of the apolipoprotein B gene is 7.6 kbp in length, but such examples are rare. Invertebrates tend to have larger exons than vertebrates because of the paucity of introns (Table 12.3). For a discussion of the origins and function of introns, see Proteins; for the mechanism of splicing, see RNA Processing.

**Gene number and density.** The total gene number of several microbial species has now been determined from genome sequencing projects (see Table 12.1 for some examples). Bacterial gene numbers vary through an order of magnitude, from 473 (*Mycoplasma genitalium*) to approximately 8000 (*Myxococcus xanthus*), with *E. coli* lying between the extremes with 4100 genes. In terms of gene numbers, the largest bacterial genomes overlap with those of lower eukaryotes. The yeast *Saccharomyces cerevisiae* has 6340 genes. The invertebrate model organisms *Drosophila melanogaster* and *Caenorhabditis elegans* are estimated to have approximately twice as many genes as *S. cerevisiae*, and vertebrates represent another layer of complexity with estimates of approximately 70 000 genes. Much of the information for these estimates has come from partial genome sequences, the analysis of CpG islands (q.v.) and EST projects (q.v.).

What is the minimum gene number required to sustain an independent living organism? Comparisons between bacterial genomes have identified a set of essential biochemical pathways, and 256 genes are required to encode their components. It is likely that more pathways are required in eukaryotic cells to set up the complex intracellular structure. Still more will be required in multicellular organisms to regulate development and the function of differentiated cells. However, the number of core biochemical pathways is likely to be similar in all metazoans, since the large gene numbers in vertebrates are thought to have arisen through two rounds of whole genome duplication, plus duplications of various chromosome segments and individual genes. The initially redundant genes have adopted specialized functions, often due to diversification of expression patterns, but the pathways are strongly conserved. There is much anecdotal evidence for this hypothesis: many genes represented singly in *Drosophila* are represented by a multigene families in vertebrates. Examples include *Drosophila hedgehog* (vertebrate *sonic hedgehog*, *desert hedgehog* and *indian hedgehog*), the *Drosophila* homeotic complex (vertebrate *Hox-A*, *Hox-B*, *Hox-C* and *Hox-D* clusters) and genes encoding the signaling proteins Ras and Raf.

The higher-order organization of genes within the genome can be addressed in terms of structural and functional organization, orientation and gene density. A unique feature of the *Mycoplasma genitalium* genome is that the orientation of most genes is away from the origin of replication; the significance of this is unclear. A major distinction between bacterial and eukaryotic genomes is the functional organization of structurally unrelated genes in bacterial operons. This type of organization is conspicuously absent from eukaryotic genomes because multiple open reading frames

cannot be translated from a common polycistronic transcript (q.v. *operon*, *polycistronic mRNA*, *ribosome binding site*). Instead, functionally related eukaryotic genes, e.g. those encoding components of a common biochemical pathway, are usually dispersed. In both bacteria and eukaryotes, clustering of structurally related genes may occur because of tandem duplications, and such genes may be functionally related and under coordinated transcriptional control (e.g. the globin gene cluster, and *Hox* clusters).

**Gene density** is highest in bacteria, with on average one gene per kbp of DNA, reflecting the close spacing of open reading frames in operons. The gene density in *S. cerevisiae* is about half the bacterial value, reflecting the separation of individual transcription units. Gene density decreases in higher eukaryotes for several reasons. Firstly, as discussed above, introns become larger and more numerous in more complex organisms. Secondly, the predominance of repetitive DNA also tends to increase. Thirdly, the intergenic distance increases because of the requirements for larger and more complex regulatory elements. There are exceptions to these generalizations, e.g. the puffer fish (*Fugu rubripes*). The *Fugu* genome is remarkable for its high complexity (>90%) and small gene size, reflecting a lack of abundant dispersed repeats and a genome-wide reduction in the size of introns. *Fugu* homologs of the large mammalian genes, such as the gene for dystrophin, are generally ~10% of the size, but retain the same intron-exon organization; the introns are much smaller, most between 50 bp and 150 bp.

Gene density is not constant throughout higher eukaryote genomes. Mammalian genomes show regional variation in gene density, reflecting the isochore organization (q.v. *isochore model*). Invertebrate genomes show a simpler organization: the *Drosophila* genome is divided into a gene-rich 115 Mbp region and a gene-poor heterochromatinized region of 50 Mbp (also q.v. *overlapping genes*, *nested genes*).

## 12.5 Repetitive DNA

**Multigene families.** A **multigene family** is a family of homologous genes within the same genome (cf. *gene family*). There is great diversity in the copy number, extent of sequence conservation, organization, distribution and functional relatedness between such genes. In some cases, family members may be extremely similar or identical (e.g. rRNA genes), and will be identified as repetitive DNA by stringent hybridization analysis. In other cases, the conservation may be very weak, and may not even be revealed by sequence comparison (q.v. *homology*, *gene superfamily*). Classical multigene families are structurally similar and conserved over the whole length of the coding sequence. They may be clustered together at a particular locus (e.g. human  $\beta$ -globin genes), dispersed (e.g. human actin genes) or both (e.g. maize zein genes). The occasional dispersed member of an otherwise clustered multigene family is termed an **orphan**<sup>1</sup>. Other multigene families may have in common only a particular subgenic region, corresponding to a conserved protein domain (e.g. homeobox genes are related by the 180 bp homeobox, which encodes a DNA-binding domain). Still others may be related by virtue of a very short amino acid motif (e.g. the MADS box, and the DEAD box RNA helicase motif). To add further complexity, many genes appear to be chimeric for relatively independent functional units corresponding to different protein domains, which potentially allow them to be members of several different families simultaneously. Such genes are presumed to have arisen by recombination between ancestral genes (q.v. *exon shuffling*) and may contain repeated segments of coding information (q.v. *exon repetition*).

**Pseudogenes.** Multigene families often contain structurally conserved genes which have diversified to fulfill different functions by accumulating mutations. In some cases, mutation abolishes gene function

<sup>1</sup>An orphan is not the same as an **orphan gene**, which is a gene identified in a genome sequencing project for which no comparable genes are known in other organisms and for which no function has been determined.



altogether, and such a nonfunctional copy of a gene is termed a **pseudogene**, often represented by the Greek letter  $\Psi$ . There are two types of pseudogene, with different origins and structures.

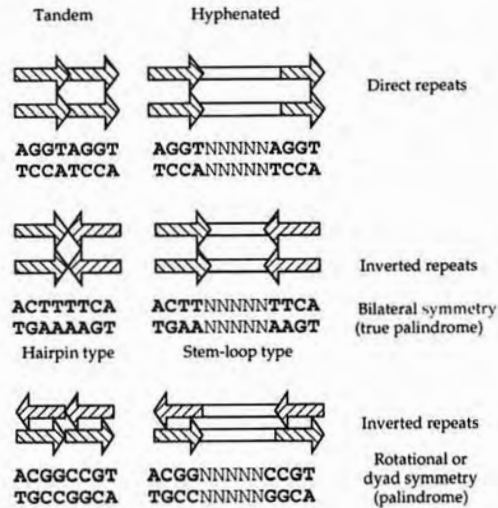
**Nonprocessed pseudogenes (conventional pseudogenes)** arise by duplication of genomic DNA (see Mutation and Selection) and often lie near a functional copy of the same gene. They contain architectural structures similar to the functional homolog, which may include introns and regulatory elements. Such pseudogenes are found in both bacteria and eukaryotes, and can be recognized because they accumulate mutations, including either regulatory mutations which abolish transcription, or nonsense mutations which cause truncation of the encoded product. Occasionally, non-processed pseudogenes may be reactivated by a favorable mutation. The same processes which generate nonprocessed pseudogenes may also generate partial genes or truncated copies.

**Processed pseudogenes (retropseudogenes)** arise by reverse transcription of mRNA and random integration of the resulting cDNA; they are usually dispersed. Processed pseudogenes are generated by adventitious activity of the reverse transcriptase and integrase enzymes encoded by *retroelements* (q.v.) and are therefore found only in eukaryotes. The structure of processed pseudogenes corresponds to the transcription unit of the original gene, i.e. lacking introns and flanking sequences. Because of the lack of flanking sequences, processed pseudogenes are generally not expressed, although they may occasionally integrate adjacent to an endogenous promoter and come under its control (a human gene encoding pyruvate hydrogenase is thought to have arisen in this manner). Processed pseudogenes of RNA polymerase III transcripts may be expressed because many class III genes have internal promoters. The superabundant human *Alu* element is an example of an expressed RNA polymerase III processed pseudogene.

**Structural and functional redundancy.** A **redundant** sequence is one which is represented more than once in the genome, i.e. a sequence which increases genome size without increasing its complexity. Redundant genes are not necessarily *functionally* redundant. Some genes are found as redundant copies as a means to produce sufficient gene product (rRNA genes fall into this category), while others may have evolved to fulfill different functions. Functional redundancy can be established by the lack of phenotype when a particular gene or other element is deleted. Total or partial functional gene redundancy is seen in many targeted mutations in multicellular organisms, even if the same gene has been shown to have a striking gain of function effects when expressed ectopically. An example is the transcription factor MyoD, which can convert many different cell types into muscle by activating the myogenic pathway. When the *myoD* gene is deleted from mice (q.v. *gene knockout*), the homozygous null individuals are normal. This is because a second transcription factor, Myf-5, is able to compensate for the loss of MyoD. Functional redundancy often reflects structural redundancy (i.e. two copies of an ancestral gene have arisen by duplication and they can compensate for each other's loss, as in the example above<sup>1</sup>). In other cases, different genes appear to have converged upon the same function, e.g. the secretion of several unrelated proteins — Chordin, Noggin, Follistatin — from the amphibian organizer; their common function appears to be the interruption of TGF- $\beta$  signaling. Functional redundancy is common for genes with important developmental roles, but less so for those with housekeeping functions. Why are so many genes functionally redundant and how are they selectively maintained? The limited studies of knockout animals have shown that targeted mutations often have different effects in different strains (i.e. different genetic

<sup>1</sup>In fact, MyoD cannot fully compensate for the loss of Myf-5 because the *myf-5* gene is expressed earlier in mouse development, and loss of function delays myogenesis until the onset of *myoD* expression, resulting in retarded muscle development and perinatal lethality. Many developmental genes show this type of **partial redundancy**, where groups of genes have overlapping functions and generate more severe phenotypes when deleted in combination than when deleted individually. The vertebrate *Hox* genes in axial skeleton and limb development are a particularly illustrative example (see Box 6.9), as is the double knockout of *myoD* and *myf-5*, which produces mice with no muscle at all.





**Figure 12.1:** Sequence architecture in repetitive DNA. Repeated sequences may be arranged in the same orientation (**direct repeats**) or in opposite orientations (**inverted repeats**). Inverted repeats may have **bilateral symmetry** (true palindrome), in which case the sequence on each strand reads the same backwards and forwards, or **dyad symmetry** (also termed a **palindrome** although somewhat inaccurately), where both strands have the same sequence when read in the same polarity. Both direct and inverted repeats may be uninterrupted (in **tandem**) or interrupted by nonspecific sequence (**hyphenated**).

backgrounds). It is therefore likely that many of the redundancies reported so far reflect the effects of very specific genetic backgrounds and environments, whereas natural selection works in an uncontrolled system where subtle differences in fitness could be exploited.

**Sequence architecture in repetitive DNA.** Repetitive DNA consists of a repeated sequence of certain size (the **repeat unit**) with a given copy number organized in a particular manner in space. Repeat units can be organized in three ways: **tandem repeats** have no spaces between individual repeat units; **hyphenated repeats** are separated by small gaps but are still grouped together; **dispersed repeats** are disseminated throughout the genome. With respect to each other, individual repeats can be arranged either in the same orientation (**direct repeats**) or in opposing orientations (**inverted repeats**) (Figure 12.1).

The simplest sequence structure in tandem repetitive DNA has a repeat unit of one nucleotide, and this is termed a **homopolymer**. There are also dinucleotide, trinucleotide, etc., tandem repeats, termed *minisatellites* (q.v.), as well as tandem repeats of large repeat units. There is obviously nothing remarkable about dispersed repeats of one nucleotide, or even of two or three nucleotides; hence dispersed repeats are only significant when they involve a reasonably long DNA sequence which occurs substantially more frequently than would be expected by chance. Short dispersed repeats often correspond to functional DNA motifs (e.g. the 30 bp *H. influenzae* DNA uptake site), whereas transposable elements and redundant genes are much larger.

Short functional motifs may be identified by their special architecture. Many recognition sites for transcription factors and other DNA-binding proteins consist of a pair of direct or inverted repeats, either hyphenated or in tandem, reflecting the dimeric nature of the proteins. Inverted repeats often show dyad symmetry and have the potential to form secondary structures such as *hairpins* and *stem-loops* (q.v.).

**Transposable elements as dispersed repetitive DNA.** As discussed above, some genome-wide dispersed repetitive DNA corresponds to members of multigene families, comprising both functional genes and pseudogenes. Otherwise it may represent motifs which function at the DNA level.

Most dispersed repetitive DNA, however, corresponds to either functional transposable elements or their ghosts (elements inactivated by mutation). The predominance of this sequence class varies widely in different organisms. In bacterial genomes, the copy number of transposable elements is often  $<10$ , whereas in vertebrates they are generally very widespread (although not in the puffer fish genome). In mammals, two classes of retroelement in particular are so abundant that they each define a distinct class of dispersed repeat. **SINEs** are **short interspersed nuclear elements** and correspond to copies of a processed 7SL RNA pseudogene, which in humans is termed the *Alu* element, and in mice, the B1 element. **LINEs** are **long interspersed nuclear elements** and correspond to copies of an abundant retroposon named LINE-1 (L1). The *Alu* element is about 300 bp in length, and like other transposable elements is flanked by direct repeats reflecting its mechanism of integration (see Mobile Genetic Elements). It is preferentially located in GC-rich DNA and has an estimated copy number of  $10^6$ , and an average density of 1 element per 4 kbp of DNA. Conversely, the L1 element has a maximum length of 6 kbp and a copy number of  $10^5$  (although full length elements are in the minority,  $<5000$  copies). Both L1 and *Alu* elements are associated with genes, but their distribution is reciprocal and related to the *isochore* organization of the genome (q.v.), perhaps due to preferential target sites for integration. Neither element has been found in the coding region of a gene — they are often found in introns and flanking regions, and *Alu* elements are occasionally to be found embedded within the 3' untranslated region of a gene and may be transcribed by RNA polymerase II as part of the gene.

**Satellite DNA.** The most repetitive component of higher eukaryotic genomes, identified by its rapid reassociation, consists of very short DNA sequences tandemly repeated many times. The predominance of highly repetitive DNA varies between species, but typically represents 10–30% of the genome. Because of its low complexity, it is sometimes termed **simple sequence DNA**, and due to its unusual nucleotide composition, it often separates as one or more 'satellite' bands from bulk genomic DNA during buoyant density gradient centrifugation, and is also known as **satellite DNA**. **Cryptic satellite DNA** has a buoyant density which is comparable to bulk genomic DNA and does not form a satellite band; this is identified using other methods, such as restriction mapping. Satellite DNA is distributed as large clusters (100–3000 kbp), often residing in heterochromatin at centromeres where it may play a role in chromosome function. Much of the centromeric DNA of human chromosomes comprises a cryptic satellite DNA called **alphoid DNA** ( $\alpha$ -satellite DNA), although a second component,  **$\beta$ -satellite DNA**, is abundant in the centromeres of at least eight human chromosomes. There is chromosome-specific sequence divergence within the  $\alpha$ - and  $\beta$ -satellite DNA families.

In insects, satellite DNA comprises many very short sequences (5–15 bp) with pronounced strand asymmetry. Mammalian satellite DNA is more complex in its organization. The simple sequence repeats show some variability but often form blocks which themselves show tandem repetition, again with a degree of variation. The satellite DNA therefore comprises simple sequence blocks with a hierarchical organization and is presumed to arise through continued cycles of mutation and expansion, probably involving unequal crossing over and gene conversion (see Mutation and Selection).

**Minisatellites and microsatellites.** Most satellite DNA exists as large clusters of repeats found in the centromeric regions of chromosomes or around *nucleolar organizers* (q.v.), but it is also found in smaller clusters (100 bp – 10 kbp) termed **minisatellites**, which are often located at the telomeres. There are two forms of minisatellite DNA. At the very ends of the chromosome arms is the telomeric DNA itself. In most eukaryotes, this consists of several kilobases of characteristic tandem pentanucleotide or hexanucleotide repeats (see Table 5.5), and its function is to prevent chromosome erosion through subsequent rounds of DNA replication (q.v. *telomeres*, *telomerase*). A second class of **hypervariable minisatellite DNA** is located in subtelomeric regions. The repeat unit of the hypervariable DNA differs from site to site, but contains a common GC-rich core consensus. The copy

number at each site is highly polymorphic (hence the alternative name **VNTR sequence**, for *variable number of tandem repeats*). The function of hypervariable minisatellite DNA (VNTR DNA) is unknown, but it may promote recombination (cross-overs tend to cluster in the subtelomeric regions of chromosomes). The preferential telomeric location means that minisatellite DNA is not generally useful for markers in genome-wide genetic mapping, but it has been widely exploited for use as diagnostic markers in DNA typing (Box 12.2).

**Microsatellites** occur in smaller clusters (<200 bp) and are characterized by very short repeat units (1–4 bp). They are highly polymorphic and distributed throughout the genome, so they make ideal genetic markers. Of the two possible homopolymers, poly(A)/poly(T) is far more common than poly(C)/poly(G), and the dinucleotide microsatellite poly(CG)/poly(GC) is rare due to the depletion of CpG motifs (q.v. *5-methylcytosine*). Tri- and tetranucleotide microsatellites are comparatively rare, but are more useful as markers than the commonly occurring dinucleotide microsatellites because there is less strand slipping during PCR genotyping (q.v. *molecular marker, strand slipping*).

Generally, minisatellite and microsatellite DNA is extragenic, but occasionally it occurs within the coding region of a gene and gives rise to a highly polymorphic protein; in some cases this can be pathological (q.v. *triplet repeat syndromes*). Other, more stable forms of repetitive DNA are also seen in genes. The  $\alpha 2$  collagen gene contains many exons comprising repeats of a basic 9 bp unit, generating the characteristic amino acid sequence with glycine at every third residue. Larger repeat units are also seen, the largest corresponding to entire *protein domains* (q.v.).

## 12.6 Isochore organization of the mammalian genome

**Biphasic organization of chromatin.** Studies of transcriptional activity and chromatin organization in mammalian chromosomes have shown that chromatin which is potentially or actually transcriptionally active adopts a different structure to repressed chromatin, reflecting differences in nucleosome modification and organization, and loss of higher-order folding (see Chromatin). Two of the consequences of this are that active chromatin replicates early in the S-phase and becomes generally sensitive to DNase I. Analysis of the distribution of active and repressed chromatin domains by cytogenetic mapping of early replicating DNA (replication banding) or DNase I-sensitive regions (D-banding) has revealed a striking correlation between the patterns observed and those generated by standard *chromosome banding* techniques (q.v.) such as G-banding. Such techniques are thought to discriminate between DNA regions of different base composition, suggesting a correlation between physico-chemical and functional properties of the mammalian genome. Further studies have indeed shown that GC-rich, pale-staining G-bands are enriched for genes, whereas the AT-rich dark-staining G-bands are relatively gene poor. In the human genome, *Alu* elements may also be preferentially located in the pale bands, whereas L1 elements may be preferentially located in the dark G-bands.

**Isochores and higher order genome organization.** The **isochore model** divides the mammalian genome into regions, >300 kbp in length, characterized by relatively homogeneous base composition. On average, the GC content of mammalian genomes is ~40%, but this varies in a regional manner between 37% and 55%. Fragmented DNA can be separated by buoyant density gradient centrifugation into five **isochore classes**: LI and L2 (AT-rich) and H1, H2 and H3 (GC-rich). All mammals show similar isochore representations.

Through determining the GC-content of cloned genes and segregating YACs into isochore classes, it has been possible to investigate gene distribution in the isochore classes. The AT-rich isochores make up 65% of the human genome but contain less than 30% of the genes. The highest gene density is seen in the H2 and H3 isochores. In the H3 isochores, a density of 1 gene per 10 kbp DNA is predicted, which is the same as the average density of the *Drosophila* genome, and only five times less dense than the *S. cerevisiae* genome. Conversely, in the L2 isochores, a density of 1 gene per 100 kbp is predicted. Strikingly, preliminary analysis of data from the human genome project



has revealed that the GC-rich isochores may be relatively enriched for small genes with few, generally small, introns and small open reading frames, whereas the AT-rich isochores tend to contain larger genes with many large introns.

Hybridization analysis has confirmed that the distribution of isochores reflects the banding pattern of mammalian chromosomes, with AT-rich isochores approximately corresponding to dark G-bands and GC-rich isochores corresponding to light G-bands. The very gene-dense H3 isochores are preferentially located in subtelomeric regions, although there are exceptions: in the human genome, chromosome 19 appears to be composed almost entirely from H3 isochores, whereas chromosome 13 is mostly L1 and L2 isochores. The observation that *Alu* and LINE elements are differentially distributed has also been confirmed: *Alu* elements are preferentially located near or within the genes of GC-rich isochores, whereas LINE elements are associated with genes in AT-rich isochores. Interestingly, *Alu* and LINE elements may play a structural role at isochore boundaries, as they are often found in adjacent clusters at these sites.

## 12.7 Gene mapping

**The purpose of gene mapping.** Gene mapping is the assignment of a gene to a locus on a chromosome. By mapping the relative positions of many genes and other markers, it is possible to generate a **chromosome map** or a map of the entire genome. Gene mapping has been used to help understand and exploit the inheritance of biological traits, particularly by using *genetic linkage* (q.v.) to associate the inheritance of one trait with another (or with a suitable *genetic marker*, q.v.), and by correlating differences in phenotype with differences in chromosome structure. Commercially, this has enabled animal and plant breeders to produce crops or herds with improved qualities, and in the sphere of human affairs, it has allowed genes for genetic diseases, whose biochemical basis was unknown, to be **mapped** onto (assigned to a locus upon) a particular chromosome or chromosome band, and tracked through pedigrees by linkage to a marker (q.v. *gene tracking*).

With recent advances in recombinant DNA technology (see Recombinant DNA, The Polymerase Chain Reaction (PCR)) it has become possible to generate detailed gene maps by the organized cloning and characterization of genomic DNA fragments. Ultimately, it is possible to obtain the entire DNA sequence of a genome, an invaluable resource which provides information not only about gene structure, but also higher-order genome sequence organization and the evolution of genes and genomes. Currently, there is an intense collaborative international effort to generate genetic and physical maps of several model organisms, with the ultimate aim of determining the complete genome sequences (Table 12.4). The technology associated with physical gene mapping is used for *positional cloning* (q.v.), and the availability of every gene sequence has many commercial applications, including drug development. The discipline of mapping, analyzing and sequencing genomes is called **genomics**.

**Three types of gene map.** Essentially, there are three types of gene map: genetic (linkage) maps, cytogenetic maps and physical (molecular) maps.

A **genetic** or **linkage map** assigns genes to **linkage groups** (groups of genes which tend to be inherited together because they are found close together on the chromosome). Genetic maps are calibrated in arbitrary units, reflecting the likelihood of marker separation by recombination (e.g. in *meiotic mapping*) or by chromosome fragmentation (e.g. in *radiation hybrid mapping* and *HAPPY mapping*). The principle behind genetic mapping is that the chance of two loci appearing together on the same fragment of DNA decreases with increasing distance between them. Genetic mapping in bacteria involves a similar principle and is made possible by natural mechanisms of horizontal gene transfer (see Gene Transfer in Bacteria). Genetic maps have a resolution between cytogenetic and physical maps, but those based on recombination frequencies are distorted by areas of relatively high or relatively low recombination (q.v. *recombination hotspots and coldspots*). Additionally, the

**Table 12.4:** Species chosen for genome mapping and sequencing projects

Organism	Comments
Bacteria	<i>E. coli</i> and <i>B. subtilis</i> are well-characterized model organisms which have been used for pioneering studies of gene structure and regulation. The sequence of the <i>E. coli</i> genome was recently published and that of <i>B. subtilis</i> is likely to be complete within a few years. A number of other bacterial genomes have been completely sequenced: the first was <i>Haemophilus influenzae</i>
<i>Saccharomyces cerevisiae</i>	The yeast <i>S. cerevisiae</i> and its distant relative <i>S. pombe</i> have been extensively used in the analysis of eukaryotic gene structure and regulation, and have been particularly useful in the study of core pathways such as DNA replication and the cell cycle. The complete sequence of <i>S. cerevisiae</i> has been published recently — this yeast has about 6300 genes; many more than originally estimated
<i>Caenorhabditis elegans</i>	The nematode worm is the simplest invertebrate in the model organisms (genome size 100 Mb). It has been extensively used in the study of development, especially neural development, because of its invariant lineage (the developmental pathway from a single egg to the 959 somatic cells of the adult is fully characterized) and the completely mapped interconnections of its nervous system. There are already dense genetic and physical maps, and the genome sequencing project is nearing completion
<i>Drosophila melanogaster</i>	The fruit fly is the most extensively studied of all the model organisms, and is renowned for pioneering studies of developmental genetics. The genome has been densely mapped using morphological markers, and <i>Drosophila</i> is widely used to teach the principles of linkage analysis. The genome size is 165 Mb and the sequencing project is in its early stages
<i>Fugu rubripes</i>	The puffer fish has a small genome compared with most vertebrates (400 Mbp), but the gene number is comparable, i.e. the size reduction reflects loss of noncoding DNA (particularly dispersed repetitive DNA and introns). Gene and regulatory sequences appear to be highly conserved, so the puffer fish genome is potentially a powerful tool for identifying genes in other vertebrates whose genomes have a much lower gene density
<i>Danio rerio</i>	The zebrafish has a genome about two-thirds the size of a typical mammalian genome, and genetic and physical maps are under construction. The zebrafish is a particularly useful developmental model because of its genetic manipulability, e.g. haploid embryos are viable up to the hatching stage. Recently, the zebrafish has been used for saturation mutagenesis screening of developmentally relevant genes
Mouse	For a number of reasons, the mouse (and to a lesser extent, the rat) is the most powerful model organism directly relevant to humans. The genomes of mice and humans are comparable in size and complexity, and large blocks of synteny are conserved. The genetic amenability of the mouse is unparalleled within the mammals, as the species is small and easy to breed. It is possible to carry out mutagenesis screens and genetic crosses to generate marker maps and to map disease loci. It is also possible to modify the germline to study gene function <i>in vivo</i> (q.v. <i>transgenic animals</i> , <i>gene knockout</i> , <i>ES cells</i> )
Other vertebrates	Dense marker maps and gene maps are being generated for a number of other vertebrates, including many commercially important livestock (cattle, pigs, sheep) and domestic animals (cats, dogs). These will be used to improve the yield and qualities of livestock by using markers to assist selection for quantitative traits, and as data for <i>comparative genomics</i> (q.v.)
<i>Homo sapiens</i>	The <b>Human Genome Project</b> , which is expected to be completed by the year 2005, aims to generate a complete genetic and physical map of the genome and to determine the entire genome sequence. The availability of this information will hopefully provide many benefits in medicine, diagnostics and drug development, not least of which will be the determination of the bases

Continued



	of most inherited disorders and diseases. There are fears that the information could also be misused, e.g. to discriminate against individuals with a particular genotype, perhaps even leading to a resurgence of eugenics. Much progress has been made in the mapping of individual genes (which comprise barely 3% of the genome) by mapping <i>expressed sequence tags</i> (q.v.)
<i>Arabidopsis thaliana</i>	<i>Arabidopsis</i> has a relatively compact genome for a plant, which places it in the same category as <i>Drosophila</i> , i.e. a relatively simple metazoan model organism. Indeed like the fruit fly, this plant has been used extensively to study development, and genetic and physical maps are under construction
Other plants	There are genome mapping projects in progress for several commercially important plant species, e.g. rice, wheat and maize, which will be used to map and select for improved traits such as resistance to pests, herbicides etc., as well as increased yield. Many of these are quantitative traits and marker-assisted selection will help to track such traits in breeding programs

frequency of recombination is not the same in all species; thus genetic map units reflect different physical distances in different species.

A **cytogenetic map** is created by correlating phenotypes with observable chromosome rearrangements and deficiencies. This type of map has a low resolution and is applicable to the few species, e.g. mammals and *Drosophila*, which display either natural or artificially inducible reproducible chromosome banding patterns (also q.v. *morbid map*). Cytogenetic maps of simple genomes (viruses, plasmids, etc.) are used to locate regions which differ in their physical properties (e.g., q.v. *denaturation mapping*).

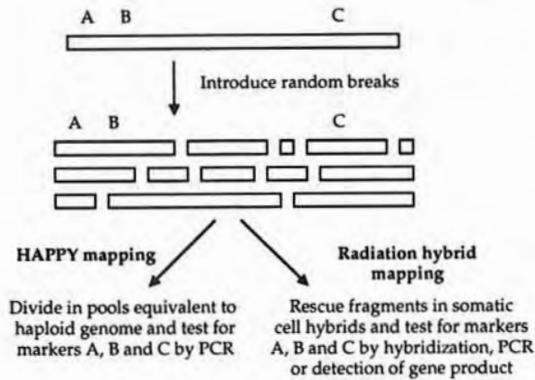
A **physical or molecular map** is created by ordering cloned fragments of genomic DNA and is calibrated in real units (base pairs, kilobase pairs, megabase pairs). The physical map has the highest resolution and is the ultimate aim of mapping projects. However, since only a small percentage of higher eukaryote genomes are expressed, some physical mapping methods have been tailored to the identification of transcribed sequences (i.e. genes).

In the past, elegant strategies have been used to generate linkage maps of many different organisms, but this approach is being replaced progressively by brute force physical mapping methods which allow entire genome sequences to be obtained relatively easily. Genetic mapping is already obsolete in bacteria and unicellular eukaryotes, where physical maps and complete genome sequences can be assembled relatively quickly. In the near future, genetic mapping is also likely to become obsolete for organisms such as *C. elegans* and even *Drosophila* (the species in which genetic mapping was pioneered), but it is still required for initial low-resolution mapping of large genomes, such as the human genome. The advent of *molecular markers* (q.v.) has allowed the three types of map to become progressively integrated. Polymorphic molecular markers such as restriction fragment length polymorphisms, and sequence-tagged sites containing minisatellite DNA, can be used both as genetic markers in linkage analysis and as nucleic acid probes to identify physical clones. Also, *in situ* hybridization allows physical clones to be precisely assigned to chromosome bands on cytogenetic maps.

## 12.8 Genetic mapping

**Genetic mapping using natural and artificial breakpoints.** The principle of genetic mapping is that the further apart two syntenic loci lie on a chromosome, the more likely they are to be separated by chromosome breakage, assuming that chromosome breakpoints arise randomly.

There are several mapping techniques which exploit this principle by artificially introducing chromosome breaks, including **radiation hybrid mapping** (where chromosome breaks are introduced by irradiating somatic cell hybrids) and **HAPPY mapping** (where genomic DNA is sheared by vortexing or sonication). In each case, physical mapping technology is then applied to detect



**Figure 12.2:** Genetic linkage mapping using artificially introduced breakpoints. In radiation hybrid mapping, chromosome breaks are introduced by irradiation and chromosome fragments are rescued in somatic cell hybrids (**radiation hybrids**) where they can be tested by hybridization, PCR detection or protein detection. In HAPPY mapping (haploid PCR mapping), genomic DNA is randomly sheared and divided into pools containing haploid genome equivalents, and markers are assayed by PCR.

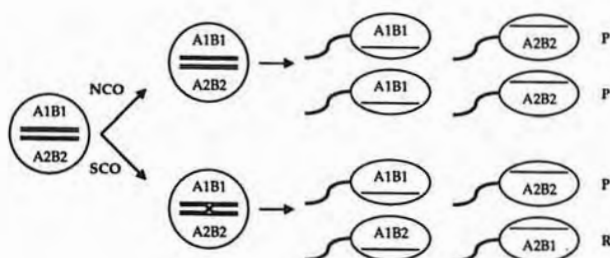
linkage: the presence or absence of two markers on the same DNA fragment is assessed by hybridization, PCR or detection of the gene product (Figure 12.2). The experiment is repeated a number of times and the degree of linkage is calculated as the frequency with which markers are separated. Radiation hybrid maps are calibrated in **centiRays** ( $\text{cR}_x$ , where  $x$  is the dosage of X-rays in Rads), 1 cR being equivalent to a 1% frequency of separation.

The most common method of linkage mapping, however, exploits the natural chromosome breakpoints which arise due to crossing-over (homologous recombination) during meiosis. In this case, physical distance is estimated by the frequency of recombination between markers (i.e. the greater the distance, the more likely that a cross-over will occur between them) and requires heterozygosity at both loci so that parental and recombinant haplotypes can be distinguished. An example is shown in Figure 12.3. There are numerous limitations and problems associated with meiotic linkage analysis, including sample size, genotype detection, the number of available markers, lack of informative segregations (in humans) and intrinsic inaccuracy (Box 12.3).

**Sample size in linkage analysis.** If enough offspring from a given cross can be scored or typed (have their genotype determined), the **recombination frequency** ( $r$ ) or **cross-over value** between any pair of heterozygous loci can be calculated directly using the following formula:

$$r = \frac{\text{Total of recombinant products of meiosis}}{\text{Total products of meiosis}}$$

Recombination frequencies are then used to calibrate genetic maps: a recombination frequency of 1% (i.e. only one out of every 100 offspring inherits the recombinant haplotype) corresponds to a genetic distance of one **genetic map unit** or **centiMorgan** (named after Thomas Hunt Morgan, in whose laboratory genetic mapping of *Drosophila* genes was pioneered). Recombination frequencies can be calculated for genetically amenable species such as yeast and *Drosophila*, where large-scale crosses can be set up, and thousands or millions of progeny typed to derive statistically significant linkage information. However, in species with small litter sizes and large generation intervals (including humans), such crosses are impracticable and data must be accumulated from the analysis of preexisting pedigrees. There are several problems specific to human linkage mapping. One is that human families, especially those in which a disease allele is segregating, are rarely large enough to generate statistically significant linkage information. Another is that because human matings are not experimentally planned, pedigrees are often imperfect, i.e. meioses are uninformative and recombinants cannot be identified.



**Figure 12.3:** The principle of genetic mapping using recombination frequencies. Consider two linked loci, A and B. An individual (left of figure) heterozygous at both loci (i.e. with genotype A1A2 B1B2) may have inherited the A1–B1 haplotype from one parent and the A2–B2 haplotype from the other — these are the **parental (nonrecombinant) haplotypes**, i.e. the **inputs** to meiosis. The individual then undergoes meiosis and transmits one of the chromosomes to each of his offspring; the four products of meiosis are shown. If there is no cross-over (NCO) between A and B, the offspring inherit one of the parental haplotypes (A1–B1 or A2–B2) as **outputs** of meiosis, and are termed **parentals** or **nonrecombinants** (P). Conversely, if there is a single cross-over (SCO) between A and B, the offspring may inherit a recombinant haplotype (A1–B2 or A2–B1) as outputs of meiosis, and are termed **recombinants** (R). The frequency at which recombinant genotypes arise is related to the distance between the loci because crossovers are more likely to occur between loci which are widely separated. However, the maximum recombination frequency is 50% because only two of the four chromatids are involved in a single cross-over, so only 50% of the meiotic products are recombinant (see also Box 12.3). Note that if the individual was homozygous at either locus, the meiosis would be **uninformative** because it would be impossible to discriminate between parentals and recombinants. In diploid species, the success of meiotic mapping depends on the ability to determine the haplotype of the gametes from the phenotype of the offspring. This may be impossible if both parents contribute the same alleles, or if there are dominance relationships between the alleles (see main text).

The solution is to calculate statistical likelihoods for linkage based on the imperfect information available. Because of the small sample sizes, genetic distances in humans are measured as **recombination fractions ( $\theta$ )**, calculated in the same way as recombination frequencies but expressed as a value between 0 and 1 rather than a percentage (i.e. one genetic map unit is equivalent to a recombination fraction of 0.01). The meioses from individual pedigrees are assessed, and probabilities that the data support linkage or independent assortment are calculated. Statistical likelihoods for linkage are expressed as a logarithm of the ‘odds for linkage’, usually termed a **lod score (Z)**:

$$Z = \log_{10} \frac{p(\text{linkage between two loci, assuming RF} = \theta)}{p(\text{the two loci are not linked, i.e. RF} = 0.5)}$$

The general strategy is to calculate Z for a range of values of  $\theta$  and plot a graph of Z against  $\theta$ . Lod scores  $>3$  are taken to be positive evidence of linkage, representing odds of 1000:1 in favor. The maximum lod score  $Z_{\max}$  indicates the most likely genetic distance between loci. Lod scores of  $<-2$  are taken to be strong evidence against linkage, and the use of negative lod scores to refute linkage relationships is termed **exclusion mapping**.

Lod scores are logarithmic values, so they carry the important advantage that data can be summed across pedigrees. Although this is a powerful method for increasing the amount of linkage information available, the map resolution is still low compared with genetic maps of simpler organisms generated by recording recombination frequencies. Another way to increase the sample size in human crosses is to type sperm instead of offspring, but since sperm do not display the disease phenotypes of people, only markers, not real diseases, can be mapped in this way.

One disadvantage of the lod score system is that statistical likelihoods for linkage in any given pedigree are calculated with the assumption that the mode of inheritance of the trait is known. This is applicable to highly penetrant Mendelian traits, but not where the mode of inheritance is ambiguous.

**Table 12.5:** Alternative forms of linkage mapping in humans, which do not require a mode of inheritance to be specified. They represent searches for chromosome segments (containing markers) that are shared by individuals with a given trait

Method	Basis and limitations
Sib pair analysis	Pairs of siblings, both affected by a particular disease (and hence presumably carrying the same disease allele), are tested for the presence of a panel of markers. A marker linked to the disease allele would tend to be shared by more than 25% of affected sib pairs (25% of sib pairs would share any two alleles, even if they were assorting independently). The greater the correlation between the marker and the phenotype, the closer the linkage
Autozygosity mapping	Used to map recessive traits in <b>consanguineous families</b> (families where inbreeding has occurred). <b>Autozygosity</b> means homozygosity at certain loci because alleles are <b>identical by descent</b> (inherited from the same source). The manifestation of a recessive trait in a consanguineous family is often indicative of autozygosity, and autozygosity for markers demonstrates linkage
Disease association	This form of mapping correlates the presence of a marker to a disease phenotype. Disease association is significant if the <b>relative risk (RR)</b> is substantially greater than or substantially less than 1. $RR = wz/wy$ , where $w$ is the frequency with which affected individuals carry the marker, $y$ is the frequency with which unaffected individuals carry the marker, and $z$ is the frequency with which unaffected individuals do not carry the marker

A number of less powerful but 'model-free' mapping strategies may be used in these circumstances, as shown in Table 12.5.

**Genetic markers.** Classical genetic mapping involves determining linkage between genes for which morphologically distinct alleles are available. In species such as yeast and *Drosophila*, which have been extensively used for genetic mapping studies, there is an immense collection of morphological variants showing Mendelian inheritance, and strains carrying multiple variants can be bred and used to map distances between genes directly. In humans, genetic mapping has been used in the past to map the position of disease genes: the diseases are the 'morphological variants' and the normal and disease alleles are the 'morphologically distinct alleles' at each locus. However, it is rare to find family pedigrees where two diseases are segregating at the same time, so direct mapping by gene-gene linkage analysis is not possible. Additionally, there may not be enough Mendelian traits available to map the entire genome.

Other complications associated with the use of morphological linkage markers — uninformative crosses and dominance effects between alleles — are specific to diploid species. The products of meiosis are gametes which need to be combined with a second gamete to generate diploid individuals. It is therefore necessary to infer the haplotype of the gamete (parental or recombinant) from the phenotype of the offspring. This may be impossible if gametes from both parents carry the same alleles (an uninformative cross), or if an allele at one locus displays dominance. In *Drosophila*, for example, genetic mapping is usually carried out using a heterozygous line and a **test stock** which is homozygous for recessive alleles at all loci under study. This allows the genotype of each gamete from the heterozygous line to be determined unambiguously in a recessive background. However, human matings cannot be controlled in this manner, so dominance and uninformative meioses are problems which can stall linkage analysis.

These problems have been solved to a large extent by the development of **molecular markers** — Mendelian characters which are abundantly distributed, easily detected, highly polymorphic and codominant, so that most individuals are heterozygous, unrelated individuals rarely carry the same alleles, and genotype can be determined directly from phenotype (Table 12.6). The most



**Table 12.6:** Molecular markers for genetic mapping

Marker	Definition and uses
<b>Morphological phenotype</b>	Morphological phenotypic variants. Useful for mapping only in organisms with many morphological varieties demonstrating Mendelian inheritance. Subject to complications of dominance relationships, nonallelic interactions and environmental effects. Human disease genes are generally dimorphic phenotypic variants and are mapped against other markers
<b>Protein polymorphisms</b>	Differences in protein mobility during electrophoresis or isoelectric focusing. Codominant and moderately polymorphic but limited in number. Many protein polymorphisms cannot be detected by electrophoresis because the amino acid substitutions do not alter the physico-chemical properties of the protein. Also, the gene encoding the polymorphic protein may itself not be mapped
<b>Restriction fragment length polymorphisms (RFLPs)</b>	Differences in DNA sequence which create or abolish restriction sites. Codominant and abundant, but usually only dimorphic so that uninformative meioses and crosses are common. Many sequence polymorphisms are not detected because variations lie outside restriction sites
<b>Simple sequence length polymorphisms (SSLPs). Variable number of tandem repeats (VNTRs). Short tandem repeat polymorphism (STRPs)</b>	Differences in restriction fragment lengths or PCR product lengths due to variable number of tandem repeats in minisatellite or microsatellite DNA. Very useful markers because they are codominant and highly polymorphic. Minisatellite DNA is found predominantly at telomeric regions and is therefore not ideal for a genome-wide linkage analysis. Microsatellite DNA is distributed throughout the genome and can be rapidly genotyped using PCR. Tri- and tetranucleotide repeat polymorphisms are preferred because strand slipping during PCR amplification complicates the use of dinucleotide repeats. Unique sequence DNA containing microsatellite DNA ( <b>polymorphic sequence tagged sites</b> ) can be used to identify clones on physical maps
<b>Randomly amplified polymorphic DNA (RAPDs)</b>	DNA fragments amplified with arbitrary primers. Abundant, but not always possible to discriminate between homozygotes and heterozygotes, and sometimes difficult to reproduce results. Use so far restricted to mapping plant genomes

useful molecular markers are unique (yet polymorphic) DNA sequences, which can be assigned to physical and cytogenetic maps by using them as probes to isolate DNA clones and to identify chromosome bands by *in situ* hybridization.

Not only have such markers allowed individual genes to be mapped (**disease-marker mapping, gene-marker mapping**), and tracked through pedigrees for diagnostic applications (**gene tracking**), but they have been used to create a dense **framework of markers** covering the entire genome of several species (**marker-marker mapping**) which can be used for subsequent gene mapping studies. In mice, dense marker maps have been generated by extensive backcrossing between two strains (e.g. *Mus musculus* and *Mus spretus*), as most types of marker vary between the two strains and will be heterozygous (hence suitable for linkage analysis) in the backcross generation. In humans, disease-marker mapping necessitates the use of families in which the disease is segregating, regardless of their pedigree structure and ultimate suitability for linkage analysis. Marker-marker mapping is not constrained in this manner, and a panel of ideally structured families have been assembled for linkage analysis at the Centre pour l'Étude des Polymorphismes Humaines (CEPH) in Paris. Established cell lines have been generated from every member of these **CEPH families**, providing a constant source of DNA for the research community.

**Mapping quantitative traits.** The linkage mapping strategies discussed above concern Mendelian traits or markers, which can be tracked through pedigrees or crosses as discrete variants to derive



linkage information. However, many important biological characters are inherited in a quantitative manner (see Biological Heredity and Variation), and to map the **quantitative trait loci (QTL)** involved such simple experiments cannot be used.

In organisms such as *Drosophila*, which are amenable to genetic analysis, QTLs have been identified by laborious breeding and linkage analysis experiments. Artificial selection over many generations can produce populations representing the extreme phenotypes for a given quantitative character, and each strain is presumed to be enriched for alleles favoring one extreme phenotype or the other. The availability of balancer chromosomes which suppress recombination and carry dominant selectable markers allows fly strains to be generated carrying specific combinations of chromosomes from each selected line. These strains can be scored for their quantitative phenotype, and hence the contribution made by each chromosome can be determined. Further experiments using recessive markers allows the location of QTLs to be narrowed down to particular chromosome bands, whereupon they fall within the resolution of physical mapping. A number of genes affecting bristle number have been mapped in this way, and include many of the genes involved in neurogenesis (e.g. the *achaete-scute* complex, *Notch*).

For commercially important organisms such as pigs, cattle, rice and wheat, and for human beings, extensive breeding programs of this nature are impractical. However, with the availability of dense marker maps, it is now becoming feasible to identify QTLs by cosegregation. In farm animals and plants, this approach to QTL mapping involves crossing two strains or breeds which differ considerably for a given quantitative trait and then scoring progeny for the trait and for a panel of genetic markers. Correlation between phenotypic performance and a given marker indicates linkage, but a simple performance-relationship correlation of this nature cannot discriminate between loose linkage to a strong QTL (major gene) and strong linkage to a weak QTL (minor gene), and is therefore no use for *positional cloning* (q.v.). A second approach, known as **interval mapping**, involves calculating the likelihood that a QTL exists at different positions along the chromosome using a similar strategy to *Lod score* analysis (q.v.). This allows the QTL to be narrowed down further and brings it within range of a *chromosome walk* (q.v.).

The mapping of QTLs contributing to multifactorial congenital diseases in humans (**susceptibility loci**) is also facilitated by searching for cosegregating markers. Pedigree data is collected from affected families, and genes shared by affected individuals can be identified by marker cosegregation. Several QTLs involved in susceptibility to insulin-dependent diabetes have been isolated using complex segregational analysis with microsatellite markers. In some cases, it is possible to isolate families who, because of their particular genetic background, demonstrate near Mendelian inheritance for an otherwise quantitative character. It is likely that the genetic background provides a high level of susceptibility, and that the presence or absence of a particular allele at a major susceptibility locus is enough to trigger the threshold causing the disease. In these cases it is possible to identify QTLs with standard lod score analysis or sib pair analysis, and the former strategy was used to identify the *BRCA1* gene, a major susceptibility locus for breast and ovarian cancer (see Oncogenes and Cancer).

## 12.9 Physical mapping

**Low-resolution physical mapping.** In mammals and in *Drosophila*, both of which have cytogenetic maps based on chromosome banding patterns, initial physical mapping may involve the localization of genes or other markers to a particular chromosome or region thereof (Table 12.7). Such mapping strategies are of low resolution, typically assigning loci to DNA fragments spanning several megabases. However, *in situ* hybridization to interphase chromatin, or DNA which has been artificially extended, can allow mapping to a resolution of under 10 kbp.

**High-resolution physical mapping.** The strategy for generating a high-resolution physical genome map is to divide the genome into a number of fragments, determine their order and then

**Table 12.7:** Techniques for low-resolution physical mapping

Mapping strategy	Basis
<i>Localization to individual chromosomes</i>	
Somatic cell hybrids	<b>Somatic cell hybrids</b> are cells made by fusing cultured cells of different species, e.g. by treatment with polyethylene glycol. In the mapping of human genes, rodent/human hybrid cells are used. Typically, initial hybrids are unstable and most of the human chromosomes fail to replicate, generating stable hybrids with a full set of rodent chromosomes and one or a few human chromosomes. A collection of such hybrids, a <b>hybrid cell panel</b> , can be assembled so that any given human DNA fragment can be mapped unambiguously to a given chromosome, either by PCR or hybridization assay or, exceptionally, by assay for the gene product. <b>Monochromosomal hybrids</b> (those containing a single human chromosome) can be generated by fusing human microcells to normal rodent cells, allowing unambiguous localization of human DNA using a panel of just 24 cell lines. <b>Microcells</b> are cell-like particles containing a single chromosome within a small nucleus, surrounded by minimal cytoplasm and a cell membrane; these are generated by prolonged inhibition of mitosis followed by centrifugation
Dosage mapping	Analysis of cell lines or somatic cell hybrids with multiple copies of a given chromosome allows genes to be mapped to the over-represented chromosome due to dosage detected by quantitative PCR, hybridization or expression of product
<i>Localization to chromosome subregions</i>	
Deletion or translocation mapping	Analysis of hybrid cell panels containing donor chromosomes with translocations or deletions. This method of physical mapping involves assay of DNA sequence by hybridization or PCR, or for a gene product. The cytogenetic mapping technique of the same name involves deducing gene position by correlating a phenotype to a cytogenetically visible chromosome rearrangement
<i>In situ</i> hybridization	Hybridization of a nucleic acid probe to a chromosome spread allows localization to a specific chromosome band. Traditional <i>in situ</i> hybridization using radioactive probes has been replaced by <b>fluorescence in situ hybridization (FISH)</b> using nonradioactive fluorescent probes. Apart from its speed and efficiency, FISH has the advantage that probes with different fluorochromes can be used to identify different targets at the same time with different colors, allowing gene order to be determined (also q.v. <i>chromosome painting</i> ). FISH to metaphase chromosomes gives a resolution of 1–10 Mbp; however, the same technique can be applied to interphase chromatin, and to artificially extended chromatin fibers ( <b>direct visual in situ hybridization, DIRVISH</b> ) and naked DNA ( <b>DNA fiber FISH</b> ) with a resolution of <10 kbp. The extensive looping of unfolded DNA demands that mapping studies involving extended fiber FISH are backed up by statistical analysis of the results

characterize those fragments individually in terms of the loci they contain and, ultimately, their sequences (Box 12.4). This is achieved by preparing a *genomic DNA library* (see Recombinant DNA) and exploiting the overlapping regions of individual clones to establish their order. In this way a **clone contig map** can be generated which recreates contiguous regions of the chromosome (Table 12.8). For small genomes such as those of bacteria, genomic mapping can be carried out using  $\lambda$  replacement vectors or cosmid vectors (see Recombinant DNA). Although in principle larger eukaryotic genomes can be mapped in the same way, the number of individual clones required for a representative library is prohibitive, and the abundance of repetitive DNA causes ordering

**Table 12.8:** Techniques used to assemble contigs of genomic clones

Technique	Principle
Chromosome walking	In principle, a random clone is used to screen a genomic library to identify overlapping clones. These clones are then used as probes in the same way to identify a series of overlapping fragments. In practice, clone-to-clone walking may be complicated by the presence of repetitive DNA, especially when using large capacity vectors such as YACs. Repetitive DNA may be suppressed by prior hybridization to known repetitive sequences (e.g. <i>Alu</i> sequences in the human genome), and instead of whole clones, labeled end fragments can be used as probes instead. Another problem associated with YACs is chimerism (Table 12.9), so YAC walking is often complemented by FISH to chromosome preparations
Hybridization mapping	A set of random clones is used to screen the library to identify overlapping clones. From the remaining negative clones, another random set is chosen and the screen repeated. This strategy is repeated until all clones have been hybridized
Oligonucleotide mapping	Oligonucleotides are hybridized to a series of cosmid clones. Cosmids hybridizing to the same oligo are likely to overlap. The oligonucleotides can also be designed around consensus motifs to derive sequence information
Restriction fragment fingerprinting	Clones are digested with a panel of restriction endonucleases and a <i>restriction map</i> (q.v.) for each clone generated. Maps for individual clones are then compared to identify overlapping regions
Sequence tagged sites (STS)	Unique sequences are identified in the genome by testing randomly subcloned DNA fragments for their ability to, e.g., identify a single band in a genomic Southern blot or amplify a single product by PCR. These are called sequence tagged sites (STSs). Genomic clones are then tested with these markers: two or more clones containing the same STS must overlap. The identification of potential STSs may be speeded up by testing cDNA rather than genomic clones as the former are more likely to contain unique sequence DNA
Repetitive DNA fingerprinting	Mammalian genomes contain much repetitive DNA which can be exploited to assemble clone contigs. Repetitive DNA fingerprinting involves the digestion of YAC or cosmid clones with restriction endonucleases, and Southern blotting with repetitive DNA probes, e.g. specific for the <i>Alu</i> element. Clones with similar banding patterns are likely to overlap. A similar PCR-based technique involves amplification of the genomic DNA between head-to-head <i>Alu</i> elements using a single <i>Alu</i> -specific primer, and the analysis of amplification products by electrophoresis — again clones with similar band patterns are likely to overlap

problems. Larger capacity vectors, such as **yeast artificial chromosomes (YACs)**, are therefore required for the initial stages of mapping (the properties of some artificial chromosome vectors in current use are compared in Table 12.9). Once such a map has been generated, the individual YACs can be subcloned into cosmids for finer-scale ordering. Finally, the cosmid inserts can be fractionated using restriction endonucleases, subcloned and individually characterized.

A major improvement in the speed and efficiency of physical mapping has been achieved by the introduction of **gridded libraries**. Here, the library clones (e.g. YACs of the entire genome, or cosmids from specific YACs) are picked and placed into individual wells in multiwell plates. Then, instead of transferring the clones to a filter by conventional methods for screening (q.v. *plaque lift*, *colony blot*), clones are spotted onto filters in a grid pattern, so that each can be given a precise reference coordinate. This technique reduces ambiguities in clone identification and allows the gridded libraries to be distributed to other laboratories for collaborative work.

**Restriction maps.** Restriction enzymes can be used to generate physical maps based on the positions of restriction sites (**restriction maps**). The resolution of the map depends upon the frequency

**Table 12.9:** A comparison of high-capacity artificial chromosome cloning vectors used for genomic mapping. The **human artificial episomal chromosome (HAEC)**, based in the Epstein-Barr virus, is one of a range of MACs under development

Vector	System	Cloning capacity	Comments
YAC (yeast artificial chromosome)	Yeast centromere and origin, plus telomeres	200 kpb–2 Mbp	High percentage of chimeras (contiguous genomic fragments not usually associated in the genome); unstable (spontaneous deletions). Difficult to separate from yeast chromosomes  Stable, but not maintained outside bacteria unless integrated into host DNA
PAC (P1 artificial chromosomes)	Bacteriophage P1	100–300 kpb	
BAC (bacterial artificial chromosomes, fosmids)	F plasmid	300 kpb	
MACs (mammalian artificial chromosomes)	Episomal vectors based on the Epstein-Barr virus	>330 kpb	

of the restriction site in the DNA fragment, which reflects both its size and base composition. Enzymes with 4–6 bp recognition sites can be used to generate restriction maps of small DNA molecules such as plasmids, PCR fragments and  $\lambda$  inserts. Rare cutters are enzymes which have large restriction sites (8–10 bp) and/or recognize underrepresented sequences, such as CpG in mammalian genomes. Restriction maps of entire chromosomes can be prepared using such enzymes, although the DNA fragments produced must be separated by *pulsed-field gel electrophoresis* (q.v.) or similar methods. Separated restriction fragments can be tested for markers by hybridization (q.v. *Southern blot*) or PCR. The use of a panel of restriction enzymes allows the ordering of fragments to form a contig. For a discussion of the use of restriction enzymes in molecular biology, see Recombinant DNA (an example showing restriction mapping of a plasmid vector is also shown in this chapter).

**Gene mapping and identification in eukaryote genomes.** In the large genomes of higher eukaryotes, most of the DNA is not expressed and much of it is repetitive. It has therefore been necessary to design strategies specifically for the identification of genes.

In principle, a gene can be identified either because its sequence is conserved with a previously identified gene, because it has a distinct structure in genomic DNA, or because it is expressed to generate an RNA transcript. All three strategies have been used (*Table 12.10*). Hybridization approaches that select vertebrate genomic clones enriched for genes include the use of CpG island probes to identify the 5' end of genes, and the use of genomic clones from the puffer fish *Fugu rubripes*, which has a genome complexity of over 90%. Both strategies also have their disadvantages: only half of the estimated 70000 mammalian genes are associated with CpG islands, and not all will have *Fugu* homologs with sufficient identity to cross-hybridize.

Another approach to gene identification is specifically to clone and characterize expressed DNA, rather than sifting through genomic DNA. This can be done by extensive characterization of cDNA libraries or by exon trapping and cDNA capture strategies (*Table 12.10*), but each suffers from the disadvantage that nothing is revealed about gene structure and regulation, and that transiently expressed genes, or genes expressed at only minimal levels or in specific cells, will be missed. By



**Table 12.10:** Approaches to gene identification in genomic DNA. For strategies used to identify specific genes, q.v. *positional cloning*

Approach	Strategy
<i>Exploit sequence conservation</i>	
Cross-species homology	Genes are often conserved between species, whereas noncoding DNA is not. A subclone can thus be used to probe genomic Southern blots from different species to reveal conserved gene sequence (q.v. <i>zoo blot</i> )
Database homology search	The sequence from a subclone can be compared to the sequences held in a sequence database. Regions of homology to a previously cloned gene may be identified
Puffer fish comparative screening	For low-complexity vertebrate genomes, hybridization to puffer fish genomic clones may help identify genes because the puffer fish genome has high complexity
Exon prediction	There are computer programs which can predict the locations of putative exons based on the presence of open reading frames and splice consensi
<i>Exploit unique structure</i>	
Identification of CpG islands	Regions containing <i>CpG islands</i> (q.v.) often mark the 5' end of genes in higher eukaryotes. These can be identified by hybridization, by restriction mapping using enzymes with CpG motifs in their recognition sequence, or by PCR
<i>Exploit gene expression</i>	
Expression hybridization	Hybridization of labeled genomic probe to northern blots to identify transcribed regions and to cDNA libraries to isolate expressed genes
Exon trapping	Genomic clones are inserted into an intron flanked by two exons in an expression vector and the construct is transfected into cells. If the genomic clone contains an exon, splicing will generate a mature transcript with three exons (the two vector exons and the central trapped exon). If it does not, the transcript will contain two exons. RNA is isolated from the cells is analyzed by RT-PCR for the acquisition of an exon ( <b>exon amplification</b> )
cDNA selection and capture	cDNA is hybridized to genomic clones (either immobilized on a filter or in solution but labeled with biotin so they can be purified). Heteroduplexes of genomic DNA and cDNA are purified by washing or streptavidin capture and the cDNA is amplified by PCR and characterized. Amplified cDNA can be put through several rounds of genomic hybridization to enrich for positive sequences

concentrating on the expressed sequences of the mammalian genome, it is possible to home in on the <5% of DNA which is carrying out most of the genome function, and has the most commercial and medical relevance. Large-scale cDNA cloning followed by random sequencing of short cDNA fragments produces a collection of **expressed sequence tags (ESTs)** of 200–300 bp which can be mapped onto physical chromosome maps by hybridization or PCR. Together with genes identified by other methods (see above), this factory-style approach to gene identification potentially allows all genes to be assigned a chromosomal locus (**transcriptional mapping**), providing a complete gene map. EST projects have been set up to characterize cDNAs from various human tissues, and much of the information produced is redundant. However, the sheer number of sequences generated (>200000) may be enough to map the majority of genes. ESTs can be mapped to chromosomes in radiation hybrids, to individual YACs in contig maps, and even to each other, to produce full-length cDNA contig sequences (**EST walking**).

Although EST characterization can identify many expressed sequences, it provides little quantitative information. Several techniques have been developed recently which allow the simultaneous

quantitation of gene expression at many loci (e.g. *SAGE*, *oligonucleotide chips*, q.v.). These, together with coordinated approaches to determining gene function by mutation, and the interaction between gene products using the two-hybrid system, comprise the rapidly expanding field of **functional genomics**, which exploits the information gathered from genome sequencing projects and uses it to assign functions to DNA sequences on a genome-wide scale. For further discussion, see *Proteins: Structure, Function and Evolution*.

**Comparative genome mapping.** **Comparative genomics** is the branch of genome science which deals with comparisons between the genomes of different species. Such comparisons serve two purposes: to provide information concerning gene and genome evolution (i.e. highlighting similarities and differences between species at the genomic level), and to facilitate gene cloning by **comparative** or **synteny mapping**. Comparative mapping in vertebrate species is particularly valuable because it may provide novel animal models for human genetic diseases, and also novel therapies, as well as providing information concerning the patterns of vertebrate evolution. The puffer fish genome is a particularly useful comparative mapping tool, because it is very compact and gene-dense. Linkage is often conserved between the fish and other vertebrates, but genes are easier to isolate from the puffer fish genome, and can then be used as probes to detect conserved mammalian genes.

Comparative maps of vertebrates are characterized by various levels of conserved chromosome segments. Low-resolution comparative mapping can be carried out by **chromosome painting (zoo-FISH)**, where DNA isolated from a single chromosome of one species can be amplified, labeled with a fluorescent probe and hybridized *in situ* to metaphase chromosome preparations of another. Comparisons between human and cat chromosomes reveal extensive regions of synteny, in some cases with entire chromosome–chromosome conservation, whereas the regions conserved between humans and mice are much smaller (the X-chromosome is particularly strongly conserved because of the constraints of *dosage compensation* (q.v.) for X-linked loci). At a finer scale, comparative mapping can reveal conserved linkage between markers within syntenic segments. The types of markers used for comparative studies are genes (which vary little within species) rather than the hypervariable microsatellite markers used in pedigrees (these are polymorphic within species, but hardly ever conserved between species). The conservation of functional DNA sequences shown by comparative sequencing can help to identify genes and regulatory elements in extragenic DNA. Comparative mapping in mammals is enhanced by the use of **anchor reference loci**. These are located within chromosome segments, showing conserved linkage in all mammals with genome maps under construction (such segments are **SCEUSs: smallest conserved evolution unit segments**). Unique sequences (sequence-tagged sites) identified within SCEUSs and in other regions can be used to generate marker maps to which all future mammalian genome maps can be aligned. These sequence-tagged sites conserved in all mammals are termed **CATS (conserved anchor-tagged sequences)**.

**Box 12.1:** Reassociation kinetics in the determination of genome properties

**Cot analysis.** Before genome analysis by sequencing was feasible, **reassociation kinetics** (the analysis of the behavior of single-stranded nucleic acids annealing in solution) was used to investigate genome properties. Although this technique is now mainly of historical interest, the principles remain useful for understanding genome architecture and *nucleic acid hybridization* in general (q.v.). Double-stranded DNA can be denatured or melted (separated into single strands) by heating, and if gradually cooled, will reassociate (renature, reanneal) to form duplex molecules. The reassociation of single-stranded DNA in solution follows second-order kinetics because there are two strands, and the rate at which this occurs can be expressed as shown in Equation 12.1, where  $C$  is the concentration of single-stranded DNA at time  $t$ , and  $k$  is the **reassociation rate constant**. The proportion of single-stranded molecules remaining at any time, given a starting concentration of  $C_0$ , can thus be determined by integration, as shown in Equation 12.2. This identifies the product of  $C_0$  and  $t$  as the parameter which controls the rate of reassociation.

$$\frac{dC}{dt} = -kC^2 \quad (12.1)$$

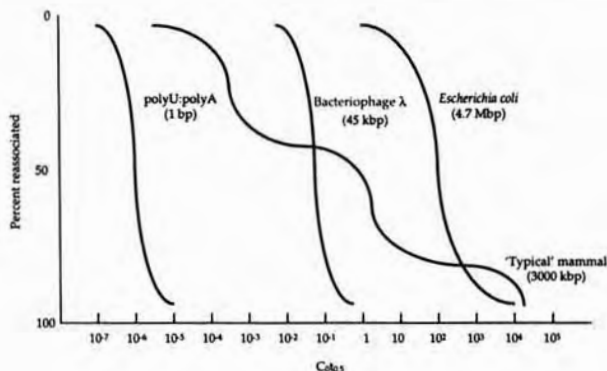
$$\frac{C}{C_0} = \frac{1}{1 + kC_0t} \quad (12.2)$$

The point at which half the DNA has reassociated ( $t_{0.5}$ ) is chosen as a reference. At this point,  $C/C_0 = 0.5$ , and by rearranging Equation 12.2, it can be shown that  $C_0t_{0.5} = 1/k$ .  $C_0t_{0.5}$  is described as the **Cot value**, and is proportional to genome complexity. This is because as complexity increases, the relative concentration of any individual sequence decreases and takes longer to find a complemen-

tary strand. The reassociation reaction thus takes longer to reach the half-way point.

**Cot curves.** Data from genomic Cot analysis are usually plotted as  $\log_{10}C_0t_{0.5}$  against the fraction of reassociated DNA ( $1 - C/C_0$ ) to give a **Cot plot** or **Cot curve**. For the simple genomes of bacteria and viruses, reassociation occurs over two orders of magnitude of Cot values, and Cot curves are linear over approx. 80% of their lengths. As the complexity of the genome increases,  $Cot_{0.5}$  increases and curves are displaced to the right. The Cot plots of *E. coli* and bacteriophage  $\lambda$  DNA are shown below, together with polyuridilate/polyadenylate, an artificial 'genome' with the minimum complexity of 1. Eukaryotic genomes subjected to similar analysis show reassociation over a much broader range of Cot values. The eukaryotic Cot plot can often be resolved into three overlapping curves, representing genome fractions with different sequence complexities. These are sometimes termed the **fast**, **intermediate** and **slow components**, and correspond to highly repetitive, moderately repetitive and unique sequence DNA. The slow component gives the best estimate of true genome complexity, because most genes are found in unique sequence DNA. A proportion of DNA also reanneals immediately. This **zero time binding DNA** is also known as **snap-back** or **fold-back DNA** because it represents regions of dyad symmetry which can hybridize by intramolecular base pairing. The Cot plot of a typical mammal is superimposed over those of the three simple genomes below.

**Rot analysis.** Reassociation kinetics has also been used to investigate the properties and abundance of RNA components. RNA reassociates with comple-



Continued

mentary DNA in solution, following similar kinetics to DNA reassociation. The driving parameter of the reaction is the initial concentration of RNA and time, which is described as  $R_0t$  or the **Rot value**. RNA reassociation with cDNA can be used to determine RNA complexity, i.e. the representation of different RNA molecules in the cell. Reassociation experiments of this type produce a broad Rot curve spanning several orders of magnitude which can often be resolved into three components, the abundant component, the intermediate component and the rare component. The **abundant component** hybridizes at low Rot values and is often termed the **simple component** because it comprises less than 50 distinct mRNAs. These may be represented up to 10000 times each in the cell, accounting for up to 50% of the entire mRNA population. The abundant component often represent tissue-specific transcripts: examples include  $\alpha$ - and  $\beta$ -globin, actin, myosin and albumin mRNAs. The **intermediate component** hybridizes at Rot values between  $10^{-2}$  and  $10^2$ , and comprises 100–1000 transcripts with a representation of a few thousand copies each. The **rare or scarce component** hybridizes at high Rot values and is often termed the **complex component** because it comprises tens of thousands of

transcripts, each represented less than 100 times. Most housekeeping transcripts are found amongst this component.

**Drivers and tracers.** The behavior of specific reassociating components can be investigated by including a small amount of radioactively labeled material in the reassociation reaction, such a component being described as a **tracer**. The kinetics of the reaction are governed by the reassociation components present in excess, such a component being described as a **driver** (a reaction where DNA is present in excess over RNA is described as a **DNA-driven** reaction and follows a Cot curve; the converse is an **RNA-driven** reaction which follows a Rot curve). An RNA tracer placed in a DNA reassociation reaction allows the expressed DNA component to be identified, and this type of experiment was used to show that most genes lie in unique sequence DNA. Where it is necessary to identify a component which does not hybridize in a reassociation reaction, the reaction can proceed to saturation, which can be observed as a plateauing of the Cot or Rot curve. Such **saturation kinetics** experiments can be used, e.g., to identify the proportion of DNA not represented by a specific population of RNA.

### Box 12.2: DNA typing

**The basis of DNA typing.** **DNA typing** or **DNA profiling** involves using minisatellite DNA (VNTR DNA) to generate a collection of DNA fragments which, when separated by electrophoresis, provides an unambiguous profile of any individual (such a profile is sometimes termed a **DNA fingerprint**). Minisatellite DNA is highly polymorphic (in terms of the number of repeat units per site), and there are many minisatellites in the genome, preferentially located in subtelomeric regions. Unrelated individuals are therefore extremely unlikely to generate identical profiles if enough sites are typed simultaneously, but because minisatellites are transmitted as Mendelian traits, related individuals should have similar profiles, and the number of matching DNA fragments will correspond to how closely related they are.

**Applications.** The ability of DNA typing to generate individual-specific DNA profiles is applied in criminal investigations. DNA can be isolated from tissues and body fluids left at the scene of a crime (usually blood, semen or hair) and compared with control samples taken from suspects. Similarly, DNA can be

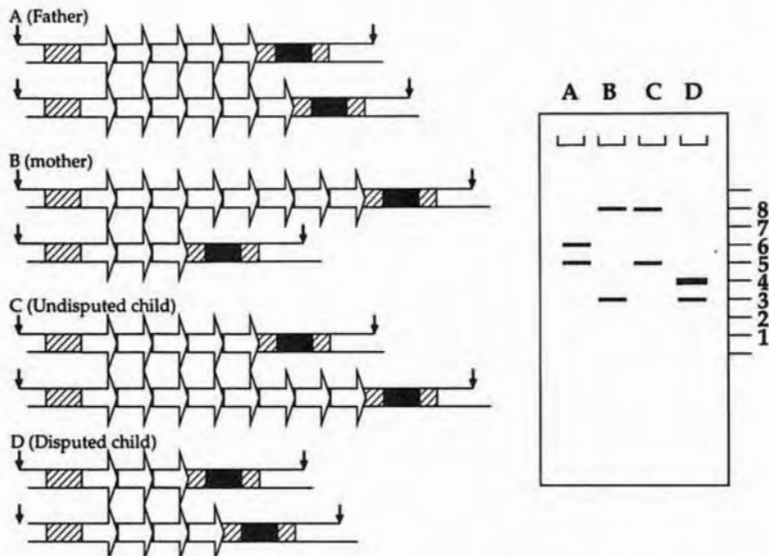
obtained from animals and plants and compared to stored references to determine their origin, e.g. in the case of stolen endangered birds and their eggs. The tendency for VNTR alleles to be shared by related individuals can also be exploited. This can help establish paternity (see the example below), confirm a pedigree, or show that individuals are related (e.g. in immigration disputes).

**DNA typing methodology.** The original DNA typing procedure involved cutting DNA with restriction enzymes and typing by *Southern hybridization* (q.v.) using a locus-specific probe. The size of the restriction fragments would depend upon the number of repeat units in each minisatellite. These techniques require a relatively large amount of fresh (i.e. undegraded) DNA, whereas evidence for forensic testing is usually available only in small quantities and is often old and hence degraded. These problems can be solved to a certain degree by using the *polymerase chain reaction* (q.v.) to amplify across the microsatellite repeats. **PCR typing** produces similar profiles but is applicable to minute samples (e.g. dried spots of blood, single hairs) and tolerates a



degree of DNA degradation. However, care must be taken to avoid contamination from exogenous

sources, and samples and controls are routinely tested in different laboratories.



The figure shows a simple example of DNA typing used in a paternity dispute. The VNTR loci of the father (A), mother (B) and two children (C, D) are shown. The paternity of Child C is undisputed, but A suspects he is not the father of Child D. DNA obtained from blood is cut with restriction enzymes (small arrows show restriction sites flanking the VNTR sequences), resolved by electrophoresis and used in a Southern blot with a probe corresponding to the invariable region flanking the VNTR (dark bar). Electrophoretic bands of different sizes are generated for each allele, depending on the number of repeats (shown as numbers). The profile shows that while Child C has inherited one VNTR sequence from each parent, Child D has inherited one VNTR sequence from the mother and an unrecognized allele which is not derived from A (shown as a black band). A is therefore likely to be correct in the assumption that he is not the father of Child D.

### Box 12.3: Limitations to the accuracy of linkage mapping

**Summary of limitations.** Linkage mapping is inherently inaccurate when either very large or very small interlocus distances are considered. This is because:

- (1) regardless of the distance between loci, the maximum recombination frequency is 50% due to the statistical distribution of cross-overs with respect to the four chromatids of the bivalent;
- (2) the effects of multiple cross-overs cause a progressive underestimation of recombination frequency as the interlocus distance becomes larger;
- (3) the effects of nonreciprocal recombination interfere with the analysis of recombination frequencies as the interlocus distance becomes very small.

Linkage mapping using recombination frequencies is therefore accurate over short distances (i.e. where multiple cross-overs are unlikely to occur), but not where the two loci under study are so close that recombination is likely to include them in a segment of heteroduplex DNA. In addition,

- (4) regional differences in recombination frequency, and the presence of recombination hotspots and coldspots, means that there is not a constant relationship between genetic and physical distances. Recombination frequencies also differ between sexes and species.

**Maximum recombination frequency.** Loci on different chromosomes assort independently, and the recombination frequency is 50% because the

parental combination of chromosomes will be obtained just as frequently as the recombinant combination due to random orientation of homologous pairs at the metaphase plate. For syntenic loci, independent assortment is impossible, but recombinant haplotypes can be generated by crossing over. As the distance between loci increases, the chance of a cross-over also increases, but the recombination frequency never rises above 50% because even if the distance is so large that a cross-over is guaranteed, a single cross-over involves only two chromatids of a bivalent and only half the products of meiosis are recombinant. If double cross-overs are considered, the distribution of different types of double cross-over again ensures a maximum recombination frequency of 50%: only double cross-overs involving all four strands generate four recombinant chromatids (100% recombination), but double cross-overs involving only two strands are statistically equally as likely to occur, and these generate four parental chromatids (0% recombination). Double cross-overs involving three strands generate two parental and two recombinant chromatids (50% recombination). The overall recombination frequency is thus 50%.

**Multiple cross-overs.** The linear range over which genetic distance is proportional to physical distance is short (about 15 map units). As interlocus distances become greater, the distances determined by recombination frequency progressively underestimate the real genetic distance between markers. This is due to the effect of **multiple cross-overs**.

Linkage mapping works on the basis that a single cross-over between two heterozygous markers changes a parental genotype into a recombinant genotype, a change which can be scored by typing the products of the cross. Up to a certain point, increasing distance between loci increases the likelihood of interlocus single cross-overs, allowing the frequency of recombination accurately to predict physical distance. However, with further separation, double cross-overs begin to occur. Such events should be counted as two single cross-overs, but in a two-locus cross the double cross-over types are indistinguishable from parentals and would be counted as such. The number of scored cross-over events is thus smaller than the true value, and the recombination frequency, and hence the distance between loci, is underestimated.

As interlocus distance increases further, higher orders of multiple cross-overs can occur. However, triple cross-over types will be scored as recombinants and quadruple cross-over types as parentals because the genotypes will be indistinguishable.

The fundamental weakness of genetic mapping is that *any even number of cross-overs will be typed as parental and any odd number as recombinant*. Eventually, the interlocus distance becomes so large that the probability of generating an even number of cross-overs is equal to that of generating an odd number and the frequency of recombination is 50%. At this point, even loci on the same chromosome behave as if they are assorting independently.

**Multiple point crossing and mapping functions.** In principle, the underestimation of genetic distances can be corrected by attempting to detect multiple cross-over types or by predicting true distances from the underestimated ones.

In genetically amenable species, the detection of multiple cross-over types can be achieved by including more loci in the cross, as in the classical *Drosophila* three-point test cross, so that the mapped region is broken into smaller interlocus distances. In humans, multipoint mapping of disease genes onto a framework of markers is more advantageous than two-point lod score analysis, because with many different markers, there is less chance of uninformative meioses. The use of more than two loci also allows gene order along the chromosome to be determined unambiguously. In some fungi (e.g. *Aspergillus*, *Ascobolus*) the four products of meiosis are retained together as a **tetrad** in a sac-like structure termed an **ascus**. Here, it is possible to derive two-point linkage data corrected for double cross-overs because the products of each meiotic chromatid can be identified (**tetrad analysis**).

The correction of inaccurate genetic map distances is facilitated by a **mapping function**, a mathematical relationship between recombination frequency and genetic distance. There are three types of mapping function, Haldane's, Kosambi's and Ott's. Haldane's function is the simplest as it assumes random distribution of cross-overs and no *interference* (see below). It can be expressed as follows, where  $d$  = genetic distance and  $r$  = recombination frequency:

$$d = \frac{-\ln(1-2r)}{2}$$

**Interference.** Where one cross-over occurs, does it influence the initiation of a second? **Interference** describes such an influence, **positive interference** where one cross-over inhibits another and **negative interference** where one stimulates another. In *Drosophila*, interference can be estimated by asking whether the observed number of double cross-overs for a given three-point cross is that expected

from the frequency of each single cross-over class. The product law states that the probability of two events occurring together is equal to the product of the probabilities of each single event occurring alone. The expected frequency of double cross-overs is thus the product of the observed frequencies of the single cross-overs. Interference is calculated as follows, where  $I$  is the **index of interference**, and  $c$  is the **coefficient of coincidence**:

$$I = 1 - c$$

and

$$c = \frac{\text{Observed frequency of double crossovers}}{\text{Expected frequency of double crossovers}}$$

In eukaryotes, it is generally observed that recombination shows positive interference, i.e. one cross-over inhibits the initiation of a second cross-over nearby. Recombination between very close markers, however, shows evidence of negative interference, i.e. several cross-overs appear to be clustered. However, this is an illusion created by the processing of heteroduplex DNA and is explained by *gene conversion* (q.v.) rather than strand exchange.

**Regional recombination frequency variation, hotspots and coldspots.** Linkage mapping relies on the random distribution of cross-overs, but there is regional variation within the genome and sites where recombination is promoted (**recombinators, recombinogenic elements, recombination hotspots**) and inhibited (**recombination coldspots**).

In humans, there is regional variation in recombination frequency within every chromosome. Cross-overs tend to be much more frequent in telomeric chromosome regions than around the centromere.

There is also a higher frequency of recombination in females compared to males, and males show an obligatory cross-over in the pseudoautosomal region of the X:Y pair, so the recombination frequency is always 50%. Note that the Y-chromosome does not have a meiotic map, because it is never involved in cross-over events, although genetic maps can be generated by radiation hybrid mapping.

Recombination hotspots are often endonuclease target sites because cleavage provides single-stranded DNA for the initiation of *homologous recombination* (q.v.). Such sites include the *chi* site in *E. coli* and related sites in other bacteria. Recombination hotspots cause the overestimation of genetic distance because if recombination is initiated preferentially at certain sites, loci flanking those sites would undergo recombination more often than average. However, on the scale of the whole genome, recombination hotspots are relatively evenly distributed. Their effects are only evident when small distances are considered, leading to the phenomenon of **polarity** in yeast crosses, where different alleles are involved in gene conversion events with different frequencies, depending on the extent of *branch migration* (q.v.) from the initial site of the cross-over.

Recombination coldspots are often sites where homologous DNA fails to synapse effectively. This usually occurs where there is steric hindrance, e.g. in an individual heterozygous for a chromosome inversion or translocation where synapsis at the breakpoints is prevented. Loci flanking such a site appear closer than they really are because the frequency of recombination between them is low. The inability of complex rearrangement isomers to synapse can be exploited to prevent recombination

#### Box 12.4: DNA sequencing

**DNA sequencing methods.** Until 1977, determining the sequence of bases in DNA (**DNA sequencing**) was a laborious process which could only be applied to small molecules such as tRNA. Two different techniques for rapid, large-scale DNA sequencing were developed at this time, both of which involved the generation of *nested sets* of DNA fragments, differing in length in steps of a single nucleotide. Four sequencing reactions are carried out in parallel, each of which generates nested fragments ending at a defined base. The side-by-side electrophoresis of these reactions allows the

sequence to be read directly from the electrophoresis gel or autoradiograph.

**Maxam and Gilbert sequencing** involves the chemical degradation of a restriction fragment with reagents that modify defined bases. The **Sanger sequencing** method involves DNA synthesis, and each reaction includes a small amount of one of the four **2',3'-dideoxynucleoside triphosphates** (**dideoxynucleotides, ddNTPs**). These are **telogens**, i.e. nucleotides which cause chain termination because they lack a 3' hydroxyl group for extension, and hence the technique is often termed the

**dideoxy method** or the **chain terminator method**. Maxam and Gilbert sequencing was initially the most popular because it could be carried out using restriction enzymes and common laboratory reagents, while the Sanger method required specialized reagents and the use of M13 vectors. With the advent of *phagemids* (q.v.), and the increasing commercial availability of fine reagents such as dideoxynucleotides, the Sanger method has gained popularity. It is the most suitable method for automation in large-scale sequencing projects, and most general sequencing is now carried out in this way. Maxam and Gilbert sequencing is used for some specialist applications such as *DNase footprinting* (q.v.).

Several novel sequencing methods have been developed more recently, of which two may be likely to be used in the future. The first, **scanning tunneling microscopy** and, similar in principle, **atomic force microscopy** can map the surface structure of a DNA molecule, and with technological improvements may be able to discriminate between individual bases. The second, **hybridization sequencing**, uses arrays of immobilized oligonucleotides to generate hybridization maps. Hybridization sequencing would require improvements in automated oligonucleotide synthesis as grids representing, e.g., all possible octamers would require the synthesis and positioning of over 50000 separate molecules. In the future, it may be possible to align thousands of oligonucleotides on small chips for hybridization, and direct pattern analysis by computer to determine sequences rapidly.

**Maxam and Gilbert sequencing.** In this technique, four reactions are carried out in which an end-labeled restriction fragment is incubated with reagents that modify or remove a specific type of base. Dimethylsulphate methylates guanine, acid removes any purine, hydrazine modifies any pyrimidine and hydrazine with NaCl specifically modifies cytosine. The modified bases are then removed by piperidine, and the strand is cleaved at the abasic site. The reagents are used at concentrations which cause each DNA strand to be modified at only one position, so that a nested set of fragments with a common labeled end and different but base type-specific unlabeled ends is generated. The four reaction products are run side by side on the electrophoresis gel and the sequence read from the autoradiograph.

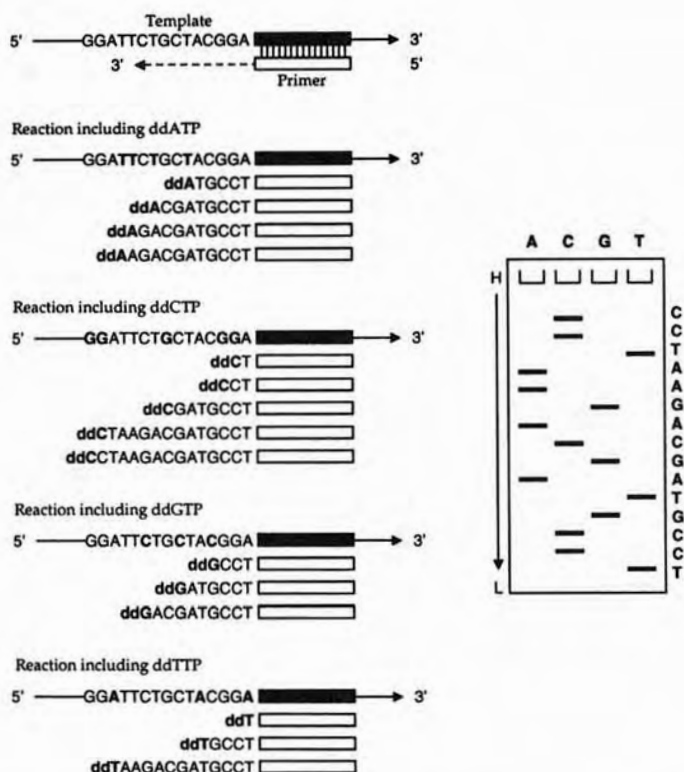
**Standard chain terminator sequencing.** In this technique, primers are annealed to single-stranded DNA and extended by DNA polymerase — a high-processivity, recombinant form of T7 DNA polymerase, termed **Sequenase**, is often used.

Isotopically labeled nucleotides are incorporated during primer extension and there are four reactions, each containing small amounts of one of the four dideoxynucleotides. Each reaction thus generates a nested set of labeled products, beginning with the sequencing primer and ending at a specific base. These are resolved by electrophoresis in adjacent lanes, allowing the sequence to be read from the autoradiograph (see figure). Originally, the technique demanded single-stranded templates in M13-based vectors. Such templates still generate the best results, but it is possible to carry out **double-stranded sequencing** by first denaturing a dsDNA template such as a plasmid. The sequencing primer anneals to one strand only, so only one of the strands is used as a template.

**Recent innovations in chain terminator sequencing.** Many of the recent innovations in chain terminator sequencing result from the drive to automate the process and increase the rate at which sequence information can be gathered. **Multiplex sequencing** allows many clones to be sequenced in the same reaction and run out in the same gel lanes: each clone is sequenced without incorporating a label using a unique primer which can later be used to identify the sequencing ladder specific to that clone by *Southern hybridization* (q.v.). **Dye-terminator sequencing** uses dideoxynucleotides labeled with fluorescent dyes. Four dyes are used, one for each base, and each emitting a different wavelength of light. This allows all four reactions to be run in a single lane and the sequence to be read by a detector at the bottom of the gel during electrophoresis (**real-time sequencing**). This not only increases throughput, but because the information is fed directly from the detector into a computer, it also reduces clerical sequence errors. **Cycle sequencing** is a hybrid between chain terminator sequencing and PCR. A double-stranded template is used, and a single sequencing primer, but the reaction is carried out by thermal cycling using a thermostable DNA polymerase (see The Polymerase Chain Reaction (PCR)). Cycle sequencing reduces artefacts caused by secondary structure in the template, and because the products accumulate in a linear fashion (q.v. *asymmetric PCR*), small amounts of template can be used. Cycle sequencing is, however, less accurate than standard sequencing because thermostable polymerases such as *Taq* DNA polymerase are error-prone.

**Sequencing strategy in genome projects.** For large-scale sequencing projects, such as genome sequencing, large genomic clones are often sub-cloned randomly into phagemid vectors, generating





a large number of overlapping clones. These are picked and sequenced arbitrarily (**shotgun sequencing**) and the sequence information is fed into a computer, which can detect overlaps and align the inserts to form a contig map with single nucleotide resolution. Any remaining gaps may be filled by **primer walking** on the original genomic clone, where a primer is designed to extend from a sequenced region into the gap, resulting in the recovery of the missing sequence.

#### Handling and storage of sequence information.

**Bioinformatics** refers to the rapidly expanding computer technology dedicated to the storage, analysis, comparison and organization of DNA and protein sequence information. There are three coordinated databases: EMBL (distributed by the European Informatics Institute), GenBank (distributed by the US National Centre for Biotechnology) and DDBJ (the DNA database of Japan), holding the

majority of DNA sequence information, which are continuously updated as new sequences are published. Special databases are also available to store the rapidly expanding collection of expressed sequence tags from genome projects (see main text), and model organism databases show genome maps, gene sequences and expression patterns. The databases are available to research scientists throughout the world via the internet, and can be used in conjunction with sophisticated analysis packages for mapping, sequence alignment, prediction of protein sequences, homology and function. At present, most sequence information is freely available, but the growing commercial interest in exploitation of genome science has resulted in the inevitable but alarming possibility that corporations such as drug companies and farming industries could control the use of genome sequences to protect their interests and investments.

## References

- Brown, T.A. (Ed.) (1991) *Molecular Biology Labfax*. BIOS, Oxford.
- Primrose, S.B. (1995) *Principles of Genome Analysis*. Blackwell Science, Oxford.
- Further reading**
- Charlesworth, B., Sniegowski, P. and Stephan, W. (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371: 215–220.
- Cooke, J., Nowak, M.A., Boerlijst, M., and Maynard-Smith, J. (1997) Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends Genet.* 13: 360–363.
- Desouza, S.J., Long, M.Y. and Gilbert, W. (1996) Introns and gene evolution. *Genes to Cells* 1: 493–505.
- Dujon, B. (1996) The yeast genome project — what did we learn? *Trends Genet.* 12: 263–270.
- Elgar, G., Sandford, R., Aparicio, S., MacCrae, A., Venkatesh, B. and Brenner, S. (1996) Small is beautiful — comparative genomics with the puffer fish (*Fugu rubripes*). *Trends Genet.* 12: 145–150.
- Fickett, J.W. (1996) Finding genes by computer — the state-of-the-art. *Trends Genet.* 12: 316–320.
- Gabor Miklos, G.L. and Rubin, G.M. (1996) The role of the genome projects in determining gene function: Insights from model organisms. *Cell* 86: 521–529.
- Gardiner, K. (1996) Base composition and gene distribution: Critical patterns in mammalian genome organisation. *Trends Genet.* 12: 519–524.
- Gerhold, D. and Caskey, C.T. (1996) It's the genes! EST access to human genome content. *BioEssays* 18: 973–981.
- Goffeau, A., Barrell, B.G., Bussey, R.W., Davis, R.W., Dujon, B., Feldmann, H., Gailbert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996) Life with 6000 genes. *Science* 274: 546–567.
- Haley, C.S. (1995) Livestock QTLs — bringing home the bacon? *Trends Genet.* 11: 488–492.
- Lander, E.S. (1996) The new genomics: Global views of biology. *Science* 274: 536–539.
- O'Brien, S.J., Weinberg, J. and Lyons, L.A. (1997) Comparative genomics: Lessons from cats. *Trends Genet.* 13: 393–399.
- Okubo, K. and Matsubara, K. (1997) Complementary DNA sequence (EST) collections and the expression of information of the human genome. *FEBS Letts.* 403: 225–229.
- Postlethwait, J.H. and Talbot, W.S. (1997) Zebrafish genomics: From mutants to genes. *Trends Genet.* 13: 183–190.
- Rowen, L., Mahairas, G. and Hood, L. (1997) Sequencing the human genome. *Science* 278: 605–607.
- Stuber, C.W. (1995) Mapping and manipulating quantitative traits in maize. *Trends Genet.* 11: 477–481.
- Tang, C.M., Hood, D.W. and Moxon, E.R. (1987) *Haemophilus* influence: The impact of whole genome sequencing on microbiology. *Trends Genet.* 13: 399–404.
- Weeks, D.E. and Lathrop, G.M. (1995) Polygenic disease: Methods for mapping complex disease

## Websites

## Genome mapping databases

## Bacteria

*H. influenzae* — <http://www.tigr.org/tdb/mdb/hidb/hidb.html>

*M. genitalium* — <http://www.tigr.org/tdb/mdb/mgdb/mgdb.html>

*E. coli* — [http://www.ncbi.nlm.nih.gov/Complete\\_Genomes/Ecoli/](http://www.ncbi.nlm.nih.gov/Complete_Genomes/Ecoli/)

*B. subtilis* — <http://acnuc.univ-lyon1.fr/nrsub/nrsub.html>

## Yeast

*S. cerevisiae* — <http://speedy.mips.biochem.mpg.de/mips/yeast/>

## Invertebrates

*C. elegans* — <http://probe.nalusda.gov:8000/other/aboutacdb.html>

*D. melanogaster* — <http://www.embl-ebi.ac.uk/flybase/>

## Plants

*A. thaliana* — <http://genome-www.stanford.edu/Arabidopsis/>

## Mammals

Humans (On-line Mendelian Inheritance in Man) — <http://gdbwww.gdb.org/omimdoc/omimtop.html>

Mouse — <http://www.jax.org/>

## Nucleic acid and protein sequence databases

GenBank (DNA and protein sequences) — <http://www.ncbi.nlm.nih.gov>

EMBL (DNA sequences) — <http://www.ebi.ac.uk>

DDBJ (DNA database of Japan) — <http://www.nig.oc.jp/home.html>

## Chapter 13

# Mobile Genetic Elements

### Fundamental concepts and definitions

- **Mobile genetic elements** are segments of DNA able to move from site to site in the genome, or between genomes in the same cell. Found in both prokaryotes and eukaryotes, they are a diverse group differing in structure, mechanism of mobilization, distribution, freedom of movement and level of autonomy.
- The mobility of some elements has a clear cellular function, but most are believed to be selfish DNA or accidentally mobilized DNA. In higher eukaryotes, a significant proportion of the genome is made up of functional and 'ghost' transposable elements (those inactivated by mutation). Some bacterial mobile elements have developed a symbiotic relationship with the host cell. Mobile elements play an important role in genome evolution.
- Mobile elements can be placed into three broad categories according to their level of independence. At one end of the spectrum, mobile elements able to exist in two states — either integrated into the host genome or as autonomous extrachromosomal replicons — are termed **episomes** (e.g. bacteriophage  $\lambda$ , the F plasmid). At the other, elements whose mobility is controlled by the cell to mediate specific genomic rearrangements are termed **cassettes**<sup>1</sup> (e.g. yeast mating type cassettes, trypanosome VSG cassettes). The mobility of episomes and cassettes is facilitated by *homologous* or *site-specific recombination* (q.v.) which restricts target site choice, often to a single locus. Finally, those elements which cannot replicate outside the host genome, but which control their own mobility within it, are termed **transposable elements** (e.g. P-elements, retroviruses). They mobilize by **transposition**, a form of recombination requiring no homology between the element and its target, and therefore allowing a degree of freedom in target site choice. The transposable elements are the largest category of mobile elements and demonstrate great diversity in structure, transposition mechanism and level of autonomy.
- Some mobile elements are viruses. Bacteriophage  $\lambda$  is an example of a viral episome, while bacteriophage Mu is a transposable element — it must integrate into the host genome in order to replicate. The eukaryotic retroviruses are also transposable elements.
- Transposable elements fall into two major classes based on their active transposition mechanism. **Class I transposable elements (retroelements)** utilize an RNA intermediate during the transposition process — an integrated element is transcribed, then reverse transcribed to generate a cDNA copy which is integrated at a new site. **Class II transposable elements (transposons)** transpose directly as DNA, either by excision and integration or by a replicative process whereby one copy of the element is left at the original site and another is integrated at a new site.
- Transposable elements are **active** or **autonomous** if they encode the functions required for their own transposition (or at least its initiation, leaving the cell to finish the process). For transposons, the minimal requirements are the enzyme **transposase** and the *cis*-acting sites it recognizes, allowing the element to determine the boundary between itself and host DNA. For retroelements, reverse transcriptase is required in addition to the transposase (which is generally termed **integrase**<sup>2</sup>). Other transposable elements, described as **defective**, lack enzyme functions but retain the *cis*-acting sites and may be mobilized in *trans* by an active element. Elements mobilized in *trans* are described as **nonautonomous** or **passive**. The active and defective elements form families of related sequences.

- A large and structurally diverse family of nonautonomous retroelements represent sequences with no similarity to known autonomous elements, and appear to be cellular RNAs transposed adventitiously when reverse transcriptase and integrase are supplied in *trans*. These generate the **processed pseudogenes** which are abundant in many vertebrate genomes.
- A second derivative type of element lacks appropriate *cis*-acting sites for transposition and is therefore 'grounded', but encodes the *trans*-acting functions and can provide them to *cis*-competent elements in the same cell, an example being the **wings-clipped** derivative of the *Drosophila* P-element.

<sup>1</sup>A **cassette** is a segment of exchangeable genetic information. Variability at most genetic loci is allelic, but cassettes are nonallelic in that they originate at different loci. Examples of cassettes include the yeast mating type cassettes, the variable surface antigen genes of trypanosomes, and the antibiotic resistance genes which are found at bacterial integrons. As an extension of the term, a cassette is also an exchangeable oligonucleotide used for *in vitro* mutagenesis (q.v. *cassette mutagenesis*).

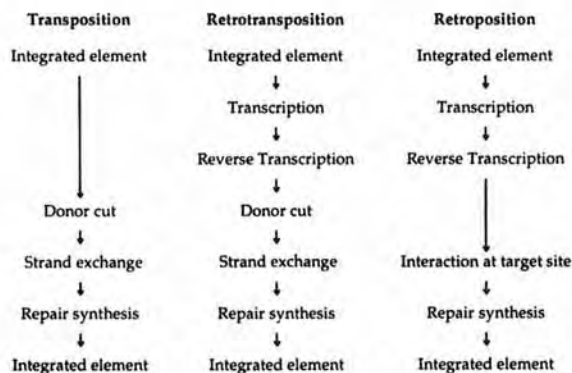
<sup>2</sup>**Integrase** enzymes encoded by class I transposable elements are homologous in structure and function to **transposase** enzymes encoded by class II transposable elements, both having nickase and transesterification activities. Confusingly, the *site-specific recombinase* enzyme encoded by bacteriophage  $\lambda$  is also termed integrase, but it is *not* a transposase. The two enzymes are functionally similar, but transposase has great specificity for the transposable element although not its target site, whereas site-specific recombinases recognize both the donor and recipient sequences in recombination.

### 13.1 Mechanisms of transposition

**Overview of transposition mechanisms.** Transposable elements can be divided into two families according to their general mechanism of transposition. **Transposons** mobilize directly as DNA, whereas **retroelements** employ an RNA intermediate. Within each family, several specific types of transposition mechanism can be used. Before describing individual mechanisms in detail, however, some general points, applicable to all transposable elements, are considered.

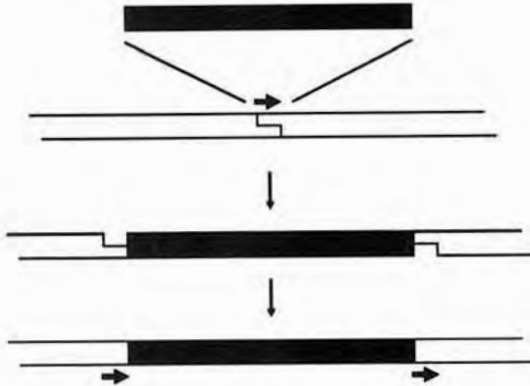
(1) **Proliferation.** All transposition events have the potential to increase the number of copies of the transposable element in the genome; this can be accomplished actively or passively.

(2) **Stages of transposition.** Transposons actively mobilize in three stages (Figure 13.1): (i) the **donor cut stage**, an endonucleolytic cleavage reaction catalyzed by transposase which separates the transposable element from the host DNA; (ii) the **strand exchange stage**, a pair of transesterification reactions also catalyzed by transposase which join the 3' ends of the transposable element to host DNA at the target site; (iii) the **repair stage**, DNA synthesis undertaken by the cell which fills any remaining gaps. Retrotransposons precede these three stages by transcription and reverse transcription to



**Figure 13.1:** Overview of the three major active transposition mechanisms which are utilized by class II, class I.1 and class I.2 elements, respectively.





**Figure 13.2:** Mechanism for the generation of target site duplications (TSDs), a hallmark of transposition but not of other mechanisms of recombination, which conserve the recombining sequences exactly. Staggered nicks are introduced at the target site and the transposable element is joined to the 5' overhangs. Repair synthesis over the resulting gaps generates direct repeats of the sequence between the nicks. Almost all elements generate TSDs, but those that do not (e.g. IS91) probably introduce a blunt-ended break at the target site.

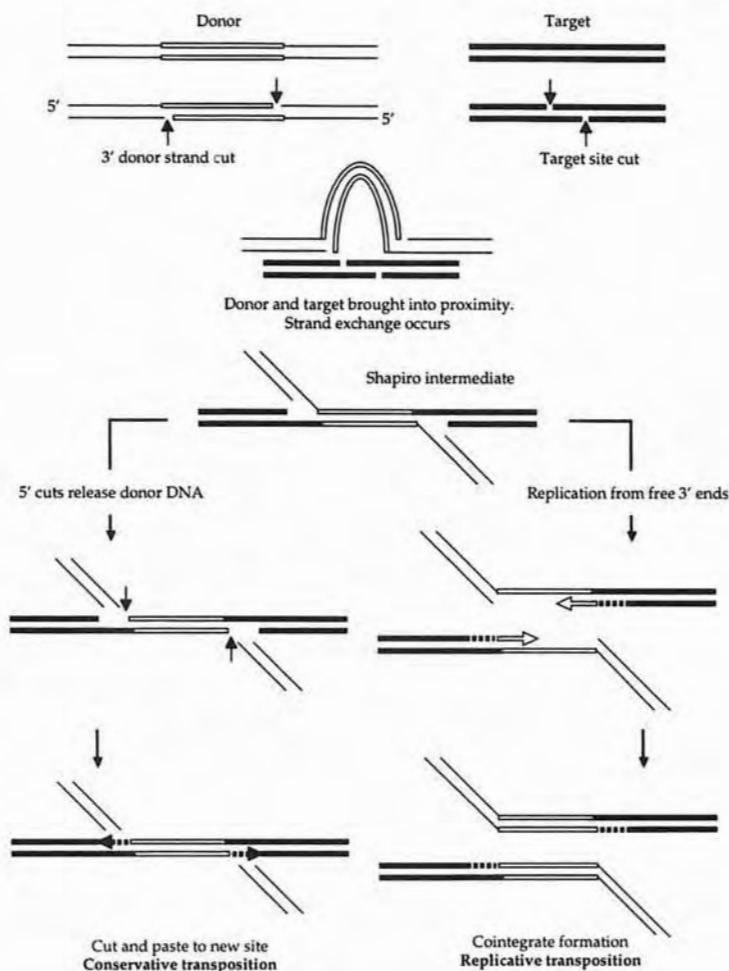
generate a cDNA for processing and integration. Nonviral retroelements integrate into host DNA essentially utilizing only the DNA repair stage of transposition.

(3) *Target site duplication.* For most transposable elements, the strand exchange stage of transposition occurs at a staggered break in the target DNA, so that the freshly integrated element is initially flanked by single-stranded DNA tails. Repair synthesis over the gaps generates direct repeats of the sequence at the target site, so-called **target site duplications (TSDs)** (Figure 13.2). Most transposable elements generate TSDs of specific length, but the heterogeneous class I.2 retroelements generate variable sized TSDs even within the same genome, probably because they utilize broken DNA ends adventitiously.

(4) *Target preference.* Transposable elements vary in their preference for target sites, some integrating at specific sequences, others appearing to integrate at random. The *E. coli* transposon Tn7 usually integrates at a specific site, but can switch to random transposition if the site is unavailable. Most elements avoid integrating into preexisting elements of the same type, i.e. they display **cis-immunity** (c.f. *twintrons*). Where individual transposition events can be studied, many elements demonstrate topographical preference. Some eukaryotic transposons move only a short distance (1–10 kbp) from their donor site; this is **regional reintegration**. Bacterial replicative transposons prefer to move to plasmids rather than from plasmids to the bacterial chromosome.

**Conservative and replicative transposition.** Transposition mechanisms can be either conservative or replicative. **Conservative transposition** (also called **nonreplicative** or **simple transposition**, or the **cut-and-paste mechanism**) involves the excision of the element from one site and its integration at another. The active transposition mechanism does not increase the copy number of the element, although this may be achieved passively (see below). **Replicative transposition** (also called **non-conservative** or **complex transposition**) involves duplication of the element, one copy remaining at the donor site and one copy integrating at the target site. Retroelement transposition is necessarily replicative because the element yielding the transcript which is converted into a cDNA for integration remains in the genome.

For transposons, the difference between conservative and replicative transposition reflects the timing of the donor cut reaction, as shown in Figure 13.3. In both cases, transposition is initiated by a donor cut at the 3' end of the element, providing free ends for interaction with the target site. In



**Figure 13.3:** A unified model of conservative and replicative transposition. Donor DNA is shown as thin lines, target DNA as thick lines, the transposon as an open box. Strand breaks are introduced at the positions indicated by pointers, and replication is shown by in-strand arrows. In this model, the stages of conservative and replicative transposition are shown as being common until the formation of the **Shapiro intermediate**, a structure where donor and target DNA are joined together. If a 5' donor cut occurs, the donor DNA is released and transposition is conservative. If not, replication proceeds across the joint and transposition is replicative. The order of events is not necessarily the same for all elements. For instance, the 5' donor cut may precede strand transfer, so that a Shapiro intermediate is never formed. This mechanism may be used by transposons which always follow the conservative pathway. Additionally, the target site cut is thought to occur during strand exchange transesterification, so that there is never a double-strand break in the target DNA (c.f. *homing introns*, *passive transposition*).

conservative transposition, the donor DNA is released by an additional cut at the 5' end of the element, leaving a gap in the donor DNA. For replicative transposition, only the 3' end of the transposon is cut, so that the donor and target DNA remain physically joined. Repair replication across the transposon, primed by the free 3' ends of host DNA, then duplicates the element. The unified model of DNA transposition is based on the study of well-characterized transposable elements including bacteriophage Mu, Tn5 and retroviruses. There are different versions of the model which place the major reactions in different orders. Most important is the timing of the 5' donor cut: if it occurs before strand exchange, the element is released as a free molecule, whereas if it occurs after

strand exchange, the element is transferred directly from site to site. For those elements such as bacteriophage Mu which transpose either conservatively or replicatively, the 5' donor cut is optional, and probably occurs after strand transfer. For transposons such as P-elements and *Ac-Ds* elements which always transpose conservatively, the 5' donor cut probably precedes strand transfer. Little is known of the mechanism of 5' end cleavage.

**Resolution.** Following replicative transposition between, for example, two plasmids, the circles become joined to form a composite structure termed a **cointegrate**, with copies of the transposable element marking the boundary between each contributing replicon. Transposons which move in this manner also carry functions which separate the cointegrate into its two constituent replicons, a process termed **resolution**. Each transposon encodes **resolvase**, a site-specific recombinase which acts on a short sequence within the transposon, the **internal resolution site** (*res*). Recombination between two *res* elements in the cointegrate catalyzed by resolvase separates the two molecules. Resolvase exhibits strong directionality in its activity, i.e. it rarely converts separate molecules into a cointegrate. This specificity may be caused by topological constraints (q.v. *site-specific recombination, DNA topology*).

**Passive transposition.** Passive transposition describes any process where the mobilization of a transposable element is not self-controlled. There are two types of passive transposition.

(1) **In-trans passive transposition** occurs by the same mechanism used by active elements and is facilitated by enzymes (e.g. transposase, reverse transcriptase) supplied in *trans* by an active element. Only nonautonomous elements move by in-trans passive transposition.

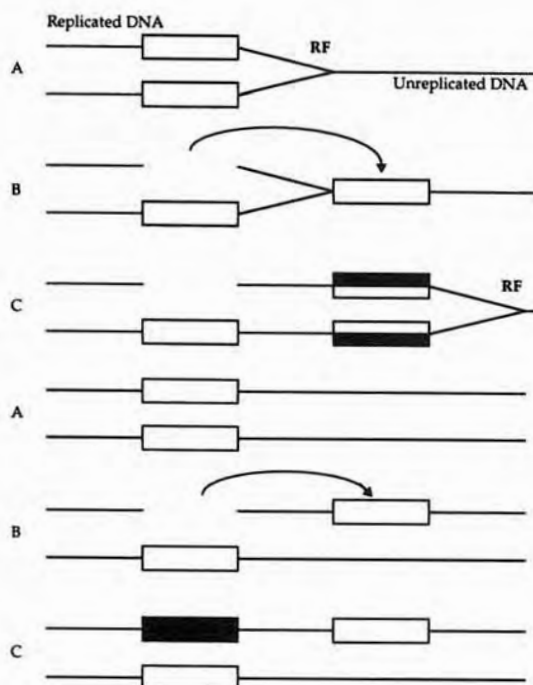
(2) **Cell-controlled passive transposition** is an apparent transposition event caused by cellular DNA replication (Figure 13.4). Both autonomous and nonautonomous elements move by this process.

All elements must increase their copy number by transposition to avoid extinction. For those elements that transpose conservatively, amplification must be achieved passively by the cell. Cell-controlled passive transposition occurs in two ways. (1) Transposition from replicated to nonreplicated DNA — this method is utilized by, for example, the *Ac-Ds* elements of maize. (2) Repair of excision breaks by gene conversion — this method is utilized by *Drosophila* P-elements, for example, and is also the sole transposition mechanism of *homing introns and inteins* (q.v.). Additionally, copy number may be increased by increasing the copy number of the host replicon — this occurs for bacterial transposons which integrate into plasmids and bacteriophage genomes.

**Retrotransposition and retroposition.** Class I.1 retroelements mobilize by transcribing an integrated element and converting the transcript into a cDNA copy which is then integrated at a new site (the process of cDNA synthesis is discussed in more detail as part of the *retrovirus* (q.v.) infection cycle). For retroviruses, mobilization occurs in the context of the viral infection cycle, i.e. the transcript is packaged into the viral particle along with reverse transcriptase and is introduced into the cell, whilst for other retrotransposons, the entire cycle occurs within the confines of the cell.

Once cDNA has been generated, retroelements of the viral family follow a pathway which is similar to transposition by transposons, but because it includes an extra reverse transcription stage, it is termed **retrotransposition**. The cDNA usually has a small amount of host DNA sequence flanking it (often 2 bp), and this is removed from the 3' end of each strand by integrase in a reaction analogous to the 3' donor cut stage of transposition. The integrase then facilitates a transesterification reaction between the free 3' ends of the cDNA and the host DNA at the target site. Integration is followed by repair synthesis, which generates 2 bp target site duplications and presumably removes the remaining 2bp of 'old' donor site DNA from the 5' ends of the element.

Class I.2 elements mobilize using a mechanism unrelated to retroviral replication. The cDNA is generated either by self-priming or by DNA-priming at the target site. Integration is achieved by cDNA 'repair synthesis' across the target site using RNA as the template. Because this pathway



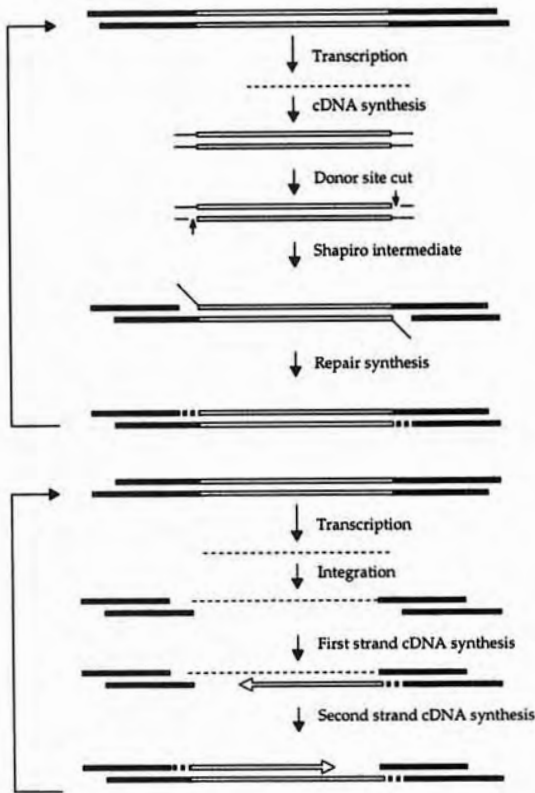
**Figure 13.4:** Increasing copy number during conservative transposition by cell-controlled passive transposition. Two mechanisms are shown, each divided into pretransposition stage (A), active transposition stage (B) and passive transposition stage (C). Homing introns and inteins use the second mechanism without actively transposing: intron- or intein-encoded endonuclease cleaves homologous, intronless DNA and the sequence is copied to the intronless allele by gene conversion.

lacks most of the steps in classic transposition but involves a reverse transcription step, it is termed **retrotransposition**. Retrotransposition and retroposition mechanisms are shown in Figure 13.5.

**Regulation of transposition.** Behaving as selfish DNA, successful transposable elements mobilize often and increase their copy number wherever possible. However, unchecked transposition would destroy the host genome and any transposable elements contained within it. Most transposable elements therefore regulate their own transposition. Such regulatory mechanisms control not only the frequency of transposition, but also its timing in relation to the cell cycle and the nature of the target site. An exception to this rule is the transposing bacteriophage Mu, to whom the survival of the host, *E. coli*, is inconsequential. The virus destroys the host genome by repeated transposition. The multiple copies are then packaged along with host DNA into phage particles.

Transposition may be regulated in a number of ways, and the *Drosophila* P-element and bacterial transposon Tn10 are well understood in this respect. Many transposons encode a transposase repressor colinear with part of the functional transposase. P-element transposition is restricted to germ cells in this manner. The transposase gene is interrupted by three introns, and differential splicing occurs in germ cells and somatic cells. In the germ cells, all three introns are removed and the mature transcript encodes a polypeptide of 87 kDa with transposase activity. In somatic cells, intron 3 is not removed and a truncated polypeptide is generated due to an in-frame stop codon within the intron. This truncated protein may act as a repressor of transposition in several ways — it could bind to the substrate of the genuine transposase and block its activity, or it could combine with the enzyme itself and inhibit it. Similar mechanisms may be used by other transposons, e.g. the





**Figure 13.5:** Overview of retrotransposition (upper) and retrotransposition (lower) mechanisms. The integrated element is shown as an open box, RNA as a thin line and DNA as a thick line (that generated by repair synthesis as a broken line, corresponding to position of target site duplications).

*Ac* element of maize, although in this case there is no tissue-specific splicing. *Tn5* also generates a truncated repressor, in this case by utilizing an alternative translational start site.

*Tn10* represses its transposition by transcription from a promoter called  $p_{OUT}$  having opposite polarity to the transposase promoter  $p_{IN}$ . Transcription from  $p_{OUT}$  represses transposition in several ways: by countertranscription against the  $p_{IN}$  product, by production of antisense RNA, and by preventing the transposase binding to its recognition site. *Tn10* transposase is an example of an enzyme which shows *cis*-preference, i.e. it acts only on the element which encoded it. Although the basis of *cis*-preference is not clear, it represents another mechanism of transpositional control: increasing the *Tn10* copy number does not increase the amount of available transposase and therefore does not lead to an increase in transpositional activity.

Both the binding of transposase and transcription from  $p_{IN}$  are inhibited by DNA methylation (q.v.). *Tn10* contains several Dam methylation sites, and its transposition is therefore effectively restricted to a window just following replication when one of the strands is unmethylated. This encourages the element to transpose only when there is a second copy of the genome in the cell, so that any damage caused by excision may be repaired by recombination. DNA methylation is used to suppress transposition in eukaryotes and appears to be a copy-number-dependent defense system against overproliferation (see DNA Methylation and Epigenetic Regulation).

### 13.2 Consequences of transposition

**Genetic consequences of transposition.** Transposable elements were first recognized by their ability to cause mutations, and **insertional mutation** (gene disruption) is the most direct consequence of transposition. However, transposable elements can also influence gene expression by interfering with the relationship between the gene and its regulatory elements, or by modulating DNA structure.

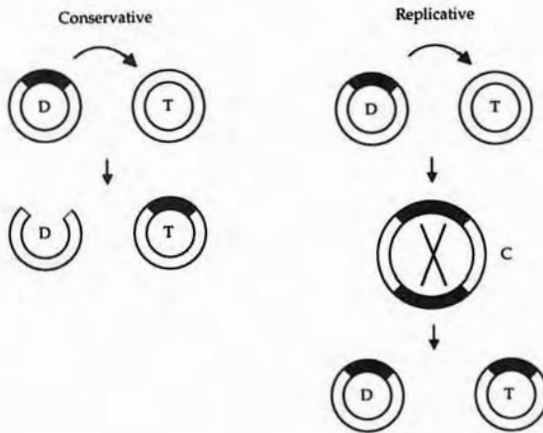
Insertional mutagenesis can affect single genes or many genes at once (q.v. *hybrid dysgenesis*). The classic effect of transposition is the generation of an **unstable mutant allele**, a *mutant allele* (q.v.) that reverts to wild-type with high frequency, but whose reversion rate is not influenced by mutagens (because the reversion represents an excision event). In bacteria, insertional mutations in operons are often polar because transcription through the element is blocked, or because downstream genes cannot be translated efficiently (q.v. *polar mutation*). Bacterial transposons often possess transcriptional terminators or complex secondary structures at their borders to prevent invasive transcription from neighboring promoters. This protects the element from ectopic activation of transposition, which could be lethal to the cell.

The modulation of gene expression without direct insertional inactivation occurs if the element carries an endogenous gene whose expression is controlled within the element, or if it carries an endogenous regulatory element which influences the expression of adjacent genes. Such phenomena are often seen in retroviruses and have led to the isolation of many *oncogenes* (q.v.). An element may influence neighboring gene expression with one of its own regulatory elements, e.g. strong promoters are often found in the *long terminal repeats* (q.v.) of retroviral-type elements. Additionally, the element may possess a sequence which creates a hybrid regulatory element upon insertion, e.g. many IS elements carry sequences which resemble the -35 *box* motifs (q.v.) of bacterial promoters and may activate repressed operons such as *lac*, which lack such motifs.

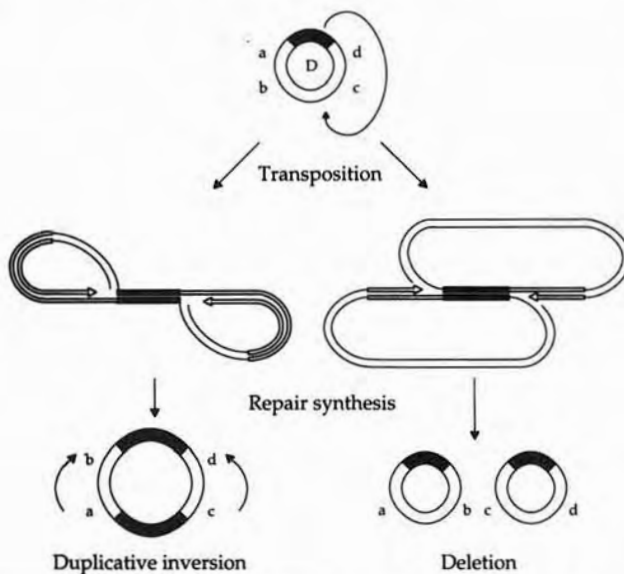
Gene expression may be influenced by modulation of the structural properties of DNA. The insertion of IS1 can activate cryptic operons by altering the topology of the local DNA (similar effects are caused by point mutations in the gene encoding DNA gyrase). DNA methylation is used by some eukaryotes to restrict transposon proliferation, but if a transposon inserts near a repetitive endogenous sequence, this can be targeted by the same system, resulting in epigenetic silencing (see DNA Methylation). Silencing of local genes can also be brought about by the *Drosophila* retrotransposon *gypsy*, which can induce loss of function effects even when it inserts several kbp away from the gene. The *gypsy* element contains a repetitive sequence which acts as a *boundary element* (q.v.), thus defining a region of repressed chromatin.

**Structural consequences of precise transposition.** Transposition events mediate a range of structural rearrangements in the flanking host DNA. Such events are simplest when the donor and target sites lie on different molecules, e.g. on separate plasmids (**intermolecular transposition**). Conservative transposition results in the integration of the element into the target replicon, leaving a gap in the donor replicon (which may be lost if the gap is not repaired). Replicative transposition generates a **cointegrate**, a replicon fusion containing two copies of the element, and this may be resolved by site-specific recombination, producing an unchanged donor replicon and a target replicon with a new insertion. These alternative schemes are represented in Figure 13.6. A new range of structures can be generated if the element moves to a new site within the same replicon (**intramolecular transposition**). During replicative transposition, the particular relationship of the strands as they are exchanged determines whether a deletion or a duplicative inversion is generated, as shown in Figure 13.7. Conservative intramolecular transposition involving, for example, an IS element does not cause any rearrangement (although it generates a break at the excision site). Compound transposons, however, can generate a range of deletion and inversion derivatives containing only one IS element and lacking the internal region of the transposon.

The **precise excision** of a transposon during conservative transposition *does not* return the structure of the donor site to its preintegration state. This is because, if the break is repaired by direct



**Figure 13.6:** The consequences of intermolecular transposition. D, donor molecule; T, target molecule; C, cointegrate. The cross shows site-specific recombination (resolution).



**Figure 13.7:** Structural consequences of intramolecular replicative transposition. The donor molecule (D) has four loci a, b, c, d, where c is origin of replication. The middle row shows the alternative Shapiro intermediates which can form when the donor and target sequences lie within the same circular molecule (compare these to Figure 13.3). In one configuration, DNA synthesis across the intermediate generates a duplicative inversion. In the other, DNA synthesis across the intermediate divides the original replicon in two. The circle lacking the origin of replication, c, is lost. The other circle persists as a deletion mutant.

ligation, the target site duplications remain as a **footprint** of the transposition event. Depending on the size and position of the footprint, precise excision from a gene may still restore gene expression. For example, *Spm* elements generate a 3 bp duplication which restores the original reading frame. Even footprints which are not multiples of three nucleotides may eventually revert due to secondary mutations. Many integration events occur in noncoding DNA, and in this case the footprint may have no effect.

**Consequences of aberrant transposition.** So far, only 'legitimate' transposition events have been considered. Aberrant or imperfect transposition processes include **one-ended transposition**, where only one end of the transposon is integrated at the target site, **partial transposition**, when only part of the transposon actually moves, leaving a large footprint at the donor site, and **cryptic-site transposition**, where flanking host DNA is mobilized along with the transposon because it is recognized by the transposase. Two transposons can also facilitate adventitious **cooperative transposition** to mobilize a segment of intervening DNA, the efficiency of this process being inversely proportional to the distance between the individual elements. Compound transposons of bacteria perform this process naturally, but small circular replicons can sometimes transpose the 'wrong' central region, i.e. the rest of the replicon instead of the central region of the transposon. This is termed **inverse transposition**.

**Consequences of host-cell activity.** The host cell can facilitate passive transposition of integrated elements as part of normal DNA metabolism. However, cellular DNA metabolism can also mediate aberrant rearrangements which occur by three major processes: (1) attempts to repair gaps left by excision; (2) recombination within elements; (3) recombination between elements.

Excision of a transposon leaves a double-stranded break at the donor site. The cell often attempts repair by homologous recombination using as the template either a sister chromatid (or a sister genome in bacteria) or, in eukaryotes, a homologous chromosome. Alternatively, the broken ends may be directly end-joined, a process which may be perfect (resulting in a simple target site duplication footprint) or may be preceded by some degradation of free ends, generating a deletion. Repair mediated by homologous recombination often leads to passive transposition in eukaryotes because an allele containing an integrated element may be used as the template. Alternatively, if the homologous chromosome lacks a transposon, an uninterrupted allele may be used as a repair template, resulting in proper *reversion*, i.e. removal of the element and its target site duplications. Occasionally, only part of the information in a transposon-bearing allele is used for homologous recombination, resulting in a 'partial' passive transposition, giving the effect of an imperfect excision.

An alternative response to the appearance of a double-strand break is the formation of **P-DNA**<sup>1</sup>, inverted repeats of host DNA at the donor site. P-DNA is generated by strand-to-strand sister chromatid ligation at the broken chromosome ends following replication, generating a hairpin which effectively joins the two chromatids together. Subsequent segregation of the dicentric compound chromosome causes random breakage at the end, generating an inverse duplication on one chromosome and a deletion on the other. The new breaks are then joined following replication and the process is repeated, a so-called **breakage-fusion-bridge cycle**, which eventually ceases when telomeres are added to the broken ends (q.v. *illegitimate recombination*).

Recombination within transposable elements can result in their (passive) excision from the donor site. This occurs by homologous recombination between inverted terminal repeats or direct repeats, the former often resulting in perfect excision (but leaving a target site duplication footprint), the latter often leaving a larger footprint comprising a single copy of the direct repeat. The latter process is probably responsible for the prevalence of 'solo' retrotransposon LTRs in many eukaryotes, e.g. the  $\delta$  elements of yeast, which are the LTRs of the Ty retrotransposons. Recombination between target-site duplications can also occur and causes reversion. Thus partial excision, perfect excision and reversion are all possible results of recombination within transposable elements, and the length of the repeat region appears to play an important role in the mechanism favored by any particular element.

Similar processes occur which are independent of host and transposon-encoded recombination

<sup>1</sup>P-DNA usually refers to inverted repeats of DNA at transposon excision points caused by hairpin-mediated repair. The same term, however, may be used to describe the tertiary structure formed by certain DNA sequences in alcoholic solvents and also to any DNA synthesized during the pachytene stage of *meiosis* (q.v.).



systems, and generally leave a footprint of the element involved. In this case, intrastrand hairpin or stem-loop structures may form between inverted terminal repeats: replication across the looped out region results in a deletion in the daughter strand. This is an extreme example of 'strand slipping' as a form of *illegitimate recombination* (q.v.).

Finally, transposable elements with moderate-to-high copy numbers represent portable recombinogenic sequences. Recombination between dispersed homologous elements on the same chromosome can cause deletion, inversion or (for linear chromosomes) circularization. Recombination between two elements on different chromosomes can cause terminal deletions, chromosome fusions (cointegration), and translocations.

**Global aspects of transposition.** The consequences of transposition have been discussed in terms of single transposition events. However, there are several tens to over a million potential transposable elements in any given genome, and it is necessary to consider the wider implications of transposition. The role of transposable elements in the evolution of gene structure and genome organization is discussed in other chapters (*see Genomes and Mapping, Proteins: Structure, Function and Evolution*)

In *Drosophila*, **hybrid dysgenesis** describes the simultaneous appearance of many deleterious mutations in the progeny of a specific cross. **Dysgenic offspring** are often sterile and have a high frequency of chromosome rearrangements and mutations, but the mutations are unstable and can revert to wild-type in subsequent generations. This phenomenon is caused by the simultaneous activity of many transposable elements. In *Drosophila*, two transposable element families cause hybrid dysgenesis, P-elements and I-elements. **PM dysgenesis** occurs when P-strain males (males carrying P-elements) are crossed with M-strain females (females lacking P-elements). The reciprocal cross does not produce dysgenic offspring; nor does a cross between P-strain males and P-strain females. The cytoplasm of P-strain eggs contains a repressor of P-element transposition, whereas the cytoplasm of the M-strain does not. These are termed the P- and M-cytotypes. When P-strain sperm DNA is released into an M-strain egg, the many P-elements in the sperm genome are activated simultaneously and jump to new sites, causing numerous mutations. Conversely, P-elements released into a P-strain egg are immediately repressed by the repressor already present. The repressor is thought to be the truncated transposase product encoded by the P-element itself (*see above*).

Bursts of transpositional activity are seen in plants which globally regulate transposon mobility by DNA methylation. Many different families of elements are mobilized simultaneously, indicating that most elements are regulated by a common mechanism. Unlike the situation in *Drosophila*, the periods of transpositional activity in plants appear to be coordinated and reproducible effects which are regulated in a developmental context. It is not clear whether this reflects a specific role for transpositional bursts in development.

**Transposable elements as research tools.** The remarkable properties of transposable elements have made them desirable tools for the molecular biologist. Their ability to jump between genomes has enabled them to be used as gene delivery vectors, and their tendency to disrupt genes has allowed them to be exploited as mutagens. They have been used to introduce genome rearrangements, as mobile reporter systems and as tags to clone nearby genes. The many ways in which transposable elements can be exploited are summarized in *Table 13.1*.

### 13.3 Transposons

**Structure and subclassification of transposons.** The term *transposon* was coined to describe a bacterial transposable element carrying genes for antibiotic resistance as well as those concerned with transposition functions. More recently, the term has been adopted to encompass all transposable elements which mobilize directly as DNA, i.e. all class II transposable elements.

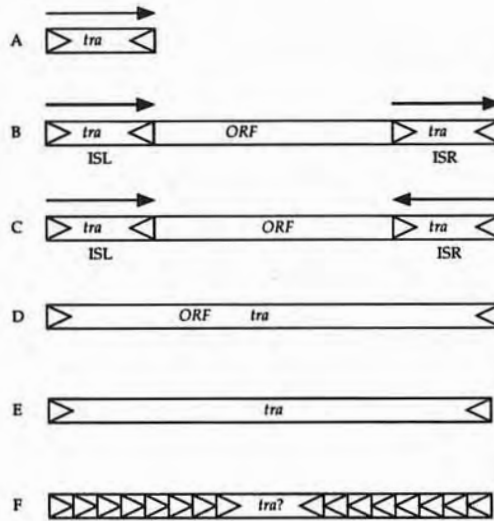
**Table 13.1:** Some ways in which transposable elements are exploited for further details see Recombinant DNA, Mutation and Selection

Use	Examples
Gene delivery vector	Introduction of genes into bacterial chromosomes by cloning in a transposon and introducing by transformation Introduction of genes into <i>Drosophila</i> by injecting plasmids containing recombinant P-elements into germ cells Introduction of genes into plants by delivering recombinant Ac-Ds elements on Ti plasmid vectors
Gene mutation	Generating transgenic mice using recombinant retroviral vectors Generation of random mutants by general insertional mutagenesis Isolation of specific mutants by PCR using transposon and gene-specific primers Generation of targeted mutations by transgene-mediated repair following excision
Structural rearrangements	Generation of nested deletions in bacteria by intramolecular transposition Generation of specific genome rearrangements by delivering FLP/FRP recombinase system into genome
Gene expression	Induce overexpression of neighboring genes using strong promoter Assay for endogenous regulatory functions using <i>entrapment vectors</i> (q.v.)
Cloning	Cloning genes directly from mutant organisms using transposon as a flag — <i>transposon tagging</i> (q.v.) Include plasmid origin of replication in transposon to facilitate <i>plasmid rescue</i> (q.v.) <i>In vivo</i> cloning with bacteriophage Mu — isolate genes directly from mutant cells by packaging into phage heads
Gene mapping	Mapping plasmids by insertional mutagenesis Introducing portable sequencing primer template to facilitate large-scale sequencing experiments

All transposons have a conserved structure comprising one or more open reading frames (one encoding transposase) flanked by **inverted terminal repeats (ITRs)**. The terminal repeats are necessary for transposition and are the substrates recognized by the transposase. However, they are often not sufficient, and further internal sequences are required.

Three major groups of transposons are recognized in bacteria: **class I transposons** include the IS elements (simple transposons) and composite transposons; **class II transposons** are the complex transposons; **class III transposons** are the transposing bacteriophages related to Mu. A fourth class has been proposed to represent transposons which do not fit into any of the above classes. All eukaryotic transposons resemble bacterial IS elements in that they encode only the functions required for transposition. Most have a canonical structure comprising a central region flanked by short ITRs. A special class of transposons termed **foldback elements** consist mainly of large inverted repeats. Little is known of how these structures mobilize, but they are not thought to utilize an RNA intermediate. Transposon structure is summarized in *Figure 13.8*.

**IS elements.** Bacterial transposable elements were discovered as a consequence of their ability to cause unstable but strongly polar mutations in *E. coli*. Hybridization analysis showed that a small family of inserted DNA sequences was responsible for many of the observed mutations, and these were termed **insertion sequences (IS elements)**. About 100 different IS elements have been identified, the majority in enteric bacteria and their plasmids. They are all small (<2.5 kbp) and consist of a central unique region flanked by imperfect inverted terminal repeats which differ in size, sequence and relatedness. The central region may contain 1–3 open reading frames, one of which encodes transposase, but there are no genes for nonessential functions. For this reason, they are known as **simple transposons**. Different IS elements also differ with respect to their target-site preference and



**Figure 13.8:** The structure of different classes of transposons. (A) A bacterial IS element. (B) A compound transposon with IS elements in same orientation. (C) A compound transposon with IS elements in opposite orientations: ISL and ISR are assigned according to the polarity of the genetic map of the internal region. (D) A complex transposon. (E) A typical eukaryotic transposon. (F) A eukaryotic foldback element. *tra* is a gene for transposase; only one of the *tra* genes needs to be functional in B and C. *ORF* is a gene for nonessential function, e.g. antibiotic resistance.

**Table 13.2:** Properties of some IS elements

Name	Element size	ITRs (L/R)	TSDs	Target-site preference
IS1	768	20/23	9	AT rich with terminal G or C
IS2	1327	32/41	5	Hotspots in P2 genome
IS3	1258	29/40	3	?
IS4	1426	16/18	11–13	AAAN <sub>20</sub> TTT
IS10	1329	17/22	9	GCTNAGC
IS91	≈1800	8/9	0	?

ITRs, inverted terminal repeats; TSDs, target site duplications.  
 Note that the left and right (L/R) inverted terminal repeats usually differ in length as well as sequence. All sizes are in base pairs.

the size of the target-site duplications they generate. These properties reflect the specificities of the transposase encoded by each element. Table 13.2 lists the properties of some IS elements.

**Composite and complex transposons.** Transposons were originally defined as bacterial transposable elements which carried not only those genes required for transposition, but also nonessential genes, e.g. for antibiotic resistance. Such elements were distinguished from IS elements by the designation Tn. Now the term transposon covers all bacterial mobile elements, IS elements are identified as simple transposons whereas the other elements fall into the two categories described below.

**Composite or compound transposons** comprise a central region containing nonessential genes flanked by two IS elements. The transposase functions, and the substrates for transposition, are provided by one or both IS elements which cooperate to transpose the intervening DNA. The IS elements may lie in the same or opposite orientations, so that the composite transposon appears to have either long direct or long inverted repeats (Figure 13.8). For some transposons, the flanking IS elements are identical, whereas in others they are different, often because one of the elements has undergone

**Table 13.3:** Features of bacterial composite and complex transposons

Name	Length	Marker(s)	ITRs
<i>Composite (class I) transposons</i>			
Nonessential genes flanked by IS elements			
Tn5	5.4 kbp	Kanamycin <sup>r</sup>	Long (1.5 kbp). Inverted IS50 elements, only IS50R is functional
Tn9	2.6 kbp	Chloramphenicol <sup>r</sup>	Short (18/23 bp). Direct IS1 elements, both functional
Tn903	3.1 kbp	Kanamycin <sup>r</sup>	Long (1.1 kbp). Inverted IS903 elements, both functional
<i>Complex (class II) transposons</i>			
Essential and nonessential genes flanked by inverted terminal repeats			
Tn3	5 kbp	Ampicillin <sup>r</sup>	38 bp
Tn501	8.2 kbp	Mercury <sup>r</sup>	35/38 bp
Tn7	14 kbp	Trimethoprim <sup>r</sup> , Streptomycin/spectinomycin <sup>r</sup>	≈30 bp

ITRs, inverted terminal repeats. <sup>r</sup> = Resistance. Note that class I transposons include both the composite transposons and the IS elements.

mutation and provides only the *cis*-acting functions. The IS elements in composite transposons can also mobilize individually. The maintenance of the transposon (**transposon coherence**) reflects selective pressure for the phenotype conferred by the intervening DNA (also q.v. *inverse transposition*).

**Complex transposons** are more like individual IS elements in their organization, but contain nonessential genes as well as those required for transposition. Both types of gene are found together in the central region, which is flanked by short inverted terminal repeats. Table 13.3 lists the properties of a representative sample of composite and complex transposons.

**Transposon families in eukaryotes.** In eukaryotes, transposons are termed class II transposable elements (class I elements mobilize using an RNA intermediate). This nomenclature is independent from that distinguishing the different classes of bacterial transposon.

Transposons have been identified in many eukaryotic species although they are poorly characterized in vertebrates and genome sequencing has shown they are absent from yeast. Most eukaryotic transposons are similar in structure to bacterial IS elements, comprising a central region encoding a putative transposase and other functions required for transposition, flanked by short inverted terminal repeats. A distinct family of eukaryotic elements, foldback elements, has much larger ITRs and a small central region containing several ORFs which may encode transposition functions. The properties of major eukaryotic class II transposable element families are listed in Table 13.4.

The first transposable elements to be recognized were maize transposons, of which there are 10 or more distinct families. As with other transposable elements, they were initially identified because of their unstable mutagenic effects on endogenous genes, and were termed **controlling elements**. The genetic analysis of controlling elements allowed them to be placed into families and to be classed as autonomous or nonautonomous. Molecular analysis has shown that this dichotomy reflects the existence of full-length active elements and variable-length defective elements which have lost essential transposition functions. An example is the *Ac-Ds* family: the autonomous activator (*Ac*) elements are 4.5 kbp in length and encode a functional transposase, whereas dissociator (*Ds*) elements are defective derivatives of variable size. The *Ac-Ds* family possess 10–11 bp imperfect terminal repeats and generate 8 bp target-site duplications. A similar situation is found for the *Drosophila* P-elements, which are the best-characterized and most widely exploited of all eukaryotic transposons. P-elements were identified through the study of *hybrid dysgenesis* (q.v.),



**Table 13.4:** Properties of some well-characterized eukaryotic transposons

Species	Class II element families	ITRs	TSDs
<i>Drosophila melanogaster</i>	P	31	8
	<i>hobo</i>	12	8
	<i>mariner</i>	28	2
	FB	Large	9
<i>Caenorhabditis elegans</i>	Tc1	54	2
<i>Zea mays</i>	<i>Ac/Ds</i>	10/11	8
	<i>Spm/dSpm</i>	13	3
	<i>Mu/Mn</i>	≈200	9
<i>Antirrhinum majus</i>	Tam1	13/14	3
	Tam3	12	8/5
Several eukaryotes	TU/Puppy	Large	8

ITRs, inverted terminal repeats; TSDs, target site duplications.

which is largely caused by P-element insertion. Some strains of *D. melanogaster* contain 40–50 P-elements, whereas others lack them altogether. Autonomous P-elements are about 3 kbp long and are flanked by 31 bp perfect ITRs. They prefer the target site GGCCAGAC, and generate 8 bp target-site duplications. About two-thirds of the P-elements in a given strain are defective, and range in size from several hundred bp to 2.9 kbp.

In both cases, the transposase gene appears to be the only open reading frame and is interrupted by several introns. In P-elements, splicing is regulated differently in the germ cells and the somatic lineage, so that transposition occurs only in the germ cells. In *Ac* elements, splicing is constitutive and transposition is not restricted to the germline. Thus, one consequence of transposition in maize is variegation, where different mutations occur in different cells and become clonally propagated. Both P-elements and *Ac-Ds* elements have been exploited as gene transfer vectors and for the investigation of gene expression and regulation. Like bacterial transposons, these elements have been used as mutagens and can be exploited to clone genes by 'tagging' (Table 13.1).

**Mobile introns and inteins.** Mobile introns are a class of transposable element found in yeast mitochondria. These introns contain an open reading frame encoding a site-specific endonuclease, similar to bacterial restriction endonucleases except that the target sequence is much longer, typically >20 bp. The size of the target sequence means that it is rare, and in fact it is only found once in the genome, at the site where the intron is inserted. The endonuclease therefore cannot cleave host DNA unless there is an intronless allele of the same gene present in the cell. If such an allele is encountered, a double-stranded break is introduced which may be repaired by homologous recombination with the intron-containing allele, a process which results in the intron being copied and inserted into the intronless allele by gene conversion (this is an example of *passive transposition*, see above). These remarkable elements may be termed **homing introns** because of their ability to home in on a particular site within the target gene. Even more intriguing, however, are the **homing inteins**, which are polypeptides spliced out of nascent proteins after translation (q.v. *protein modification*). These may also encode endonucleases, which in this case cleave within the open reading frame of an inteinless allele of the same gene and initiate the transfer of the intein segment by gene conversion.

**Bacterial integrons.** Integrons are bacterial operons incorporating various structural genes. The integron consists of a regulatory region, an open reading frame encoding a site-specific recombinase and a target site for the recombinase, *attI*, which can integrate a variety of antibiotic resistance gene cassettes carrying a **59-base element**, which is also recognized by the recombinase. Unlike other types of cassette system, where the target locus is never 'empty' and one cassette replaces another by nonreciprocal recombination, integrons may be empty or may integrate one or more cassettes in

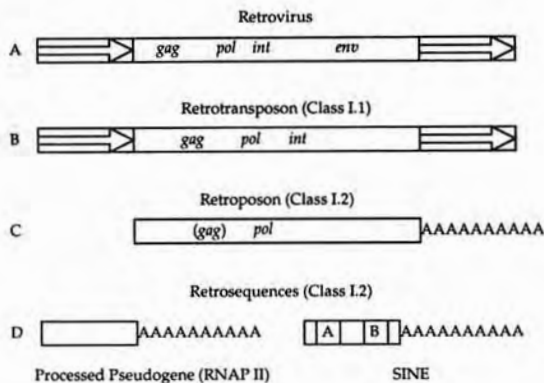
tandem. In this case, recombination can occur between 59-base elements of different cassettes to eliminate particular genes or establish stable junctions. Transcription through *attI* facilitates the expression of all inserted genes in a common operon. Because, like bacteriophage  $\lambda$ , the integron cassettes mobilize by site-specific recombination rather than transposition, they are not true transposons; the term **site-specific transposon** is used to describe such elements.

### 13.4 Retroelements

**Structure and subclassification of retroelements.** Retroelements (Class I elements) are genetic elements which mobilize by reverse transcribing an RNA intermediate and then integrating the cDNA into the genome. Such elements appear to be exclusive to the eukaryotes (cf. *retrotransposons*) and, in yeast and vertebrates, are the predominant form of mobile DNA. Retroelements can be divided into two major families. The **viral family (class I.1 transposable elements)** includes the *retroviruses* (q.v.) and nonviral elements which resemble them in structure and mechanism of transposition (these are termed **retrotransposons**). The **nonviral family (class I.2 transposable elements)** includes autonomous transposable elements which do not resemble retroviruses (termed **retroposons**) and DNA sequences which appear to have been mobilized by accident, due to the adventitious action of reverse transcriptase obtained in *trans*. These are known as **retrosequences**, and those corresponding to cellular mRNA are **retrogenes**, also termed *processed pseudogenes* (q.v.). The classification of

**Table 13.5:** Subclassification of retroelements

Retroelement	Definition
<b>Retroelement</b>	A transposable element mobilizing by reverse transcription of an RNA intermediate
<b>Viral family</b>	Retroelements resembling retroviruses in structure and mechanism of transposition
<b>Class I.1 elements</b>	
<b>Retrovirus</b>	A virus which must integrate into the host genome to replicate, and replicates using an RNA intermediate. Characterized by the possession of flanking long terminal repeats (LTRs) and three open reading frames <i>gag</i> , <i>pol</i> , <i>env</i> , with <i>pol</i> encoding reverse transcriptase/integrase. Only found in eukaryotes
<b>Pararetrovirus</b>	A retrovirus-like particle which cannot transpose (e.g. due to lack of reverse transcriptase). A defective retrovirus
<b>Retrotransposon</b>	A mobile element with the characteristics of a retrovirus but lacking the ability to form infectious particles (although some kind of <i>intracellular</i> particle may be formed)
<b>Nonviral family</b>	Retroelements which do not resemble retroviruses either in structure or mechanism of transposition
<b>Class I.2 elements</b>	
<b>Retroposon</b>	A mobile element with dissimilar structure to retrovirus (i.e. lacking LTRs and retrovirus internal organization) but which encodes reverse transcriptase/integrase
<b>Retrosequence</b> (passive retroposon)	A mobile element similar in structure to a retroposon (no LTRs, polyadenylate tail) but lacking reverse transcriptase/integrase function and requiring these in <i>trans</i>
<b>Retrogene</b>	A retrosequence which resembles a cDNA copy of an endogenous gene, a <i>processed pseudogene</i>
<b>SINE</b>	Short interspersed nuclear element, a retrosequence originating from a class III eukaryotic gene (e.g. <i>Alu</i> , <i>B1</i> ) which is repeated hundreds of thousands of times in the eukaryotic genome
<b>Retron</b>	A bacterial operon of unknown significance encoding reverse transcriptase and containing an untranslated region which is first transcribed and then partially reverse transcribed to generate a composite DNA-RNA structure. Retrons are not mobile because their products are unable to integrate, but they may be classed as retroelements because the sequence is mobilized by reverse transcription



**Figure 13.9:** Structures of eukaryotic retroelements. (A) Retroviruses possess long terminal repeats (arrows) as well as *gag*, *pol/int* and *env* ORFs. (B) Retrotransposons have a similar structure although they lack a functional *env* ORF. (C) Retroposons lack retroviral structure and have a polyadenylate tail, but usually possess ORFs for *gag* and *pol/int*. (D) Retrosequences are nonautonomous retroelements, also known as processed pseudogenes, which are copies of cellular RNA polymerase II transcripts. They possess a polyadenylate tail and are not expressed when integrated, otherwise they are unrelated. SINEs are abundant processed pseudogenes of RNA polymerase III transcripts. They have bipartite internal promoters (shown as blocks A and B) which allow them to be expressed.

**Table 13.6:** Retroelement families in some eukaryotic species

Species	Class I.1 elements (retrotransposons)	Class I.2 elements (active retroposons)	Class I.2 elements (passive retroposons — SINEs and other processed pseudogenes)
<i>Saccharomyces cerevisiae</i>	Ty		
<i>Drosophila melanogaster</i>	<i>copia</i> , <i>gypsy</i> , <i>blood</i> , B104	F, I, G, <i>jockey</i> , D, <i>Doc</i>	
Mammals	IAP, THE1?	LINE	<i>Alu</i> , B1, ID
<i>Dictyostelium discoideum</i>	Tdd-1		
<i>Zea mays</i>	Bs1	Cin4	

different retroelements is summarized in Table 13.5. The structures of the various eukaryotic retroelements are compared in Figure 13.9 and some examples are listed in Table 13.6.

**Class I.1 retroelements.** Retrotransposons share many structural features with retroviruses, including direct long terminal repeats, each flanked by short inverted repeats, and a central region containing several open reading frames (see Viruses for detailed discussion of retroviral genome structure and replication cycle). Similarities within the LTRs often extend to recognizable U3, R and U5 regions, similar promoter and polyadenylation sites, and a P-site to which the tRNA primer of reverse transcription anneals. The central region of retrotransposons contains open reading frames homologous to *gag* and *pol*, the latter encoding reverse transcriptase and integrase. The *env* gene is usually missing or disrupted, however, and this may explain the inability of retrotransposons to form infectious particles. The mechanism of retrotransposon activity is otherwise similar to that of retroviruses. The retrotransposon transcript is localized to a ribonucleoprotein particle, undergoes reverse transcription and then integrates into the genome, usually generating 2 bp target-site duplications. The mechanism of *retrotransposition* is discussed above.

**Class I.2 retroelements.** Class I.2 retroelements represent all the retroelements not resembling retroviruses, i.e. the nonviral family. Autonomous class I.2 elements are **retroposons**. They lack the LTRs and canonical organization of retroviruses but may contain open reading frames with homology to retroviral *gag* and *pol* genes. Distinct families of retroposons exist in *Drosophila*, and the LINE-1 element of mammals represents an important class of moderately repetitive DNA (LINEs are *long interspersed elements*). Unlike other transposable elements, class I.2 elements generate target-site duplications varying in size, perhaps due to their adventitious exploitation of double-stranded breaks in DNA. Cellular mRNAs often undergo passive transposition to form **processed pseudogenes** (these are **passive retroposons**). There appears to be little control over which sequences are transposed, and although processed pseudogenes may comprise up to 10% of the higher eukaryotic genome, individual sequences are represented at a low frequency. scRNA pseudogenes (also known as **SINEs**, for *short interspersed elements*) are an exception to this rule, as certain sequences of this class are highly repetitive, the best example being the human *Alu* element (a processed pseudogene of the 7SL RNA gene) with a copy number approaching one million. The *Alu* element is reorganized compared with the endogenous gene, but retains its internal RNA polymerase III promoter architecture and is therefore expressed, which probably contributes to its preferential amplification. Some mobile introns are class I.2 retroelements as they mobilize by *retrohomology*, involving intron-encoded reverse transcriptase.

**Bacterial retrons.** Transposable retroelements are restricted to eukaryotes, although many bacteria have been shown to synthesize reverse transcriptase. The sole purpose of the enzyme appears to be the production of multiple copies of a single-stranded DNA sequence (**msDNA**; multiple-copy, single-stranded DNA) which is covalently linked to RNA (**msdRNA**; msDNA-associated RNA). The msDNA-RNA structures have been identified in several bacteria. They differ in sequence but share common secondary and tertiary structures. The distinguishing features of msDNA-msdRNA structures are as follows: (1) the 5' end of the msDNA is covalently joined to an internal guanine residue in the msdRNA by a 2'→5' phosphodiester bond; (2) the 3' end of the msDNA and the 3' end of the msdRNA form a short hybrid duplex; (3) both msDNA and msdRNA form multiple stem loop structures.

The msDNA, msdRNA and reverse transcriptase are encoded at a single locus and are cotranscribed. The locus is thus termed a **retron** (reverse transcriptase operon). Some retrons have been found as parts of prophage (these are termed **retronphage**) and can mobilize as part of the phage. Retrtons are generally nonmobile, however, and there is no evidence that the primary transcript of the retron can facilitate its integration into the genome like a retrovirus.

## References

- Berg, D.E. and Howe, M. (1989) *Mobile DNA*. American Society of Microbiology, Washington, DC.
- Saedler, H. and Gierl, A. (1996) Transposable elements. *Curr. Top. Microbiol. Immunol.* Vol. 204.
- Further reading**
- Azpiroz-Leehan, R. and Feldmann, K.A. (1997) T-DNA insertion mutagenesis in *Arabidopsis*: Going back and forth. *Trends Genet.* 13: 335-340.
- Craig, N.L. (1997) Target site selection in transposition. *Annu. Rev. Biochem.* 66: 437-474.
- Cummings, M.P. (1994) Transmission patterns of eukaryotic transposable elements — arguments for and against horizontal transfer. *Trends Ecol. Evol.* 9: 141-145.
- Finnegan, D.J. (1990) Transposable elements and DNA transposition in eukaryotes. *Curr. Opin. Cell Biol.* 2: 475-480.
- Gierl, A., Saedler, H. and Peterson, P.A. (1989) Maize transposable elements. *Annu. Rev. Genet.* 23: 71-85.
- Goff, S.P. (1992) Genetics of retroviral integration. *Annu. Rev. Genet.* 26: 527-544.
- Kleckner, N. (1990) Regulation of transposition in bacteria. *Annu. Rev. Cell Biol.* 6: 297-327.
- Lambowitz, A.M. and Belfort, M. (1993) Introns as mobile genetic elements. *Annu. Rev. Biochem.* 62: 587-622.
- Mizuuchi, K. (1992) Transpositional recombination: Mechanistic insights from bacteriophage Mu and other elements. *Annu. Rev. Biochem.* 61: 1011-1051.
- Sundaresan, V. (1996) Horizontal spread of transposon mutagenesis — new uses for old elements. *Trends Plant Sci.* 6: 184-190.
- Yoder, J.A., Walsh, C.P. and Bestor, T.H. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* 13: 335-340.



## Chapter 14

# Mutagenesis and DNA Repair

### Fundamental concepts and definitions

- A **mutation** is a heritable change in genotype caused by a change in the sequence of nucleotide residues in a given region of the genome. Such changes arise in four ways, which are described as forms of **mutagenesis**.
  - (1) Misincorporation of nucleotides during DNA replication.
  - (2) Damage to DNA (which may be spontaneous or may be induced by agents in the environment) and its repair by the cell.
  - (3) Illegitimate recombination (e.g. end joining, strand slipping during replication, unequal crossing-over) (q.v. *illegitimate recombination, unequal exchange*).
  - (4) The activity of mobile genetic elements (*see* Mobile Genetic Elements).
- Misincorporation and damage to single bases often generates **point mutations**, where one nucleotide is exchanged for another (q.v. *transition, transversion*). Point mutations within genes can alter properties of the encoded protein, while those outside may affect DNA-protein interactions. Damage to bases can also cause lethal replication blocks. Other forms of DNA damage (such as chromosome breaks), illegitimate recombination and mobile element activity often cause more dramatic **macromutations** which affect many contiguous nucleotides. Mutations can therefore affect gene function by several different mechanisms, and are often deleterious. (*see* Mutation and Selection).
- The cell devotes much of its resources to the maintenance and repair of DNA. It possesses mechanisms to ensure the fidelity of DNA replication, to repair damaged nucleotides in solution, to repair or replace damaged bases in DNA and to regulate its activity in response to DNA damage.
- All cells possess a **spontaneous mutation rate**, defined as the number of mutation events which normally occur in the genome over a particular time. This rate reflects the fidelity of DNA replication and the efficiency of DNA repair. Mutations in genes which control these processes may increase or decrease the spontaneous mutation rate, and such genes are known as **mutator** and **antimutator genes**, respectively (c.f. *mutation frequency*).

### 14.1 Mutagenesis and replication fidelity

**Misincorporation of nucleotides during DNA synthesis.** Although DNA bases form specific pairs, this is not a highly selective process and energetic considerations predict an error frequency of 1–10%. The components of the replication machinery enhance the accuracy of DNA replication by several orders of magnitude using three major mechanisms:

- (1) base selection by DNA polymerase;
- (2) proofreading exonuclease activity of DNA polymerase;
- (3) the activity of accessory proteins stabilizing the replication complex (e.g. single-stranded binding protein, which limits secondary structure formation in template DNA).

DNA polymerases are highly selective when dNTPs bind to the template-primer-enzyme ternary complex, and mispaired nucleotides bind less well than correctly paired ones. The level of discrimination ranges from nil to 400-fold depending upon the particular polymerase, the type of mispair and its sequence context. This reflects the influence of both base pair and stacking forces in the discrimination step. In addition, the rate of the conformational change which occurs after binding, and which positions the dNTP at the active site of the enzyme, is up to 5000 times greater for bases which form Watson-Crick pairs. The rate at which the phosphodiester bond forms is also

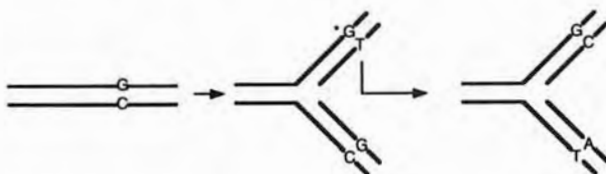
greater for correctly paired bases. The relative importance of these three discriminatory steps varies with each individual polymerase.

Proofreading is facilitated by a 3'→5' exonuclease activity intrinsic to the DNA polymerase. There is a slight pause after each addition step which allows the enzyme to remove mispaired bases, but the major opening for proofreading occurs at the beginning of the subsequent polymerization cycle. The rate of incorporation is very slow when the primer terminus and template are mispaired, and this provides a good opportunity for exonuclease activity. Thus proofreading is an opportunistic activity controlled by the inability of the enzyme to efficiently extend a mispaired terminus.

In *E. coli*, these factors reduce the cumulative error frequency to one misincorporation in every  $10^7$  base pairs, which is still 3–4 orders of magnitude higher than the observed spontaneous mutation rate. The greater fidelity is accounted for by *post-replicative mismatch repair* (q.v.), which corrects mismatched nucleotides on the newly synthesized strand following replication.

**Transient, spontaneous chemical changes in DNA.** The error frequency associated with DNA synthesis (before proofreading and mismatch correction) reflects the inaccuracy of the polymerase as it selects nucleotides. In part, this is an intrinsic property of the enzyme (i.e. a propensity to insert occasional mismatching nucleotides), but also reflects the ability of DNA bases to undergo spontaneous, transient changes involving the rearrangement of hydrogen atoms. Such changes are **tautomeric shifts** and the structural isomers formed are **tautomers** (Figure 14.1). All bases exist as a mixture of tautomeric forms in equilibrium (although for those occurring in DNA and RNA, one tautomer is very stable and shifts to the unstable form are short lived). The alternative tautomers may pair with different bases in the opposite strand, e.g. the stable *keto*-tautomer of guanine pairs only with cytosine, but the rare *enol*-tautomer pairs with thymine. Most tautomeric shifts are harmless because they occur in double-stranded DNA and quickly return to the original state. However, if a shift occurs in the template during replication, a misincorporation can occur. If this is left uncorrected, a mutation appears at the next round of replication. The original misincorporated base does not constitute a mutation because it may be replaced with the correct nucleotide before being used as a template (Figure 14.1).

**Frameshift mutagenesis.** Selectivity and proofreading increase **substitution fidelity**, i.e. the accuracy of incorporating one nucleotide rather than another. The other type of infidelity demonstrated by DNA polymerases is **frameshift infidelity**, which occurs when the template and primer strands slip out of register during replication. This process, sometimes termed **slipping**, **stuttering** or **chattering**, generates small insertions or deletions. In coding regions slipping causes *frameshifts* (q.v.), but the term has been adopted to describe all mutations of this nature regardless of where they occur. Frameshift mutagenesis depends less on the properties of the DNA polymerase than it does on the nature of the template: such mutations are stimulated by repetitive sequences (which promote slipping) and regions of complementarity (which facilitate the formation of hairpins and other secondary structures). These regions are *mutation hotspots* (see Mutation and Selection).



**Figure 14.1:** Mutagenesis by tautomeric shift. The normal *keto*-tautomer of guanine is initially paired with cytosine, but it shifts to the rare *enol*-tautomer (indicated by \*) while acting as a template, causing thymine to be incorporated in the opposite strand. This generates a *mismatch* but not a mutation, because the thymine can be replaced by mismatch repair. If the thymine persists to the subsequent round of replication, adenine is incorporated in the granddaughter strand and the mutation is stabilized.

Frameshift errors are increased by low processivity, suggesting that they occur preferentially when the polymerase dissociates from the template (this offers a possible explanation for the observed differential fidelity of leading strand and lagging strand synthesis; *see* Replication). Frameshift errors are reduced by single-stranded binding proteins which increase the stability of single-stranded DNA and prevent the formation of secondary structures.

## 14.2 DNA damage: Mutation and killing

**DNA damage.** DNA is under assault at all times and sustains many types of damage, a damaged region of DNA being referred to as a **lesion**. The effects of DNA damage are twofold. (1) If a lesion alters base pairing specificity, it is known as a **misinstructional lesion** and may generate a mutation during the next round of replication by the mechanism shown in *Figure 14.1*. (2) If a lesion renders the DNA unable to specify a complementary base, it is known as a **noninstructional lesion** and may cause a **replication block** (where the replication fork is stalled), which is lethal. DNA damage may thus cause *mutation or killing*. To avoid the lethal effects of replication blocks, cells have evolved **damage tolerance mechanisms** which allow replication to continue in the presence of a lesion (q.v. *recombination repair*, *SOS response*). Mutations occurring at the sites of lesions are **targeted mutations**, whereas those occurring elsewhere, e.g. due to replication errors, are **untargeted mutations** (cf. *gene targeting*). A lesion which heralds a mutation is a **premutagenic lesion**.

**Classes of damage to DNA.** DNA damage can be defined as any structure which is not part of a normal intact DNA molecule and which may, if left unattended, result in a mutation or a replication block. Some classes of DNA damage are listed in *Table 14.1*.

**Spontaneous and induced lesions.** DNA may sustain damage intrinsically (due to natural properties of the molecule itself) or extrinsically (due to the effects of outside agents). **Spontaneous lesions** arise due to intrinsic properties of the DNA molecule or interaction with DNA-damaging chemical agents, such as free radicals, produced as a by-product of metabolism. Mutations generated by these lesions, together with those occurring through replication and recombination errors, and the activities of mobile genetic elements, comprise the **spontaneous mutations** underlying the *spontaneous mutation rate* (q.v.) of a given organism. **Induced lesions** are caused by the application of DNA-damaging agents termed **mutagens**. However, the molecular mechanisms by which certain mutagens and natural metabolic by-products cause DNA damage are identical, and here, discrimination between spontaneous and induced lesions becomes blurred.

Certain chemical reactions occur spontaneously within DNA to generate lesions. These include **deamination** (the loss of amino groups from bases), and **base loss** by hydrolysis of the *N*-glycosidic bond, a process described as **depurination** or **depyrimidination** depending upon the type of base. Deamination generates misinstructional lesions; the consequences of this are shown in *Table 14.2*. Base loss by hydrolysis generates a noninstructional **AP site** (*Table 14.1*), which blocks replication. During the *E. coli SOS response* (q.v.), however, translesion synthesis may occur across AP sites and adenosine residues are preferentially inserted, the so-called **A-rule**. As well as these intrinsic spontaneous reactions, a spectrum of oxidative DNA lesions is caused by reactive oxygen species, such as free radicals. Deamination, base loss and oxidative damage may also be induced by mutagens.

Agents in the environment which damage DNA and lead to either mutation or killing are **genotoxic** (*Table 14.3*). Mutagens can be chemical, physical or biological agents (*Table 14.4*). Chemical and physical mutagens interact with DNA in such a way that bonds are broken or rearranged, or new bonds formed, often involving the addition of new chemical groups. Biological mutagens are *mobile genetic elements* (q.v.) which interrupt DNA by transposing into it; they also cause various types of structural rearrangements (*see* Mobile Genetic Elements).

**Table 14.1:** Classes of damage in DNA, and how they arise

Type of damage	Definition	Origin
<b>Illegitimate base</b> (illegitimate bases may be normal bases which have been damaged or inappropriate bases in DNA such as uracil)	A base other than A, C, G or T found in DNA	Spontaneous deamination Base damage by physical or chemical agents Uracil incorporated during replication Base analogs incorporated during replication Addition of bulky chemical adducts to bases
<b>AP site</b> (apurinic or apyrimidinic site)	An <b>abasic site</b> (a site where the base has been removed from the DNA backbone)	Spontaneous base loss induced hydrolysis AP sites generated by the enzymatic removal of illegitimate bases
<b>Gap</b>	A site where one or more nucleotide residues is missing from one strand of a duplex	Gaps resulting from exonuclease activity (often repair associated) Gaps resulting from incomplete replication (including excision of primers, and replicative bypass)
<b>Nick</b>	A lesion in the backbone of one strand of a duplex, without loss of bases. Nicks may be characterized by further damage to DNA ends which prevent simple religation	Physical/chemical damage to the DNA backbone $\beta$ -elimination of AP sites Nicks remaining following repair synthesis Endonuclease activity associated with many cellular processes
<b>Break (double-strand break)</b>	A lesion in both strands of a duplex, such that the molecule is cleaved	Physical/chemical damage to the DNA backbone Endonuclease activity associated with certain cellular processes
<b>Cross-link</b>	A lesion where DNA strands become covalently joined. Such a link may involve two parts of the same strand ( <b>intrastrand cross-link</b> ) or two separate strands, often the two strands of a duplex ( <b>interstrand cross-link</b> )	UV-induced base dimerization Chemically induced cross-linking

**Table 14.2:** The consequences of deamination

Legitimate base	Complementary partner	Deamination product	Complementary partner	Consequence
Adenine	Thymine	Hypoxanthine	Cytosine	A–G transition
Cytosine	Guanine	Uracil	Adenine	G–A transition
Guanine	Cytosine	Xanthine	Unstable	Replication block
Thymine	Adenine	<i>None</i>		
5-Methylcytosine	Guanine	Thymine	Adenine	G–A transition

The base pairing specificities of the four DNA bases (plus 5-methylcytosine) and their deamination products are shown.



**Table 14.3:** Classification of genotoxic agents

Genotoxic agent	Definition
Carcinogen	An agent promoting neoplastic transformation of eukaryotic cells. Many carcinogens are also mutagens, but the converse is not necessarily true
Clastogen	An agent inducing chromosome fragmentation, i.e. double-strand breaks
Mutagen	An agent promoting mutagenesis either by directly interacting with DNA or by generating metabolic products which do so
Oncogen	An agent inducing tumor formation
Supermutagen	An extremely potent mutagen, such as ethylmethane sulfonate
Telogen	An agent which causes premature termination of replication, e.g. 2',3'-dideoxynucleotides
Teratogen	An agent inducing developmental abnormalities. Not necessarily a genotoxic agent, i.e. includes molecules which mediate their effects without altering DNA structure

### 14.3 DNA repair

**Repair mechanisms.** If all DNA damage were left unrepaired, cells would quickly die due to the accumulation of lethal mutations and the inhibition of essential processes relying on the integrity of DNA (i.e. replication and transcription). Cells have evolved numerous mechanisms to deal with DNA damage, and these can be divided into three groups:

- (1) direct reversal repair mechanisms;
- (2) damage excision and repair using complementary sequence;
- (3) inducible damage tolerance.

Despite these safeguards, cells nevertheless fail eventually (for example, as seen in aging). However, deterioration is significantly slowed by damage avoidance and repair mechanisms.

The types of lesion which can be directly reversed are limited — the resources required to synthesize enzymes capable of the specific reversal of every possible lesion would be vast. The major **DNA repair** process in all living organisms is therefore **excision repair**, where the damaged DNA is removed and new DNA synthesized over the resulting gap using the undamaged strand as a template. This emphasizes an important advantage of the double-stranded DNA genome over the single-stranded genomes of many viruses: if one strand becomes damaged, the other strand can always be used to recover the missing information. Direct reversal and excision are therefore **accurate, nonmutagenic or error-free repair** systems.

In situations where (a) both strands of a duplex are damaged, (b) damage occurs on a single-stranded region of DNA (such as the template strand during replication), or (c) where the excision repair system is saturated, the cell can tolerate noninstructional lesions rather than suffer the lethal effects of replication blocks. Under these circumstances, the replication fork may bypass a lesion leaving a gap which may be filled by recombination with a homologous duplex, leaving the original lesion *in situ*. Alternatively, an inducible system of damage tolerance may come into effect whereby the DNA polymerase synthesizes over the lesion with reduced fidelity. This **error-prone or mutagenic repair** is responsible for many UV-induced mutations in *E. coli*.

DNA repair mechanisms are essential for survival and integrity, and in animals the repair of DNA is particularly important because mutations which affect cell growth cause cancer (see Oncogenes). A number of human diseases resulting from the loss of DNA repair function or of a cellular response to DNA damage are listed in Table 14.5.

### 14.4 Direct reversal repair

**Direct reversal of DNA damage.** There are three classes of lesion subject to **direct reversal repair**: (1) simple nicks can be directly religated; (2) certain UV photoproducts can be repaired by

**Table 14.4:** Different classes of mutagen

Mutagens	Description
<i>Chemical mutagens</i>	
Deaminating agents	Molecules inducing base deamination, including nitrous acid which is nonspecific and sodium bisulfite, which specifically deaminates cytosine
Alkylating agents	Molecules reacting with nucleophilic centres in nucleic acids and substituting them with alkane groups and their derivatives. Such chemicals are often potent mutagens; examples include ethylmethane sulfonate (EMS) and dimethylnitrosamine. Alkylating agents cause various types of DNA lesion — base modification may change the base pairing potential of the alkylated base causing a misincorporation during replication; alternatively a bulky adduct may block replication. Bifunctional agents (those which can alkylate two nucleophilic centres) may form cross-links
Donors of bulky addition products	Molecules which, when metabolized, form reactive species which add bulky chemical groups to bases. Examples include aflatoxin B1 and benzopyrene. The consequence of such a reaction is a bulging distorted helix which blocks replication
Base analogs	Molecules resembling bases which form nucleotides that can be incorporated into a growing nucleotide chain (e.g. bromodeoxyuracil). Unlike the major bases, base analogs may demonstrate aberrant base pairing properties (e.g. they may pair with more than one base because they are stable in more than one tautomeric form). Base analogs generate point mutations, often in a highly specific manner
Intercalating agents	Molecules with a planar component which fits in between the bases of DNA (e.g. ethidium bromide, acridine orange). Intercalating agents increase the <i>length</i> of the DNA strand by unwinding it. This induces frameshifts, blocks replication and inhibits <i>nucleotide excision repair</i> (q.v.) by sequestering the enzymes into inactive complexes
Cross-linking agents	Molecules facilitating the covalent attachment of DNA strands. Such agents include bifunctional alkylating agents, nitrogen and sulfur mustards and platinum derivatives. Planar molecules, <b>psoralens</b> , also cross-link DNA bases when exposed to UV light
<i>Physical mutagens</i>	
Ionizing radiation	Ionizing radiation causes a variety of DNA lesions including damaged bases, damaged sugar rings, nicks and breaks. The effects of ionizing radiation may be direct (caused by ionization of atoms within the DNA molecule) or indirect (caused by the generation of other reactive molecules in the cell, predominantly <b>reactive oxygen species</b> , which then interact with DNA)
UV radiation	UV-induced lesions are termed <b>photoproducts</b> . The predominant effect of UV irradiation is the generation of <b>photodimers</b> involving adjacent bases. The most common photodimer is the <b>cyclobutyl pyrimidine dimer</b> , occurring between any two adjacent pyrimidines, T=T being most frequent, followed by C=T, T=C and C=C. Another frequently observed photodimer is the <b>(6-4) lesion</b> , and photodimers involving purines are also formed. UV radiation may also induce damage to single bases, often by hydrating them. A unique <b>spore photoproduct</b> is generated by UV irradiation of <i>B. subtilis</i> spores
<i>Biological mutagens</i>	
Restriction enzymes	Enzymes synthesized by bacteria which introduce breaks into 'invading' DNA such as phage genomes (q.v. <i>restriction and modification systems</i> , <i>restriction endonuclease</i> )
Mobile genetic elements	DNA sequences which can move around in the genome. Examples include viruses (e.g. bacteriophage Mu, retroviruses), episomal plasmids (e.g. F factor) and transposable genetic elements (e.g. P-elements, Ty-elements). These may be mutagenic due to their insertion and interruption of genes, or they may carry genes/regulatory elements affecting endogenous gene expression. Dispersed copies may facilitate illegitimate recombination

**Table 14.5:** Human inherited diseases associated with deficiencies in DNA repair

Disease	Clinical phenotype	Genetic basis
Ataxia telangiectasia	Facial blood vessel dilation, neuromuscular degeneration, immunodeficiency	Sensitivity to ionizing radiation suggests DNA repair defect, and gene has been mapped to chromosome 11q in humans, although its function has not been determined
Bloom's syndrome	Photosensitivity, immunodeficiency, growth retardation, predisposition to cancer	High frequency of sister chromatid exchange, hypersensitivity to mutagens, delay in joining Okazaki fragments. DNA ligase deficiency
Cockayne's syndrome (CS)	Severe photosensitivity, neurological disorders, dwarfism, optic defects, disproportionately sized facial features and limbs	Similar to XP although dermatosis and skin cancers are rare. Three complementation groups and three genes identified, one is <i>XPB</i> , others specific to CS
Fanconi's anemia	Congenital skin and skeletal abnormalities, loss of blood cells	Sensitivity to cross-linking agents suggests DNA repair role. Disease characterized by multiple chromosome mutations reflecting chromatid breaks. Four complementation groups, and a gene <i>FACC</i> have been identified
Hereditary nonpolyposis colorectal cancer (HNPCC)	Colorectal tumors	Mutations in genes involved in mismatch repair
Trichothiodystrophy (TTD)	Brittle hair, scaly skin, growth defects, mental retardation. Photosensitivity in about 50% of cases	Many TTD cells cannot be complemented by XP-D cells suggesting that TTD may involve the <i>XPD</i> gene. Suggested that different mutations affect the <i>XPD</i> protein in different ways — some its role in repair, others its role in basal transcription (q.v. <i>TFIIIF</i> )
Xeroderma pigmentosum (XP)	Severe photosensitivity, skin cancers, ocular defects, neurological disorders	Seven complementation groups named XP-A to XP-G and XP-V. Human genes <i>XPA</i> , <i>XPB</i> , <i>XPC</i> , <i>XPD</i> , <i>XPF</i> and <i>XPG</i> have been cloned. Primary defect appears to be in nucleotide excision repair (see Table 14.6)

photoreactivation; (3) certain alkylated bases can be repaired by removal of adducts. In addition, there are enzymes which repair damaged nucleotides before they are inserted into DNA (e.g. MutT in *E. coli*).

**Direct repair of nicks.** Nicks generated by DNA damaging agents such as ionizing radiation can be directly repaired by DNA ligase if the 3' hydroxyl and 5' phosphate groups are intact. Frequently, however, such damage involves the modification of these end groups, and further processing is required before ligation.

**Photoreactivation.** Photoreactivation or photorestitution (PR) is a light-dependant DNA repair mechanism in which certain types of pyrimidine dimer are cleaved (monomerized). This repair pathway is found in many bacteria, e.g. *Escherichia coli* and *Salmonella typhimurium* but not in *Bacillus subtilis*, where the major product of UV irradiation, the spore photoproduct, is thought to be monomerized by an analogous but nonlight-dependent pathway. In *E. coli*, the enzyme responsible for photoreactivation is variously termed **DNA photolyase**, **deoxyribodipyrimidine photolyase** or **photoreactivating enzyme**, and is encoded by the *phr* gene. The photolyase binds to DNA containing a pyrimidine dimer and, when exposed to light with a wavelength between 300 and 500 nm, converts the dimer into pyrimidine monomers.

Photolyases with similar properties are found in many but not all lower eukaryotes, but there is little evidence to support their widespread existence in higher eukaryotes (although a light-dependant enzyme which can repair (6-4) photoproducts has been described in *Drosophila*).

Photoreactivation should not be confused with other, nonenzymatic mechanisms of monomerization: **direct photoreversal** occurs during continued UV-irradiation of DNA and reflects the establishment of an equilibrium between the monomer and dimer states of adjacent pyrimidines; **sensitized photoreversal** occurs in the presence of tryptophan, which donates an electron to facilitate monomerization. Furthermore, UV irradiation may perturb other cellular components, inhibiting growth and allowing time for the excision repair of photodimers, a phenomenon termed **indirect photoreactivation**.

**Direct repair of alkylated bases.** In *E. coli*,  $O^6$ -alkylated guanine and  $O^4$ -alkylated thymine residues, as well as methylphosphotriesterases (methylated phosphodiester bonds), are repaired by the **Ada enzyme** (known as  $O^6$ -methylguanine DNA methyltransferase I,  $O^6$ -MGT I before its broader substrate specificity was discovered). Ada is bifunctional: one activity transferring an alkyl group from the *keto*-group of the base, another transferring a methyl group from the methyl phosphotriester, directly reversing the effect of the methylating agent. Ada is known as a **suicide enzyme** because it transfers the alkyl groups onto itself and becomes inactive in the process. A second alkyltransferase,  $O^6$ -MGT II, does not catalyze the transfer of methyl groups from phosphotriesters. Similar enzymes are found in other bacteria and in eukaryotes, although their specificity for the various substrates of the Ada protein varies.

**The adaptive response.** In *E. coli*, there is a repair system stimulated by low concentrations of alkylating agents in the environment. This **adaptive response** is SOS-independent (q.v. *SOS mutagenesis*) and involves the induction of the *ada*, *aidB*, *alkA* and *alkB* genes. The *alkA* gene encodes 3-methyladenine DNA glycosylase, which carries out *base excision repair* (see next section), recognizing a broad spectrum of alkylated bases. The *ada* gene encodes  $O^6$ -MGT I, the Ada protein discussed above, which effects the direct reversal of alkylation damage to DNA by transferring alkyl groups from bases and methylphosphotriester bonds to the protein itself. The methyl groups removed from phosphotriester bonds are transferred onto a C-terminal cysteine residue. This not only inactivates the protein, but also converts it into a strong positive regulator of its own gene and of *aidB*, *alkA* and *alkB*. Inactivation of the protein caused by its repair activity thus stimulates its replenishment by inducing gene expression. Whereas alkylation of DNA bases occurs frequently, the generation of phosphotriester bonds by methylation occurs only when the cell is exposed to dangerous levels of alkylating agents, and a more concerted effort is required to remove them. The adaptive response may be terminated by proteolytic cleavage of the methylated inactive Ada protein, since certain cleavage products can act as inhibitors of adaptive response gene expression. The functions of *aidB* and *alk2* are not known. Both *ada* and *ogt* (which encodes the  $O^6$ -MGT II) are transcribed constitutively at a low level, but only the Ada protein is a genetic regulator, and consequently only the *ada* gene is inducible by alkylation in the cell.



## 14.5 Excision repair

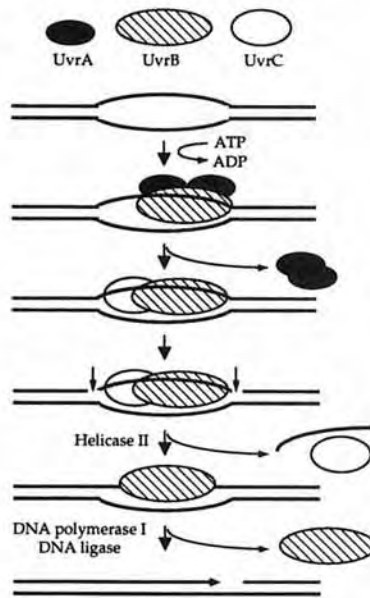
**Base excision repair.** Specific types of damaged or inappropriate base can be removed from DNA by enzymes termed **DNA glycosylases** (Table 14.6) in a process termed **base excision repair (BER)**. The base is removed by hydrolysis of the *N*-glycosidic bond (Figure 16.2) which attaches it to the sugar ring in the DNA backbone. The damaged moiety is excised as a **free base**, generating a second type of DNA damage, an *AP* site (q.v.).

DNA glycosylases have the following properties:

- (1) They recognize a single, specific type of damaged or inappropriate base in DNA or a small group of related chemical adducts.
- (2) They remove the base by hydrolysis of the *N*-glycosidic bond. Certain glycosylases may also introduce a nick 3' to the site of the damage through associated **AP lyase activity**, although this activity is not required for subsequent stages of repair and its physiological significance is unclear.
- (3) Almost all DNA glycosylases act upon single bases and have no specificity for larger, more complex lesions involving multiple bases. However, a DNA glycosylase specific for pyrimidine dimers is found in bacteriophage T4 and in *Micrococcus luteus*.
- (4) Most DNA glycosylases recognize damaged or inappropriate DNA bases, but some which recognize legitimate bases function during *mismatch repair* (q.v.).

**Table 14.6:** DNA glycosylases and their substrate specificities, together with the *E. coli* genes which encode them and functionally homologous eukaryote genes where known. Each enzyme is named according to the convention 'substrate-DNA glycosylase (DG)'. The substrate is usually abbreviated (e.g. uracil = Ura)

Enzyme	Species	Gene	Specificity	Comments
Ura-DG	<i>E. coli</i> <i>S. cerevisiae</i> Humans	<i>ung</i> <i>UNG</i> <i>UDG</i>	Uracil, 5-hydroxyU	
3-meA-DG	<i>E. coli</i>	<i>tag</i> <i>alkA</i>	3-meA 3-meA, 7-meG, 3-meG, O <sup>2</sup> -meC O <sup>2</sup> -meT, 5-foU 5-hydroxyU hypoxanthine	<i>alkA</i> part of <i>adaptive</i> <i>response</i> (q.v.) to sublethal doses of alkylation agents
	<i>S. cerevisiae</i>	<i>MAG</i>	3-meA, 7-meG, hypoxanthine	
fapy/8-oxoG-DG	Humans <i>E. coli</i> <i>S. cerevisiae</i> Humans	<i>MPG</i> <i>fpg/mutM</i> <i>OGG1</i> <i>OGG1</i>	3-meA, 8-oxoG Ring-opened purines (e.g. fapy, 8-oxoG)	Repairs oxidative damage to purines
Endonuclease III (thymine glycol-DG)	<i>E. coli</i>	<i>nth</i>	Oxidized pyrimidine derivatives	Repairs oxidative damage to pyrimidines
MutY	<i>E. coli</i>	<i>mutY</i>	Adenine in A:G and A:8-oxo-G mispairs	Mismatch repair
Endonuclease VIII	<i>E. coli</i>	<i>nei</i>	Oxidized pyrimidine derivatives	
Exonuclease IX UV-endonuclease	T4 <i>Micrococcus</i> <i>luteus</i>	<i>denV</i> ?	Thymine dimers	
GT mismatch	Humans		Thymine in GT mismatches	Mismatch repair



**Figure 14.2:** Nucleotide excision repair in *E. coli*. UvrA loads UvrB onto damaged DNA and then dissociates; there is hydrolysis of ATP during this process. UvrC then binds to the complex and nicks are introduced 5' and 3' to the lesion (arrows). UvrD helicase then removes the oligonucleotide fragment and UvrC, leaving UvrB bridging the gap. The gap is repaired by DNA polymerase I and DNA ligase; UvrB is expelled.

**Completion of repair following base excision.** After excision of the base by the DNA glycosylase, an AP (apurinic/apyrimidinic) site is generated. This is recognized by a second class of repair-specific enzyme termed an **AP endonuclease** which introduces a nick 5' to the AP site. In *E. coli* there are two AP endonucleases:

(1) **Exonuclease III.** Originally characterized as a 3' to 5' exonuclease with associated phosphatase activity (hence its name), the major physiological role of this enzyme is its 5' AP endonuclease activity. Expressed constitutively, this is the major AP endonuclease and is encoded by the *xth* gene. There are functional homologs in other species: *ExoA* (*S. pneumoniae*), *APE/HAP* (human), *APEX* (murine), *BAP* (bovine) and *Rrp1* (*Drosophila*).

(2) **Endonuclease IV.** Normally constituting only 10% of AP endonuclease activity, this enzyme is inducible by superoxide radicals reflecting the importance of oxidative damage repair. Encoded by the *nfo* gene, a functional homolog, *APN1*, is found in *S. cerevisiae*.

Following the 5' incision to the AP site, a further enzyme, **deoxyribophosphodiesterase (dRpase)**, is required to hydrolyze the 5' residue, resulting in a single nucleotide gap. DNA polymerase may initiate repair synthesis from this gap, replacing the missing nucleotide residue with a repair patch of one nucleotide. Alternatively, DNA 3' to the damage site is excised, followed by more extensive replacement synthesis. Finally the nick remaining after repair synthesis is closed by *DNA ligase* (q.v.).

**Nucleotide excision repair.** Both direct reversal repair and BER involve enzymes recognizing specific DNA lesions. **Nucleotide excision repair (NER)** is a system which recognizes a diverse spectrum of DNA damage, including cross-links, bulky adducts and lesions involving multiple bases. These are removed by a single, multipotent enzyme complex generating a gap which is repaired by DNA polymerase and DNA ligase. It is not clear how NER enzymes recognize the different lesions: some, but certainly not all, cause helix distortion, yet natural perturbations such as

kinks, hairpins and mismatches make poor substrates. The damaged DNA is released by hydrolysis of phosphodiester bonds either as an intact nucleotide or, more usually, as an oligonucleotide fragment. The repair of pyrimidine dimers in DNA occurs in both light and darkness: in light, dimers are repaired by *photoreactivation* (q.v.), hence **light repair**, whereas **dark repair**, which occurs by a different mechanism, is nucleotide excision repair.

**Nucleotide excision repair in *E. coli*.** There are three major NER genes in *E. coli*: *uvrA*, *uvrB* and *uvrC*, whose products combine to form an ATP-dependent endonuclease (variously termed the **UvrABC endonuclease** or **excinuclease**, the latter because it is involved in excision; Figure 14.2). Two molecules of UvrA bind to one of UvrB and the complex binds to the damaged DNA (the function of UvrA may be to load the UvrB protein). The UvrA molecules then dissociate (ATP hydrolysis is required), and UvrC binds causing a conformational change allowing UvrB to generate a nick 3' to the lesion. The UvrC protein then generates the 5' nick. The binding of UvrC and the subsequent **bimodal incisions** require ATP binding but not hydrolysis. The 5' incision is usually 8 nt upstream of the lesion, whereas the position of the 3' incision is more variable depending on the nature of the lesion. Both 5' and 3' positions may be affected by sequence context. The resulting oligonucleotide fragment is excised by DNA helicase II encoded by the *uvrD* gene. This also removes the UvrC protein, leaving UvrB bridging a gapped duplex, perhaps protecting the single-stranded region from further damage. DNA polymerase I binds to the exposed 3' hydroxyl group and synthesizes over the gap, removing UvrB in the process, typically generating a repair patch of about 12 nt (this type of repair is sometimes referred to as **short patch NER**). The final nicks are sealed by DNA ligase.

The UvrABC endonuclease is expressed constitutively at a low level, but the *uvrA* and *uvrB* genes are also SOS inducible (q.v. *SOS response*). A second nucleotide excision repair process, which also appears to involve the *uvr* genes and is also SOS inducible, results in the excision and replacement of up to 2 kb of DNA. The basis of this so called **long patch NER** is less clearly understood.

Nucleotide excision repair preferentially targets the transcribed strand of active genes. As transcription is arrested by DNA lesions, the interruption of transcription may induce excision repair. A protein termed **transcription repair coupling factor** (TRCF) encoded by the *mfd* gene is able to bind to stalled RNA polymerase–DNA lesion–RNA complexes and displace the transcription machinery. TRCF also binds to UvrA, indicating that once bound to damaged DNA, it might specifically recruit UvrA–UvrB complexes to the template.

**Nucleotide excision repair in eukaryotes.** Nucleotide excision repair in eukaryotes involves a large number of genes. In yeast, these have been assigned to the *RAD3* epistasis group (q.v. *epistasis*) because most have been identified in genetic screens for radiation-sensitivity. The human homologs of yeast *RAD* genes were initially named *ERCC* (excision repair complementation competent) genes, but many have been attributed to specific disease phenotypes such as xeroderma pigmentosum and have been renamed (Table 14.7). Other NER genes have been identified because they are also essential for transcription, demonstrating the close link between transcription and repair in eukaryotic cells. The basis of this link is the general transcription factor *TFIIH*, which forms an essential part of both the basal transcriptional apparatus of the cell, and the complex of repair proteins, the **repaosome**. In the transcription complex, the *TFIIH* core is associated with a cyclin-dependent kinase complex (CDK7/cyclin H and the assembly factor MAT1 in mammals; q.v. *CAK*), whereas in repair, it is associated with other repair proteins. Hence, repair is preferentially directed to the transcribed strand of actively transcribed genes (**transcription coupled repair**) and is more active when transcription is in process (**transcription dependent repair**). The mechanism of nucleotide excision repair appears to be quite similar in eukaryotes and bacteria, with bimodal incision followed by excision and resynthesis. The repair patch following nucleotide excision repair in humans is slightly longer than that in *E. coli* (30 nt) and is confusingly termed **long patch repair**, to distinguish

**Table 14.7:** Eukaryotic genes involved in nucleotide excision repair

<i>S. cerevisiae</i> ( <i>S. pombe</i> ) gene	Human gene	Function
<i>RAD1</i> ( <i>rad16</i> )	<i>XPF</i>	Rad1 and Rad10 combine to form endonuclease with duplex-3' single-strand junction specificity
<i>RAD10</i> ( <i>swi10</i> )	<i>ERCC1</i>	Endonuclease with duplex 5'-single-strand junction specificity
<i>RAD2</i> ( <i>rad13</i> )	<i>XPG</i>	
<i>RAD4</i>	<i>XPC</i>	Unknown, may help to convert transcription complex into repairosome
<i>RAD14</i>	<i>XPA</i>	Damage-preferred DNA-binding protein
<i>RAD3</i> ( <i>rad15</i> )	<i>XPB</i>	Component of TFIIH with 5'→3' helicase activity
<i>SSL2/RAD25</i> ( <i>ERCC3sp</i> )	<i>XPB</i>	Component of TFIIH with 3'→5' helicase activity
<i>SSL1</i>	<i>P44</i>	Components of basal transcription factor TFIIH
<i>TFB1</i>	<i>P62</i>	
<i>TFB2</i>	<i>P52</i>	
<i>TFB3</i>	<i>MAT1<sup>a</sup></i>	
<i>TFB4</i>	<i>P34</i>	Perturb chromatin structure
<i>RAD7</i>		
<i>RAD16</i>		
<i>RAD23</i>		
<i>PSO2, PSO3</i>		Repair of psoralen-induced damage

<sup>a</sup>In yeast, core TFIIH has seven components, whilst the human core TFIIH has six. MAT1, a human cyclin-dependent kinase (CDK) assembly factor, is not found in the core during repair but is part of the CDK complex. The yeast homolog TFB3 (p38) remains in the core during repair

it from the 1–2 nucleotide short patch repair which occurs in base excision repair. There is no evidence for a >1500 bp long patch repair in eukaryotes such as that found in *E. coli*.

## 14.6 Mismatch repair

**Base mismatches and mismatch repair.** Mismatch repair is an excision repair process which corrects base pair mismatches in duplex DNA. A **mismatch** is a nonWatson–Crick base pair, although mismatch repair systems will also correct small (<4 nucleotide residue) insertions and deletions and hence contribute to frameshift fidelity as well as substitution fidelity. Mismatch repair differs from other forms of DNA repair in that the excised material is a legitimate (undamaged) base normally found in DNA, and therefore there must be a mechanism to discriminate between the correct and incorrect strands so that the mutagenic nucleotide and not the correct nucleotide on the template strand is removed.

Mismatches can arise in a number of ways: (1) by misincorporation during replication; (2) by the formation of heteroduplex DNA during recombination; (3) by the deamination of 5-methylcytosine, to generate thymine (see DNA Methylation and Epigenetic Regulation). In addition, short insertions or deletions can arise by: (1) strand slipping during DNA replication, and (2) the formation of a heteroduplex with unequal numbers of residues

Mismatch repair does not affect recombination between very similar sequences, but reduces recombination between significantly diverged sequences because overlapping excision tracts are



generated. The repair system therefore acts as a species barrier between organisms that are significantly diverged, but which are not reproductively isolated.

**Long patch mismatch repair in *E. coli*.** Mismatches incorporated during replication are excised specifically from the nascent strand, allowing resynthesis of the correct sequence over the resulting gap. This is **post-replicative mismatch repair**, and increases substitution and frameshift fidelity by several orders of magnitude. The repair is directed to the daughter strand because it is transiently undermethylated. Adenine residues in the sequence GATC are methylated at the N<sup>6</sup> position by DNA adenine methylase (*Dam methylase*, q.v.), but following replication, there is a lag before the enzyme completes methylation of the newly synthesized daughter strand (see DNA Methylation). During this window, it is possible for the cell to discriminate between the parental and daughter strands and therefore direct repair to the daughter strand. If a heteroduplex arises by another mechanism, e.g. by recombination, repair is still targeted to the undermethylated strand. If both strands are unmethylated, repair occurs but is unbiased in its direction, and if both strands are methylated, repair is also unbiased, but occurs at a much lower efficiency.

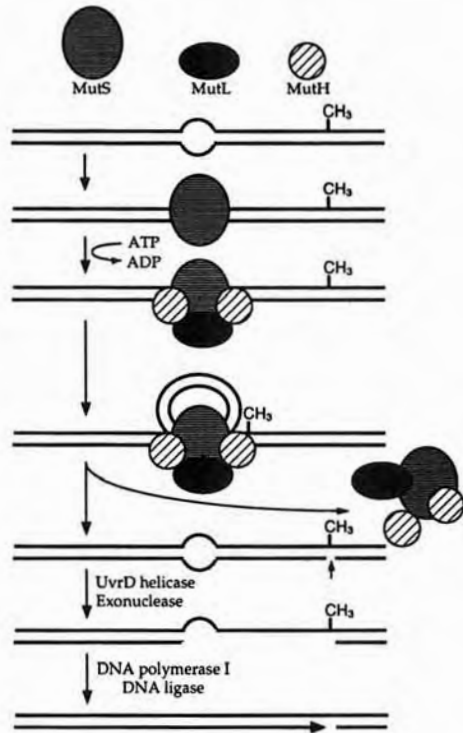
Four genes (*mutH*, *mutL*, *mutS* and *uvrD*) are absolutely required for long patch mismatch repair. They are *mutator genes* (q.v.) because mutations which affect their function alter the fidelity of DNA replication and thus alter the spontaneous mutation rate. MutS protein recognizes mismatched DNA with an efficiency depending upon mismatch type and sequence context (G:T and A:C > G:G and A:A > T:T, C:T and G:A > C:C). The unmethylated GATC sequence acts as a substrate for MutH, which is an endonuclease. The nicking of the daughter strand at this position, immediately 5' to the guanosine residue, allows the excision of a patch which includes the mismatched nucleotide. MutS, MutL, ATP and divalent cations are required in addition to MutH for endonuclease activity, and UvrD is the DNA helicase required for excision.

Because MutH can function only in the presence of a mismatch, it is likely that the repair proteins, GATC site and mismatch must physically interact. This model is supported by demonstrating that, regardless of whether the mismatch lies 5' or 3' to the nearest GATC site, the repair excision takes the shortest possible route, starting at the GATC motif and ending about 100 nt beyond the mismatch. As UvrD is a 5' to 3' helicase; this requires the enzyme to be loaded onto the methylated strand for 5' mismatches and onto the unmethylated strand for 3' mismatches, a mechanism which would certainly require interaction between the two sites. Exonucleases with different directionalities are also required, according to the site of the mismatch. A model for mismatch repair is shown in Figure 14.3.

**Long patch mismatch repair in other organisms.** Systems homologous to *E. coli* MutHLS mismatch repair are found in other bacteria and eukaryotes, although in many bacteria and all eukaryotes repair is not methylation-dependent. Strand discrimination may involve the recognition of nicks introduced during replication prior to their repair by DNA ligase (such nicks would occur naturally in the lagging strand and be introduced deliberately into the leading strand; see Replication). Consequently, eukaryotic homologs of the *E. coli* *mutS* and *mutL* (but not *mutH*) have been isolated, as shown in Table 14.8. In yeast and human cells, there are at least two homologs of each *E. coli* gene, whose products form heterodimers. The eukaryotic genes *MLH1* and *MSH2* are thought to be required predominantly for the repair of insertion/deletion mismatches because mutations in either lead to minisatellite instability, which in humans is associated with **hereditary nonpolyposis colorectal cancer (HNPCC)**.

**Short patch mismatch repair.** Both prokaryotes and eukaryotes have mismatch repair systems characterized by short excision patches, typically less than 10 nt in length. In *E. coli* there are two systems which are independent of *mutHLS* mismatch repair.

(1) **MutY-dependent repair** replaces adenosine residues in A:G and A:C mismatches. MutY encodes a DNA glycosylase (q.v. *base excision repair*) whose primary role is to remove adenine residues opposite 8-oxo-7, 8-dihydrodeoxyguanine residues, but which can also act upon mismatches.



**Figure 14.3:** A model for mismatch repair in *E. coli*. MutS binds to a mismatch and recruits MutH and MutL, a process which is dependent on ATP hydrolysis. DNA is spooled past the repair complex *from both sides*, creating a loop, until the first GATC site is encountered. Here MutH introduces a nick. UvrD helicase unwinds the DNA from the GATC site and ejects the repair proteins; the DNA is degraded by an exonuclease. Finally the gap is repaired by DNA polymerase and ligase. UvrD is a 3'–5' helicase, so for the repair shown above it must be loaded onto the unmethylated strand and a 3'–5' exonuclease such as exonuclease I must be used. If the nearest GATC site was on the other side of the mismatch, the helicase would have to be loaded onto the methylated strand and a 5'–3' exonuclease, such as exonuclease VII, would be required.

**Table 14.8:** Eukaryotic homologs of *E. coli* long patch mismatch repair genes

<i>E. coli</i> gene	<i>S. cerevisiae</i> gene	Mammalian gene	Comments
<i>mutL</i>	<i>MLH1</i>	<i>MLH1</i>	HNPCC-related
	<i>PMS1</i>	<i>PMS2</i>	
<i>mutS</i>		<i>PMS1</i>	HNPCC-related
	<i>MSH2</i>	<i>MSH2</i>	
	<i>MSH6</i>	<i>GTBP</i>	
		<i>MSH1</i>	Mitochondrial
	<i>MSH2</i>	<i>Duc-1, REP3</i>	
	<i>MSH3</i>	<i>MSH3</i>	
	<i>MSH4</i>		

(2) **Very short patch (VSP) mismatch repair** corrects T residues in G:T mismatches. These occur within the Dcm methylase target sequence CC(A/T)GG (*see* DNA Methylation); the internal cytosine residue is modified to 5-methylcytosine and may be converted to thymine by deamination. VSP repair requires *mutS* and *mutL* but not *mutH* or *uvrD* gene products. It also requires *vsr*, which encodes an endonuclease specific for GT mismatches in the sequence CT(A/T)GG.

In mammals a short patch mismatch repair system has been identified which corrects T residues in G:T mismatches, preferentially in CpG motifs. This is also a DNA glycosylase, which acts like *E. coli* VSP to prevent mutation resulting from the deamination of 5-methylcytosine (see DNA Methylation).

## 14.7 Recombination repair

**Evidence for recombination repair.** Recombination repair is any DNA repair or damage tolerance strategy involving *homologous recombination* (q.v.). The existence of such repair mechanisms is indicated by differences in sensitivity to ionizing radiation demonstrated by haploid yeast cells arrested in G<sub>1</sub> and G<sub>2</sub> (and similarly for stationary phase and log phase bacterial cells); in each case, the cell with more copies of its genome is more resistant. More direct evidence for recombination repair comes from studies of recombination deficient mutants of *E. coli* and *S. cerevisiae*, which are also radiation-sensitive.

**Mechanisms of recombination repair.** There are two substrates for recombination repair: double-strand breaks and single-strand gaps. For the mechanisms involved in homology-dependent repair see Figure 25.3 in Recombination. Double-stranded breaks may be repaired in two ways, either by a cross-over repair (**double-strand break repair**) or a noncross-over repair (**synthesis-dependent strand-annealing repair**), the latter occurring in specialized systems such as P-element transposition and the switching of mating types in yeast. Single-stranded gaps can arise when replication is blocked at a noninstructional lesion and reinitiates downstream, leaving a gap opposite the damaged template. The gap is filled with the homologous region of a sister duplex (**daughter strand gap repair** (DSGR) or **postreplicative repair** (PRR)), and repair synthesis of the gapped daughter strand occurs using the free, complementary strand of the sister duplex as a template. Note that this process does not actually repair the original lesion on the template strand. As the name implies, it is the daughter strand which is repaired; the original lesion remains in the genome unless it is removed by excision. DSGR is a major damage tolerance mechanism, which in *E. coli* is induced as part of the *SOS response* (q.v.).

## 14.8 The SOS response and mutagenic repair

**Control of the SOS response.** The *E. coli* *SOS response* is an inducible response to DNA damage resulting in increased capacity for DNA repair, inhibition of cell division and altered metabolism. The response is mediated by the activation of about 20 *SOS genes* (Table 14.9) normally expressed at low levels due to transcriptional repression by the *LexA repressor*, which binds to operator sequences (**SOS boxes**) upstream of each gene. LexA has differing affinities for different operator sites; thus the *SOS genes* are induced at different response thresholds leading to the **split phenotypes** observed when only part of the *SOS response* becomes activated. There is a low affinity *SOS box* upstream of the *lexA* gene itself, so the repressor inhibits its own synthesis, but the system is leaky enough to facilitate repression of the other *SOS genes*. When DNA becomes extensively damaged, single-stranded regions are exposed. Single-stranded DNA interacts with RecA to produce an activated complex termed **RecA\***, which facilitates cleavage of the LexA repressor by increasing the rate of LexA autoproteolysis. The cleaved LexA protein can no longer bind to DNA and the *SOS genes* are derepressed. When damaged DNA is no longer present in the cell, RecA becomes inactivated and no longer facilitates LexA cleavage. A high level of uncleaved LexA rapidly accumulates in the cell from the existing *lexA* mRNA pool and the *lexA* gene and other *SOS genes* are shut down. This is a classic negative feedback loop.

**SOS mutagenesis.** The increased capacity for DNA repair accompanying the *SOS response* is due in part to the upregulation of constitutive genes involved in nucleotide excision and recombination-based repair processes (*uvrA* and *uvrB* — but not *uvrC* — have *SOS boxes*, as do *recA* and *recN*).

**Table 14.9:** *E. coli* SOS genes and their functions

SOS gene	Function
<i>dinB</i>	<i>din</i> genes are damage-inducible genes whose functions have yet to be determined
<i>dinD</i>	
<i>dinF</i>	
<i>dinG</i>	
<i>dinH</i>	
<i>dinI</i>	Subunit of DNA polymerase III
<i>dnaN</i>	
<i>lexA</i>	
<i>nrdAB</i>	
<i>polB (dinA)</i>	
<i>recA</i>	Repressor of SOS genes
	DNA metabolism
	DNA polymerase II
	(1) Cleaves LexA to facilitate SOS response
	(2) Necessary for recombination repair
	(3) Cleaves UmuD protein and . . .
	(4) . . . assembles UmuD <sub>2</sub> C complex onto lesion to facilitate error-prone repair
<i>recN</i>	Recombination repair
<i>recQ</i>	
<i>ruvAB</i> operon	
<i>sulA</i>	
<i>umuDC</i> operon	
<i>uvrA</i>	SOS-specific error-prone repair
<i>uvrB</i>	
<i>uvrD</i>	
	Nucleotide excision repair
	Nucleotide excision repair
	DNA helicase, involved in repair

While both these mechanisms are accurate (**error-free repair**), a third system, induced uniquely during the SOS response and involving the *umuC* and *umuD* genes, enhances the repair of DNA but also leads to mutagenesis (**error-prone** or **mutagenic repair**). This **SOS mutagenesis** occurs specifically at premutagenic noninstructional lesions and results from synthesis across the unreadable template with reduced fidelity (**translesion synthesis**). Mutations in the *umuDC* operon render cells non-mutable by UV radiation and other mutagens which generate replication blocks, but more sensitive to killing by these agents, commensurate with the loss of translesion synthesis ability. These mutants are still sensitive to mutagens which generate misinstructional lesions, leading to base mispairing.

The exact mechanism of translesion synthesis is unknown; UmuDC proteins may interact with DNA polymerase III and reduce its selectivity and/or proofreading ability. UmuDC-dependent error-prone repair is strongly repressed under normal circumstances because RecA\* protein is required for three stages of its activation.

(1) RecA\* cleaves LexA repressor to derepress *umuDC* transcription.

(2) RecA\* cleaves UmuD (an inactive precursor) to yield active UmuD' (uncleaved UmuD is not only inactive in error-prone repair, but also inhibits the process by sequestering UmuD' into inactive heterodimers).

(3) RecA\* may help to position the UmuD'UmuC protein complex at the site of the lesion.

**Effects of the SOS response upon bacteriophage infection.** Some viruses have evolved mechanisms to sense damage of the host cell and exploit the RecA system. The activated RecA\* complex induces cleavage, not only of LexA and UmuD, but also of the CI repressor of bacteriophage  $\lambda$  because all three proteins share a consensus autoproteolytic motif. The CI repressor is responsible for maintaining lysogeny, and its inactivation heralds the lytic cycle (see Viruses, Box 30.1). In this way, stimulation of the SOS response through, for example, UV irradiation induces the excision of  $\lambda$  prophage. The level of DNA damage necessary for  $\lambda$  induction is much higher than that required to elicit the SOS response, due to the relatively low susceptibility of CI repressor to RecA\*.



Bacteriophage with damaged genomes generally survive poorly when introduced into a host cell. However, if the host has been UV-irradiated prior to infection, the phage survive. This phenomenon, termed **Weigle reactivation (W-reactivation)**, is accompanied by a higher phage mutation frequency (**Weigle mutagenesis**) and reflects induction of the SOS response, leading to rapid repair of the phage genome, some of which will be UmuDC-dependent error-prone repair.

**Error-prone repair in other species.** SOS-like error-prone repair systems are found in few bacteria other than *E. coli*, and most are therefore resistant to UV-induced mutagenesis, although more susceptible to killing. Genes which are functionally homologous to *umuC* and *umuD* are found on several conjugative plasmids. There is also a limited amount of evidence for mutagenic repair in eukaryotes.

## Reference

Friedberg, E.C., Walker, G.C. and Siede, W. (1995) *DNA Repair and Mutagenesis*. ASM Press, Washington, DC.

## Further reading

- Chu, G. and Mayne, L. (1996) Xeroderma-pigmentosum, Cockayne syndrome and trichothiodystrophy — do the genes explain the diseases? *Trends Genet.* 12: 187–192.
- Friedberg, E.C. (1996) Relationships between DNA repair and transcription. *Annu. Rev. Biochem.* 65: 15–42.
- Lindahl, T., Karran, P. and Wood, R.D. (1997) DNA excision repair pathways. *Curr. Opin. Genet. Dev.* 7: 158–169.
- Modrich, P. and Lahue, R. (1996) Mismatch repair in replication fidelity, genetic recombination and cancer biology. *Annu. Rev. Biochem.* 65: 101–133.
- Sancar, A. (1996) DNA excision repair. *Annu. Rev. Biochem.* 65: 43–81.
- Walker, G.C. (1995) SOS-regulated proteins in translesion DNA synthesis and mutagenesis. *Trends Biochem. Sci.* 20: 416–420.
- Wood, R.D. (1996) DNA repair in eukaryotes. *Annu. Rev. Biochem.* 65: 135–167.

**This Page Intentionally Left Blank**

## Chapter 15

# Mutation and Selection

### Fundamental concepts and definitions

- A **mutation** is a stable, heritable change in genotype caused by an alteration to the nucleotide sequence in a particular region of the genome (c.f. *epimutation*, *paramutation*). A gene, genome, cell or individual carrying a given mutation is a **mutant**.
- Mutations can be localized (i.e. affecting a single nucleotide or a small cluster of nucleotides) or can involve large segments of the genome. In the former category, **gene mutations** occur within a gene and can affect the nature of the gene product or interfere with its expression, whereas **extragenic mutations** usually have no effect (unless they disrupt a regulatory element). Large-scale mutations involve tens to many thousands of nucleotides and affect whole genes or groups of genes. In eukaryotes, the largest mutations are visible at the cytogenetic level and are termed **chromosome mutations** (see Chromosome Mutation).
- Gene mutations convert one allelic form of a gene into another. For many gene loci, there is a **wild-type allele** which predominates in the population because it confers the greatest **fitness** (ability to survive and reproduce). This generally encodes the normal, functional product associated with the gene, and the **wild-type phenotype** reflects this normal gene activity. Other, rare alleles are designated **mutant alleles**, and the quantity and/or structural properties of the encoded product may differ, generating distinct **mutant phenotypes**. Gene mutations away from the wild type are usually deleterious or selectively neutral; few are beneficial.
- Instead of a single wild-type allele, several alleles conferring equal fitness may exist in equilibrium within the population, and the locus is described as **polymorphic**. Polymorphism may also involve alleles of unequal fitness if the population is in a transition state following the appearance of a recent beneficial mutation, or if there is balancing selection for more than one allele. There are different forces which act to maintain or change the frequency of alleles in a population, including mutation pressure, migration, random genetic drift and natural selection. Natural selection eliminates alleles reducing fitness, so there is a clear bias in the spectrum and location of surviving mutations, leading to evolutionary conservation of functionally important DNA sequences. An allele which is deleterious or neutral in one environment may be beneficial in another, or may depend on its frequency in the population. The wild-type allele at a given locus in a population today is likely to have been a rare mutant allele in the evolutionary past.
- Unicellular organisms pass newly arising mutations to all progeny receiving the affected chromosome, but in multicellular organisms, mutations can occur in the germline (gamete-forming tissue) or somatic tissue. Whereas **germinal** or **germline mutations** can be transmitted to progeny, **somatic mutations** will not pass to future generations except where it is possible to clone from somatic cells. Depending on when and where a somatic mutation arises, it may affect a single cell, it may generate a clone of mutant cells in a wild-type background, or, if the mutation occurs early in development, it may generate a *mosaic* (q.v.). Somatic mutations which affect growth control genes may cause cancer (see Oncogenes and Cancer).

### 15.1 Structural and functional consequences of mutation

**Structural categories of mutation.** Mutations fall into four structural categories based on the nature and amount of DNA sequence involved (Table 15.1). The four types of mutation are caused by distinct mechanisms and have different consequences.

**Table 15.1:** Categories and causes of mutations**Point mutation (simple mutation, single-site mutation)**

**Definition.** A mutation at a single site involving a single residue or a small number of residues. **Base substitutions** are point mutations where one base is replaced by another: **transitions** occur when a purine is replaced by another purine, or a pyrimidine by another pyrimidine; **transversions** occur when a purine is replaced by a pyrimidine or *vice versa*. **Small insertions and deletions** (<5 nt) may also be regarded as point mutations.

**Causes of base substitutions** (see below for causes of insertions and deletions):

- (i) DNA polymerase spontaneous errors.
- (ii) Replication over a template containing a single base misinstructional lesion, e.g. a damaged or inappropriate base, a base analogue, or a rare *tautomer* (q.v.).
- (iii) Replication over a template containing a single base noninstructional lesion, e.g. an AP-site (q.v. the *A-rule*).
- (iv) Deamination of 5-methylcytosine to thymine.
- (v) Mismatch repair of heteroduplex DNA.

For further discussion, see Mutagenesis and DNA Repair.

**Effects.** Depends on location — see Table 15.2.

**Complex mutation (multisite mutation)**

**Definition.** A cluster of several discrete point mutations.

**Causes:**

- (i) Mismatch repair of heteroduplex DNA (either allelic gene conversion or nonallelic, involving repetitive DNA).
- (ii) Activity of low-fidelity DNA polymerase (e.g. retroviral reverse transcriptase).
- (iii) Error-prone replication during *SOS response* (q.v.)

**Effects.** As for point mutations — see Table 15.2.

**Macromutation**

**Definition:** A mutation which occurs at a single site but affects many consecutive nucleotide residues. The largest macromutations may be observed at the cytogenetic level. Mutations involving DNA loss are **deletions** or **deficiencies**. Mutations involving DNA gain are **insertions** or **additions** (however, if the inserted material is already found elsewhere in the genome, the mutation may be termed a **duplication**, and if the additional material is already found as tandem repeats, the mutation may be termed an **amplification** or **repeat expansion**). **Substitutions** are mutations where a section of DNA is replaced with an equivalent amount of novel information. **Rearrangements** involve the reorganization of genetic material, but no loss or gain. Rearrangements include **inversions**, where a segment of DNA is reversed in orientation, **translocations**, where DNA is transferred from one position in the genome to another, and **cointegrations** or **fusions**, where two chromosomes are covalently joined, or a linear chromosome is circularized.

**Causes:**

- (i) chromosome breaks and random rejoining;
- (ii) homologous recombination between nonallelic sequences: unequal crossing over or unequal sister chromatid exchange, or recombination between dispersed repeats;
- (iii) site-specific recombination;
- (iv) the activity of mobile genetic elements;
- (v) slipping of the template or primer strand during replication of tandem repeats (deletions and insertions only);
- (vi) stabilization of secondary structures in the template or primer strand during replication of inverted repeats (deletions and insertions only);
- (vii) unbalanced meiosis involving structurally rearranged chromosomes (deletions and insertions only).

For further discussion, see Chromosome Mutation and q.v. *illegitimate recombination*.

**Effects:** Depends on site and size of mutation — see Table 15.3.

**Chromosome imbalance**

**Definition:** A numerical chromosome disorder, either deficiency or addition, caused by loss or gain of whole chromosomes (**aneuploidy**) or sets of chromosomes (**polyploidy**).



**Table 15.1:** continued**Causes:**

- (i) unscheduled or absent DNA replication;
- (ii) unscheduled or absent mitosis/meiosis;
- (iii) failure of chromosome segregation (often due to pre-existing structural abnormality);
- (iv) fertilization involving unbalanced (aneuploid) gametes.

For further discussion, see Chromosome Mutation.

Effects: Multiple gene dosage effects — see Table 15.3.

**Point mutations** occur at a single site and involve a small number of nucleotide residues. If point mutations occur within a gene (see Table 15.2), the consequences depend on any change to the structure or expression of the encoded polypeptide, and range from neutral to severely deleterious. Extragenic point mutations are often neutral, although those occurring in regulatory elements may alter the level or scope of gene expression.

**Complex mutations** are rare. They are clustered groups of point mutations, and usually reflect *gene conversion* (q.v.) events in heteroduplex DNA occurring during homologous recombination, which may or may not be associated with DNA repair. The recombining sequences may be allelic (classical gene conversion) or nonallelic (often involving direct or inverted DNA repeats). The latter may result in *concerted evolution* (q.v.).

**Macromutations** are structural chromosome mutations — large deletions, insertions or rearrangements, the largest visible at the cytogenetic level. They often result in gene disruption or loss, and sometimes cause gene fusion effects, position effects or gene dosage effects (see Table 15.3).

**Chromosome imbalances** are numerical chromosome mutations — loss or gain of entire chromosomes or, occasionally, chromosome sets. Chromosome imbalances cause multiple gene dosage effects and, in mammals, are usually lethal.

**Functional consequences of mutation — general principles.** Many mutations do not cause a change in phenotype because the **mutation site** (the position of the mutation in the genome) does not influence gene function or expression, or higher-order genome function (DNA replication, mitotic segregation, etc.). Such mutations are described as being selectively **neutral** because they do not influence the Darwinian fitness of the individual; they include most extragenic mutations. Most mutations which have a phenotypic consequence fall within genes or the regulatory elements which control them. There are three different target sequences for such mutations: the coding region of the gene, noncoding sequences within the transcription unit, and regulatory sequences outside the transcription unit.

Many point mutations within genes are neutral because they do not alter either the structure or expression of the encoded product (see Table 15.2). Point mutations which do modify the gene product or its expression in some way are usually deleterious or neutral — a few may be beneficial, but this depends on the selective constraints on the structure of the polypeptide and the environment in which the polypeptide functions (q.v. *natural selection*, *molecular clock*). Macromutations occurring within genes or involving genes are generally deleterious because they cause large-scale disruptions (see Table 15.3). The consequences of many different types of mutation are exemplified by the study of hemoglobin disorders (Box 15.1).

Whatever the consequences of a mutation *per se*, whether these effects are expressed at the level of the phenotype depends on several additional factors.

- (1) **Dominance.** In diploids, the mutant allele may be recessive to the wild-type allele and its effects will not manifest in the heterozygote.
- (2) **Genetic background and environment.** The mutant allele may not be penetrant even in the

**Table 15.2:** The consequences of point mutations on polypeptide structure and gene expression

Position and class of mutation	Definition and consequences
<p><i>Site: coding region</i></p> <p>Silent, same-sense or synonymous mutation</p>	<p><b>Definition:</b> A base substitution which does not affect the sense of a codon and thus has no effect on polypeptide structure. Silent mutations are possible because of the <i>degeneracy</i> (q.v.) of the genetic code. Example: ATT → ATC; both codons specify isoleucine.</p> <p><b>Effects:</b> Synonymous mutations are usually neutral. They may generate a phenotype if (i) there is <i>codon bias</i> (q.v.) because of a rare tRNA species, (ii) if the mutation disrupts an internal regulatory element, or (iii) if the mutation happens to generate a cryptic splice site.</p>
Missense or nonsynonymous mutation	<p><b>Definition:</b> A base substitution which alters the sense of a codon, resulting in the replacement of one amino acid residue with another in the encoded polypeptide. If a missense mutation is <b>conservative</b>, the new amino acid has similar chemical properties to the original, whereas if <b>nonconservative</b>, the original amino acid is replaced by one with different chemical properties. Example: AAG → GAG converts basic lysine to glutamic acid.</p> <p><b>Effects:</b> The effect of a missense mutation depends on whether it is conservative or nonconservative, and the importance of the residue which is associated with polypeptide function. Conservative exchanges may be neutral unless they occur at a critical residue (such as the active site of an enzyme), whereas nonconservative exchanges often disrupt folding and/or change the properties of the encoded polypeptide, generating a mutant phenotype. Some missense mutations are subtle and may only reveal their effects under extreme conditions, such as elevated temperature (q.v. <i>conditional mutant</i>).</p>
Nonsense mutation	<p><b>Definition:</b> A base substitution which converts a sense codon into a nonsense (termination) codon. Classified as <b>amber</b>, <b>opal</b> or <b>ochre mutations</b> depending upon the type of nonsense codon generated. Example TGC (cysteine) → TGA (stop).</p> <p><b>Effects:</b> Nonsense mutations cause premature termination of protein synthesis and generate truncated proteins. The severity of the effect depends on position in the coding region. 5' nonsense mutations, which severely truncate the encoded polypeptide, cause loss of function. 3' nonsense mutations may not affect polypeptide structure to a great degree, but can reduce mRNA stability. Nonsense mutations in eukaryotic genes occasionally cause <i>exon skipping</i> (q.v.) during splicing.</p>
Readthrough mutation	<p><b>Definition:</b> A base substitution which converts a nonsense (termination) codon, into a sense codon allowing readthrough and hence extension of a polypeptide. Example TAG (stop) → CAG (glutamine).</p> <p><b>Effects:</b> Readthrough mutations may modify the properties of the polypeptide and often affect mRNA stability. Generally, polypeptides are not extended greatly because adventitious termination codons are found downstream of the natural one.</p>

Continued

Frameshift mutation	<p>Definition: A small insertion or deletion of <math>3n \pm 1</math> nucleotides which disrupts the reading frame, generating a novel polypeptide sequence distal to the mutation.</p> <p>Effects: Like nonsense mutations, the consequence of a frameshift depends on its position, with 5' mutations having more severe effects than 3' mutations. Most open reading frames are peppered with out-of-frame termination codons, so frameshifts tend to cause premature termination and truncation of the polypeptide. Also q.v. <i>cotranslational frameshifting</i>.</p>
Nonframeshifting indel	<p>Definition: A small insertion or deletion of <math>3n</math> nucleotides (<b>indel</b> is generic for insertion or deletion).</p> <p>Effects: This type of mutation does not disrupt the reading frame and is often tolerated, although deletion or interruption of critical residues may abolish gene function.</p>
<p>Site: <i>intragenic noncoding regions</i></p> <p>Intron</p>	<p>Effects: Point mutations in introns are often neutral. A phenotype may be generated if (i) the mutation disrupts an intronic regulatory element such as an enhancer, (ii) the mutation abolishes a splice site (prevention of splicing allows readthrough into the intron, often causing truncation, and sometimes exon skipping), or (iii) the mutation creates a cryptic splice site, causing the intron sequence to be included in the mature transcript (this often causes truncation due to an intronic termination codon, or a frameshift which uncovers an out-of-frame exonic termination codon).</p>
<p>Untranslated region</p>	<p>Effects: Many UTR mutations are neutral, but in some cases, mutations may disrupt posttranscriptional regulation by affecting protein synthesis, RNA stability or RNA localization. Mutations of the polyadenylation site are often deleterious because posttranscriptional processing is disrupted.</p>
<p>Site: <i>extragenic</i></p> <p>Regulatory elements</p>	<p>Effects: Most extragenic mutations are neutral, but those which modify gene regulatory elements, such as promoters and enhancers, may disrupt or modify transcription patterns. Mutations in transcriptional enhancers are thought to be important for the diversification of expression patterns following <i>tandem gene duplication</i> (q.v).</p>

THE CAT SAW THE DOG AND RAN (END)	Wild type
THE CAT SAW THE DOG AND MAN (END)	Missense
THE CAT SAW THE DOG END	Nonsense
THE CAT SAW THE DOG AND RAN AND .....	Readthrough
THE CAT TSA WTH EDO GAN DRA NEN D .....	Frameshift
THE CAT SAW THE BIG DOG AND RAN (END)	Nonframeshift Indel

**Figure 15.1:** Point mutations in the coding region and their effect on the interpretation of genetic information. Each three letters (nucleotides) form a word (a codon) which makes sense in the context of the sentence (the polypeptide). Different types of point mutation disrupt the sentence in different ways. Nonsense and frameshift mutations are potentially most disruptive at the beginning of a polypeptide. In this example, the nonsense mutation occurs near the natural termination codon and most of the sense of the information is retained. Conversely, the frameshift occurs at the beginning of the sentence and most of the meaning is lost.

**Table 15.3:** The consequences of macromutations and chromosome imbalance on gene structure, expression and function

Physical effects of macromutation	Functional consequences
Whole gene deletion	Abolishes gene function. Effect depends on dosage sensitivity of the locus (q.v. <i>haploinsufficiency</i> ). May occur as part of a larger deletion, resulting in a <i>contiguous gene syndrome</i> (q.v.)
Partial deletion (truncation)	Effect depends upon extent of deletion. Extensive truncation abolishes gene function. Less severe truncations may generate products with novel functions, e.g. the <i>v-erb</i> oncogene (see <i>Oncogenes and Cancer</i> )
Partial deletion (internal)	Effect depends upon extent of deletion. Large deletions abolish gene function. Smaller deletions in exons may generate a modified product, but alteration of reading frame often occurs resulting in truncation and instability. In eukaryotes, deletions within introns are often neutral unless a regulatory element is modified
Whole gene duplication (gene amplification)	Increases level of gene product. Effect depends on dosage sensitivity of the locus. For many genes, amplification has no consequences and can be beneficial in certain environments (q.v. <i>gene amplification</i> ). Other genes are dosage-sensitive because their products are titrated against the products of other genes (q.v. <i>dosage effects</i> in main text)
Partial duplication (internal)	Effect depends on extent of duplication and number of copies. Duplication of exons may modify polypeptide function. Smaller (intraexonic) duplications are often deleterious because they alter the reading frame and generate truncated, unstable proteins. Expansion of intergenic tandem trinucleotide repeats occurs in some human disease genes ( <i>Box 15.2</i> )
Disruption by deletion/rearrangement/insertion	Usually loss of function for the interrupted gene. Occasionally, a fusion gene may be generated with novel properties (see main text). There may be position effects if a gene is brought under heterologous regulation (see main text)
Chromosome loss or gain	Changes the dosage of many gene products, generally resulting in severe multiple dosage effects (see <i>Chromosome Mutation</i> )
Whole genome duplication	Changes the dosage of all genes. Well tolerated in plants, but not in higher animals where sex-chromosome dosage and the dosage of imprinted genes plays a critical role in development (q.v. <i>polyploidy</i> , <i>parental imprinting</i> )

homozygous state if its effects can be compensated by nonallelic interactions (e.g. redundant genes, external suppressor mutations) or by environmental factors (e.g. cross-feeding).

- (3) *Mosaicism*. In multicellular organisms, the effect of a somatic mutation depends on the proportion of mutant cells and their distribution.
- (4) *Monoallelic expression*. In mammals, imprinted genes and (in females) genes on the inactive X-chromosome are epigenetically repressed. The effect of a mutation may thus depend upon its parental origin and the distribution of clones of cells containing a particular inactivated X-chromosome (see *DNA Methylation and Epigenetic Regulation*).

**Functional consequences of point mutation.** Most point mutations with pathological consequences occur in the coding region of the gene and are nonsynonymous (*Table 15.2*). These cause either specific localized changes in polypeptide structure by replacing single amino acids (missense mutations), or more profound changes involving many amino acids (nonsense mutations, frameshifts and readthrough mutations) (*Figure 15.1*). Synonymous mutations are usually neutral, but may generate a phenotype if they adventitiously modify gene expression, e.g. by creating a cryptic splice



site. Deleterious point mutations in noncoding transcribed sequences usually identify functionally important motifs such as regulatory elements and splice sites. Point mutations outside the gene can modify promoter or enhancer function by creating or abolishing binding sites for transcription factors, or altering their spatial relationships.

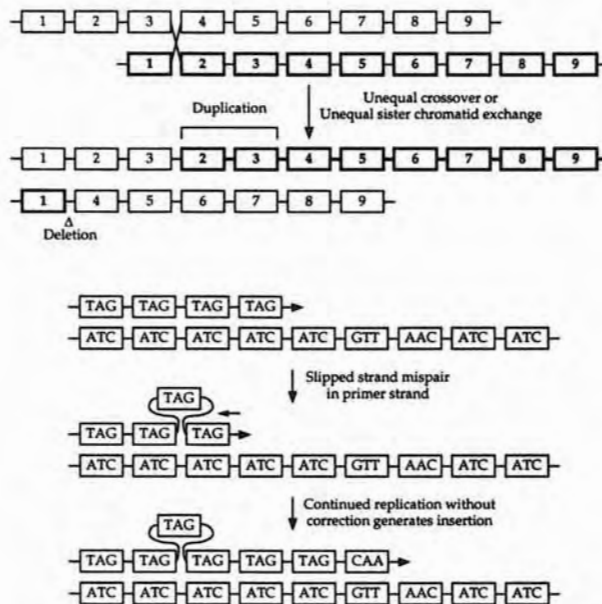
Point mutations usually affect only single genes. However, mutations in master regulatory elements such as the promoter and operator sites of bacterial *operons* (q.v.), and eukaryotic *locus control regions* (q.v.) can have more dramatic effects. Mutations in operon structural genes may occasionally affect the expression of downstream genes by disrupting protein synthesis on the polycistronic mRNA (q.v. *polar mutations*).

**Functional consequences of macromutations.** Macromutations and chromosome imbalances tend to have more severe effects than point mutations because they involve greater amounts of DNA (Table 15.3). Deletions and duplications may involve parts of genes, entire genes or many genes at the same time. Partial gene deletions, insertions (e.g. of transposable elements) and duplications often cause frameshifts as well as loss or gain of sequence, resulting in abnormal and unstable proteins which are readily degraded. In eukaryotes, this often occurs even if the indel spans an isolated exon, because exon boundaries within the same gene do not necessarily respect the same reading frame parameters (q.v. *exon deletion, exon repetition*).

As well as these consequences of **gene disruption**, which usually result in loss of function, macromutations can generate phenotypes in several other ways. **Gene fusion effects** are caused by deletions, duplications, inversions and translocations which interrupt two genes and bring the separate elements together. If the fusion is in-frame, and the gene is expressed, a hybrid gene product can be generated with novel properties (e.g. q.v. *Burkitt's lymphoma, hemoglobin Kenya*). Deletions and duplications may also cause **dosage effects**, the pathological consequences of altering the number of copies of a gene in the cell. This may be harmless or indeed beneficial for some loci (q.v. *gene amplification*), but deleterious dosage effects occur if the normal gene copy number produces the correct amount of product to maintain a competitive balance with the products of other genes. This is seen where stoichiometric amounts of polypeptides form a multimeric protein (e.g.  $\alpha$ - and  $\beta$ -globin in hemoglobin, see Box 15.1), and in delicately balanced regulatory systems such as the control of *sex determination* (q.v.) in *Drosophila*. Additionally, systems dependent on *quantitative* regulation are dosage-sensitive — this applies to many cellular signaling pathways (see *Oncogenes and Cancer*). Chromosome imbalances affect the dosage of many genes at once, and these mutations generate the most severe dosage effects (also q.v. *sex-chromosome aneuploidy, dosage compensation mechanisms*). Finally, inversions, translocations and large deletions may bring a gene into an unusual regulatory environment so that, for example, it comes under the control of a heterologous enhancer or is sequestered into inactive heterochromatin. The level or pattern of gene expression may thus be altered, due to these *cis-acting position effects*.

**Mutations involving repetitive DNA.** Deletions, insertions and rearrangements may be caused by random chromosome breaks and erroneous repair by *end-joining* (q.v.), but *repetitive DNA* (q.v.) is also frequently to blame for macromutations. Repetitive DNA can arise by several routes (e.g. replicative transposition, breaking and rejoining sister chromatids, unscheduled replication), but once present, it can act as a nonallelic homology domain, allowing unorthodox homologous recombination and unorthodox DNA replication by strand-slipping.

In large direct repeat units, such as the globin gene clusters, chromosomes can become misaligned, and homologous recombination can occur between nonallelic repeats, either between sister chromatids of the same chromosome (**unequal sister chromatid exchange**) or between homologous chromosomes (**unequal crossover**). In either case, the result is a reciprocal deletion from one recombining partner and insertion into the other, but only unequal crossover also involves recombination of flanking markers (Figure 15.2). Unequal crossing over in the globin clusters can generate hybrid globin



**Figure 15.2:** Insertions and deletions promoted by direct repeats in DNA. (A) Direct repeats are hotspots for unequal exchange (unequal crossing over or unequal sister chromatid exchange). Two chromatids can misalign, and crossing over generates reciprocal duplication and deletion products. (B) Short direct repeats are hotspots for replication slipping, i.e. where the template and primer strands slip out of register. Backward slipping of the primer strand generates an insertion. Forward slipping (not shown) generates a deletion. Both slipping and unequal exchanges are implicated in the pathology of triplet repeat syndromes (Box 15.2).

polypeptides, and in some cases causes hereditary persistence of fetal hemoglobin (Box 15.1).

Recombination also occurs between dispersed repeats, such as transposable elements. The consequences of these events depend largely on the relative locations and orientations of the recombining partners. Recombination between direct repeats on the same chromosome results in the deletion of the DNA between the repeats. Conversely, if the repeats are inverted, recombination inverts the intervening DNA. If the dispersed repeats are located on different (nonhomologous) chromosomes, recombination can cause reciprocal translocations (linear chromosomes) or cointegration (circular chromosomes and plasmids).

Short tandem repeats, such as those occurring in microsatellite DNA, are often subject to strand slipping during replication, i.e. the template and primer strands slip out of register, so that equivalent repeat units on the two strands are staggered (Figure 15.2). This is thought to be the process by which microsatellite DNA polymorphism is generated, as there is no recombination between flanking markers (i.e. no crossing-over). Slipped-strand replication is also stimulated by intrastrand secondary structures, such as hairpins and cruciforms, which stabilize strand misalignments. Inverted repeats are thus hotspots for deletions and insertions, although the formation of secondary structures is inhibited by *single-stranded DNA binding proteins* (q.v.), which therefore greatly increase the *frameshift fidelity* (q.v.) of DNA polymerases. Short tandem repeats are distributed throughout higher eukaryote genomes, usually in extragenic DNA, but occasionally within the coding regions of genes. Expansion of these intergenic repeats is implicated in a number of human diseases (Box 15.2).

Physical interactions between repetitive DNA sequences also allow nonallelic *gene conversion* (q.v.) events to occur. The heteroduplex DNA generated as a recombination intermediate is sub-

**Table 15.4:** Three systems for the classification of mutant alleles

System	Criteria
Loss or gain of function	Applicable to both haploid and diploid organisms. The level and scope of function of wild-type and mutant alleles are compared. Alleles are classed as loss of function if the product is less active than the wild type, or gain of function if it is more active than the wild type or has acquired novel functions
Dominance relationships	Relevant only in diploid organisms. The phenotypes of wild-type, heterozygous and homozygous mutant individuals are compared. Alleles are classed as dominant, partially dominant, codominant or recessive depending on the degree to which the mutant phenotype is expressed in the heterozygote (also see <i>Table 1.2</i> )
Müller classification	Relevant only in diploid organisms. This system was developed in <i>Drosophila</i> , a species with readily available panels of deletion mutants. The phenotype of an individual homozygous for a deletion at the locus of interest is compared to that of a homozygous mutant and a deletion/mutant heterozygote. Alleles may be classed as follows: Amorphic — no activity, Hypomorphic — reduced activity compared to wild type, Antimorphic — opposite activity to wild type, Hypermorphic — greater activity compared to wild type, Neomorphic — novel activity compared to wild type.

jected to mismatch repair, resulting in sequence homogenization of the repeats. This is the major source of clustered point mutations in eukaryotic genomes (but also q.v. *somatic hypermutation*, *SOS response*), and may be one mechanism of *concerted evolution* (q.v.). Physical interactions between repetitive DNA sequences also allow epigenetic modification of gene expression (q.v. *paramutation*, *homology-dependent silencing*, *cosuppression*).

## 15.2 Mutant alleles and the molecular basis of phenotype

**Wild-type and mutant alleles.** Alleles are variant forms of genes, initially defined by their phenotypic effects, but ultimately by their nucleotide sequences. The **wild-type allele** at any locus is the predominant allele in the population, it generally confers the greatest fitness, and produces a fully functional gene product. A **forward mutation** is a mutation away from the wild type, generating an alternative **mutant allele** whose product may differ from the wild type in *quality*, *quantity* or *distribution*. A mutation back to the wild type is a **reversion**.

It is convenient to classify mutant alleles by comparing their phenotypes to that of the wild-type allele. Particularly relevant in diploids is the way in which the mutant and wild-type alleles interact in the heterozygote. Three systems of classification have been developed to define the properties of mutant alleles (*Table 15.4*).

In principle, a forward mutation may affect gene function or expression in three ways: it may cause reduction or abolition of gene activity (a **loss of function allele**); it may cause an increase in gene activity, or confer a novel function upon the encoded polypeptide (a **gain of function allele**); or the mutant allele may be phenotypically indistinguishable from the wild type, even though different in nucleotide sequence (the wild-type and mutant alleles would be classed as **isoalleles**). Mutations which generate isoalleles are neutral; they are often synonymous substitutions. Mutations which cause loss or gain of gene function may be neutral, beneficial or deleterious. It is important to distinguish the consequences of losing or gaining the function of one particular gene from the consequences of losing or gaining overall fitness, e.g. in humans, loss of function at one locus results in a totally harmless inability to roll the tongue (neutral), whereas gain of function in an oncogene is a (deleterious) step towards cancer.

**Alleles with less activity than the wild-type allele.** Alleles which have reduced activity compared to the wild type are **loss of function alleles**. These are generated either by downregulating gene expression and thus reducing the *quantity* of gene product (**down** or **downpromoter mutations**), or by altering the product so that it functions less well than the wild type, i.e. reducing the *quality* of the gene product. **Null alleles (amorphs)** are **total loss of function alleles**, where gene expression is abolished or the mutant gene product is totally unable to function. Null alleles are often generated by full gene deletions, by mutations which destroy regulatory elements, or by point mutations causing truncation of the encoded polypeptide. **Leaky alleles (hypomorphs)** are **partial loss of function alleles**, where gene function is reduced but not completely abolished, enabling the organism to carry out those activities encoded by the gene, although at reduced efficiency compared to the wild type. Leaky alleles are often generated by missense base substitutions, permitting minimal gene function, or by regulatory down mutations which reduce gene expression but do not abolish it. The severity of the mutant phenotype depends upon the residual function of the mutant polypeptide: alleles which generate a severe phenotype are described as **strong alleles**, while those which generate a mild phenotype are **weak** or **moderate alleles**. The severity of a leaky mutation may differ in different environments (q.v. *conditional mutant*) and may therefore show incomplete penetrance when the environment is not constant.

In diploids, loss of function alleles are usually *recessive* (q.v.) to the wild type, i.e. the effects of the mutation are not seen in the heterozygote. This is because for most loci, one wild-type copy of the gene (50% dosage of its product) is sufficient for the needs of the cell.

There are two situations where loss of function mutations may exhibit dominance over the wild-type allele. **Haploinsufficiency** occurs when two functional copies of the gene are required to maintain the wild-type phenotype, i.e. 50% dosage of the product is insufficient for physiological gene function. Loss of function mutations at haploinsufficient loci demonstrate *partial dominance* (q.v.) over the wild-type allele (i.e. the effect of the mutation is apparent in the heterozygote — at 50% dosage — but more severe in the homozygote — at nil dosage). An example is hypercholesterolemia, a partially dominant human disease caused by a 50% reduction in the level of the low density lipoprotein (LDL) receptor.

Loss-of-function mutations may show complete dominance over the wild-type allele if the mutant products interfere with wild-type function. This usually occurs when the gene product is a multimer, and the mutant can sequester wild-type polypeptides into inactive complexes. For example, receptor tyrosine kinases are dimeric, and mutant nonsignaling receptors can effectively block signaling from wild-type receptors by forming inactive heterodimers. Alleles of this nature are described as **dominant negatives** (sometimes classed as **antimorphs** because they oppose or antagonise the wild-type allele).

**Alleles with greater activity than the wild-type allele.** Alleles which have increased activity compared to the wild type are termed **gain of function alleles** or **hypermorphs**. Such alleles increase the activity of the gene product either by increasing its quantity (**up** or **uppromoter mutations**), encoding a product with superior or novel qualities compared to the wild type, or causing the gene product to be expressed or activated outside its usual scope (e.g. **constitutive mutants** — where the wild-type product is regulated, the mutant is active all the time; **ectopic expression mutants** — regulatory mutants which cause a gene to be expressed outside its normal spatial or temporal domains). Where a phenotype is apparent, gain-of-function alleles are usually dominant to the wild-type allele, but it may be possible for a wild-type polypeptide to mask the effect of a qualitative gain of function mutant in a multimeric protein. **Dominant positives** are analogous to dominant negatives, i.e. the mutant polypeptide exerts its effects at the expense of the wild-type polypeptide in a multimeric protein, but in this case the mutant overcomes some restriction or regulation experienced by the wild-type product, as seen, for example, in constitutively signaling receptor tyrosine kinases. A **neomorph** possesses novel activity compared to the wild type. Ectopic expression



mutants are often neomorphic because the effects of synthesizing a polypeptide in a region from which it is usually excluded are unpredictable. Gain-of-function homeotic mutations, such as *Drosophila Antennapedia*, which causes legs to sprout from the segment where antennae should develop, are neomorphic (q.v. *homeotic genes*).

**Alleles with the same activity as the wild-type allele.** Many mutations have no phenotypic effect and are selectively neutral. Different alleles which have the same phenotype and thus cannot be discriminated at the morphological level are termed **isoalleles**. Isoalleles may be generated by mutations which do not alter either the quantity, quality or distribution of the gene product (e.g. synonymous nucleotide substitutions), or by mutations which do alter the structure or expression of the encoded polypeptide but lack a phenotype because the effects of these changes are negligible (e.g. regulatory mutations which cause moderate but asymptomatic changes to the rate of gene expression, or conservative missense mutations in functionally unimportant polypeptide domains).

While isoalleles are selectively neutral, not all neutral alleles are isoalleles: mutations which do cause a change in phenotype may still be neutral if the different phenotypes do not affect fitness (e.g. mutations which cause changes to eye color). The effects of mutations can thus be considered at several levels: (i) the effect on nucleotide sequence; (ii) the effect on gene activity; (iii) the effect on phenotype; (iv) the effect on overall fitness. Only mutations which alter fitness are subject to selection. Mutations which have no effect on gene activity (e.g. synonymous mutations and most mutations outside the coding region of the gene) are the most likely to be neutral in all environments. Those causing changes in gene activity or in phenotype may be neutral in some environments but not in others.

Neutral alleles are not subject to selection, and thus several can exist in equilibrium within a population at relatively high frequencies (q.v. *polymorphism*). Phenotypically distinct neutral alleles can be detected as morphological variants, but isoalleles can only be discriminated at the molecular level. In some cases, **protein polymorphisms** may be detected by the differential behavior of protein alloforms on electrophoretic gels (also q.v. *protein truncation test*). **DNA sequence polymorphisms** may be detected by changes to the length of restriction fragments or PCR products, either because a restriction site has been created or destroyed (**restriction fragment length polymorphism**), or because there has been an expansion or contraction in the number of tandem repeat units in satellite DNA (**simple sequence length polymorphisms**). Alternatively, the behavior of single-stranded DNA or heteroduplex DNA can be exploited to detect mutations (q.v. *mutation screening*). The only way unambiguously to detect and characterize all polymorphisms is through DNA sequence analysis. It has been estimated that the **mean heterozygosity** of human DNA is 0.004, i.e. on average, one in every 300 bases is polymorphic.

### 15.3 The distribution of mutations and molecular evolution

**Mutation spectra and regional distribution of mutations in the genome.** The spectrum and distribution of mutations in a population of genomes is nonrandom due to the existence of **mutation hotspots** — sites which are particularly susceptible to certain types of DNA damage or rearrangement. Tandemly repetitive DNA is a hotspot for slipped-strand replication or unequal exchange, inverted repeats are hotspots for deletions induced by secondary structures, and 5-methylcytosine residues are hotspots for C→T transitions through deamination. The instability of 5-methylcytosine partially explains the unexpected predominance of transitions over transversions in mammalian DNA. Each base has two choices for transversion but only one for transition, so random changes should produce transversions with twice the frequency of transitions. In fact, the opposite is true: transitions are twice as common as transversions. The frequency of different base substitutions varies widely. Due to the instability of 5-methylcytosine, C→T transitions are nearly ten times more likely to occur than any other substitution, at least in organisms with methylated DNA. There is also

**Table 15.5:** Terms used to describe the distribution of alleles in populations and the forces which change them

Term	Definition
Mutation rate	The number of mutations occurring over a period of time, e.g. mutations per gene per generation
Mutation frequency (allele frequency)	The number of individuals in a given population carrying a particular mutant allele
Mutation pressure	The force which increases the frequency of a particular allele by recurrent mutation
Mutation load (genetic load)	The effect of deleterious alleles on a population
Migration	The force which changes allele frequencies by importing and exporting individuals from a population
Natural selection	The force which changes allele frequencies by eliminating alleles causing loss of fitness in a given environment
Random genetic drift	The force with changes allele frequencies by random sampling

bias in the frequency of the other 11 possible base substitutions, reflecting underlying bias in DNA repair mechanisms, especially mismatch repair and the repair of common misinstructional lesions caused by base damage (see Mutagenesis and DNA Repair).

When mutation hotspots and base substitution biases are taken into account, any region of the genome should, in principle, be equally susceptible to mutation, i.e. mutation is a stochastic and undirected process providing 'adaptive randomness' as a substrate for natural selection. The concepts of **programmed mutations**, induced by the cell as part of the developmental program, and **directed mutations**, occurring in response to particular selection pressures, are discussed in Box 15.3.

**Allele frequencies in populations.** There are numerous factors which influence the frequency of alleles in a given population, including population size and structure, mating patterns, mutation rate, migration, natural selection, random genetic drift and gene conversion (Table 15.5). Mutation and migration are the two factors which can introduce new alleles into a population. Newly arising (or arriving) alleles may be deleterious, neutral or (rarely) advantageous compared with the current wild-type allele. In the simplest case, alleles which reduce fitness would be eliminated by natural selection (**negative selection**) and would be maintained at a low frequency by the rate of recurrent mutation (**mutation pressure**) and immigration. Alleles which increase fitness would spread throughout the population (**positive selection**) and would eventually displace the previous wild-type allele. During this displacement process, population analysis would reveal **polymorphism** at the locus (a situation where there are two or more alleles, each with a frequency greater than 0.01). This could be termed a **transient polymorphism** because the alleles are progressing towards **fixation** (a frequency of 0 or 1).

If new alleles are selectively neutral, changes in allele frequency depend on chance events, i.e. random sampling of gametes (**random genetic drift**), rather than natural selection. Drifting alleles eventually reach fixation, but this takes a long time in large populations, and would be revealed as a **neutral polymorphism**. The effects of drift are more pronounced in small populations, population bottlenecks and new populations (the founder effect), where they can cause rapid and dramatic changes in the representation of particular alleles.

More complex interactions occur if the fitness of a heterozygote is outside the range specified by the two homozygotes. Where one allele is fitter than another in the population, and the fitness of the heterozygote falls between the fitnesses of the homozygotes, selection is directional and will lead to fixation of the fittest allele. If the heterozygote is fitter than either homozygote (overdominance), the heterozygote will be selected and both alleles will be maintained in a **balanced polymorphism**. An example is overdominant selection for the normal (HbA) and sickle-cell (HbS) alleles of  $\beta$ -globin. These are polymorphic in some African countries because, while HbS is deleterious in the homozygous state,

it confers malarial resistance in HbA/HbS heterozygotes, a genotype which is thus fitter than HbA homozygotes. Other forms of balancing selection involve alleles whose fitness is frequency-dependent. The heterozygote may also be less fit than either homozygote (underdominance). Where this occurs, there will be divergent selection, but if random mating continues, the heterozygous population will be replenished and the alleles will be maintained as an **unstable polymorphism**.

**Selection pressure and the molecular clock.** When regional differences in the distribution of mutation sites are taken into account, the susceptibility of DNA to mutation should be generally equal throughout the genome. However, mutations occurring in noncoding DNA are predominantly neutral, whereas many mutations occurring in coding DNA have deleterious effects and are therefore eliminated by natural selection. The observed frequency of *surviving* mutations in coding DNA is thus much less than in noncoding DNA, with the result that coding DNA sequences are conserved over evolutionary time.

The **selection pressure** which maintains DNA coding sequences controls the distribution and spectrum of mutations observed. Most surviving mutations in coding DNA are base substitutions because more dramatic mutations — frameshifts, large deletions — are almost always deleterious and are eliminated. Base substitutions occur more frequently at **degenerate sites** (sites where substitutions will not alter the sense of the codon, e.g. often in the third position of the codon) than at **nondegenerate sites** (sites where missense mutations arise, causing amino acid replacements). The rate of nonsynonymous substitution for an evolving protein is indicative of the intensity of the selective pressure to maintain its structure. Some proteins change very little because almost all amino acids play an important role in maintaining the structure and function of the polypeptide (e.g. histone proteins). Others are evolving very quickly because the polypeptide structure is not important for protein function (e.g. the insulin linker chain, whose role is to separate the A and B chains, and which is discarded following polypeptide cleavage).

The rate of synonymous substitution in a gene is independent of selective pressure because most such mutations are neutral. Thus, the rate of synonymous substitution for histone and albumin proteins is about the same, although the nonsynonymous substitution rate for albumin is several hundred-fold greater than that of the histones. This has given rise to the concept of the **molecular clock**, the measurement of evolutionary time as the rate of neutral evolution. The molecular clock is not constant, however. Different genes vary in the neutral substitution rate as well as their amino acid replacement rate, e.g. insulin has a neutral substitution rate twice that of hypoxanthine phosphoribosyltransferase (HPRT), even though both have more or less the same nonsynonymous substitution rate and are therefore under equal selective pressure. A number of factors could influence neutral evolution, one of which is DNA repair bias. As discussed above, some mismatches and types of base damage are more likely than others to be repaired, so the rate of neutral evolution could be influenced by base composition in the gene. Differences in repair efficiency between genomes is also involved (this contributes to the rapid evolution of animal mitochondrial genomes, which lack *nucleotide excision repair*, q.v.). The molecular clock also runs at different rates in different species lineages. Generally, the clock is slower for organisms with longer generation intervals, because the effects of newly arising mutations are tested in the subsequent generation. The fewer generations per unit of real time, the fewer new mutations can be tested.

## 15.4 Mutations in genetic analysis

**Classical genetic analysis and reverse genetics.** The classical approach to the dissection of a biological system is to isolate mutants deficient for that system and then determine the structure and precise function of the mutated genes. The alternative approach, which involves isolation of the gene on the basis of its structure — usually by determining the sequence of its encoded protein — and then mutagenizing the gene to study its function, is sometimes termed **reverse genetics** (q.v. *in*



*vitro* mutagenesis, gene targeting, gene knockout). Classical genetic analysis requires the generation of mutants, screening of populations to identify mutants for the system of interest, and then the mapping and cloning of the responsible gene. Once interesting mutants are available, a second screen for mutations which modify the phenotype of the first mutation can identify genes whose products interact with those identified in the first screen (Box 15.4).

**Genetic screens.** Mutations may be generated by exposing a population of organisms to physical or chemical mutagens: nitrosoguanidine is often used for bacteria, ethylmethanesulfonate or X-rays for *Drosophila*. Alternatively, transposons can be used to generate mutations: P-elements have been widely used in *Drosophila* and Ac-Ds elements in plants (q.v. *P-element mutagenesis, transposon tagging*). A population of randomly mutated individuals is generated in this way, and it is then necessary to identify and isolate mutants for the system of interest.

In many cases, mutants can be identified by laborious visible screening for a particular morphological phenotype (e.g. in the large-scale screens for developmental mutants in plants, *Drosophila* and zebrafish, for cell-cycle mutants in yeast, and for interesting expression patterns in *enhancer trap* and *gene trap* (q.v.) lines of *Drosophila* and mice). Where a biochemical or physiological mutation is sought, it is valuable to enrich the population for desired mutants by selection. For gain of function mutations, such as gain of antibiotic resistance in bacteria, **positive** or **direct selection** is used — in this case by culturing the bacteria in the presence of antibiotics to kill nonresistant (nonmutant) cells. For loss of function mutations, such as *auxotrophy* (q.v.) in bacteria, **negative selection (counterselection)** may be used. This strategy kills wild-type cells by exploiting any sensitivities which have been lost by the mutant cells. In the case of auxotrophy, counterselection is often carried out by **penicillin enrichment**: the mutants are unable to proliferate on minimal medium due to their metabolic deficiency and are therefore resistant to penicillin, which is lethal to proliferating cells because it prevents synthesis of cell wall components (also q.v. *positive-negative selection*). Penicillin enrichment does not identify specific metabolic mutants, and this selection is carried out indirectly using a two-stage process of **replica plating**. This involves taking a cloth print of a master plate of bacterial colonies (by laying a piece of velvet over the colonies and picking up some bacterial cells from each colony) and placing this cloth onto a fresh plate so that cells are deposited onto the agar and the same pattern of colonies is generated. If the master plate is supplemented with an appropriate metabolic end product so that both prototrophs and auxotrophs can grow, but the replica plate contains minimal medium, the auxotrophs will not grow on the replica plate. They can then be identified on the master plate as colonies with no counterparts on the replica plate (also q.v. *recombinant selection*).

The same principles of genetic screening can be applied to higher organisms, but only where large numbers of individuals can be mutagenized and bred, e.g. *Drosophila*, yeast, plants, animal cells in culture. These screens are often complicated by diploidy, and to identify recessive mutations, the mutants must be bred to homozygosity over several generations or studied in a haploid background (e.g. by using aneuploid cell lines). In some diploid species, haploid individuals are viable (e.g. many plants, zebrafish). Once a mutant has been identified the gene can be isolated and cloned, and its biochemical role determined. The principles for doing this are discussed elsewhere (see Recombinant DNA).

**Conditional mutants.** The genetic analysis of essential systems, e.g. DNA replication, development and the cell cycle (see individual chapters on these topics), is made difficult because mutations are often lethal. A cell which cannot undergo replication will die, as will an organism blocked at an early stage in development. In diploids, recessive lethal mutations can be maintained in heterozygotes and the basis of lethality studied by crossing heterozygotes and analyzing the homozygous mutant progeny. However, in both haploid and diploid organisms, **conditional mutants** are widely exploited in the analysis of essential systems. A conditional mutant carries a (usually missense) mutation whose effects manifest only under certain **restrictive conditions**. Under normal permis-



**sive conditions**, the wild-type phenotype is displayed. Important classes of conditional mutation include **temperature-sensitive mutations**, which display the mutant phenotype under conditions of elevated temperature, and **cold-sensitive mutations**, which display the mutant phenotype at low temperature. In each case, the properties of the mutant are likely to involve an increased tendency of the protein to denature at restrictive temperatures.

**Genetic pathway analysis.** Many genes form parts of **genetic pathways** and **genetic networks**, which take a substrate and convert it into a product through several intermediate stages, each controlled by a different gene. In some cases, the substrate is information, e.g. in the form of a signal arriving at the cell surface which must be transduced to the nucleus, or in the form of gene regulation, which proceeds through a cascade of regulatory switches to downstream targets. In other cases, the substrate is a physical molecule, a metabolite which must be converted into a useful product. Mutational analysis can determine the genes involved in the pathway, their order of activity, and where branching and convergence of pathways occur.

Metabolic pathways are in some ways the easiest to dissect because the initial substrate, intermediates and final product are physical molecules rather than states of information processing. Typically, metabolism involves a series of chemical reactions each catalyzed by a specific enzyme, and each enzyme is encoded by a gene. Mutations which disrupt the function of the enzymes lead to a **metabolic block** characterized by (a) failure to produce the end-product of the reaction, and (b) the accumulation of a metabolic intermediate. Either or both of these unusual states can generate a phenotype and may often be harmful to the organism. Bacteria can synthesize many essential organic molecules using a simple carbon source, water and minerals, and can therefore grow on a **minimal medium** containing these substrates. A bacterial cell with the wild-type metabolic properties of the species is **prototrophic**. An **auxotroph** is a bacterial mutant deficient for a metabolic enzyme, with the result that auxotrophs cannot grow on the medium which is sufficient for the growth of wild-type cells, but need **supplemented medium**, containing the end product of the disrupted metabolic pathway. If the phenotype of the metabolic block arises principally from failure to produce the end product, mutations in any of the genes in the pathway can produce the same phenotype (**locus heterogeneity**). Initially, the number of steps in the pathway can be estimated by **complementation analysis** (q.v.), which involves bringing two mutations together in *trans*, and seeing if the products produced by each genome can compensate for deficiencies in the other. Gene order can be established by the analysis of metabolic intermediates and cross-feeding to establish whether the intermediates produced by one mutant cell can be used by another mutant cell to generate the end product.

The analysis of information transfer pathways is more complex because mutations can cause gain of function effects (constitutive pathway activation) as well as loss of function effects, which are the most common metabolic disorders. The availability of dominant gain of function mutations is useful, however, because pathway order can then be established by crossing two mutations into one strain. An early-acting loss of function mutation which blocks information transfer will be hypostatic to a later-acting gain of function mutation which causes constitutive information transfer, whereas a later-acting loss of function mutation will be epistatic to an early-acting gain of function mutation (q.v. *epistasis*, *hypostasis*).

**Box 15.1: Mutation and pathology in human disease — hemoglobin disorders**

**Normal and abnormal hemoglobins.** Hemoglobin is the oxygen-carrying protein of erythrocytes which allows these cells to transport oxygen through the circulatory system. Hemoglobin is a tetrameric protein, containing two  $\alpha$ -type globin chains and two  $\beta$ -type globin chains associated by hydrogen bonds. Each globin polypeptide is conjugated to a heme molecule whose function is oxygen-binding. The human globin genes are found in two clusters. The  **$\alpha$ -globin cluster** contains the  $\zeta$ -globin gene, two identical  $\alpha$ -globin genes, and a gene whose function is unknown,  $\theta$ -globin. The  **$\beta$ -globin cluster** consists of the  $\epsilon$ -globin gene, two  $\gamma$ -globin genes whose products differ from each other at a single amino acid residue, and the  $\delta$ -globin and  $\beta$ -globin genes. Both clusters also contain pseudogenes.

The globin genes of both clusters are expressed in a temporal sequence so that the type of hemoglobin synthesized changes throughout development (the molecular basis of developmental regulation in the  $\beta$ -globin cluster is discussed in Box 29.3). During the first 6–8 weeks of life, hemoglobin is synthesized in the yolk sac and comprises a tetramer of two  $\zeta$ -globin chains and two  $\epsilon$ -globin chains (**embryonic hemoglobin** or **Hb Gower I**). However, starting at about week 2, synthesis of the embryonic globins begins to decline and synthesis of  $\alpha$ -globin and the  $\gamma$ -globins begins. Until birth, hemoglobin is synthesized mainly in the liver and comprises a tetramer of two  $\alpha$ -globin chains and two  $\gamma$ -globin chains (**fetal hemoglobin**, **HbF**).  $\alpha$ -globin continues to be synthesized into adult life, but between about 30 weeks gestation and 12 weeks after birth,  $\gamma$ -globin synthesis declines and  $\beta$ -globin and  $\delta$ -globin synthesis increases. The primary site of erythropoiesis shifts from the liver to bone marrow. **Adult hemoglobin** is mainly HbA ( $\alpha_2\beta_2$ ) with HbA2 ( $\alpha_2\delta_2$ ) representing a small (~2%) fraction of the total.

Hemoglobin disorders come in three forms: **hemoglobinopathies** (qualitative structural alterations to the globin chain resulting in the production of unusual globin polypeptides); **thalassemias** (quantitative reductions in globin synthesis, leading to imbalance between the  $\alpha$ -globin and  $\beta$ -globin chains); developmental disorders (disruption to the developmental time course of globin expression). These disorders, which range from asymptomatic to lethal, demonstrate the pathological effects of many different types of mutation.

**Variant globins generated by point mutations.** Many different globin variants are generated by

missense mutations, and substitutions have been identified in over 50% of the residues in both  $\alpha$ - and  $\beta$ -globin chains. Many substitutions are neutral in their effects on hemoglobin function, but some are pathological because they disturb its tertiary structure or ability to undergo conformational change, and thus alter the oxygen affinity of the molecule or interfere with its ability to bind the heme group. The most common pathological substitution converts codon 6 of the  $\beta$ -globin chain from GAG to GTG (replacing glutamic acid with valine), generating a form of hemoglobin (**HbS**) with increased intermolecular adhesion in its deoxygenated state. HbS thus crystallizes at low oxygen tension, causing the formation of inflexible, sickle-shaped erythrocytes that block capillary beds and damage internal organs. The destruction of these cells results in the severe **sickle-cell anemia** associated with individuals homozygous for this mutation. The HbS  $\beta$ -globin allele is generally rare, but polymorphic in some African countries because heterozygotes for normal and HbS  $\beta$ -globin chains are resistant to the most severe form of malaria (q.v. *overdominant selection*). Although heterozygotes are carriers of HbS, they rarely show disease symptoms — only in conditions of extreme low oxygen tension (**sickle-cell trait**). However, by exposing collected cells to such conditions, carriers can be identified.

Other point mutations occurring in the globin coding regions have been classified as frameshifts, nonsense mutations and readthrough mutations. Frameshifts and nonsense mutations tend to generate variant hemoglobins if they occur at the 3' end of the coding region but thalassemias if they occur at the 5' end, due to severe truncation and loss of function. Hemoglobin Cranston contains a variant  $\beta$ -globin chain, generated by a 3' end frameshift (the insertion of two nucleotides, GA, between codons 144 and 145). This causes readthrough of the termination codon, generating a polypeptide which is 10 residues longer than normal. Hemoglobin Constant Spring is generated by a readthrough mutation which converts the termination codon of the  $\alpha$ -globin chain from UAA to CAA. This variant is 31 residues longer than the normal  $\alpha$ -globin chain.

**Variant globins generated by recombination between repetitive DNA sequences.** Misaligned sequence exchange between repeated sequences (unequal crossing over, or unequal sister chromatid exchange) can generate both small rearrangements within individual globin genes, and large rearrangements involving entire globin clusters. Unequal

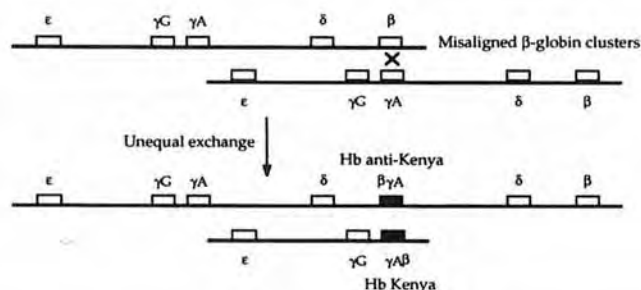
exchange occasionally occurs between directly repeated copies of the sequence GCTGCACGTG, found in codons 91–94 and 96–98 of the  $\beta$ -globin gene. This results in a deletion from one strand, and an insertion in the other, of 15 nucleotides, thus preserving the reading frame and generating the variant hemoglobin Gun Hill. The misalignment of entire genes followed by unequal exchange generates hybrid globin fusion chains. The most common rearrangements involve unequal crossing over between the  $\gamma^A$ -globin and  $\beta$ -globin genes, to generate hemoglobin Kenya (with the N-terminal region of  $\gamma^A$ -globin and the C-terminal region of  $\beta$ -globin; see figure below). This event also deletes the  $\delta$ -globin gene and causes hereditary persistence of fetal hemoglobin (see below). Different Hb Kenya subtypes reflect alternative points of crossing over. For every Hb Kenya chromatid, there is another chromatid carrying the reciprocal exchange products: Hb anti-Kenya is a  $\beta\gamma^A$ -globin fusion (the chromatid also contains a duplicated  $\delta$ -globin gene).

**Mutations causing thalassemias.** Thalassemias result from loss of globin gene expression, caused either by large-scale deletions or more subtle mutations. The solubility and oxygen-carrying capacity of hemoglobin depends on the stoichiometric amounts of  $\alpha$ - and  $\beta$ -globin in the molecule. In  $\alpha$ -thalassemia, only  $\beta$ -globin is available, and in  $\beta$ -thalassemia, only  $\alpha$ -globin is available. In each case, the remaining globin chains attempt to form tetramers, but these lack oxygen-carrying capacity and form insoluble complexes.

Because there are two redundant  $\alpha$ -globin genes, severe  $\alpha$ -thalassemia occurs only when three or all four alleles are lost. Most  $\alpha$ -thalassemias are caused either by unequal crossing over between the pair of  $\alpha$ -globin genes, or deletions generated by chromatid breaks. Occasionally, loss of gene function occurs through a more subtle mutation, e.g. a frameshift, but this is seen more frequently in  $\beta$ -thalassemia because there is only one  $\beta$ -globin gene.  $\beta$ -thalassemias are frequently caused by point muta-

tions. These include mutations in the  $\beta$ -globin promoter, 5' nonsense and frameshift mutations, mutations in the polyadenylation site and mutations in introns which prevent splicing. More unusual examples include a missense mutation in codon 26 of the  $\beta$ -globin gene, which exchanges glutamic acid for lysine. Although this is a nonconservative mutation, it would be expected to produce a variant hemoglobin rather than to cause thalassemia. However, the mutation also introduces a cryptic splice site into the  $\beta$ -globin gene, which causes aberrant splicing to occur, reducing the amount of wild-type  $\beta$ -globin to 50–60% of normal levels. Single thalassemias are generated by point mutations and macromutations affecting individual globin genes, but multiple thalassemias can be generated by deletions of the globin *locus control regions* (q.v.) which are responsible for the high level coordinated transcription of all genes in a cluster. Deletion of the  $\beta$ -globin LCR, for instance, causes  $\epsilon\gamma\delta\beta$ -thalassemia.

**Hereditary persistence of fetal hemoglobin (HPFH).** In normal adults, fetal hemoglobin comprises <1% of the total hemoglobin because there is a switch from  $\gamma$ -globin to  $\beta$ -globin gene expression during development. In cases of HPFH, however, this normal developmental switch fails and HbF can account for 20–100% of total hemoglobin in the cell. Deletional forms of HPFH coincide with  $\beta$ -thalassemia, i.e. deletion of all or part of the  $\beta$ -globin and  $\delta$ -globin genes and their control elements prevents the switch occurring. In such cases, HbF can compensate for the absence of the adult hemoglobins and the condition is benign (however, HPFH does not occur in all cases of deletional  $\beta$ -thalassemia). Other instances of HPFH occur in the absence of deletion and reflect promoter mutations which prevent switching. Furthermore, there are reported cases of HPFH where no mutation in the  $\beta$ -globin cluster could be detected, providing evidence for a regulatory element outside the cluster.



**Box 15.2: Pathogenic triplet repeat expansion in human disease**

**Triplet repeat syndromes.** A number of human diseases, including Huntington's disease, myotonic dystrophy and fragile-X syndrome, display complicated pedigrees characterized by incomplete penetrance, variable expressivity and anticipation (the tendency for a disease to become more severe and show an earlier age of onset through successive generations). In 1990, sequence analysis of the fragile X syndrome gene *FMR1* identified a polymorphic region of tandemly repeated triplets, (CGG) $_n$ , in the 5' untranslated region. In normal individuals, the repeat copy number was less than 50, whereas in affected individuals, the region had expanded and contained over 1000 copies. Since then, more than 10 additional **triplet repeat syndromes** have been identified.

**Three classes of triplet repeat syndrome.** Fragile-X syndrome is typical of class II triplet repeat syndromes, where the triplet repeat sequence is GC-rich and found in noncoding DNA. The extent of pathological triplet expansion is great (from a normal 50–100 copies to over 2500 copies), and this causes transcriptional repression through chromatin re-modelling and DNA methylation. Huntington's disease is a class I triplet repeat syndrome: the triplet (CAG) $_n$  lies within the coding region of the gene and encodes a run of glutamine residues. The extent of pathological triplet expansion is moderate (from a normal 10–30 copies to 40–100). Transcription is unaffected by the expansion, and the effects of the mutation presumably act at the

level of protein function. In myotonic dystrophy, a (CTG) $_n$  repeat is found in the 3' UTR and expands from under 50 copies to more than 3000. This appears to have no effect on either transcription or polypeptide structure — the pathological mechanism may involve the titration of specific RNA-binding proteins which recognize the triplet repeat sequence in the transcript. Several CUG-binding proteins have been identified and repeat expansion may sequester them and prevent them interacting with their normal target RNAs.

**Mechanism of repeat expansion.** The unstable nature of the intergenic tandem repeats adequately explains the complicated pedigree patterns observed for the triplet repeat syndromes. It is likely that moderate increases in repeat number are caused by strand-slipping during DNA replication, but sudden expansions must involve a different mechanism, such as unequal crossing over or unequal sister chromatid exchange. The basis of anticipation reflects a strong bias for repeat expansion once a critical **premutagenic** threshold number of repeats has been reached. In myotonic dystrophy, this threshold lies within a narrow range of 40–50 repeats, whereas for fragile-X syndrome, the threshold is 50–200 copies. Once this copy number has been reached, there is a high probability of continued, pathological expansion and a low probability of contraction. The nature of this bias is unknown.

**Box 15.3: Random, directed and programmed mutations**

**Random vs directed mutations in bacteria.** Do mutations occur randomly, providing the raw material for natural selection, or do they arise in response to selective pressure? The random, undirected nature of mutation was first shown by the **fluctuation test** of Luria and Delbruck in 1943. Bacterial cultures were maintained in optimal conditions for several generations and then shifted to a harsh environment (by infecting them with bacteriophage T1). Phage-resistant mutants were isolated after several more generations. If mutations occurred randomly (i.e. independent of the phage infection), some cultures would be expected to contain few phage-resistant cells (because the mutation occurred late in the experiment), while others would contain many (because the mutation occurred early, well before

exposure to the phage). Conversely, if mutations arose in response to phage infection, all cultures would be expected to contain similar numbers of resistant cells. The results showed a large fluctuation in the numbers of resistant cells between cultures, suggesting that mutations had arisen randomly.

In 1988, Cairns provided evidence that *lac*<sup>-</sup> bacteria (bacteria auxotrophic for lactose utilization) reverted to wild type at a greater rate than normal if provided with lactose. Other researchers have shown similar results for other auxotrophs. This would suggest that bacteria can direct mutations to particular genes where the environment favors such a mutation, i.e. the mutation occurs in response to selective pressure. Controversy has surrounded the concept of directed mutations since the results of these experi-



ments were published, and there has been vigorous debate concerning both the mechanism and evolutionary consequences of such a process.

**Random vs directed gene amplification in mammalian cells.** When cultured mammalian cells are treated with drugs that inhibit specific enzymes, resistant cells can be isolated and grown. The resistant colonies can be exposed to progressively greater concentrations of the drug, far in excess of levels that would be lethal to the entire starting population of cells, yet resistant cells can still be isolated. For example, the drug methotrexate is a folic acid analog, and acts as a competitive inhibitor of the enzyme dihydrofolate reductase (DHFR). Wild-type cells are sensitive to  $0.1 \mu\text{g ml}^{-1}$  methotrexate, but by stepwise selection, cells can be isolated which tolerate up to  $1 \text{ mg ml}^{-1}$  methotrexate, five orders of magnitude greater than the wild-type lethal dose. Analysis of the DNA from resistant cells shows that some have simple mutations in the *DHFR* gene, which reduce the sensitivity of the enzyme to the inhibitor, and some have mutations in other genes which limit methotrexate uptake. Most, however, have amplified the *DHFR* locus, and in the highly resistant cells subjected to several rounds of selection, the locus can be amplified thousands of times. In chromosome preparations from such highly selected cells, the amplified region can be seen as an extended chromosome band (a **homogeneously staining region**) or as small extra chromosomes termed **double minutes**. Occasionally, copies of the amplified region translocate to another chromosome.

Superficially, the amplification of drug resistance genes in response to increasing drug dosage looks like a candidate for directed mutation in response to selective pressure. However, only a very few cells from the initial wild-type population ( $1:10^7$ ) survive the first round of selection, and these represent individuals where the *DHFR* locus has been amplified by chance. The amplification of any region of the mammalian genome occurs at a low spontaneous frequency, and the selection procedure enriches the culture for *preexisting* cells with the beneficial amplification. Once isolated and used to generate a resistant colony, some of those cells will undergo further amplification and can be selected in a second round of drug treatment, and so on in a stepwise manner.

The amplification unit (**amplicon**) is often very large ( $>100 \text{ kbp}$ ) and contains much flanking DNA in addition to the *DHFR* locus; DNA from other regions of the genome, including other chromosomes, may also be included. The repeats are not homogeneous: they are different lengths and undergo rearrangements. They are unstable once selective

pressure is removed. The mechanism of gene amplification is therefore not entirely clear, although it may involve very promiscuous and dynamic recombination events or unscheduled replication. The co-amplification of unselected flanking sequences can be exploited for mammalian expression cloning (q.v. *amplification vectors*). Gene amplification is also seen in cancer cells, as a predominant mechanism for the overexpression of proto-oncogenes (see *Oncogenes and Cancer*).

**Programmed amplification in development.** As well as the random amplifications which occur in all cells and can be selected by drug treatment, or by somatic natural selection in cancer, certain amplification events occur in a programmed manner as part of development. Targets for programmed amplification include the rRNA genes of many amphibians, which become excised from the genome as small DNA circles, and the chorion genes of *Drosophila*, which become selectively amplified within the genome. For a further discussion of programmed amplification, see *Development: Molecular Aspects*.

**Somatic hypermutation.** A clear example of programmed mutation is **somatic hypermutation**: the alteration of germline immunoglobulin DNA by the introduction of changes to the nucleotide sequence during B-cell development. In humans and mice, somatic hypermutation occurs specifically in B-cells where the immunoglobulin genes have already been rearranged and expressed: it is the mechanism of **affinity maturation**, i.e. the increase in affinity of an antibody for its specific antigen. In sheep, hypermutation of unrearranged immunoglobulin genes occurs to provide a more diverse primary repertoire of antibodies. It is likely that the initial role of somatic recombination was to generate primary diversity, as lower vertebrates with little combinatorial or junctional diversity carry out somatic hypermutation (q.v. *V(D)J recombination*).

The mechanism of somatic hypermutation is unknown, but a consensus site for hypermutation recruitment has been identified, and several lines of evidence suggest a link with transcription. Most work has concentrated on the mouse Igk locus, and Igk transgenes have been widely exploited in the study of this process because they act as hypermutation substrates. The hypermutation domain of Igk begins within the leader intron upstream of the rearranged V segment, and extends across the V and J segments and into the J-C intron (q.v. *immunoglobulin genes*). However, the mutations are largely restricted to the DNA corresponding to the variable domains and are clustered in the hypervariable regions, corresponding

to the parts of the antibody which actually contact the antigen. The consensus nucleotide sequence RGYW is thought to be a partial hypermutation recruitment site because many (but not all) such sites are local hypermutation hotspots. Investigation of the distribution of serine codons suggests that the germline immunoglobulin genes have evolved to target hypermutation to hypervariable regions. Serine is encoded by two unrelated codon families: AGY (which is part of the hypermutation consensus) and TCN (which is not). There is biased serine codon usage in the immunoglobulin loci, as AGY codons tend to occur in hypervariable regions, and TCN codons elsewhere, whereas TCN codons are distributed throughout the V-regions of the T-cell receptor genes, which do not undergo hypermutation.

The hypermutation domain is located within the

Ig $\kappa$  transcription unit, and hypermutation shows distinct strand polarity. These data suggest that hypermutation is coupled to transcription, perhaps in the same way as *transcription-coupled DNA repair* (q.v.). Further support for a link with transcription comes from transgenic experiments, which have shown that the  $\kappa$  light chain enhancer is required for hypermutation, but that the promoter and most of the V-segment can be replaced by heterologous sequence and still act as a hypermutation substrate. *Trans*-acting factors with an explicit role in hypermutation have not been identified, although a possible candidate is TFIIF — the basal transcription factor with a central role in transcription-coupled DNA repair. A current model suggests that TFIIF could recruit an error-prone DNA polymerase to the locus, which would introduce nucleotide substitutions in the following round of DNA replication.

#### Box 15.4: Second site mutations

**Suppression and enhancement. Second site mutations** are mutations occurring in addition to an initial **primary site mutation** which may modify the phenotype determined by the primary site mutation. When the effect of the first mutation is ameliorated by the second, the phenomenon is termed **suppression**, whereas if it is augmented, the phenomenon is termed **enhancement**. The effects of a primary mutation can also be suppressed by the environment, e.g. streptomycin can mimic *informational suppression* (see below) by reducing the fidelity of translation; this is termed **phenotypic suppression** (also q.v. *phenocopy*). At the level of the phenotype, the consequences of suppression are identical to those of reversion. However, only with suppression can the components of the effect be separated by recombination, i.e. a cross-over between the mutations.

**Suppressors in the same gene. Intragenic or internal suppressors** are second site mutations occurring in the same gene as the primary mutation, in the *cis*-configuration (c.f. *allelic complementation*), and which restore the wild-type phenotype by making good some structural deficiency, e.g. where a primary frameshift is caused by a single nucleotide insertion, a nearby single nucleotide deletion would act as a suppressor to restore the original reading frame. In the special case of **intracodon suppressors**, the second site mutation is in the same codon as the primary mutation and compensates for the effect of the primary mutation by restoring the original sense of the codon or converting a nonconser-

vative change to a conservative change.

**Suppressors in different genes. Intergenic** (also **extragenic** or **external**) **suppressors** are second site mutations occurring in a different gene to the primary mutation. The effect of an intergenic suppressor is not compensatory at the gene level, but at the level of its product (i.e. they suppress functions in *trans*). In some cases intergenic suppressors may compensate physiologically (e.g. a loss of function mutation which prevents synthesis of an essential enzyme, such as one required for tyrosine synthesis, could be compensated by a mutation in a second gene which allows more efficient uptake of tyrosine from the environment). In other cases, intergenic suppressors identify genes encoding interacting proteins, and screens for unlinked suppressor mutants have been widely exploited for this purpose (also q.v. *two hybrid system*). Intergenic suppression is a form of *nonallelic interaction* (q.v.), and also q.v. *complementation*.

**Informational suppressors. Informational suppressors (supersuppressors)** are a class of intergenic suppressors which compensate for missense, nonsense and even small frameshift mutations by introducing a compensatory change in the *anticodon loop* (q.v.) of the corresponding tRNA molecule, thus causing the mutated coding region to be read as it was originally intended. Nonsense suppressors are classed as **amber**, **ochre** and **opal suppressors** depending upon which type of termination codon they interpret as a sense codon. Because tRNA

genes are generally present in many copies, the occurrence of one informational suppressor mutation does not result in the misinterpretation of all stop codons; thus normal termination of wild-type genes also takes place and the organism is viable.

**Enhancer mutations.** A second site mutation which increases the severity of the original mutant

phenotype, a process described as **enhancement**, is termed an **enhancer mutation**. Like suppressors, enhancer mutations can be *cis*-acting and intergenic or *trans*-acting and intergenic, the latter identifying possible interacting gene products. Enhancer mutations should not be confused with *enhancers* (q.v.), which are *cis*-acting regulatory elements.

## References

- Cooper, D.N. and Krawczak, M. (1993) *Human Gene Mutation*. BIOS Scientific Publishers, Oxford.
- Humphries, S. and Malcolm, S. (1994) *From Genotype to Phenotype*. BIOS Scientific Publishers, Oxford.
- Li, W.-H. and Grauer, D. (1991) *Fundamentals of Molecular Evolution*. Sinauer, Sunderland, MA.
- Britten, R.J. (1986) Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**: 1393–1398.
- Cao, A., Galanello, R. and Rosatelli, M.C. (1994) Genotype–phenotype correlations in  $\beta$ -thalassemias. *Blood Rev.* **8**: 1–12.
- Drake, J.W. (1991) Spontaneous mutation. *Annu. Rev. Genet.* **25**: 125–146.
- Miller, J.H. (1983) Mutational specificity in bacteria. *Annu. Rev. Genet.* **17**: 215–238.
- Patel, P.I. and Lupski, J.R. (1994) Charcot-Marie-Tooth disease — a new paradigm for the mechanism of inherited disease. *Trends Genet.* **10**: 128–133.
- Richards, R.I. and Sutherland, G.R. (1997) Dynamic mutation: Possible mechanisms and significance in human disease. *Trends Biochem. Sci.* **22**: 432–436.
- Sniegowski, P.D. and Lenski, R.E. (1995) Mutation and adaptation — the directed mutation controversy in evolutionary perspective. *Annu. Rev. Ecol. Systematics* **26**: 553–578.
- Spencer, D.M. (1996) Creating conditional mutations in mammals. *Trends Genet.* **12**: 181–187.
- Wagner, S.D. and Neuberger, M.S. (1996) Somatic hypermutation of immunoglobulin genes. *Annu. Rev. Immunol.* **14**: 441–457.

## Further reading

**This Page Intentionally Left Blank**



## Chapter 16

# Nucleic Acid Structure

### Fundamental concepts and definitions

- DNA and RNA are **nucleic acids**, polymers composed of nucleotide subunits. Each nucleotide comprises a nitrogenous base linked to a phosphorylated sugar. The sugar residues are covalently joined by 5'→3' phosphodiester bonds, forming a polarized but invariant backbone with projecting bases.
- The nature and order of the bases along the polymer comprises the genetic information carried by nucleic acids. The projecting bases interact specifically with other bases to form complementary pairs, allowing nucleic acids to form duplexes, act as templates and recognize homology, three processes which underpin the essential biological processes of *replication*, *recombination* and *gene expression* (q.v.).
- Duplex nucleic acids adopt different conformations depending on the base sequence, topological constraints, environmental conditions and interaction with proteins. Such **conformational polymorphism** is as important for the function of nucleic acids as the base sequence itself.
- DNA is the genetic material of cells and exists primarily in a double-stranded form — this makes it particularly suitable as a repository of genetic information, a **blueprint**, because it can preserve its integrity by acting as a template for its own repair (see Mutagenesis and DNA Repair). Cellular RNA is transcribed from the DNA and exists predominantly in a single-stranded form, although it usually folds to form complex secondary and tertiary structures. There are several classes of RNA which have distinct functions, mostly concerning the expression of genetic information (Table 16.1). Viral genomes can be composed of either DNA or RNA (see Viruses).

### 16.1 Nucleic acid primary structure

**Nucleotide structure.** Nucleotides are the basic repeating units of nucleic acids and are constructed from three components: a base, a sugar and a phosphate residue. Nucleotides also have many other functions in the cell, e.g. as energy currencies, neurotransmitters and second messengers (see Signal Transduction).

**Bases** are derivatives of the *basic* nitrogenous heterocyclic compounds **pyrimidine** and **purine** (Figure 16.1). DNA and RNA both contain four **major bases**, three of which (the purines **adenine** and **guanine**, and the pyrimidine **cytosine**) are present in both nucleic acids, whilst **uracil** is specific to RNA and **thymine** to DNA. DNA probably evolved to contain thymine to prevent mutations caused by deamination of cytosine to form uracil (however, q.v. *5-methylcytosine*). Both DNA and RNA also contain infrequent **minor bases** (e.g. inosine), which may be incorporated as such or may result from modification after polymerization (q.v. *DNA modification*, *tRNA*, *RNA editing*, *base analogs*). Bases can exist as alternative *tautomeric forms* (q.v.) with different hydrogen bonding potentials, and these are frequent sources of mutations (see Mutagenesis and DNA Repair). The common bases in DNA and RNA are relatively stable in one tautomeric form (the **dominant tautomeric form**), which is probably why they have been selected to carry genetic information.

Both DNA and RNA contain five carbon (**pentose**) **sugars** where the intramolecular formation of a hemiketal group generates a **furanose** ring structure (so-called because of resemblance to the heterocyclic compound **furan**) (Figure 16.1). The essential difference between DNA and RNA is the type of sugar each contains: RNA contains the sugar **D-ribose** (hence **ribonucleic acid**, RNA)

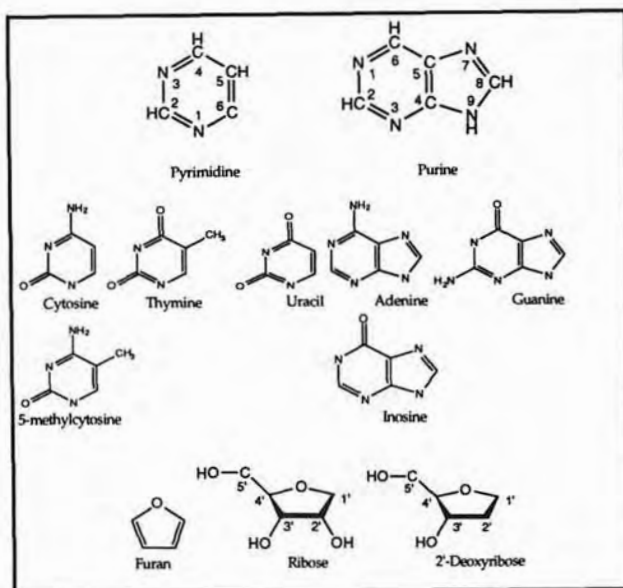
**Table 16.1:** Major and minor functional classes of cellular RNAs

RNA class	Function
<i>Major classes</i>	
<b>mRNA (messenger RNA)</b>	The RNA transcribed from protein-encoding genes which carries the message for translation. Some mRNA-like transcripts are untranslated, e.g. <i>XIST</i> , <i>H19</i> (q.v. <i>parental imprinting</i> )
<b>hnRNA (heterogenous nuclear RNA)</b>	Prespliced mRNA. The unmodified transcripts of eukaryotic genes, so called because of its great diversity of size compared to tRNA and rRNA.
<b>tRNA (transfer RNA)</b>	The adaptor molecule which facilitates translation. tRNA also primes DNA replication during retroviral replication (q.v. <i>retroviruses</i> )
<b>rRNA (ribosomal RNA)</b>	Major structural component of ribosomes, required for protein synthesis
<i>Minor classes</i>	
<b>iRNA (initiator RNA)</b>	The short RNA sequences used as primers for lagging strand DNA synthesis (q.v. <i>replication</i> )
<b>snRNA (small nuclear RNA) or U-RNA (uridine-rich RNA)</b>	Low molecular weight RNA molecules found in the nucleoplasm which facilitate the splicing of introns and other processing reactions. Rich in modified uridine residues
<b>snoRNA (small nucleolar RNA)</b>	Low molecular weight RNA found in the nucleolus, probably involved in the processing of rRNA
<b>scRNA (small cytoplasmic RNA)</b>	Low molecular weight RNA molecules found in cytoplasm with various functions. Examples are 7S RNA which is part of the <i>signal recognition particle</i> (q.v.) and <b>pRNA (prosomeal RNA)</b> , a small RNA associated with approximately 20 proteins and found packaged with mRNA in the <i>mRNP</i> or <i>infosome</i> (q.v.), which may have a global regulatory effect on gene expression
<b>Telomerase RNA</b>	A nuclear RNA which contains the template for <i>telomere</i> (q.v.) repeats and forms part of the enzyme <i>telomerase</i> (q.v.)
<b>gRNA (guide RNA)</b>	An RNA species synthesized in trypanosome kinetoplasts which provides the template for <i>RNA editing</i> (q.v.)
<b>Antisense RNA (mRNA-interfering complementary RNA, micRNA)</b>	Antisense RNA is complementary to mRNA and can form a duplex with it to block protein synthesis. Naturally occurring antisense RNA is found in many systems but predominantly in bacteria, and is termed mRNA-interfering complementary RNA (q.v. <i>plasmid replication</i> , <i>F transfer region</i> , <i>bacteriophage λ</i> , <i>regulation of protein synthesis</i> , <i>gene therapy</i> )
<b>Ribozymes</b>	RNA molecules which can catalyze chemical reactions (RNA enzymes). Usually autocatalytic (q.v. <i>self-splicing introns</i> ), but ribonuclease P is a true catalyst (q.v. <i>tRNA processing</i> ). Other RNAs work in concert with proteins, e.g. MRP endonuclease in mitochondrial DNA replication (see <i>Organelle Genomes</i> )

Most RNAs are linear, but some can be branched (e.g. lariats during intron processing) and some may be circular (e.g. viroids, and possibly *SRY* mRNA; q.v. *sex-determination*).

whereas DNA contains its derivative **2'-deoxy-D-ribose**, where the 2' hydroxyl group of ribose has been replaced by a hydrogen (hence **deoxyribonucleic acid, DNA**). This minor structural difference confers very different chemical and physical properties upon DNA and RNA, the latter being much stiffer due to steric hindrance and more susceptible to hydrolysis in alkaline conditions, perhaps explaining in part why DNA has emerged as the primary genetic material.

**Nucleosides** consist of a base joined to a pentose sugar at position C1'. The sugar C1' carbon atom is joined to the N1 atom of pyrimidines and the N9 atom of purines (*Figure 16.2*); this is a **β-N-glycosidic bond**. The nomenclature of nucleosides differs subtly from that of the bases (*Table 16.2*).

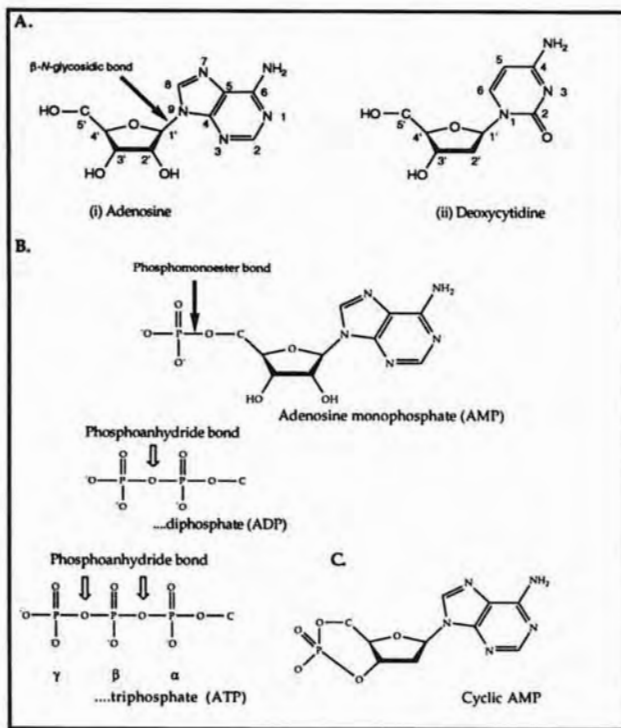


**Figure 16.1:** Bases and sugars in nucleic acids. The major bases cytosine, thymine and uracil are derivatives of the heterocyclic compound pyrimidine, whereas the bases adenine and guanine are derivatives of the heterocyclic compound purine. Note that thymine and uracil have very similar structures — both bases pair in the same manner with adenine (q.v. *complementary base pairing*) so that thymine in DNA is replaced by uracil in RNA. The minor bases of DNA, 5-methylcytosine and inosine, are also shown. The sugars D-ribose and 2'-deoxy-D-ribose are called furanose sugars because of their similarity to the heterocyclic compound furan. Conventional ring numbering systems are shown. The sugar numbering system uses primed numbers to avoid confusion with the base numbering system.

**Nucleotides** are phosphate esters of nucleosides. Esterification can occur at any free hydroxyl group, but is most common at the 5' and 3' positions in nucleic acids. The phosphate residues are joined to the sugar ring by a **phosphomonoester bond**, and several phosphate groups can be joined in series by **phosphoanhydride bonds** (Figure 16.2). Nucleoside 5'-triphosphates are the substrates for nucleic acid synthesis. Two hydroxyl groups can also be esterified by the same phosphate moiety to generate a **cyclic nucleotide**, e.g. cyclic AMP (cAMP, adenosine 3'-5'-cyclic phosphate; see Signal Transduction).

**Nucleic acid primary structure.** Nucleic acids are long chains of nucleotide units, or **polynucleotides**. The substrates for polymerization are nucleoside triphosphates, but the repeating unit, or monomer, of a nucleic acid is a monophosphate (**nucleoside monophosphate residue**, **nucleotidylate residue**, **nucleotide residue**). During polymerization, the 3' hydroxyl group of the terminal nucleotide residue in the existing chain makes a nucleophilic attack upon the (innermost)  $\alpha$ -phosphate of the incoming nucleoside triphosphate to form a **5'→3' phosphodiester bond**. This reaction is catalyzed by enzymes termed *DNA* or *RNA polymerases* (Box 26.1) and pyrophosphate is produced as a by-product (q.v. *DNA replication*, *transcription*). Serial polymerization generates long polymers variously called **chains** or **strands**, containing an invariant **sugar-phosphate backbone** with **5'→3' polarity** and projecting nitrogenous bases. The primary chemical structure of DNA and RNA is shown in Figure 16.3 along with common shorthand notations (also q.v. *PNA*).

**Oligonucleotides** are short nucleic acids (i.e. <100 nt in length). Oligoribonucleotides occur naturally and are used as primers during DNA replication and for various other purposes in the cell. Synthetic oligonucleotides can be made by chemical synthesis and are essential for many laboratory techniques (e.g. q.v. *DNA sequencing*, *polymerase chain reaction*, *in situ hybridization*, *nucleic acid probe*, *nucleic acid hybridization*, *gene therapy*).



**Figure 16.2:** Nucleosides and nucleotides. (a) The chemical structures of two nucleosides showing the glycosidic bond joining the sugar to the base: (i) adenosine, a ribonucleoside containing a purine; (ii) deoxycytidine, a deoxyribonucleoside containing a pyrimidine. (b) The ribonucleotide adenosine monophosphate (adenylic acid) and the phosphate residues of the diphosphate and triphosphate derivatives showing the nomenclature of the phosphate groups. The positions of the monoester and phosphoanhydride bonds are indicated. (c) The cyclic nucleotide cyclic adenosine monophosphate.

## 16.2 Secondary structure of nucleic acids

**Two forms of base interactions.** Nucleic acid secondary structures are generated by two kinds of noncovalent interactions between bases: base pairing and base stacking. **Base pairing** involves hydrogen bonds and is the predominant force causing nucleic acid strands to associate, but the structures are stabilized by hydrophobic interactions between adjacent bases brought about by electrons in  $\pi$  rings. It is these  $\pi$ - $\pi$  interactions which are described as **base stacking forces**. The secondary structure of DNA is characterized by intermolecular base pairing to generate **double-stranded** or **duplex** molecules (**dsDNA**). Secondary structures in RNA, which exist primarily in single-stranded form, generally reflect intramolecular base interactions.

**Complementary base pairing.** Edwin Chargaff first showed that although the relative abundances of the four bases varied considerably in DNA isolated from different organisms, the ratio of adenine to thymine or guanine to cytosine was always the same. Thus **Chargaff's rules** state that  $A = T$  and  $G = C$ . The molecular basis of Chargaff's rules is **complementary base pairing** between adenine and thymine and between guanine and cytosine in double-stranded DNA. This is facilitated by the formation of stable and specific configurations of **hydrogen bonds** between the bases. The regular structure of the DNA double helix, deduced by James Watson and Francis Crick (see below), arises because a purine in one strand is always paired with a pyrimidine in the other. Specific pairing is achieved by reciprocal positioning of hydrogen bond acceptors and donors, two bonds in A:T pairs



**Table 16.2:** Nomenclature of the bases and their nucleoside and nucleotide derivatives (shorthand notations in parentheses)

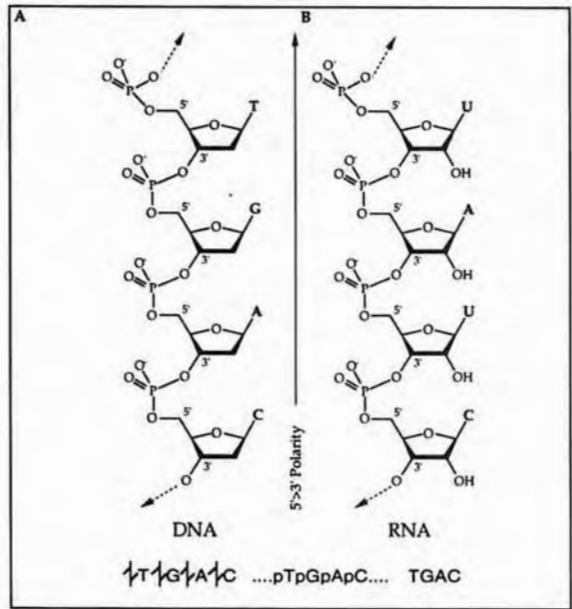
	DNA	RNA
Bases	Adenine (A) Cytosine (C) Guanine (G) Thymine (T)	Adenine (A) Cytosine (C) Guanine (G) Uracil (U)
Nucleosides	Deoxyadenosine (dA) Deoxycytidine (dC) Deoxyguanosine (dG) (Deoxy)thymidine (dT)	Adenosine (A) Cytidine (C) Guanosine (G) Uridine (U)
<i>Nucleotides (adenine derivatives used as the example)</i>		
Nucleoside monophosphates	Deoxyadenosine 5'-monophosphate or deoxyadenylic acid (dAMP, dpA)	Adenosine 5'-monophosphate or adenylic acid (AMP, pA)
Nucleoside diphosphates	Deoxyadenosine 5'-diphosphate (dADP, dppA)	Adenosine 5'-diphosphate (ADP, ppA)
Nucleoside triphosphates	Deoxyadenosine 5'-triphosphate (dATP, dpppA)	Adenosine 5'-triphosphate (ATP, pppA)
Nucleotide residue	Deoxyadenylate	Adenylate

Note that in shorthand notation, nucleoside and nucleotide derivatives of deoxyribose are distinguished by the prefix 'd'. Where clarity is especially important, ribonucleosides and ribonucleotides can similarly be identified with the prefix 'r', e.g. ATP = rATP. Only the second shorthand notation can discriminate between 5' and 3' phosphates, with 5' phosphate residues placed before the base (e.g. pA is adenosine-5'-monophosphate) and 3' phosphates placed after the base (e.g. Ap is adenosine-3'-monophosphate). The deoxy-prefix can be omitted from the names of thymidine derivatives because, as a predominantly DNA-specific base, it is usually evident that sugar is deoxyribose. However, the full nomenclature is preferred for the sake of convention and because thymine is a minor base in RNA. Where context is obvious, both DNA and RNA sequences are represented as a simple series of bases.

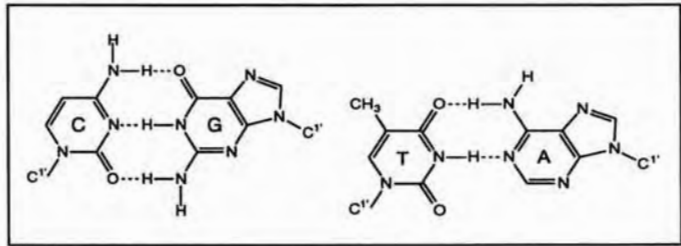
Ambiguous bases: R = A or G (any puRine); Y = C or T/U (any pYrimidine); K = G or T (Keto); M = A or C (aMino); S = G or C (strong — three bonds); W = A or T (weak — two bonds); B = G, T or C (i.e. *not* A); D = G, A or T (i.e. *not* C); H = A, C or T (i.e. *not* G); V = A, C or G (i.e. *not* T).

(donor-acceptor verses acceptor-donor) and three in G:C pairs (acceptor-acceptor-donor verses donor-donor-acceptor) — see Figure 16.4. These **Watson-Crick base pairs** form the basis of most secondary structure interactions in nucleic acids, and as well as explaining Chargraff's rules, they simultaneously demonstrate how DNA can act as a template for *replication* and *transcription* (q.v.), undergo *recombination* (q.v.) and preserve its genetic integrity *by repair* (q.v.). In RNA, uracil replaces thymine, but since uracil has a similar chemical structure to thymine and forms the same hydrogen bonds with adenine, both nucleic acids hybridize according to the same general rules. Ubiquitous as these interactions are, however, there are alternative base pairing schemes playing important roles in the formation of secondary and tertiary structures. These are discussed below.

**Alternative forms of base pairing.** Watson-Crick base pairs are predominant in the structure and function of nucleic acids. However, there are 28 possible arrangements of at least two hydrogen bonds between bases which provide the basis for a diverse set of interactions. The most significant of these alternative configurations are the **Hoogsteen base pairs**, which contribute to tRNA structure and allow the formation of triple helices. A modification to Watson-Crick base pairs are the **wobble pairs**, which allow bases in the 5'-anticodon position of tRNA to pair ambiguously with the mRNA (q.v. *genetic code*, *wobble hypothesis*). The wobble base pairs are formed because bases are offset from their normal Watson-Crick positions, and one of the hydrogen bonds is lost.



**Figure 16.3:** Primary chemical structure of a short sequence of (a) DNA and (b) RNA with 5'→3' phosphodiester bonds indicated. Three shorthand notations for nucleic acid primary structure are also shown.



**Figure 16.4:** Watson-Crick base pairs. Hydrogen bonds are shown as dotted lines. Three hydrogen bonds form in G:C base pairs and two in A:T (or A:U) base pairs. The G:C pairs are therefore the most stable (q.v. GC-content, thermal melting profile).

**The DNA double helix.** The structure of **double-stranded DNA (dsDNA)** was solved by James Watson and Francis Crick in 1953 using X-ray diffraction of DNA fibers prepared by Maurice Wilkins and Rosalind Franklin. Duplex DNA generally exists as a right-handed **double helix**, with two **antiparallel** strands (i.e. with opposite 5'→3' polarities) wrapped round a common axis. The repeating sugar-phosphate units form the helical **backbone**, and the bases project inwards forming hydrogen bonds across the helical axis. The structure is stabilized by hydrogen bonding between base pairs and ( $\pi$ - $\pi$ ) stacking interactions between adjacent bases. The bases are shielded from the environment (protecting the genetic information from physical and chemical attack; see *Mutagenesis and DNA Repair*), and because of base interactions, the outside of the helix is not smooth but has two grooves, a wide **major groove** and a narrow **minor groove** which facilitate sequence-specific protein-DNA interactions (see *Nucleic Acid-Binding Proteins*).

These features of the helix were deduced from long DNA fibres, and are thus general properties which do not take into account local perturbations caused by unusual sequence architecture. Furthermore, whereas this form of the DNA double helix, known as B-DNA, is prevalent *in vivo*, other forms of the helix with distinct structures also exist (see below).

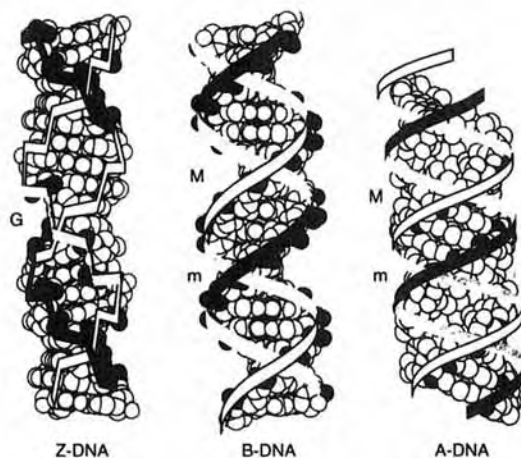
**Helical conformations.** The first investigations of DNA secondary structure demonstrated that alternative helical conformations (**conformers**) formed at different humidities (helical conformations are described in terms of gross morphological features, bond angles and helix parameters, summarized in Box 16.1). **A-DNA** and **B-DNA** helices are both right handed, but A-DNA is wider, incorporates more base pairs per helical turn and is less flexible than the canonical B-DNA described above. This reflects differences in helical twist, base inclination, and displacement of the base pairs from the helical axis resulting from dehydration. The change in conformation alters the shape of the major and minor grooves, potentially influencing the nature of protein–DNA interactions. Under physiological conditions, duplex RNA and RNA:DNA hybrids are thought to adopt an A-form structure because they are inherently less flexible than DNA. The A-form of DNA is less soluble than the B-form, which is why DNA which is overdried during, for example, plasmid preparation is difficult to dissolve.

With improved methods of analysis and the ability to make customized oligonucleotides (see above), a number of different types of DNA helix have been observed. These fall into three major classes, the A-form and B-form described above, and the Z-form (Table 16.3 and Figure 16.5). Early studies of oligonucleotides with alternating purine–pyrimidine sequences revealed the left-handed helical conformation of **Z-DNA**. This structure is characterized by alternating helical parameters and torsion angles with a two-base pair periodicity, causing the backbone of the helix to zig-zag (hence the name Z-DNA). The repeating structural unit of Z-DNA is therefore two base pairs rather than one, as is the case for A-DNA and B-DNA. Although alternating purine–pyrimidine tracts such as oligo-dGdC and oligo-dAdC provide a good substrate for Z-DNA, this sequence specificity is now known to be neither necessary nor sufficient for its formation. Z-DNA structures tend to form in torsionally stressed DNA and are stabilized by dehydration; they may play an important role in the control of gene expression (see Transcription).

**Table 16.3:** Comparison of some morphological features and selected bond torsion angles and helical parameters of the three major types of DNA helix

	Conformation		
	A	B	Z
<i>Morphological characteristics</i>			
Helical sense	R	R	L
Pitch (base pairs per turn)	11	10	12
Major groove	Deep, narrow	Wide	Flat
Minor groove	Broad, shallow	Narrow	Narrow and very deep
Helix diameter	2.3 nm	1.9 nm	1.8 nm
<i>Torsional parameters</i>			
Sugar pucker	C2' <i>endo</i>	C3' <i>endo</i>	Alternating
Glycosidic bond angle	<i>anti</i>	<i>anti</i>	Alternating <i>anti/syn</i>
<i>Helical parameters</i>			
Displacement	−4.4	0.6	3.2
Twist	33	36	−49/−10
Rise	2.6	3.4	3.7
Inclination	22	−2	−7

Note that the low displacement and tilt of B-DNA indicates that the bases sit on the axis and are perpendicular to it (Figure 16.5).



**Figure 16.5:** Structure of the A-DNA, B-DNA and Z-DNA forms of the DNA double helix. The B-form is thought to be the most prevalent *in vivo*. M = major groove, m = minor groove, G = single groove in Z-DNA. (Modified from *DNA Replication*. Kornberg and Baker (1992), WH Freeman, New York.)

**Local flexibility in DNA structure.** The analysis of oligonucleotide crystals as opposed to fibres shows that there is great variation in the helical parameters of molecules with diverse base sequences. This occurs because different base sequences influence helical and torsional parameters to maximize the stability of stacking and pairing interactions. B-DNA is particularly flexible in this respect and different local conformations form to adapt to particular sequences. This indicates that DNA probably does not exist in rigid conformational forms but may change smoothly between different conformations punctuated by local polymorphisms such as bent DNA and **helical transitions** (sudden transitions between different helical conformations within a single molecule, e.g. B–Z transitions). **DNA bending** is an intrinsic property depending on stacking interactions which, according to local sequence, may be **isotropic** (unbiased) or **anisotropic** (bending in a specific direction). Intrinsic DNA bends occur in AT-rich runs and in repeats of the sequence GGCC in step with helical periodicity. DNA bending can also be induced by proteins (see Nucleic Acid-Binding Proteins) and by circularization (q.v. DNA topology). Induced bending is necessary for DNA packaging in chromosomes (see Chromatin) and for *replication*, *recombination* and *transcription* (q.v.). Proteins may also recognise DNA that is bent in a certain way (e.g. topoisomerases).

**Secondary structure in RNA and nonduplex DNA.** In RNA and single-stranded regions of DNA, secondary structure is determined by intramolecular base pairing. Since cellular DNA is usually present as a duplex, the bases are available for intramolecular interactions only rarely. Conversely, intramolecular secondary structures are abundant in cellular RNA and underlie their functional specialization. RNA secondary structures play a major role in gene expression and its regulation: base pairing between rRNA and mRNA controls the initiation of protein synthesis, base pairing between tRNA and mRNA facilitates translation, RNA hairpins and stem loops control transcriptional termination, translation efficiency and mRNA stability, and RNA–RNA base pairing also plays a major role in the splicing of introns (see Transcription, RNA Processing, Protein Synthesis). The major classes of intramolecular nucleic acid secondary structures are listed in Table 16.4. Like DNA, RNA helical conformation is modulated by local sequence character, but the relatively high percentage of modified bases further adds to the variety of structures which form.



**Table 16.4:** Intrastrand nucleic acid secondary structure elements

Secondary structure	Definition
<b>Bulges and bulge loops</b>	Deformities on one side of a duplex which has excess residues. A <b>bulge</b> is caused by a single excess residue, and <b>bulge loops</b> by more than one. These distort stacking of neighboring bases and induce a bend, increasing the accessibility of the major groove. Bulges and bulge loops in double-stranded DNA are caused by insertions (see Mutagenesis and DNA Repair)
<b>Internal loops (bubbles)</b>	Deformities caused by one or more mismatching base pairs in an otherwise duplex structure. In RNA, internal loops or bubbles have been implicated as protein recognition sites
<b>Hairpins, stem loops</b>	Secondary structures which may form in regions of dyad symmetry. The two complementary regions base pair to form a stem which may fold over on itself ( <b>hairpin</b> ) or end in a loop of unpaired nucleotides ( <b>stem-loop</b> , <b>hairpin-loop</b> ). Three and four nucleotide loops are particularly stable structures due to special base pairing and stacking interactions within the loop. Hairpins and stem loops are key functional elements in many biological systems, including, for example, tRNA and rRNA structure, transcriptional termination, control of translation and packaging of viral genomes, suggesting that they are sites for protein–RNA interaction
<b>Panhandle</b>	A discrete linear nucleic acid whose termini are complementary and form a short duplex region, the rest of the molecule forming a loop
<b>Cruciform</b>	In double-stranded nucleic acids with regions of dyad symmetry, a cross-shaped structure which forms when hairpins or stem loops arise simultaneously in both strands. An important source of replication errors, such structures are repressed by single-strand binding proteins

*Lariats* (q.v.) are often classed as secondary structures but, because they are formed by the covalent bonds joining nucleotides, they are strictly primary structures.

### 16.3 Nucleic acid tertiary structure

**Tertiary strand interactions in DNA.** Nucleic acid tertiary structures reflect interactions which contribute to overall three-dimensional shape. This includes interactions between different secondary structure elements, interactions between single strands and secondary structure elements, and topological properties of nucleic acids.

In DNA, tertiary interactions involve single strands interacting with duplexes or duplexes interacting with duplexes, resulting in the formation of triple and quadruple strand structures. Guanine can form **base tetrads**, and DNA containing runs of guanosine residues can form quadruplex structures which may contribute to *telomere* structure (q.v.). Triple-stranded DNA forms spontaneously when a single strand interacts with bases in a duplex molecule; the interactions occur through the major groove and involve the formation of nonWatson–Crick base pairs between the invading strand and one of the resident strands. Studies of the interactions of oligonucleotides with duplex DNA provided the first evidence for triple helices, and identified four common types of **base triples** where Hoogsteen base pairing is involved. **H-DNA** is a form of intramolecular triple-stranded DNA (so-called because it is protonated) which arises in paired homopurine/homopyrimidine sequences and involves Hoogsteen base pairs. The physiological role of H-DNA is unclear, although it is implicated in the regulation of some genes, e.g. *GAP-43* in mammals (also q.v. *triple-helix therapy*). Triple-stranded DNA also forms during recombination when a single strand invades a duplex. Because of topological constraints (see next section) an intact invading strand must pair with the complementary strand in the duplex without winding around it. Such a structure is a **paranemic joint**, and must be stabilized by proteins. It may involve extensive unwinding of the target duplex, or the

formation of alternative segments of left- and right-handed helix (**V-DNA**). If the invading strand has a free end, or further topoisomerase activity, the invading strand winds around its complementary partner in the duplex in the normal fashion to form a **plectonemic joint** — the resident strand is ejected as a **displacement loop (D-loop)**. Similar tertiary structures termed **R-loops** form when RNA transcribed from duplex DNA is stabilized *in situ*, as occurs, for example, during the priming of replication in the ColE1 plasmid (see Plasmids). Four-strand tertiary structures, Holliday junctions, also form during recombination (q.v. *homologous recombination*).

**Tertiary strand interactions in RNA.** RNA folds into complex structures involving tertiary interactions between strands, loops, and duplexes. For example, in tRNA there are examples of base triples, sections of triple helix, **stem junctions** (where two or more duplex regions are joined) and **pseudoknots** (where strands interact with stem-loops).

RNA folding is often controlled by *molecular chaperones* (q.v.) like protein folding. The complexity of RNA tertiary structure allows it to form biologically active molecules, and like proteins, RNA can catalyze biochemical reactions. Such catalytic RNAs are termed **ribozymes**. Some ribozymes are autocatalytic (e.g. the transcripts of *self-splicing introns*, q.v.). Others are *trans-acting*, including ribonuclease P and the family of **hammerhead ribozymes** found in some plant *viroids* (q.v.), so called because of the three-helical structure of the catalytic domain (also q.v. *gene therapy*).

**DNA topology.** Topology is the branch of mathematics dealing with the properties of geometric structures which are independent of size and shape and unchanged by deformation. If a double-stranded DNA molecule has free ends (e.g. a linear molecule), the two strands wind around each other in the most energetically favorable manner, and the molecule is said to be **relaxed**. The number of times one strand winds around the other in this relaxed state is the **duplex winding number**. If extra twists are introduced into such a molecule to make it **overwound**, then the total number of helical turns — which is the **linking number** — exceeds the duplex winding number. Conversely, if twists are removed from the molecule to make it **underwound**, the duplex winding number exceeds the linking number. In either case, the strands can rotate with respect to each other and return the molecule to its relaxed state. In a closed circle, however, there are no free ends and the linking number is a topological property — it can be changed only by breaking the circle open, not by deforming it. If DNA in a closed circle becomes overwound or underwound, the only way to relax the torsional strain thus produced is by **supercoiling**, where a twist is introduced into the helical axis itself. Supercoiling is another form of nucleic acid tertiary structure, one involving the effect of torsional stress upon shape rather than strand-strand interactions (Box 16.2).

The physiological significance of supercoiling is that unconstrained DNA is often biologically inactive. Negative supercoiling is required for many essential processes: replication, transcription and recombination included. Supercoiled DNA has stored energy which drives these reactions. In eukaryotes, which possess linear chromosomes, topological constraints are introduced by organizing chromatin into loops with ends fixed by scaffold proteins; nucleosomes introduce negative supercoils into eukaryote DNA (q.v. *chromatin loops*, *matrix associated region*, *DNA topoisomerase*, *site-specific recombination*).

**Nucleic acid quaternary structure.** In many structures, nucleic acids interact in *trans* (e.g. the ribosome and spliceosome), and this may be considered a quaternary level of nucleic acid structure. Nucleic acids also interact with an enormous number of proteins (e.g. genome structural proteins, transcription factors, enzymes, splicing factors). Many of these proteins have a significant effect on DNA or RNA conformation. Interactions with proteins may be general or sequence specific, and may involve subtle or overt changes in structure. The restriction endonucleases *EcoRI* and *EcoRV*, for instance, both introduce a pronounced kink in the DNA at their recognition sequence which may facilitate their endonucleolytic activity. Proteins of the HMG class appear specifically to bend DNA in order to facilitate interactions between components bound at distant sites. For further discussion of nucleic acid-protein interactions see Nucleic Acid-Binding Proteins.

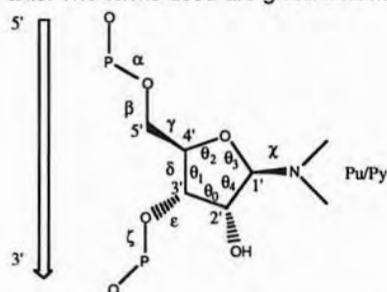
**Box 16.1:** Helix morphology, parameters and torsion angles

**Helix morphology.** DNA and RNA helices are classified according to their gross morphology. Criteria include helical diameter, **helical sense** (direction of rotation), **pitch** (number of base pairs per helical turn) and the width and depth of the major and minor grooves. The grooves were originally defined on the basis of their relative sizes in B-DNA. However, because both can change size under conformational stress, a precise definition is needed. The major groove is defined as that containing the C<sup>4</sup> of a pyrimidine or N<sup>7</sup> of purine, whereas the minor groove contains the O<sup>2</sup> of a pyrimidine or the N<sup>3</sup> of a purine.

**Torsion angles.** Details of conformational structure can be described unambiguously in terms of the torsion angles of the bonds in the sugar-phosphate backbone, in the furanose ring itself, and of the glycosidic bond, as shown in the figure below. Many of these angles are interdependent and conformational descriptions can be abbreviated to the specification of five bonds:  $\delta$  (C<sup>3'</sup>-C<sup>4'</sup>, which is related to sugar pucker),  $\chi$  (the glycosidic bond),  $\gamma$  (C<sup>4'</sup>-C<sup>5'</sup>) and the phosphoester bonds  $\alpha$  and  $\zeta$ . Additionally, these bonds can be described in terms of gross

conformation rather than absolute angle, e.g. by describing the conformation of the glycosidic bond as *anti* or *syn* and by describing the sugar pucker as C<sup>3'</sup> *endo* or C<sup>2'</sup> *endo*.

**Helical parameters.** Details of the positional relationship between stacked and paired bases are described in the terms of a universal set of **helix parameters**. There are translational (displacement) and rotational parameters which describe the relative position of bases in base pairs, the relative position of successive base pairs in base stacks, and the absolute position of base pairs relative to the helical axis. The terms used are given below.



Coordinates	x-axis	y-axis	z-axis
Absolute position relative to helical axis	x-displacement	y-displacement	
Displacement of successive base pairs	Shift	Slide	Rise
Displacement of bases within pairs	Shear	Stretch	Stagger
Absolute rotation relative to helical axis	Inclination ( $\eta$ )	Tip ( $\theta$ )	
Rotation of successive base pairs	Tilt ( $\tau$ )	Roll ( $\rho$ )	Twist ( $\Omega$ )
Rotation of bases within pairs	Buckle ( $\kappa$ )	Propeller twist ( $\omega$ )	Opening ( $\sigma$ )

**Box 16.2:** Quantifying the topological properties of DNA

**Measuring helical winding.** The **linking number** ( $L$ ) is the number of times one DNA strand wraps round the other in a duplex, and for right-handed helices,  $L$  is positive. The **duplex winding number** ( $L_0$ ) is the linking number for relaxed DNA and represents the most energetically favorable configuration. For B-DNA, the average  $L_0 = n/10.3$  where  $n$  is the number of base pairs. If DNA is relaxed DNA,  $L = L_0$ , but any deviation from this state by overwinding or underwinding creates torsional strain. In open DNA (DNA with free ends), the strain is countered by rotation of the strands relative to each other, whereas in covalently closed DNA (circular DNA or DNA with fixed ends) oppositional rotation is

prevented and torsional strain must be countered by supercoiling.

**Measuring supercoiling.** The degree of supercoiling in a given DNA molecule is expressed as the **superhelical density** ( $\lambda$ ), which is calculated as follows:

$$\lambda = \frac{\tau}{L_0} = \frac{L - L_0}{L_0}$$

The **superhelix winding number** ( $\tau$ ) is the difference between  $L$  and  $L_0$ . If DNA is overwound, positive supercoils are introduced and  $\tau$  is positive, whereas underwound DNA generates negative supercoils

*Continued*

and  $\tau$  is negative.  $\tau$  quantifies the degree of torsional strain a given molecule is under and thus its propensity to undergo supercoiling, but it does not measure the actual number of superhelical turns, because the pitch of the helix may also be changed by torsional strain. The number of superhelical turns is expressed as the **writhe number** ( $W$ ). This is related to the linking number in the equation  $L = T + W$  where  $T$ , the **twisting number**, is the total number of turns in a DNA molecule. The linking number is topological (i.e. invariable under deformation) so any change in  $W$ , the number of turns of superhelix, must be countered by an equal and opposite change in  $T$ . In a relaxed molecule,  $L = T$ , hence  $W = 0$  and all turns are helical turns. One unit of writhe is equivalent to one half superhelical turn, i.e. a turn of  $180^\circ$  in the helical axis of the DNA. Each unit of

writhe can be thought of as a point at which two duplexes cross each other when a supercoiled molecule is forced to lie on a flat surface, such a point being described as a **node**.

**Catenation and knotting.** As well as helical winding within a closed molecule, **catenation** (the interlocking of DNA circles) and the formation of **knots** are also topological properties of DNA. In neither case can such structures be resolved without breaking the DNA molecule open to untangle it, and both structures involve nodes where duplexes must cross each other when the molecules are placed flat. The total amount of linking in a given DNA molecule is thus expressed as its linking number  $L$  plus the amount of knotting and catenation, i.e. total linking =  $L + C + K$ .

## References

- Blackburn, G.M. and Gait, M.J. (eds) (1996) *Nucleic Acids in Chemistry and Biology*. Oxford University Press, Oxford.
- Dickerson, R.E. et al. (1989) Definitions and nomenclature of nucleic acid structure parameters. *EMBO J.* 8: 1-4.
- Kornberg, A. and Baker, T.A. (1992) DNA structure and function. In: *DNA Replication*. 2nd edn, pp. 1-52. W.H. Freeman, New York.
- Rich, A. et al. (1984) The chemistry and biology of left-handed Z-DNA. *Annu. Rev. Biochem.* 53: 791-846.
- Yang, Y., Westcott, T.P., Pedersen, S.C., Tobias, I. and Olson, W.K. (1995) Effects of localised bending on DNA supercoiling. *Trends Biochem. Sci.* 20: 313-319.
- Doudna, J.A. and Cate, J.M. (1997) RNA structure: crystal clear? *Curr. Op. Struct. Biol.* 7: 310-316.
- Eaton, B.E. and Pieken, W.A. (1995) Ribonucleosides and RNA. *Annu. Rev. Biochem.* 64: 837-863.
- Frank-Kamenetskii, M.D. and Mirkin, S.M. (1995) Triplex DNA structures. *Annu. Rev. Biochem.* 65: 65-95.
- Jaeger, J.A., Santa-Lucia, J. and Tinoco, I. (1993) Determination of RNA structure and thermodynamics. *Annu. Rev. Biochem.* 62: 255-287.
- Lebrun, A. and Lavern R. (1997) Unusual DNA conformations. *Curr. Op. Struct. Biol.* 7: 348-354.
- Roca, J. (1995) The mechanisms of DNA topoisomerases. *Trends Biochem. Sci.* 20: 156-160.
- Scott, W.G. and Klug, A. (1996) Ribozymes — structure and mechanism in RNA catalysis. *Trends Biochem. Sci.* 21: 220-224.
- Stark, W.M. and Boocock, M.R. (1995) Topological selectivity in site-specific recombination. In: *Mobile Genetic Elements: Frontiers in Molecular Biology* (ed. D. Sherratt), pp. 101-129. Oxford University Press, Oxford.
- Strobel, S.A. and Doudna, J.A. (1997) RNA seeing double: Close-packing of helices in RNA tertiary structure. *Trends Biochem. Sci.* 22: 262-266.
- Travers, A.A. (1995) DNA bending by sequence and proteins. In: *DNA Protein: Structural Interactions: Frontiers in Molecular Biology* (Lilley, D.M.J. ed). Oxford University Press, Oxford, pp. 49-75.
- Weeks, K.M. (1997) Protein-facilitated RNA folding. *Curr. Op. Struct. Biol.* 7: 336-342.

## Further reading



## Chapter 17

# Nucleic Acid-Binding Proteins

### Fundamental concepts and definitions

- Proteins interacting with nucleic acids are considered particularly important because they control some of the most fundamental biological processes, including replication, recombination, DNA repair, transcription, RNA processing and protein synthesis. Protein-nucleic acid interactions fulfill many roles in the cell, but these can be divided into four major categories: (i) structural and packaging roles (e.g. histones in chromatin, HU in the bacterial nucleoid, viral capsid proteins); (ii) transport and localization roles, including DNA segregation and localization in the nucleus, RNA export and localization, and plasmid transfer; (iii) metabolism and rearrangement roles (e.g. DNA and RNA polymerases, nucleases, helicases, DNA repair enzymes, recombinases, topoisomerases); and (iv) gene expression roles (e.g. RNA polymerases, transcription factors, ribosomes and initiation factors, the RNA splicing apparatus, amino acyl-tRNA synthetases). Many of these functions overlap (e.g. histones are packaging proteins which influence gene expression, RNA polymerases facilitate gene expression but are involved in RNA metabolism). All nucleic acids are associated with proteins at some time during their life, and many exist as permanent **nucleoprotein complexes**.
- DNA usually exists as a long double-stranded molecule with a relatively uniform helical structure, whereas RNA is predominantly single-stranded and adopts a range of secondary and tertiary structures. These differences are reflected in the principles and complexities of DNA- and RNA-protein interactions.
- Nucleic acid-binding proteins can be placed into three major categories according to substrate specificity: (i) nonspecific binding proteins; (ii) sequence-specific binding proteins; and (iii) proteins that bind unorthodox structures (e.g. illegitimate bases, recombination intermediates, etc. in DNA and splice lariats in RNA). The mode of interaction may involve binding to end groups, or enclosing the nucleic acid in a cleft or ring. Most nucleic acid-binding proteins, however, interact in a localized manner at an internal site. This often involves a defined secondary structure in the protein which is used as a **recognition element**, and nucleic acid-binding proteins are assigned to families according to the structure of the module containing this element. For DNA-binding proteins, the module often contains an  $\alpha$ -helix which penetrates the major groove. Conversely,  $\beta$ -sheets and other flat surfaces are found more frequently in RNA-binding proteins.
- Proteins may interact with the nucleic acid backbone or the bases, which, in double-stranded molecules, must be accessed through either the major or minor grooves. Many different types of noncovalent bonds are used in protein-nucleic acid recognition, principally electrostatic attractions and hydrogen bonds to the backbone, and hydrogen bonds to the bases. Hydrogen bonds often involve the use of ordered water molecules at the interface. Protein-nucleic acid interaction is usually characterized by conformational changes in both molecules, resulting in the maximization of complementary surfaces, including buried interfaces: van der Waals' forces are important for these interactions. Sequence-specific binding may be achieved by the recognition of bonding patterns displayed by the bases, or by recognition of the conformation of the backbone. Many sequence-specific DNA-binding proteins act as dimers. This increases the sensitivity and specificity of the interaction and may result in cooperative binding. Dimerization can also be used to increase the diversity of recognition and in a regulatory capacity.

## 17.1 Nucleic acid recognition by proteins

**General aspects of DNA recognition.** Cellular DNA exists predominantly as very long double-stranded molecules (dsDNA) whose overall structure, in contrast to that of RNA, is relatively constant (q.v. *double helix*). DNA-binding proteins recognize this overall helical geometry and interact with DNA by forging contacts with the invariable sugar-phosphate backbone, or the bases, which can be accessed via the *major* and *minor grooves* (q.v.). However, the conformation of the helix is polymorphic, differing both according to external conditions (e.g. hydration), and due to local base stacking interactions, which are sequence-dependent. Local structural polymorphism affects the relationship between bases and the helical axis (see Box 16.1 for discussion of helix parameters), and thus influences helical periodicity and the dimensions of the major and minor grooves. The base sequence also determines the intrinsic tendency for DNA to bend. These local variations control the stereochemical relationship between DNA and proteins by governing the spatial organization of bond-forming atoms and the ability of DNA to change conformation upon protein binding.

Proteins binding to double-stranded DNA (dsDNA) may be divided into three categories: (i) those interacting with DNA ends (e.g. DNA ligases, exonucleases); (ii) those enclosing DNA or binding it in a deep crevice (e.g. DNA polymerases, topoisomerases); and (iii) those interacting with the face of the helix. The first two categories consist mostly of DNA-processing enzymes. The final category is the largest, and includes most transcription factors, restriction enzymes, DNA-packaging proteins, site-specific recombinases and DNA-repair enzymes. This category thus includes both general and sequence-specific DNA-binding proteins, and proteins which recognize unorthodox structures such as damaged bases, etc. There are also proteins which interact with single-stranded DNA (ssDNA), e.g. *RecA* (q.v.), *SSB* (q.v.) and capsid proteins of the filamentous bacteriophage M13.

Proteins which interact with the face of the double helix often possess a motif which fits into the major groove. This maximizes the contact area between the two surfaces by creating a buried interface which contributes to the stability of the binding. Penetration of the major groove is important for sequence-specific proteins because it facilitates sequence readout by direct binding to bases. The major groove is more suitable than the minor groove for such interactions because it is larger (and can thus accommodate  $\alpha$ -helices,  $\beta$ -ribbons, strands and loops) and because the pattern of bonds displayed is unambiguous. Conversely, different base pairs generate the same bond pattern in the minor groove (see later in the chapter). Contacts to the phosphate backbone allow general recognition and stabilize structures lodged in the major groove, but backbone contacts are also used for sequence-specific binding, because local base sequence influences the tertiary conformation of the backbone.

DNA-binding proteins are assigned to families according to the structural motif used for DNA recognition (Table 17.1, and see following section). The motif usually contains an  $\alpha$ -helix which penetrates the major groove, but in some families, a  $\beta$ -ribbon or loop is used instead. Interaction involves complementary surfaces, which may change conformation on binding to bring the appropriate chemical groups in the protein and nucleic acid into contact. Hydrogen bonds are the predominant type of interaction involved in recognition specificity, although electrostatic bonds, van der Waals' forces and dispersive forces due to base stacking make important contributions to overall recognition and stability.

**General aspects of RNA recognition.** Unlike DNA, cellular RNA is usually single stranded, but folds to form secondary structures (bulges, hairpins, stem-loops) which act as protein-binding sites, and complex tertiary structures (triple helices, pseudoknots). The tendency for RNA to adopt higher order structure generates great conformational diversity, which allows it to perform catalytic roles in the same way as proteins (q.v. *ribozymes*) and suggests that the principles of RNA-protein recognition may be complex.

As for DNA, some RNA-binding enzymes interact specifically with end-groups, and some enclose their substrate in a channel or cleft. Most RNA-binding proteins, however, interact with

**Table 17.1:** Principle recognition structures in DNA- and RNA-binding proteins. These structures define families of proteins discussed in the following sections. The largest families (the HTH, zinc finger and basic domain families) are further divided into various subfamilies

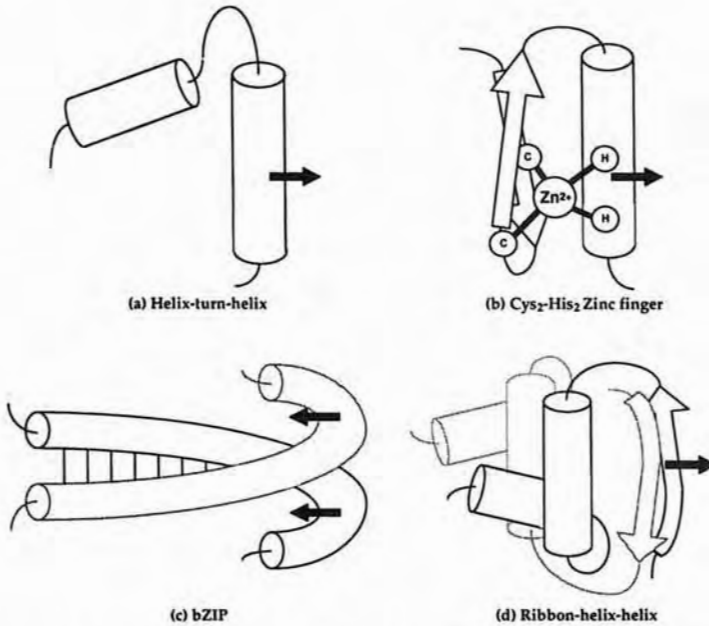
Structure	Recognition	Taxonomic distribution	Examples of solved structures
<b>DNA-binding structures</b>			
Helix-turn-helix	$\alpha$ -helix	All	<i>E. coli</i> Lac repressor, <i>Drosophila</i> Antennapedia
HMG domain	$\alpha$ -helix	Eukaryotes	Rat HMG1
Zinc finger	$\alpha$ -helix	Predominantly eukaryotes	Mouse Zif268 <i>Drosophila</i> Tramtrack
Steroid receptor family	$\alpha$ -helix	Eukaryotes	Human retinoic acid receptor, rat glucocorticoid receptor
Binuclear cluster	$\alpha$ -helix	Yeast	<i>S. cerevisiae</i> GAL4
Basic domain	$\alpha$ -helix	Eukaryotes	<i>S. cerevisiae</i> GCN4, human MyoD, mouse Max
BPV E2 motif	$\alpha$ -helix	Papillomaviruses	BPV E2 protein
Ribbon-helix-helix and other sheet structures	$\beta$ -sheet	All	<i>E. coli</i> MetJ repressor, <i>Arabidopsis</i> TATA-binding protein
Histone — core — linker	Electrostatic $\alpha$ -helix	Eukaryotes	Human TAF <sub>II</sub> 31, TAF <sub>II</sub> 80
Rel homology	Loop	Eukaryotes	Human H5 NF- $\kappa$ B p50 homodimer
<b>RNA-binding structures</b>			
RNP domain	$\beta$ -sheet	Eukaryotes	U1A snRNP
dsRBD	$\beta$ -sheet	All	<i>Drosophila</i> Staufien
K-homology	Loop	Eukaryotes	hnRNP K

<sup>a</sup>Abbreviations: HMG = high mobility group; TAF = TBP (TATA-binding protein)-associated factor; BPV = bovine papilloma virus; RNP = ribonucleoprotein; dsRBD = double-stranded RNA-binding domain; hnRNP = heterogeneous nuclear ribonucleoprotein.

RNA in a localized manner which may be general or sequence-specific. Notwithstanding the potential complexity of RNA-protein interactions, the recent solution of a number of protein-RNA structures has identified some common themes (Table 17.1, and see following sections). In particular,  $\beta$ -sheets are often used as recognition surfaces, perhaps because this allows the exposed RNA bases to be spread out to make appropriate chemical contacts. Where double-stranded RNA is found in intramolecular secondary structures, the 2' hydroxyl group reduces the flexibility of the resulting double helix, which adopts a conformation typical of A-DNA (q.v.). In this structure, the minor groove is wide and shallow whereas the major groove is narrow and deep. Binding to the minor groove is exploited by the capsid proteins of some RNA viruses. However, distortions caused by RNA looping and bending, and by protein binding, can widen the major groove, allowing canonical DNA-binding structures such as the homeodomain also to bind RNA.

## 17.2 DNA-binding motifs in proteins

**The helix-turn-helix motif.** The first DNA-binding structure to be identified, and the one which has been studied in the most detail, is the **helix-turn-helix (HTH) motif**. The prototypical HTH is found in many of the best-characterized gene regulatory proteins of *E. coli* and its phage. The following protein-DNA structures have been solved by X-ray crystallography and/or NMR spectroscopy: Lac repressor, Trp repressor, catabolite activator protein (CAP),  $\lambda$  repressor,  $\lambda$  Cro, phage 434 repressor,



**Figure 17.1:** Common DNA-binding motifs. (a) A helix-turn-helix motif from bacteriophage  $\lambda$  repressor. (b) A Cys<sub>2</sub>His<sub>2</sub> zinc finger module from *X. laevis* Xfin. (c) A basic  $\alpha$ -helix with leucine zipper/helix-loop-helix dimerization motif (bZIP/HLH) from mouse Max (lines represent leucine-leucine bonds in the coiled coil dimerization interface). (d) The ribbon-helix-helix motif from bacteriophage P22 Arc repressor. Arrows represent recognition structures, which penetrate the major groove and form hydrogen bonds with bases. Dimers are shown in two shades.

phage 434 Cro and PurR (q.v. *bacteriophage λ*, *lac operon*, *transcriptional regulation-bacteria*). Many other HTH protein structures have been solved in the absence of DNA.

The motif comprises two  $\alpha$ -helices, separated by a short  $\beta$ -turn, allowing the helices to pack together through hydrophobic interactions (Figure 17.1). The first helix stabilizes and exposes the second, which interacts with the major groove of DNA to make sequence-specific contacts with the bases. The second helix is thus termed the **recognition helix**, although both helices, and often other residues in the domain containing the HTH, make contacts with the DNA. These anchor the recognition helix in position and stabilize the DNA conformation to control the affinity of different proteins for their binding sites.

Classical bacterial HTH proteins have a highly conserved turn architecture, comprising four residues with glycine at the second position. In some, nonconventional bacterial HTH proteins (e.g. *LexA* (q.v.) and *AraC*), this constraint is relaxed. Eukaryotic HTH proteins have more variable turns or loops, and can be divided into a number of families on this basis, and due to overall sequence and structural conservation. The largest and best-characterized is the **homeodomain** family. The homeodomain was first identified in the *Drosophila* Antennapedia protein and occurs predominantly in transcription factors controlling homeostasis and regional specification in development (q.v. *Hox genes*). However, homeodomain proteins with roles in DNA repair and chromatin structure have also been described. The homeodomain is a conserved domain of 60 amino acids which adopts a structure containing four  $\alpha$ -helices. Helices II and III are separated by a  $\beta$ -turn and lie at right angles, forming a HTH motif, with helix III acting as the recognition helix.

Closely related HTH-containing structures include the POU-specific domain and the Paired domain. The POU transcription factors contain two HTH motifs, one in the homeodomain and the second forming a **POU-specific domain** which is remarkably similar to the HTH structure of the  $\lambda$



repressor. As the homeodomain and POU-specific domain are separated by a helix-spanning linker arm, POU transcription factors bind simultaneously to opposite sides of the DNA helix. Either domain can bind to DNA alone, but together, the stability and specificity of binding are increased. The POU-specific domain is so-called because it was first discovered in the vertebrate Pit-1, Oct-1 and Oct-2 transcription factors, and the *C. elegans* protein Unc-86.

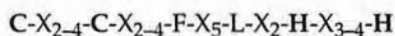
The Paired family of homeodomain proteins (which includes *Drosophila* Paired and Gooseberry, and the many vertebrate Pax — Paired box — transcription factors) may also contain two HTH motifs. The Paired family is identified by a highly conserved **Paired homeodomain**, but a subset of Paired-related proteins contain a second HTH motif and a module homologous to Hin recombinase, which together comprise the distinct **Paired domain**.

Other groups of eukaryotic HTH proteins include the large family of proteins related to the yeast heat shock regulator, and proteins of the HNF3/Fork head family of so-called **winged-helix** transcription factors. The **high mobility group (HMG) domain**, which is found in chromatin structural proteins (see Chromatin) and transcription factors such as the testis-determining factor Sry, also comprises a series of  $\alpha$ -helices separated by turns. However, the geometry of the folding is distinct from that of the canonical HTH motif, and HMG proteins may interact with the minor groove of dsDNA. They are therefore considered to be a DNA-binding protein family which is distinct from the HTH proteins.

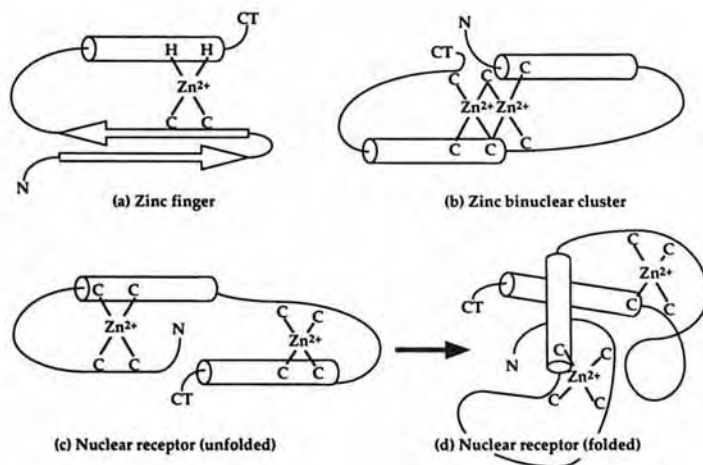
The bacterial and eukaryotic HTH motifs are similar in general topology, with the second helix serving as the recognition helix and the remainder of the protein serving to present the recognition helix to the major groove of DNA (as well as making additional DNA contacts). The recognition helices are distinct, however, with respect to their orientations in the major groove. The N-terminal region of the bacterial recognition helix makes the critical contacts, whereas it is the central region of the homeodomain recognition helix that fulfills this function. Furthermore, the orientation of the bacterial recognition helix varies widely (e.g. lying lengthways along the major groove in the case of the  $\lambda$  repressor, but almost perpendicular to it in the case of the Trp repressor), but the orientation of homeodomain recognition helices is highly conserved. Bacterial HTH proteins appear to function universally as homodimers, recognizing palindromic DNA sequences. Until recently, it was thought that homeodomain-containing proteins acted as monomers at asymmetric sites. However, there is increasing evidence for homeodomain dimerization *in vivo*, and the structures of several homeodomain homodimer-DNA and heterodimer-DNA complexes have been solved (e.g. the *Drosophila* Paired homodimer, and the heterodimer of the yeast mating type regulators MAT $\alpha$ 1 and MAT $\alpha$ 2). Homeodomains may also act as dimers with nonhomeodomain proteins (e.g. MAT $\alpha$ 2 dimerizes with MCM1, and a number of *Drosophila* homeodomain cofactors, e.g. Teashirt, have also been identified).

**The cysteine-histidine zinc finger.** The **zinc finger** was the first eukaryotic DNA-binding structure to be identified and is one of the most common protein modules known, accounting for up to 0.5% of the coding sequence of eukaryotic genomes (q.v. *protein families*). Zinc-coordinating proteins are structurally diverse and can be divided into at least six major families according to zinc coordination mechanism and domain structure (Figure 17.2 and see following sections). Conversely, zinc fingers are rare in bacteria: the *E. coli* DNA repair protein MutM is one example of a bacterial protein containing a zinc-coordinating module.

The **Cys<sub>2</sub>His<sub>2</sub> finger**, contains a pair of cysteine residues and a pair of histidine residues which coordinate a zinc ion at the base of a loop of approximately 12 amino acids. The loop also contains some conserved hydrophobic residues and a number of basic residues. The core consensus sequence is



The loop projects from the surface of the protein (hence 'finger') and basic residues at its tip interact with DNA at the major groove. The structure adopted by the finger is thought to be a



**Figure 17.2:** Structure of three major classes of zinc-coordinating structure. (a) Classical C<sub>2</sub>H<sub>2</sub> zinc finger. (b) C<sub>6</sub> Binuclear cluster of GAL4. (c) 2x C<sub>4</sub> finger of the nuclear receptors, unfolded to show topology (the transcription factor GATA-1 has a single C<sub>4</sub> finger of similar organization) and (d) folded to show domain structure, with the N-terminal recognition helix uppermost.  $\alpha$ -helices are represented by cylinders and  $\beta$ -strands by arrows. C = cysteine, H = histidine, N = N-terminus, CT = C-terminus.

$\beta$ -hairpin with an adjacent  $\alpha$ -helix (containing the histidine residues), the former interacting with the DNA backbone and holding the helix in the major groove, where sequence-specific contacts are made (Figure 17.1). Protein function is abolished by removing zinc, or by mutating one of the conserved cysteine or histidine residues and preventing zinc coordination, indicating that the zinc is essential for adoption of the correct structural conformation.

The first zinc finger motifs were identified in the *Xenopus laevis* basal transcription factor TFIIIA. This protein has nine fingers in tandem and multiple fingers connected by flexible linkers are a general feature of this family of DNA-binding proteins. The mouse transcriptional regulators Sp1 and Zif268 have three fingers, whereas the *Drosophila* transcription factors Tramtrack, Krüppel and Hunchback have two, four and six, respectively. A remarkable *Xenopus* protein, Xfin, has no less than 37 fingers!

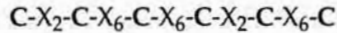
The structures of several zinc finger protein–DNA complexes have been solved and show that multiple fingers often make serial contacts with DNA bases by following the helical path of the major groove. Each finger recognizes a group of three bases and recognition appears to be mediated by hydrogen bonds (e.g. Zif268 forms 11 hydrogen bonds within its 9 bp recognition sequence). Simple rules are now emerging which may provide a complete amino acid–DNA base recognition code, at least for Zif268-like fingers (see later in chapter). In some finger proteins, although the overall geometry of DNA interaction is conserved, not every finger takes part in base-specific recognition, and some fingers act primarily as spacers and do not contact the DNA at all.

**Cysteine-only zinc fingers.** The classical zinc finger module contains a single zinc ion coordinated by two cysteine and two histidine residues. Most other zinc-containing DNA-binding domains use cysteine residues alone, and are sometimes termed **multicysteine zinc fingers**.

The largest family of multicysteine zinc finger proteins is the steroid/thyroid hormone **nuclear receptor family** of transcription factors. Each protein contains two fingers within which the zinc ion is coordinated by four cysteine residues, and each finger forms over the N-terminal end of an extended  $\alpha$ -helix (Figure 17.2). The fingers lack the conserved hydrophobic residues found in the Cys<sub>2</sub>His<sub>2</sub> finger and contribute to a single, globular, DNA-binding structure comprising

perpendicular  $\alpha$ -helices (Figure 17.2). The N-terminal helix is the recognition helix, whereas the second appears to be primarily concerned with dimerization. The steroid receptors act as homodimers and bind to palindromic target sites in DNA, whereas the thyroid hormone, vitamin D and retinoic acid receptors act predominantly as heterodimers with another member of the family, the retinoid X receptor (RXR), and bind to direct repeats. Some nuclear receptors are also capable of binding as monomers, at least *in vitro*.

Other multicysteine zinc fingers include the haemopoietic transcription factor GATA-1, proteins related to the yeast transcriptional regulator GAL4, and a human elongation factor. GATA-1 is structurally very similar to the nuclear receptor family, although it possesses only a single zinc-binding module. Conversely, the structure of GAL4-related proteins is unique and so far appears to be restricted to yeast: six cysteine residues coordinate two zinc ions, so that the central cysteine residues contribute to the coordination of both ions (Figure 17.2). This structure is sometimes termed the **zinc binuclear cluster**, and has the following consensus sequence:



The zinc coordinating structures considered above help to present an  $\alpha$ -helix to the major groove of DNA. However, in the human elongation factor, a  $\beta$ -sheet makes contact with the DNA. The p53 tumor suppressor protein uses an  $\alpha$ -helix to penetrate the major groove, but the zinc coordinating structure appears to stabilize a loop of residues which are presented to the minor groove.

**The basic binding domain.** A large family of eukaryotic transcription factors share a highly basic  $\alpha$ -helix which is used as the principle DNA recognition structure. In most cases, this helix is directly linked to a dimerization domain, either a leucine zipper (hence **basic leucine zipper, bZIP**) or a helix-loop-helix (hence **basic helix-loop-helix, bHLH**), or both. These structures form coiled coils, and hold the basic domains in a conformation which allows them to interact with opposite sides of the DNA. Functional diversity and regulation in this DNA-binding family is facilitated by the ability of different family members to form homodimers and heterodimers (see later in this chapter).

The recognition helix may form as a consequence of DNA binding: studies of the yeast bZIP protein GCN4 suggest that the basic domain adopts a disorganized, partial helical structure in solution but undergoes a conformational change that induces the formation of a typical  $\alpha$ -helix as it binds DNA. In GCN4 dimers, the recognition helices are rigid structures which glance off the DNA and contact only a few bases. In the mouse Max protein, however, the helices are bent at the center and fold around the DNA in a scissor-grip, following the major groove and establishing more extensive contacts (Figure 17.1). Mutational analysis has shown that the dimerization domains, as well as the basic domains, are required for DNA binding. This is because residues in the dimerization domains can make contacts with DNA, and because dimerization is required for high affinity DNA binding; a pair of lone basic domains artificially joined by a disulfide bond can successfully bind to DNA. Some basic domain proteins lack zipper/HLH dimerization motifs (e.g. the *Drosophila* protein Mastermind). Furthermore, the zipper is used as a dimerization motif in some HTH proteins (e.g. yeast heat shock factor) and some zinc finger-containing factors (e.g. GAL4).

**Histone-like binding motifs.** Eukaryotic DNA is packaged into repeating structural units termed nucleosomes through interaction with histones (see Chromatin). The nucleosome core particle comprises two negative superhelical turns of bent DNA wrapped round a histone octamer. The octamer consists of two copies each of the core histone proteins H2A, H2B, H3 and H4. The interaction involves AT-rich DNA presenting the minor groove to the surface of the histone octamer, and wrapping around it. Primary sequence alignments reveal only low level identity between histones (< 20%) but there is a conserved structural motif, the **histone fold**, comprising two fused helix-strand-helix structures, which mediates dimerization. Linker histones (e.g. H1, H5) are involved in

higher order chromatin structures (q.v. *30 nm fiber*) and are thought to seal the core particle by binding to DNA entering and leaving the core.

A number of transcription factors are homologous to various core histone proteins, including many TAFs (q.v.) associated with basal transcription factor IID (TFIID). For example, human TAF<sub>II</sub>31 is homologous to H3 and TAF<sub>II</sub>80 is homologous to H4. TAF<sub>II</sub>20 is homologous to H2B, but the basal apparatus appears to lack an H2A homolog. TAF<sub>II</sub>20, TAF<sub>II</sub>31 and TAF<sub>II</sub>80 are thought to assemble into a histone octamer-like quaternary structure, with TAF<sub>II</sub>20 forming homodimers. Furthermore, there is also evidence that TFIID induces negative supercoils in DNA. The negative regulator DR1, which inhibits TFIID and is blocked by TFIIA (q.v. *transcriptional initiation — RNA polymerase II*), is homologous to histone H2A. DR1 acts as a dimer with a protein called DRAP1, and the dimer resembles the histone H2A/H2B dimer.

The solved structure of the linker histone H5 identified a helix-turn-helix motif in its globular domain. As discussed above, this motif is found in many transcription factors of both prokaryotic and eukaryotic origin. The histone HTH motif is embedded within a larger domain comprising several additional helices, strands and loops. A similar topology is seen in the winged helix domain of eukaryotic HTH transcription factors belonging to the HNF3/Fork head family, and in the *E. coli* protein BirA which is the repressor of the *bio* operon.

**The papillomavirus E2 protein.** This protein has a unique structure comprising four  $\beta$ -strands and a projecting  $\alpha$ -helix. The four strands form a curved  $\beta$ -sheet which acts as a dimerization interface. The dimeric protein thus has a domed, eight-stranded  $\beta$ -barrel with two projecting  $\alpha$ -helices. DNA is bent uniformly around the protein, and the  $\alpha$ -helices sit in two successive major grooves, making base-specific contacts.

**Recognition by  $\beta$ -strands.** In most of the proteins discussed above, DNA sequence recognition involves an  $\alpha$ -helix presented to the major groove. Conversely, a **ribbon-helix-helix motif** is found in three bacterial repressors (MetJ, Arc and Mnt) and represents a structure where recognition is mediated instead by a  $\beta$ -sheet (a ribbon is a two-strand antiparallel  $\beta$ -sheet which fits into the major groove of DNA). The three repressors act as dimers, each monomer contributing a single  $\beta$ -strand stabilized by two  $\alpha$ -helices. The  $\beta$ -strands become arranged as a ribbon by quaternary interactions, lay along the major groove and facilitate recognition through the formation of multiple hydrogen bonds (*Figure 17.1*), i.e. dimerization is required to construct the DNA binding domain. A similar mode of action has been predicted for the bacterial nucleoid packaging protein HU.

In eukaryotes, *TATA-binding protein (TBP)* (q.v.), which is responsible for the basal transcription of all eukaryotic genes, also interacts with DNA through an antiparallel  $\beta$ -sheet. Remarkably, this protein is a single polypeptide but consists of two domains which are structurally very similar and form a symmetrical C-shaped molecule which sits over the DNA like a saddle. The residues of the  $\beta$ -strands on the concave surface are predominantly those which interact with DNA, whereas those in the  $\alpha$ -helices of the outer convex surface interact with other basal transcriptional components. Unlike the bacterial ribbon repressors, the  $\beta$ -strands in TBP interact with DNA through the minor groove.

**The Rel family.** The Rel family of eukaryotic transcription factors responds to a variety of environmental stimuli as well as having important roles in development. Examples of Rel transcription factors include *Drosophila* Dorsal and mammalian NF- $\kappa$ B. Rel proteins contain a conserved structure, the **Rel homology domain (RHB)**, which is about 300 amino acids in length and comprises two immunoglobulin folds joined by a flexible hinge. One of the folds is a dimerization interface and the other directs sequence-specific DNA interactions, although residues from both folds may contact DNA. Rel proteins can form homodimers and various heterodimers, but show dimerization preferences. Homodimers and heterodimers bind to different DNA sites, but the interactions are quite promiscuous. The combination of binding and dimerization promiscuity, and cell type- and



developmental-specific expression contribute to a complex regulatory network. Some Rel proteins also bind DNA as monomers.

Dimeric Rel proteins include the archetypal transcription factor NF- $\kappa$ B which is composed of two subunits — p50 and p65 (RelA). The structure of the p50 homodimer-DNA complex has been solved and shows that DNA interactions are mediated predominantly by a long loop which lies along the major groove. Five residues in the loop make contacts with five sequential base pairs. A further residue outside the loop also makes a sequence-specific contact with DNA.

**Binding motifs in DNA processing enzymes.** DNA-processing enzymes use a variety of structures to interact with their substrates. Enzymes which catalyze synthesis or topological modification often enclose DNA in a groove or ring. The first such structure to be solved was that of the Klenow fragment of *E. coli* DNA polymerase I. This enzyme has a deep groove which first binds DNA and then folds around to enclose it. *E. coli* DNA polymerase III uses a different strategy: the  $\beta$ -subunit, which acts as a sliding clamp, is a dimer of crescent-shaped subunits and forms a ring around DNA. The eukaryotic counterpart, PCNA (q.v.), functions in the same manner, but is trimeric. Clefs are also found in other polymerases, e.g. HIV reverse transcriptase, and RNA polymerase II, which presumably bind their substrates in similar ways. The high resolution structure of the HIV reverse transcriptase-nucleic acid complex shows that contacts are made to the phosphate backbone, and through a loop region embedded in the major groove.

Enzymes with simpler tasks, such as endonucleases, interact with the face of the helix as discussed for the protein families in the preceding sections. Most possess structures that penetrate the major and/or minor grooves, but their interactions are often more extensive than those of e.g. transcription factors and the binding domains partially enclose the helix. The structures of several restriction enzyme-DNA complexes have been solved and reveal a number of different DNA-binding motifs. However, a common theme is that restriction enzymes tend to introduce radical distortions to the DNA. *EcoRI* introduces a sharp kink, widening the major and minor grooves at the cleavage site and allowing a large, four-barelled  $\alpha$ -helix to sit in the major groove while extended arms follow it around the helix. Conversely, *EcoRV* uses loops to interact with the major groove of its target site, as does the nonspecific nuclease DNase I.

### 17.3 RNA-binding motifs in proteins

**The RNP domain.** The best characterized RNA binding domain, the RNP (ribonucleoprotein) domain, was first identified in the yeast *polyadenylate-binding protein* (q.v.) and has since been found in about 250 further RNA-binding proteins. There are two conserved motifs, RNP1 and RNP2, which are critical for RNA binding specificity. They form part of a larger structure comprising four  $\beta$ -strands and two  $\alpha$ -helices with the topology  $\beta$ - $\alpha$ - $\beta$ - $\alpha$ - $\beta$ . The RNP1 and 2 motifs are located on the central  $\beta$ -strands of a four-strand  $\beta$ -sheet.

U1A spliceosomal protein contains two such domains and binds to a hairpin in U1 snRNA. The structure of this RNA-protein complex has been solved. The  $\beta$ -sheet forms a nonspecific RNA-binding platform, and interactions between the U1 snRNA hairpin and a highly negatively charged loop of residues joining two of the  $\beta$ -strands in the sheet stabilizes the RNA conformation so that it is able to interact specifically with the RNAP 1 and 2 motifs.

**The dsRNA-binding domain.** The double-stranded RNA-binding domain (dsRBD) is a short (~65 residue) module found in several proteins which bind double-stranded RNA (e.g. adenosine deaminase, dsRNA-dependent kinase, *Drosophila* Staufen and *E. coli* RNase III). The topology is similar to the RNP domain (three  $\beta$ -strands which form a sheet, and two  $\alpha$ -helices, with the topology  $\alpha$ - $\beta$ - $\beta$ - $\alpha$ ), although in this case it appears that specific binding interactions occur at a cleft in the domain between the second  $\alpha$ -helix and the  $\beta$ -sheet.

**The K-homology domain.** The K-homology domain was first identified in the heterogeneous nuclear RNA-binding protein hnRNP K, which contains three copies of the domain. It is found in several other RNA-binding proteins including the product of the fragile-X gene *FMR1*. Like the dsRBD, the K-homology domain is centered around a three-strand  $\beta$ -sheet, but there are three  $\alpha$ -helices (with the topology  $\beta$ - $\alpha$ - $\alpha$ - $\beta$ - $\alpha$ ) and specific interactions are thought to occur between RNA and residues located in the linker between the first two  $\alpha$ -helices.

**Other RNA binding structures.** Some of the eukaryotic DNA-binding proteins discussed above have also been shown to bind RNA. The zinc finger protein TFIIIA interacts with 5S rRNA as a form of protein synthesis regulation, while the *Drosophila* homeodomain protein Bicoid represses synthesis of Caudal protein by binding to *caudal* mRNA. The major groove of RNA is too narrow to accept the conventional recognition helix presented by these two proteins, so it is possible that distinct residues make DNA and RNA contacts, or that RNA is severely distorted at the recognition site to allow penetration of the major groove. The recently solved structure of ribosomal protein L11 also shows a HTH RNA binding motif similar in structure to a homeodomain, and HIV Rev is known to interact with a widened major groove using an  $\alpha$ -helix. The  $\alpha$ -helix is also the primary recognition motif in the ColE1 plasmid-encoded Rom protein (see Plasmids). In this case, however, a four-helix bundle is used in the same way as a  $\beta$ -sheet to provide a flat surface for splaying and reading RNA bases. Conversely, in the HIV Tat protein, a  $\beta$ -ribbon is used to penetrate the *Tar* major groove sideways to facilitate base sequence reading. Many nucleic acid-binding structures are also used occasionally for protein-protein interactions.

## 17.4 Molecular aspects of protein-nucleic acid binding

**Direct interactions between nucleic acids and proteins.** The four ways in which nucleic acids and proteins can interact directly are: (i) protein side chains with bases (the major form of sequence specific interaction); (ii) protein main chain (e.g. amide groups) with bases; (iii) protein side chains with phosphate backbone; and (iv) protein main chain with phosphate backbone.

Many different types of noncovalent interaction occur between proteins and nucleic acids: hydrogen bonds, van der Waals' forces, hydrophobic attractions, global electrostatic forces of attraction and repulsion and specific electrostatic bonds. These are complemented by hydrogen bonds and base stacking interactions within the nucleic acid itself, and forces of attraction and repulsion between amino acid side chains.

Contact between proteins and the nucleic acid phosphate backbone often involves electrostatic attraction between the negatively charged phosphate groups and positively charged amino acid side chains such as lysine and arginine. Hydrogen bonds also form between the backbone and side chains or amide groups of the main chain. As well as direct protein-backbone contacts, long-range electrostatic attractions may be important in protein-nucleic acid interactions. The binding affinities of some transcriptional regulators have been shown to increase through the replacement of neutral residues with basic residues, even when these are too far away to make direct contacts with DNA. Backbone contacts are often used for nonspecific binding, but particular base sequences influence the tertiary structure of the backbone which can be recognized and interpreted by proteins, allowing such contacts to be used for sequence-specific interactions.

Contact between proteins and nucleic acid bases, either openly (in single strands) or through the major and minor grooves (in duplex molecules) is predominantly mediated by hydrogen bonds formed between the bases and amino acid side chains. Contacts may show a one-to-one correspondence, but several side chains can interact with one base pair, and long side chains can lie across several base pairs. Sequence-specific nucleic acid-binding proteins tend to maximize their contacts using buried surfaces, e.g. an  $\alpha$ -helix inserted into the major groove. The interaction of protein side chains with DNA bases is the most important for sequence-specific recognition but not all such contacts are specific. For example, in the interaction of the glucocorticoid receptor with DNA, extensive base-side chain contacts are made which contribute to nonspecific recognition.

**The role of water in nucleic acid-protein interactions.** Ordered water molecules are often found at nucleic acid-protein interfaces, either as space fillers or actually taking part in the formation of bonds. In nonspecific-binding proteins, water is often found as an insulator, lining internal surfaces and allowing scanning, thus promoting *nonspecific* interaction. Water can also participate in hydrogen bond formation between proteins and the backbone or bases of DNA and RNA. In some cases, water may be required for base recognition, e.g. in the Trp repressor-DNA complex and in the complex of tRNA<sup>Gln</sup> and glutamyl-tRNA synthetase. It is now thought that water may be responsible for the affinity and specificity of many binding reactions. When *Bam*HI endonuclease interacts with its target, four ordered water molecules mediate contacts between protein side chains and bases. Water can thus act to extend the network of hydrogen bonds formed between nucleic acids and proteins, and increase the specificity of the recognition. In some protein-nucleic acid complexes (e.g. the zinc finger protein Tramtrack bound to its target site), such water molecules are trapped in a particular conformation and the polar interface is rigid. In others (e.g. the homeodomain protein Antennapedia bound to its target site), water molecules can flip, allowing amino acid side chains to adopt different conformations and switch between two or more alternative bonding states. Such fluidity at the interface is important because it increases the adaptability of the protein in the face of mutation (either of the protein itself or of its target site), and because the interface is partially disordered, increasing the entropy of the system. While water plays a major role in backbone and major groove contacts, interactions at the minor groove (e.g. involving HMG proteins, or the basal transcription factor TBP) are predominantly hydrophobic, and water appears to be excluded from the interface.

**Protein modulation of nucleic acid structure.** Particular base sequences in DNA may promote intrinsic *anisotropic bending* (q.v.) and *breathing* (q.v.), but DNA is also distorted by its interaction with proteins. Proteins can alter DNA structure by melting bases, under- or overwinding and bending the helix. Most DNA-binding proteins introduce some structural changes in DNA, and undergo conformational changes themselves, resulting in a better fit between the two surfaces. However, the role of DNA conformation goes beyond the moulding of complementary surfaces, and may control the selective binding and/or activity of sequence-specific DNA-binding proteins.

The most dramatic bends and kinks are introduced by restriction enzymes, as DNA untwisting is required to widen and expose the major groove to the catalytic cleavage centre. *Eco*RI binds to its restriction site in DNA with great selectivity, whereas *Eco*RV binds to most DNA sequences with the same affinity. In the case of *Eco*RI, substrate selectivity arises at the binding stage due to sequence-specific base contacts, whereas for *Eco*RV, substrate selectivity arises at the transition stage when only certain sequences can be untwisted and kinked. Transcription factors may also exploit DNA flexibility. The looping of DNA around a transcription complex or *enhanceosome* (q.v.) is one example: this brings appropriate regulatory proteins into contact and often involves dedicated DNA-bending proteins of the HMG family. The intrinsic flexibility of DNA plays an important role in chromatin structure (q.v. *nucleosome*, *nucleosome phasing*).

**Dimerization and cooperativity.** Many transcription factors, restriction enzymes and other sequence-specific DNA-binding proteins act as dimers or higher order multimers. In some cases, this is required to create the DNA-binding interface (as in the case of ribbon-helix-helix repressors) whereas in others, each monomer has an independent recognition structure, but dimerization increases the specificity and sensitivity of binding. This occurs in several ways: (i) by increasing the size of the binding protein, and thus increasing the number of contacts made and bases recognized; (ii) by increasing the surface area of protein in contact with DNA, increasing the stability of binding; and (iii) increasing affinity by productive protein-protein interactions; such cooperative binding is seen in many transcription factors, including those of the steroid receptor family (where cooperatively can be modulated by altering the spacing of the repeat recognition sequences). Dimerization also increases the



functional diversity of the proteins (i) by using heterodimers to increase the range of target sites recognized; and (ii) by using inactive monomers in a regulatory capacity.

Most transcription factors, etc. act as dimers, although the range of interactions is diverse. Some (e.g. bacterial HTH proteins, steroid receptors) function as homodimers and recognize palindromic sites in DNA. Others (e.g. bZIP proteins) can function as selected homodimers and heterodimers within the family. Homeodomain proteins can act as selected homodimers, heterodimers within the family, and heterodimers with nonhomeodomain proteins. Thyroid receptors and related transcription factors function as heterodimers with predominantly a single partner and recognize direct repeats. Other proteins act as monomers but are internally repetitive: POU domain transcription factors contain two HTH motifs that bind to opposite sides of the helix, TATA-binding protein contains two similar DNA-binding modules, and zinc finger proteins contain from two to nearly forty DNA-binding modules.

Dimerization is central to the function of the bZIP and bHLH protein families. The bZIP proteins dimerize through **leucine zipper** motifs — amphipathic  $\alpha$ -helices which form a *coiled coil* (q.v.). The periodicity of each helix is reduced from 3.6 to 3.5 residues in the dimer, allowing leucine residues at every seventh position to interact. However, the original model, in which leucine residues were proposed to interdigitate analogous to the teeth of a zip, has been shown to be incorrect — oppositional leucine residues are now known to form direct contacts like rungs on a ladder (*Figure 17.1*). bHLH proteins possess a **helix-loop-helix** dimerization motif which is related to the leucine zipper but distinct from the helix-turn-helix structure discussed above. Some proteins contain both zipper and HLH dimerization structures, e.g. Max.

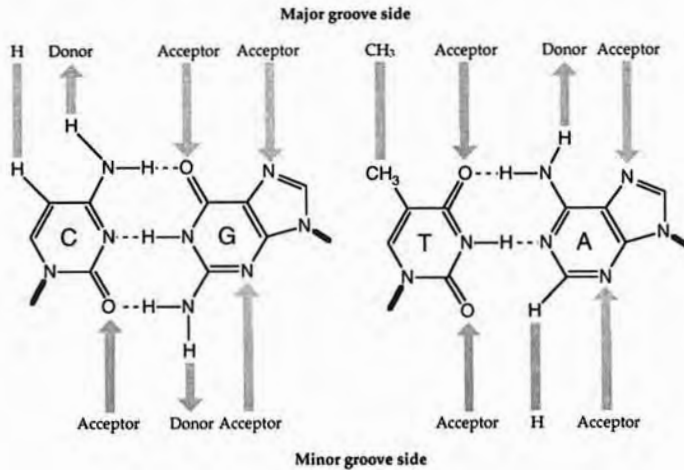
The bZIP and bHLH families demonstrate the functional and regulatory diversity that can be generated by selective dimer formation. Two bZIP proteins c-Fos and c-Jun form a heterodimeric transcription factor called AP-1. c-Jun can also form homodimers, but c-Fos cannot. The myogenic bHLH proteins MyoD1, myogenin, Myf-5 and MRF-4 act predominantly as active heterodimers. However, each can heterodimerize with a small protein called Id (inhibitor of differentiation) which lacks a basic domain and prevents DNA binding. Id thus acts like a natural *dominant negative mutant* (q.v.), by mopping up all the active bHLH proteins into inactive heterodimers.

## 17.5 Sequence-specific binding

**DNA sequence readout by proteins.** The structures of many sequence-specific protein–DNA complexes have now been solved, and while some proteins cause the melting of one or more base pairs, most recognition occurs in the context of a closed duplex. In principle, this can be achieved in two ways. **Direct readout** involves interaction between the protein and the bases themselves in the major or minor grooves: no one bond can distinguish a base pair unambiguously, so direct readout involves multiple points of contact, either between protein and DNA, or with a bridging water molecule. **Indirect readout** involves interaction between the protein and the sugar phosphate backbone. In this case, the protein can sense the sequence by its effect on the overall conformation of the DNA. One of the most surprising findings from structural analysis of DNA binding proteins has been the predominance of indirect readout in sequence-specific recognition.

**Molecular signatures in DNA.** The recognition of specific sequences by direct readout involves the formation of bonds between proteins and DNA bases. This suggests that DNA base pairs carry a ‘molecular signature’ of potential bond patterns. The signature can be determined by examining the chemical groups in the major and minor grooves, where bases are exposed to the solvent. In the major groove, the pattern of bond-forming groups is unique to each base pair, thus the major groove is used most often for nonambiguous sequence readout (*Figure 17.3*). In the minor groove, the patterns are degenerate allowing discrimination between A/T and G/C pairs only. Certain individual bond positions in the major groove are also ambiguous, thus both the major and minor grooves may





**Figure 17.3:** Molecular signatures of DNA base pairs in the major and minor grooves. The major groove allows specific discrimination between all four possible base pairs because the distribution of bonds is asymmetrical. Conversely, the distribution of bonds in the minor groove is symmetrical, so it is difficult to discriminate between A:T and T:A base pairs, and between G:C and C:G base pairs. The thick bonds show where each base is joined to the DNA backbone.

be used to recognize degenerate sequences. While some transcription factors, restriction enzymes, etc. recognize strictly invariant binding sites, others tolerate a degree of degeneracy at certain positions, e.g. discrimination between purines and pyrimidines, between keto and amino bases or between strong and weak bond bases (see legend, Table 16.2 for explanation and base symbols).

**Readout amino acids in proteins.** The discovery of families of proteins with conserved DNA binding domains naturally led to the question of whether there was a specific set of rules governing sequence specific interactions, i.e. an amino acid code for nucleotide recognition. The larger families (HTH, zinc finger, basic domain) contain examples of structurally very similar domains with differing sequence specificities. These permit a subtle mutational investigation which can be considered on three levels: (i) domain swaps; (ii) partial swaps (substitutions); and (iii) structural analysis.

**Domain swap** experiments involve the exchange of DNA binding domains between proteins. Such experiments have been used in many cases to confirm that the DNA binding domain is necessary and sufficient for both general DNA binding ability and sequence specificity. Isolated domains are often able to bind to DNA alone, and hybrid proteins take on the binding characteristics of the protein from which the DNA-binding domain originated.

Partial domain swaps, at the finest level involving the replacement of single residues, can identify the parts of the DNA-binding domain which control sequence specificity. The exchange of specific residues can alter recognition specificity, and the comparison of proteins with highly related binding domains but different recognition sequences often shows that only a few critical residues are involved. The systematic replacement of these residues with those from second protein may bring about a change in binding specificity to match that of the second protein. Systematic substitutions have been superseded recently by the advent of *phage display* (q.v.), which allows DNA-binding domains to be randomly altered at one or more positions, a repertoire of variants to be displayed on the surface of bacteriophage and selection according to binding specificity. This technique has been successfully applied to the analysis of zinc finger DNA binding proteins, as discussed below.

The structural analysis of protein-DNA complexes by X-ray crystallography or NMR spectroscopy (see Box 22.2) is the only way to determine the spatial arrangements of intermolecular bonds

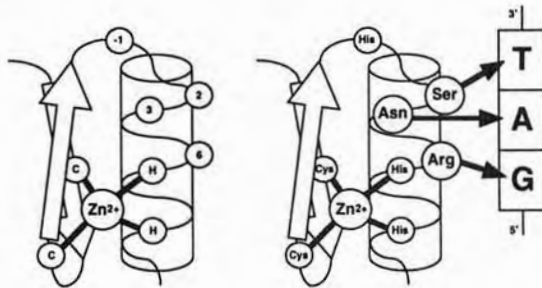
at atomic resolution. A large number of protein–DNA structures have now been solved, and such studies have been particularly useful for showing the conformational changes in both the DNA and the protein which accompany binding, and the role of water molecules in protein–DNA complexes.

**A putative zinc finger DNA recognition code.** Through systematic domain swap experimentation and the analysis of high resolution protein–DNA complex structures, it has emerged that there is no universal code governing all sequence-specific protein–DNA interactions. This reflects the diversity of recognition structures, and the different ways in which the same recognition structure, e.g. an  $\alpha$ -helix, can be presented to DNA. In some families of DNA-binding proteins, the interactions are complex and highly degenerate because the same tertiary structure can be produced in many different ways. This is particularly evident when the same structure can mediate both sequence-specific and generalized interactions with DNA (e.g. the site-specific recombinase INT and its close relative, the general DNA binding protein HU).

However, in certain protein families interactions with DNA are characteristically simple because the same few amino acid side chains are used to contact the bases. This is true of at least some Cys<sub>2</sub>His<sub>2</sub> zinc finger proteins, where a partial stereochemical code has been defined (Table 17.2). The interaction between DNA and zinc finger modules related to Zif268 has been well characterized. Three or four particular side chains in the recognition helix make the majority of base contacts, and most of these are simple one-to-one bonds. The zinc finger code is reminiscent of the *genetic code* (q.v.) in that it is degenerate but nonambiguous. At present, the code is based on a small number of

**Table 17.2:** A partial DNA recognition code for zinc finger proteins related to Zif268

Position of base in triplet			
Base	5'	Middle	3'
A		3→Asn	-1→Gln + 2→Ala
C		3→Asp, Leu, Thr, Val	
G	6→Arg; 6→Ser, Thr + 2→Asp	3→His	-1→Arg + 2→Asp
T	6→Ser, Thr + 2→Asp	3→Ala, Ser, Val -1→Asn;	-1→Gln + 2→Ser



Bases are recognized by specific amino acids, particularly those at positions -1, 2, 3 and 6 in the recognition  $\alpha$ -helix of the zinc finger module (left panel). As an example, the interaction of finger 1 from the *Drosophila* Tramtrack protein is shown (right panel). In many cases, there is 1:1 binding between base and amino acid side chain. In others, two amino acid side chains buttress each other to recognize a single base. In the table, a base is specified when position *n* of the helix → contains amino acid *x*, *y*, or *z*. If two residues buttress each other, the second residue is shown following a +. Bases are arranged as triplets, and different amino acids are used to recognize the same base in different positions within the triplet. This is because the pitch of the recognition helix is such that the residues at positions -1, 3 and 6 do not fall on the same helical face and thus lie at different distances from the DNA. Amino acids of the same class but with side chains of differing lengths are often used for base recognition (e.g. valine, leucine, isoleucine). Like the genetic code, therefore, the zinc finger code is degenerate but rarely ambiguous. It also involves base triplets, but this is a superficial property concerning the structure of the zinc finger module; the code itself does not involve 3:1 translation like the genetic code.

structures and is incomplete; it is likely to become more complex in the future as more structures are solved and assimilated, and the contributions of other side chains to base recognition are revealed.

**Sequence-specific RNA recognition.** Relatively little is known about the mechanisms of sequence-specific protein–RNA interactions, principally because so few structures of proteins bound to their RNA substrates have been solved. Those which have, three amino acyl–tRNA synthetases, the U1A spliceosomal protein bound to the U1 snRNA hairpin, and the coat protein of bacteriophage MS2 bound to an RNA hairpin, have yielded few unifying principles. Part of the reason for this is the diverse roles of RNA in the cell, and the concomitant diverse range of structures it can form.

The amino acyl–tRNA synthetases are a disparate group of enzymes which have been shown to interact with their substrates in an idiosyncratic manner. It is therefore not surprising that the detailed structures show little similarity. The enzymes can be divided into two classes according to which of the two hydroxyl groups of the tRNA terminal adenosine residue is charged with the amino acid. Class I enzymes charge the 2' hydroxyl group, and the structure of the *E. coli* Glu–tRNA synthetase has been solved. Class II enzymes charge the 3' hydroxyl group and the structures of the *S. cerevisiae* Asp–tRNA synthetase and *T. thermophilus* Ser–tRNA synthetase have been solved. Each enzyme binds in a distinct manner to tRNA. Asp–tRNA synthetase binds to the acceptor stem on the major groove side, whereas Glu–tRNA synthetase binds on the minor groove side (the enzymes also have distinct catalytic mechanisms). Both enzymes interact with the anticodon loop of tRNA to read it, but do so in different ways. Ser–tRNA synthetase does not interact with the anticodon loop, but a coiled coil structure interacts with the T $\Psi$ C and variable loops. Both direct and indirect read-out mechanisms appear to be in use.

## 17.5 Techniques for the study of protein–nucleic acid interactions

**Characterizing nucleic acid sequences that interact with proteins.** The analysis of nucleic acid–protein interactions often begins with the study of DNA or RNA sequences involved in protein binding. A number of methods have been developed to identify and characterize such sequences, and these exploit either the separation of nucleic acid–protein complexes from naked nucleic acids or the protection of nucleic acids from chemical or enzymatic degradation due to protein binding. Details of these techniques are provided in Table 17.3 and Figure 17.4.

**Methods for purifying proteins that interact with nucleic acids.** Traditional methods for purifying proteins (e.g. HPLC) give poor results for nucleic-acid binding proteins such as transcription factors because of their low abundance. However, it is possible to obtain small quantities of very pure nucleic acid-binding proteins by exploiting their affinity for specific DNA or RNA sequences. Firstly, protein extracts are mixed with total genomic DNA or nonspecific RNA (e.g. yeast tRNA) to subtract nonspecific binding proteins. Then oligonucleotides designed around a consensus DNA binding site, or specific RNA molecules, are used to extract the protein(s) of interest. An early and very simple method for isolating RNA-binding proteins was simply to filter the mixture through nitrocellulose, so that RNA would bind (retaining its associated protein) and other proteins would be washed through; RNA-binding proteins can also be isolated by cutting the bands from retardation assay gels. An efficient and widely used means for isolating DNA-binding proteins is affinity chromatography, where the specific oligonucleotide is bound to the solid matrix of a chromatography column, allowing extracts to be passed through the column, trapping interacting proteins in the process. Affinity capture using the *biotin–streptavidin* system (q.v.) can be used in the same way.

Proteins isolated by the above methods can be microsequenced by automated *Edman degradation* (q.v.) to give partial polypeptide sequences. These can be used to design degenerate oligonucleotide probes or primers, to screen a cDNA library and isolate clones corresponding to the DNA binding proteins. Alternatively, a cDNA expression library can be screened using an oligonucleotide probe carrying the nucleic acid recognition sequence (a process termed **southwestern screening** for DNA-

**Table 17.3:** Techniques for characterizing nucleic acid sequences that bind proteins — see *Figure 17.4***Methods for studying protein–nucleic acid interactions**

**Technique:** **gel retardation assay** (electrophoretic mobility shift assay, bandshift assay, gelshift assay)

**Principle:** DNA/RNA–protein complexes move more slowly through an electrophoretic gel than naked DNA/RNA

**Uses:** to identify and characterize protein binding sites in DNA or RNA

**Comments:** labeled DNA/RNA fragments are subjected to gel electrophoresis with and without prior incubation with protein extract. Successful binding is revealed as a bandshift, a difference in band mobilities between treated and untreated samples due to **gel retardation** (relatively slow movement of protein–DNA/RNA complexes). Proteins can be identified by antibodies which cause a further bandshift (or **supershift**) because of increased retardation. Precise characteristics of binding site can be investigated by adding an excess of a **competitor oligonucleotide** of specific base sequence. If the competitor binding site is suitable, the oligonucleotide will sequester all the protein away from the labeled DNA/RNA and the shifted band will disappear. If the competitor binding site is not suitable, it should have no effect. This type of experiment can be used to precisely define consensus binding sites

**Technique:** **DNase I footprinting**

**Principle:** DNA with bound protein is protected from nuclease digestion

**Uses:** to identify the exact nucleotides covered by a particular protein

**Comments:** DNA is labeled at one end only and digested with limiting amounts of DNase I to generate a nested set of labeled fragments, with and without prior incubation with protein extract. The fragments can be separated by electrophoresis to give a ladder of bands. Protein binding to the fragment protects it from nuclease activity generating a gap in the ladder. The cleavage products can be run against a sequencing reaction of the same fragment (q.v. *Maxam and Gilbert sequencing*) to identify the precise nucleotides involved. DNase I footprinting allows the simultaneous characterization of several protein-binding sites on the same DNA fragment, and competitor oligonucleotides can be used to investigate binding specificity as discussed above

**Technique:** **modification protection** (e.g. Dimethylsulphate protection assay)

**Principle:** proteins bound to DNA/RNA can protect specific bases from chemical modification (c.f. *modification interference*)

**Uses:** to identify the specific bases that interact with proteins

**Comments:** dimethylsulfate (DMS) specifically methylates guanine residues allowing the nucleic acid to be cleaved at the modified site by piperidine. Protein binding to guanine residues protects them from methylation and hence prevents cleavage. DNA/RNA is incubated with protein extract and then treated with DMS. A protected guanine residue is revealed by the absence of a cleavage product compared to naked DNA. DMS protection can also be carried out *in vivo* to investigate DNA–protein reactions in the cell (*in vivo* footprinting) rather than in the artificial environment *in vitro*. Cells are permeable to DMS, which thus methylates guanine residues in the genome. DNA is subsequently isolated, treated with piperidine and amplified by PCR to identify protected and unprotected products. If primer sites flank the protected guanine, protein binding would be detected as an extra PCR band

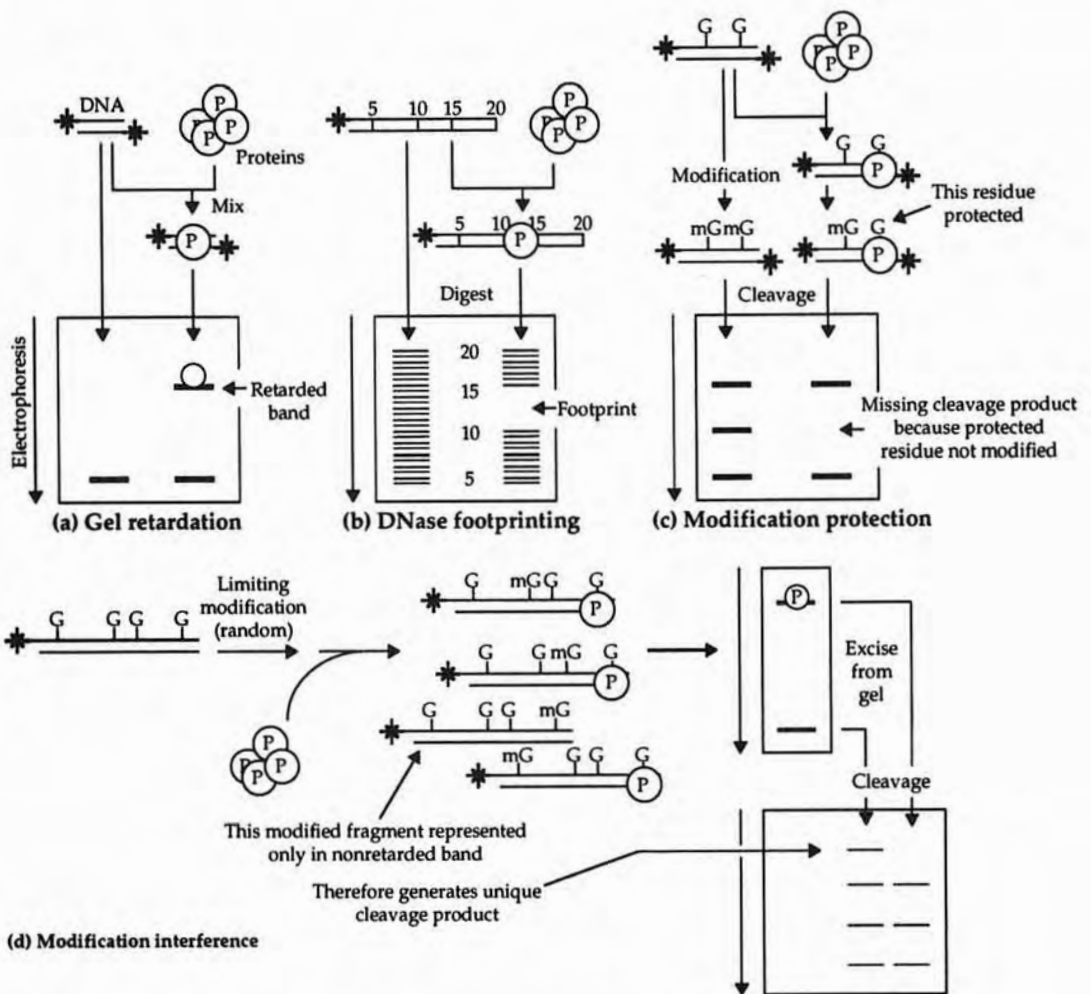
**Technique:** **modification interference**

**Principle:** a base which normally binds a protein may not do so if it is chemically modified (c.f. *methylation protection*)

**Uses:** to identify specific bases that interact with proteins

**Comments:** in the DNA **methylation interference assay**, DNA fragments are subjected to limiting DMS treatment so methyl groups are introduced at random, then mixed with protein extract and separated by gel retardation electrophoresis. If a particular guanine residue is responsible for protein binding, methylation will interfere with that binding and all fragments containing that modified residue will appear in the unretarded (naked) DNA band. The retarded and unretarded bands are excised and the DNA cleaved with piperidine. Because the unretarded band contains a unique fragment, a unique cleavage product will be generated whose size should reveal the position of the interacting residue. A similar methodology, but using different chemicals, can be used to investigate the roles of other bases in DNA. For RNA, diethylpyrocarbonate (DEPC) can be used to carboxyethylate purine bases, and hydrazine cleaves pyrimidines. Both prevent RNA–protein interaction and allow RNA cleavage by aniline





**Figure 17.4:** Analysis of protein–nucleic acid interactions. Labeled DNA or RNA is mixed with protein extracts, treated as shown and separated by electrophoresis. Details of the methods are provided in *Table 17.3*. (a) **Gel retardation assay.** Protein–nucleic acid binding is revealed as a ‘bandshift’, a difference in band mobilities due to gel retardation of protein–nucleic acid complexes. (b) **DNase I footprinting.** Precise protein-binding sites can be identified by labeling DNA at one end and digesting with limiting amounts of nuclease to generate a nested set of labeled fragments. Protein binding to the fragment protects it from nuclease activity and generating a gap in the ladder of bands. (c) **DMS protection.** Dimethylsulfate (DMS) specifically methylates guanine residues which can then be cleaved by piperidine. Protein binding protects guanine from methylation and prevents cleavage. Protein binding to specific guanine residues is thus revealed by a missing cleavage product. (d) **DMS interference.** Similar to above except that modification is carried out before incubation with protein extracts so that protein binding is prevented. The separation of naked DNA from protein–DNA complexes thus isolates the fragment containing the guanine that normally binds the protein into the nonretarded band. Isolation of the DNA from each band followed by cleavage with piperidine and further electrophoresis reveals a bond-forming guanine by generating extra cleavage products.

binding proteins and **northwestern screening** for RNA-binding proteins; q.v. *nucleic acid hybridization*). Once a cDNA has been isolated, it can be cloned into an expression vector and overexpressed in *E. coli* to generate large amounts of recombinant protein suitable for further functional studies (q.v. *expression cloning*).

**Identifying essential ribonucleoproteins.** RNA-protein interactions reflect a wide range of cellular functions and ribonucleoproteins are often complex structures containing several RNA molecules and many proteins (q.v. *ribosome*, *spliceosome*, *signal recognition particle*, *editosome*, *informosome*). It is useful to study complex ribonucleoprotein structures by determining their structural organization (i.e. which components are in contact with each other) and the functional significance of different components. Structure can be investigated by the biophysical methods discussed in Box 22.2, and by techniques such as cross-linking, co-immunoprecipitation and affinity capture to investigate interaction between different components. Functionally, it is useful to remove one component at a time to investigate its role in the complex. RNA can be removed by RNaseA digestion, but specific RNA molecules or parts thereof can be removed by first hybridizing to a DNA oligonucleotide and then digesting with RNaseH, which specifically digests RNA strands in DNA/RNA duplexes.

## References

- Latchman, D.S. (1995) *Eukaryotic Transcription Factors*. 2nd Edn. Academic Press, London.
- Liley, D. (ed.) (1995) *DNA-Protein: Structural Interactions: Frontiers in Molecular Biology*. Oxford University Press, Oxford.
- Nagai, K. and Mattaj, I. (eds) (1995) *RNA-Protein Interactions: Frontiers in Molecular Biology*. Oxford University Press, Oxford.
- Ptashne, M. (1992) *A Genetic Switch: Phage  $\lambda$  and Higher Organisms*. 2nd Edn. Cell Press, MA/Blackwell Science, MA.
- Burley, S.K., Xie, X., Clark, K.L. and Shu, F. (1997) Histone-like transcription factors in eukaryotes. *Curr. Opin. Struct. Biol.* 7: 94-102.
- Choo, Y. and Klug, A. (1997) Physical basis of a protein-DNA recognition code. *Curr. Opin. Struct. Biol.* 7: 117-125.
- Chytil, M. and Verdine, G.L. (1996) The Rel family of eukaryotic transcription factors. *Curr. Opin. Struct. Biol.* 6: 91-100.
- Harrison, S.C. (1991) A structural taxonomy of DNA-binding domains. *Nature* 353: 715-719.
- Nagashi, K. (1996) RNA-protein complexes. *Curr. Opin. Struct. Biol.* 6: 53-61.
- Pabo, C. and Saur, R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* 61: 1053-1095.
- Raumann, B.E., Brown, B.M. and Sauer, R.T. (1994) Major groove DNA recognition by  $\beta$ -sheets: the ribbon-helix-helix family of gene regulatory proteins. *Curr. Opin. Struct. Biol.* 4: 36-43.
- Rhodes, D., Schwabe, J.W.R., Chapman, L. and Fairall, L. (1996) Towards an understanding of protein-DNA recognition. *Phil. Trans. R. Soc.* 351: 501-509.
- Schwabe, J.W.R. (1997) The role of water in protein-DNA interactions. *Curr. Opin. Struct. Biol.* 7: 126-134.
- Siomi, H. and Dreyfuss, G. (1997) RNA-binding proteins as regulators of gene expression. *Curr. Opin. Genet. Dev.* 7: 345-353.
- Suzuki, M. (1993). Common features in DNA recognition helices of eukaryotic transcription factors. *EMBO J.* 12: 3221-3226.
- Suzuki, M. and Giraldo, R. (1995) Zipperless bZips and zipped homeodomains. *Proc. Jpn Acad. Sci.* 71: 39-44.

## Websites

TRANSFAC, a database which contains a list of known transcription factors and their binding sites — <http://transfac.gbf-braunschweig.de/TRANSFAC/browse/index.html>

The Homeobox page, containing a listing and classi-

fication of all homeodomain proteins — <http://copan.bioz.unibas.ch/homeo.html>

Structural classification of proteins (SCOP) database — <http://scop.mrc-lmb.cam.ac.uk/scop>

## Chapter 18

# Oncogenes and Cancer

### Fundamental concepts and definitions

- **Cancer** is a disease of multicellular organisms whose basis is abnormal, unregulated cell proliferation, often accompanied by abnormal differentiation (**neoplasia**). Cancers are generally caused by the accumulation of mutations (see below); they are common in higher vertebrates, but many other organisms do not manifest cancer because they do not live long enough. Cancer is well studied in animals, but tumors also occur in plants (q.v. *crown gall disease*).
- When cancer occurs, it is analogous to natural selection, but involving the somatic cells within an individual rather than competing organisms. Normally, somatic cells obey a developmental program which ensures controlled growth for the overall benefit and survival of the organism. A single cell which loses growth control will multiply more rapidly than its neighbors and establish a localized population of proliferating cells. By a process of **oncogenesis** (or **tumorigenesis**), this may develop into a discrete growth (a **neoplasm** or **primary tumor**) which can contain uncharacteristic differentiated cell types, and may recruit blood vessels, etc. Tumors are **benign** if they remain in one place, because they can often be removed surgically. **Malignant** tumors are **metastatic** — cells break off, disseminate and then colonize other tissues. At this stage cancer becomes difficult to control and is ultimately lethal, due to disruption or compression of vital organs.
- Cancer is caused by the disruption of cell proliferation control mechanisms, either through the activity of a virus (or occasionally bacterial signaling) at the level of protein function, or by mutation of critical growth regulation genes. Two classes of genes have been identified: **oncogenes** promote cell proliferation, and cancer results from their inappropriate activity (gain of function mutations), whereas **tumor suppressor genes (TSGs)** inhibit cell proliferation and cancer results from their reduced activity (loss of function mutations). There are many signal transduction systems and regulatory mechanisms which control cell proliferation and differentiation. In principle, disruptions to any of the pathways could cause cancer (see Signal Transduction, The Cell Cycle, Development: Molecular Aspects). Other genes often found to be involved in cancer are those which function in DNA repair or cellular responses to DNA damage. Although mutations in these genes are not tumorigenic *per se*, they cause an increase in the mutation rate, making a subsequent oncogene or TSG mutations more likely.
- The development of cancer occurs in several stages, usually starting with a mild growth disorder (**dysplasia**), which gradually becomes more serious (**tumor progression**). Similarly, primary cultured cells are initially like cells *in vivo* (i.e. they show serum dependence, contact inhibition, finite life span, etc.). They can become converted into cancerous cells lacking these inhibitions by a multistep process of **growth transformation**. Early in this process, cells become immortal but are otherwise normal. However, later, cells lose serum dependence, etc., and become progressively more virulent. This progressive increase in severity reflects the fact that single mutations in genes controlling cell growth and differentiation are not usually sufficient to cause cancer. Rather, cancer results from an accumulation of mutations whose effects are manifest gradually and in a stepwise manner (the **multiple hit hypothesis**). This reflects the existence of multiple cell proliferation control mechanisms, a safety feature necessary for the success of animals, such as humans, which contain  $>10^{13}$  cells and live a long time. Given the average mutation rate of  $10^{-6}$  per gene per generation, cancer resulting from a single mutation would affect many thousands of cells in every individual, and the species could not survive. A requirement for 4–6 independent mutations in any one cell reduces the risk considerably.

18.1 Oncogenes

**The basis of oncogene activity.** Oncogenes are genes whose *activity* has the potential to promote tumor development. They were originally discovered through the analysis of **tumor viruses**, whose infection can result in the induction of tumor growth. DNA and RNA viruses cause tumor growth by different mechanisms. The DNA tumor viruses encode products whose function is to interfere with host cell growth regulation at the protein level — these products have no counterparts in the host genome (see later). The RNA tumor viruses stimulate cell proliferation in two ways, either by carrying hyperactive derivatives of host genes whose normal function is to promote cell proliferation, or by activating the endogenous genes with strong viral regulatory elements when they integrate at an adjacent site. In either case, the corresponding host genes can be identified, and mutations affecting those genes (in the absence of viral infection) are also tumorigenic.

**Viral oncogenes and proto-oncogenes.** Some RNA tumor viruses can induce tumors in their hosts immediately after infection, and are termed **acute transforming retroviruses**. Their oncogenic potential arises from extra information in the viral genome, which usually comprises a single **viral oncogene** (designated **v-*onc***). The oncogene is often included at the expense of viral genomic information; thus the acute transforming retroviruses are defective and require a **helper virus** to supply missing functions. The first retrovirus from which an oncogene was identified, Rous sarcoma virus, is anomalous in this respect: the *v-src* gene is present in addition to the entire viral genome.

Each viral oncogene has a cellular counterpart, a **cellular oncogene** or **proto-oncogene** (designated **c-*onc***), with a similar or identical structure. The proto-oncogenes are cellular genes whose function is to process signals regulating cell growth and differentiation. Viral oncogenes are thus *transduced* (q.v.), modified copies of the cellular sequences and are inappropriately active. Proto-oncogenes can be divided into a number of categories according to the type of cellular functions carried out by their products (Table 18.1). The many signaling pathways which promote cell growth are discussed elsewhere (see Signal Transduction), and oncogene functions correspond to most components of these pathways. In principle, inappropriate activity of any signaling protein is potentially oncogenic. Of particular interest, however, are the most downstream components of these pathways, the transcription factors that directly control the genes responsible for growth control and which integrate the complex signals arriving at the cell surface. Some of these factors can act alone (e.g. Ets, Myb), whereas others act as complexes, either with other oncoproteins (e.g. Fos and Jun) or with different proteins (e.g. Myc with the ubiquitous protein Max). Many of the oncogenic transcription factors are expressed in low amounts and regulated at the protein level by phosphorylation. Mutations leading to overexpression, or those which abolish regulation, are responsible for oncogenic activity.

**Table 18.1:** Major functional categories of oncoproteins and examples of proto-oncogenes that encode them

Oncoprotein function	Examples
Secreted protein (growth factor)	<i>c-sis</i> , <i>wnt1</i>
Transmembrane receptor	<i>c-kit</i> , <i>c-erbB</i>
Adaptor protein	<i>crk</i> , <i>vav</i>
GTP-binding protein	<i>c-Hras</i>
Intracellular kinase	<i>c-abl</i> , <i>c-src</i> , <i>c-raf</i>
Transcription factor	<i>c-jun</i> , <i>c-fos</i> , <i>c-myc</i>
Transcription elongation factor	<i>ell</i>
RNA-binding protein	<i>ews</i>
Cell cycle regulator	<i>cycD1</i> , <i>cdc25A</i>

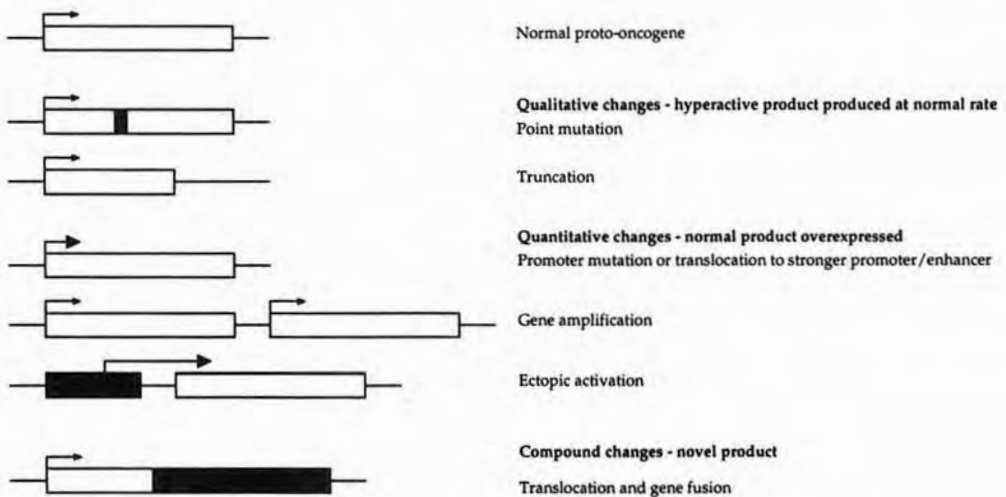
The oncogenes originally identified as viral oncogenes (e.g. *v-myc*) are identified as cellular oncogenes using a prefix c (e.g. *c-myc*).



Retroviruses are likely to preferentially transduce oncogenes because the viruses can only replicate in dividing cells, and oncogenes allow them to induce proliferation in their hosts. Some oncogenes originally identified as viral oncogenes, and the functions of the corresponding proto-oncogenes, are listed in Table 18.2.

**Oncogene activation by mutation.** Although many oncogenes have been identified by the analysis of acute transforming retroviruses, most cancers occur in the absence of viral infection. Another approach to identifying oncogenes is the direct analysis of tumor cell DNA, either structurally, by looking for differences between the DNA of normal and neoplastic cells, or functionally by transfecting tumor cell DNA into cultured cells and assaying for growth transformation. Mouse 3T3 fibroblasts are ideal for this assay because they readily undergo growth transformation, having already accumulated several mutations required for cancer development. The **3T3 assay** has identified many of the same oncogenes originally discovered through the study of retroviruses, as well as additional genes whose presence in transforming retroviruses has not been reported (e.g. *wnt-1*, *trk*, *ret*, *lck* and *bcl-2*).

Comparisons of corresponding *v-onc* and *c-onc* sequences, and of *c-onc* sequences from normal and transformed cells, has shown that the oncogenic potential of a proto-oncogene arises either from a change in structure (a qualitative change which causes inappropriate activity) or increased expression (a quantitative change which causes excessive amounts of the oncoprotein to be produced) (Figure 18.1; Table 18.3). Qualitative changes have the most severe effects where the normal oncoprotein is constrained by its structure (e.g. a receptor which needs to be activated by its ligand), whereas quantitative changes have the most severe effects when the activity of the normal oncoprotein is regulated by its abundance (e.g. a transcription factor). The simplest qualitative changes are gene mutations, either point mutations or truncations, which alter the structure of the oncoprotein and make it hyperactive even when produced in normal amounts. Quantitative changes can involve mutations in regulatory elements increasing the transcriptional activity of the gene, or mutations increasing mRNA or protein stability. Alternatively, increased gene expression may be brought about by gene amplification or by translocations which bring a proto-oncogene under the influence of a strong heterologous promoter or enhancer. In humans, oncogenic gene amplification is often associated with the EGF receptor gene, the *RAS* and *MYC* genes, or a region of chromosome 11 which contains several oncogenes, including *BCL1*, *INT2* and *ETS1*. Amplification is often a secondary response to loss of p53 protein function (see below) which allows replication even in the presence of



**Figure 18.1:** Mechanisms of oncogene activation.

**Table 18.2:** Oncogenes originally identified in acute transforming viruses and the functions of their cellular counterparts; most are named after the viruses which transduce them

Oncogene	Transducing retrovirus	Species	Principle tumor	Product <sup>a</sup>
<i>abl</i>	Ableson leukemia virus	Mouse	Lymphoma	Tyrosine kinase (cytosolic)
<i>erbA</i>	Avian erythroblastosis virus	Chick	Leukemia, sarcoma	Transcription factor (thyroid hormone receptor)
<i>erbB</i>				Receptor tyrosine kinase (EGF receptor)
<i>fms</i>	Feline sarcoma virus	Cat	Sarcoma	Receptor tyrosine kinase (M-CSF receptor)
<i>fos</i>	FBJ murine osteocarcinoma	Mouse	Chondrosarcoma	Transcription factor
<i>fps</i>	Fujinami sarcoma virus	Chick	Sarcoma	Tyrosine kinase (cytosolic)
<i>jun</i>	Avian sarcoma virus	Chick	Sarcoma	Transcription factor
<i>kit</i>	Feline sarcoma virus	Cat	Sarcoma	Receptor tyrosine kinase (Steel factor receptor)
<i>mos</i>	Moloney murine sarcoma virus Mo-MuSV	Mouse	Sarcoma	Serine/threonine kinase
<i>myb</i>	Avian myeloblastosis	Chick	Leukemia	Transcription factor
<i>myc</i>	Avian myelocytomatosis virus	Chick	Myelocytoma, sarcoma	Transcription factor
<i>raf</i>	Murine sarcoma virus	Mouse, chick	Sarcoma	Serine/threonine kinase
H- <i>ras</i>	Murine sarcoma virus (Harvey strain) (Ha-MuSV)	Rat	Sarcoma erythroleukemia	Ras GTPase
K- <i>ras</i>	Murine sarcoma virus (Kirsten strain) (Ki-MuSV)			
<i>rel</i>	Reticuloendotheliosis virus	Turkey	Leukemia	Transcription factor
<i>sis</i>	Simian sarcoma virus	Monkey	Sarcoma	Growth factor (PDGF $\beta$ chain)
<i>sis</i>	Feline sarcoma virus	Cat		
<i>src</i>	Rous sarcoma virus (RSV)	Chick	Sarcoma	Membrane-associated tyrosine kinase

Most viruses carry a single oncogene although some carry more, e.g. AEV carries two, *v-erbA* and *v-erbB*, which are unrelated. Many oncogenes are specific to one particular virus, although *v-sis* is carried by two unrelated viruses. Some viral strains carry alternative forms of the same oncogene, e.g. MuSV carries alternative forms of *v-ras*.

Abbreviations: EGF, epidermal growth factor; M-CSF, macrophage colony stimulating factor; PDGF, platelet-derived growth factor).

<sup>a</sup>For information concerning the roles of these products in cell growth, see Signal Transduction.

**Table 18.3:** Mechanisms of oncogene activation

Change	Example
Qualitative	
Point mutation	<p><b>c-ras:</b></p> <p>Ras is a GTP-binding protein involved in the transduction of signals from growth factor receptors. Receptor activation promotes GTP binding to Ras by recruiting a guanine nucleotide exchange factor. Ras-GTP can recruit the intracellular kinase Raf to the cell membrane, causing it to be activated and thus transducing the growth signal. The Ras protein possesses GTPase activity and rapidly inactivates itself, resulting in a transient burst of signaling activity. Point mutations in the <i>ras</i> gene, specifically those affecting amino acids at positions 12 and 61, decrease GTPase activity, resulting in exaggerated response to growth factors and excessive cell proliferation</p>
Truncation	<p><b>v-erbB:</b></p> <p>The <i>c-erbB</i> gene encodes the epidermal growth factor receptor, a transmembrane receptor tyrosine kinase. The viral oncogene <i>v-erbB</i> is truncated at both ends and hence lacks both the N-terminal ligand binding domain (allowing the receptor to dimerize in the absence of its ligand) and the C-terminal inhibitory domain (allowing constitutive signaling). The receptor is thus constitutively active in the absence of its ligand, leading to unregulated cell proliferation</p>
Quantitative	
Amplification	<p><b>c-myc:</b></p> <p>Cancer is often caused by simple amplification of a structurally normal proto-oncogene, especially where growth is limited by the quantity of an oncoprotein. This is seen for the <i>c-myc</i> gene, which encodes a transcription factor usually available in limiting quantities. The gene is amplified several hundred-fold in certain cancers, including breast cancer (q.v. <i>DNA amplification</i>)</p>
Increased transcription	<p><b>v-mos, c-myc, c-myb:</b></p> <p>In other cases where growth regulation reflects limited quantities of a particular oncoprotein, cancer may be caused by increasing the transcription of an oncogene. This occurs in several ways. Regulatory mutations can increase transcriptional activity of an oncogene, and this has been observed for <i>c-myc</i>. In other cases, viral transduction may bring a structurally normal proto-oncogene under the control of a strong viral promoter: the transduced viral oncogene <i>v-mos</i> is identical to <i>c-mos</i>, and oncogenicity appears to reflect overexpression from within the virus, driven by the viral LTR promoter. Proto-oncogenes may also be <i>cis</i>-activated by adjacent integration of a retrovirus. The <i>c-myc</i>, <i>c-myb</i> and <i>c-raf</i> genes are activated in this manner, and in the <i>c-myc</i> locus, retroviral integration may occasionally generate chimeric mRNA</p>
Translocation	
Position effects	<p><b>MYC:</b></p> <p>In humans, translocation t(8;14)(q24;q32) brings most of the <i>MYC</i> gene to the IgH locus and places it under the influence of the strong immunoglobulin enhancer. This translocation is believed to be the basis of Burkitt's lymphoma. There is no structural disturbance to the <i>MYC</i> protein, but the gene is upregulated. Translocations involving the immunoglobulin loci and the T-cell receptor loci are often associated with cancer because the V(D)J recombination system provides an opportunity for aberrant rearrangements, and each locus is under the control of a strong enhancer (q.v. <i>V(D)J recombination</i>).</p>
Fusion oncogenes	<p><b>ABL/BCR:</b></p> <p>The <i>ABL</i> gene is located on chromosome 9 and the <i>BCR</i> gene on chromosome 22. Reciprocal translocation generates a compound</p>

Continued

chromosome (the **Philadelphia chromosome**) where most of the *ABL* gene is spliced to a variable 5'-end segment of the *BCR* gene. The fusion product lacks the N-terminal kinase regulatory domain of *ABL*, and its active state may be stabilized by the *BCR* protein segment. The result is a constitutively active fusion tyrosine kinase. Many other oncogenic fusion proteins have been described, most of which are transcription factors

double-strand breaks. Translocations which bring oncogenes under the influence of stronger transcriptional regulation are known in at least 50 types of cancer. However, translocation can also generate **fusion oncogenes**: the products may function normally but may be overexpressed (e.g. *MYC* in Burkitt's lymphoma), or they may have unusual or inappropriate activity (e.g. *ABL/BCR* fusion proteins in chronic myeloid leukemia). Most of the fusion oncogenes which have been characterized encode fusion kinases or fusion transcription factors with altered substrate specificities.

Whichever route is chosen, mutations which activate oncogenes are all *gain of function* mutations (i.e. a gain of growth-promoting activity) and are *dominant* over the wild-type allele. In the case of retroviral transduction, the transduced oncogene must exert its effect over two wild-type alleles.

**Oncogene activation by slow RNA tumor viruses.** Whereas the acute transforming retroviruses transform their host rapidly by transducing a hyperactive or overexpressed viral oncogene into the genome, a second class of RNA tumor viruses can promote tumor development only after a long period of latency following infection. These **slow transforming retroviruses** are not defective and do not carry *v-onc* genes. The mechanism of oncogene activation is retroviral integration adjacent to a normal proto-oncogene so that it becomes activated by the strong promoter present in the viral long-terminal repeat (q.v. *retroviruses*). The latent period reflects multiple cycles of retroviral replication and integration until the virus fortuitously integrates at a site where activation of an adjacent proto-oncogene can occur. Different retroviruses preferentially activate different oncogenes, e.g. avian leukosis virus preferentially activates *c-myc*. This may reflect regional specificity of retroviral integration particular to each viral group.

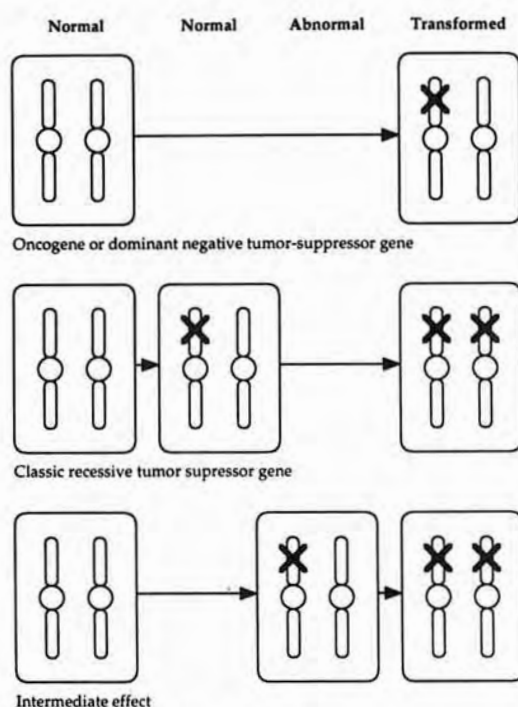
## 18.2 Tumor-suppressor genes

**The basis of tumor-suppressor gene activity.** Tumor suppressor genes (**anti-oncogenes**, **recessive oncogenes** or **onco-suppressor genes**) are genes whose *inactivity* has the potential to promote tumor growth. The existence of TSGs was predicted from the results of cell fusion experiments, where growth-transformed cells could be rescued (i.e. the tumorigenic phenotype could be corrected) by fusion to normal cells. Such experiments identify TSGs with recessive loss of function alleles (e.g. *RB-1*), but many TSGs act as *dominant negatives* (q.v.) when mutated (e.g. *TP53* and *WT-1*) (Figure 18.2). The normal function of TSGs is to *restrain* cell growth, and many of the TSG products that have been identified are regulators of the *cell cycle* (q.v.) or transcriptional repressors of growth-related genes (Table 18.4).

**Familial cancer and Knudson's two-hit hypothesis.** Most cancers caused by oncogenes are not familial but arise through somatic mutation. This is because oncogenes exert dominant effects, and dominant constitutional loss of growth control in an embryo would be lethal. **Familial cancers**, transmitted as Mendelian characters, are therefore further evidence for the existence of recessive tumor suppressor genes. Ironically, familial cancers tend to be transmitted as autosomal *dominant* traits, although with incomplete penetrance.

**Knudson's two-hit hypothesis** offers an explanation for this apparent paradox. The first familial cancer to be studied was retinoblastoma, a rare childhood tumor of the retina. Retinoblastoma occurs both as familial (60%) and sporadic cases (40%), and the familial cancer is transmitted as an autosomal dominant trait. Most sporadic cases are unilateral, whereas bilateral tumors are common





**Figure 18.2:** Mutations and the classification of tumorigenic genes.

in familial retinoblastoma. Knudson suggested that two independent mutations (two hits) were required to generate a retinoblastoma cell. In familial retinoblastoma, the first mutation is present in the germline, so only one somatic mutation is required. Since the probability of mutation is approximately  $10^{-5}$ , and there are approximately  $10^6$  cells in the retina, a retinoblastoma founder cell will be generated in most individuals with the germline mutation, i.e. the cancer will have strong but incomplete penetrance. In sporadic cases, a cell would have to undergo two independent somatic mutations, and this would occur once in every 10000 individuals ( $10^{-5} \times 10^{-5} \times 10^6 = 10^{-4}$ ). The dominant inheritance and incomplete penetrance of familial retinoblastoma thus reflects the probability of a 'second hit' somatic mutation, not a dominance effect exhibited by the germline loss of function allele, which is recessive.

The two-hit hypothesis is supported by frequent **loss of heterozygosity** in tumor cells. If retinoblastoma cells are typed for linked markers using normal cells of the same individual as a control (e.g. from blood), it is often found that markers which are heterozygous in controls are hemizygous in tumor cells. This is because second-hit mutations are often chromosome deletions, which remove the retinoblastoma locus and any associated markers all together. Although this is not always the case (the second mutation can be a point mutation in the *RB-1* gene, or a deletion which does not affect flanking markers), loss of heterozygosity can be used to identify and map Mendelian TSGs onto a panel of genome-wide markers such as microsatellite repeats (see Gene Structure and Mapping).

**DNA tumor viruses.** Most DNA viruses of eukaryotic cells establish lytic infections, which result in death of the host cell. Occasionally, however, certain DNA viruses may become latent in cells which are nonpermissive for lytic infection. Of these, the **DNA tumor viruses** (including the polyomavirus family (e.g. SV40), the human papillomavirus (HPV) family, and the adenovirus family) are capable of transforming the host cell. This is achieved by the synthesis of viral oncoproteins which interact with and inhibit host TSG proteins. The products of the tumor suppressor genes were first isolated through their interaction with these viral proteins which, unlike the oncoproteins of the RNA viruses, have no direct counterpart in the host genome. The polyomavirus family produces two

**Table 18.4:** Some well-characterized tumor-suppressor genes and their functions

TSG	Function of protein
<b>RB1</b>	The retinoblastoma protein RB1 plays a critical role in cell cycle control by binding to and inhibiting the activity of E2F family transcription factors and components of the basal transcriptional machinery of RNA polymerases I and III, notably UBF (q.v. <i>RNA polymerase I</i> ). The E2F proteins are required to activate genes for entry into S-phase (e.g. <i>c-myc</i> , <i>c-myb</i> and <i>cdc2</i> ), while RNA polymerases I and III synthesize rRNA and tRNA. Thus RB1 controls cell proliferation not only by arresting the cell cycle, but also limiting the rate of protein synthesis. RB1 is regulated by phosphorylation. It is inactive in its phosphorylated form, and it is phosphorylated several hours before the onset of S-phase by cyclin-CDK complexes containing cyclin D1 (see The Cell Cycle). Other proteins interact with and inhibit RB1, e.g. the product of the cellular oncogene <i>MDM2</i> . The products of several DNA tumor virus oncogenes also inhibit RB1, either by sequestering the protein into an inactive complex, or by targeting it for degradation (e.g. SV40 T antigen)
<b>TP53</b>	The p53 protein is a transcription factor that regulates cell proliferation and responds to signals, including DNA damage, by arresting the cell cycle or inducing apoptosis. Cell-cycle arrest is facilitated by transcriptional activation of the genes encoding the p21/p27 family of cyclin-dependent kinase inhibitors, and repression of genes such as <i>MYC</i> whose activity leads to proliferation. p53 may have a broader role in recognizing and responding to DNA damage, as it possesses a 3'-5' exonuclease activity which may reflect a DNA repair function. The p53 protein induces apoptosis in certain cell types in response to oncogenic cell behavior, partly through regulation of the <i>BAX</i> and <i>BCL2</i> genes (q.v. <i>apoptosis</i> ). Point mutations or deletions in <i>TP53</i> are seen in over 60% of all human cancers and tend to cluster in the central DNA-binding domain. Mutations are often dominant negatives, since p53 acts as a tetramer. Germline mutations of <i>TP53</i> are associated with the rare familial Li-Fraumeni syndrome, which is characterized by a diverse spectrum of tumors. Like RB1, p53 activity is inhibited by the MDM2 oncoprotein, and interacts with the oncoproteins of several DNA tumor viruses
<b>NF1</b>	NF1 is a RAS-GTPase activating protein (GAP), i.e. a protein whose function is to antagonize RAS signaling by accelerating the RAS intrinsic GTPase activity. RAS is a guanine nucleotide-binding protein; it is active when bound by GTP but inactive when bound by GDP. Active RAS recruits RAF to the cell membrane, where it is phosphorylated and then able to initiate the MAP kinase signaling cascade (see Signal Transduction). Loss of <i>NF1</i> results in the inability of the cell to shut down RAS signaling, and hence constitutive activation of the mitogenic MAP kinase pathway. This occurs even when the ubiquitous GTPase activating protein GAP <sup>p120</sup> is present. Germline mutations in <i>NF1</i> are associated with neurofibromas and chronic myelogenous leukemia
<b>WT1</b>	The <i>WT1</i> gene encodes at least four zinc finger proteins by alternative splicing. The main splice variant is a transcriptional repressor of several genes involved in growth regulation, including <i>BCL2</i> , <i>IGF-II</i> and <i>MYC</i> . Mutations in the <i>WT1</i> gene which abolish DNA-binding are often seen in cases of Wilm's tumor of the kidney and the Denys-Drash syndrome. WT1 acts as a dimer and <i>WT1</i> mutations are therefore often dominant negatives
<b>P21 and P16</b>	p21 is a general <i>cyclin-dependent kinase inhibitor</i> (q.v.) and is activated by p53 in response to DNA damage. p16 is a specific inhibitor of CDK4 and CDK6-D cyclin complexes. Loss of function mutations in the genes encoding both inhibitors prevents normal growth inhibitory signals blocking the cell cycle at G <sub>1</sub> . D cyclins are responsible for initiating the events which promote entry into the S-phase (see The Cell Cycle)
<b>BRCA1, BRCA2</b>	<i>BRCA1</i> and <i>BRCA2</i> mutations are responsible for most familial cases of combined breast and ovarian cancer, and about half of the cases where breast cancer appears alone. The proteins are very similar. Each contains a zinc finger module and each is regulated by phosphorylation during the cell cycle. Their precise functions are unknown, but a developmental role for <i>BRCA1</i> has been suggested.

**tumor antigens**, T and t, which are required for tumorigenesis. T is known to interact with and inactivate both RB-1 and p53, to release the cell cycle from inhibition. Similarly, adenoviruses synthesize two proteins, E1A and E1B, the former having been shown to interact with RB-1 and the latter with p53. The HPV E6 protein also inhibits p53 function, whereas E7 inhibits RB-1. It thus appears that P53 and RB-1 play a decisive role in the infection strategy of the DNA tumor viruses, and as suggested by other studies, are the central regulators of the cell cycle.

### Further reading

- Bishop, J.M. (1995) Cancer — the rise of the genetic paradigm. *Genes Dev.* 9: 1309–1315.
- Herschman, H.R. (1991) Primary response genes induced by growth factors and tumour promoters. *Annu. Rev. Biochem.* 60: 281–319.
- Hunter, T. and Karin, M. (1992) The regulation of transcription by phosphorylation. *Cell* 70: 375–387.
- Jacks, T. (1996). Tumor suppressor gene mutations in mice. *Annu. Rev. Genet.* 30: 603–636.
- Lees, E.M. and Harlow, E. (1995) Cancer and the cell cycle. In: *Cell Cycle Control: Frontiers in Molecular Biology*, (eds C. Hutchinson and D.M. Glover), pp. 228–263. Oxford University Press, Oxford.
- Lowy, D.R. (1993) Function and regulation of Ras. *Annu. Rev. Biochem.* 62: 851–891.
- Milner, J. (1995) Flexibility — the key to p53 function. *Trends Biochem. Sci.* 20: 49–51.
- Sherr, C.J. and Roberts, J.M. (1995) Inhibitors of mammalian G1 cyclin-dependent kinases. *Genes Dev.* 9: 1149–1163.
- Vaux, D.L. and Strasser, A. (1996) The molecular biology of apoptosis. *Proc. Natl. Acad. Sci. USA* 93: 2239–2244.
- Weinberg, R.A. (1996) E2F and cell proliferation — a world turned upside-down. *Cell* 85: 457–459.
- Weinberg, R.A. (1997) The cat and mouse games that genes, viruses, and cells play. *Cell* 88: 573–575.
- Zhang, C-C. (1996) Bacterial signalling involving eukaryotic-type protein kinases. *Mol. Microbiol.* 20: 9–15.

**This Page Intentionally Left Blank**



## Chapter 19

# Organelle Genomes

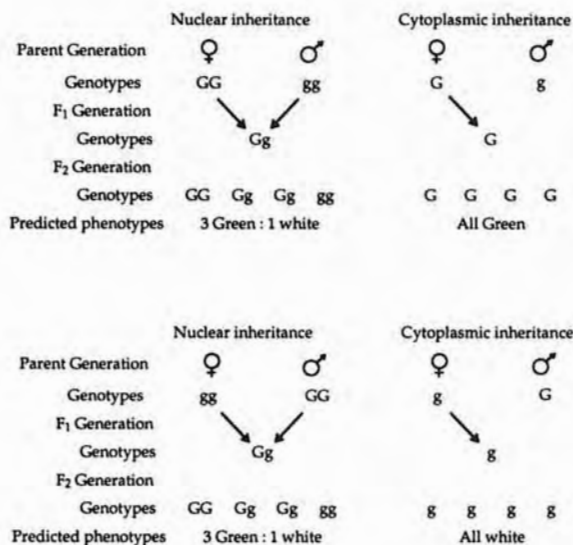
### Fundamental concepts and definitions

- In sexually reproducing eukaryotes, each parent contributes equally to the nuclear genome, generating transmission patterns typical of *Mendelian inheritance* (q.v.) (c.f. *parental imprinting*, *sex linked inheritance*). **Nuclear inheritance** is thus described as **biparental**.
- The nucleus contains chromosomes and, in some eukaryotes, plasmids. The nucleus, however, is not the only source of genetic material in the cell. Certain organelles carry their own genome (mitochondria and chloroplasts are the major examples), and cells may also harbor cytoplasmic organisms ranging from parasites and viruses to endosymbionts.
- These cytoplasmic sources of genetic information follow rules of inheritance which are different from those of nuclear genes. **Cytoplasmic** (or **extranuclear** or **extragenomic**) **inheritance** is usually **uniparental**, i.e. transmission to the zygote is from one parent only. In animals, cytoplasmic inheritance is synonymous with **maternal inheritance** because only female gametes contribute cytoplasm to the zygote. In plants, maternal inheritance is predominant, but some species show **paternal inheritance** and in others either parent is capable of cytoplasmic transmission (although not at the same time). In plants and lower eukaryotes where both gametes contribute equivalent amounts of cytoplasm to the zygote, the cytoplasmic genes transmitted by one of the parents are often selectively destroyed or inactivated, so that inheritance remains functionally uniparental. The progeny of such a cross thus have a genotype depending on only that of the contributing parent, irrespective of the genotype of the other (*Figure 19.1*).
- Organelle genomes share many structural and functional properties with those of eubacteria. Accordingly, they are believed to have arisen from endosymbiotic organisms colonizing early eukaryotic cells. There is a high level of functional integration between the organelle and nuclear gene products required for organelle function, and this is indicative of substantial gene transfer to the nuclear genome during organelle evolution.

### 19.1 Organelle genetics

**Maternal inheritance.** Maternal inheritance is the transmission of genes, and the traits they control, solely through the maternal line. This reflects the fact that in many higher eukaryotes, the mother provides all the cytoplasm in the egg whereas the father provides only the paternal nucleus. The consequences of maternal inheritance are shown in *Figure 19.1*. In a cross between two contrasting lines where one allele shows full dominance, the results of a reciprocal cross are identical for nuclear genes but different for maternally inherited genes and dependent on the genotype of the mother. Maternal inheritance should not be confused with the *maternal effect* (q.v.), which reflects maternal control of early development, but involves nuclear inheritance and shows normal Mendelian segregation ratios (see *Figure 6.1*).

**Organelle mutants.** Mutations affecting organelle genomes cause deficiencies in either photosynthesis or oxidative phosphorylation. Chloroplast mutants often show clear morphological phenotypes (e.g. leaf variegation) which have been exploited in studies of maternal inheritance. Mitochondrial mutations have been extensively studied in microbial eukaryotes, where they generate a slow growth phenotype (termed 'petite' in yeast and 'poky' in *Neurospora crassa*). In yeast, cytoplasmic inheritance is biparental (although this is not the case with other fungi) which allows the behavior of mitochondrial mutants to be studied in the heteroplasmic state (i.e. where organelles with different genotypes coexist in the same cytoplasm). Different classes of petite mutant are



**Figure 19.1:** Nuclear (Mendelian) inheritance and maternal inheritance. Color variation in plants is often caused by mutations in genes controlling chloroplast function. Some of the genes involved are carried by the chloroplast and show maternal inheritance. Others are located in the nucleus and show Mendelian inheritance. If green (normal pigment) is dominant to white (loss of pigment, a null mutation), distinct patterns of transmission are observed. The equal contribution of the parents in Mendelian inheritance means that the results of reciprocal crosses are the same (plants do not possess heteromorphic sex-chromosomes like those of mammals, and the complication of *sex-linked inheritance* (q.v.) does not arise).

classified in Table 19.1, according to their mode of inheritance and their dominance over wild-type mitochondria. Table 19.2 gives definitions of some terms used in organelle biology.

In humans, mitochondrial mutant phenotypes are complex because they show tissue-specific effects reflecting interactions between mtDNA products and cell type-specific nuclear-encoded proteins — they cause degenerative disorders termed **mitochondriopathies**. Familial cases of a severe myopathy, Kearns-Sayre syndrome, are associated with deletions in mtDNA, but are inherited in a Mendelian fashion like yeast *pet* mutations. Such diseases result from the loss of nuclear genes which control mtDNA stability. Other diseases, such as Leber's hereditary optic neuropathy, show maternal inheritance and reflect specific mutations in the mtDNA, analogous to yeast *mit* mutants. In all human mitochondriopathies, cells are heteroplasmic for wild-type and mutant mtDNA. Thus, like yeast *rho* mutants, the mutant organelles are dominant to normal organelles. The mechanism by which heteroplasmy and dominance is maintained is unclear, but selfish replication of the mutant mtDNA is likely, as most mutant mtDNAs retain both origins of replication (see below).

## 19.2 Organelle genomes

**General features of organelle genomes.** The molecular characterization of organelle genomes has been facilitated by a range of techniques which exploit structural and function differences between nuclear and organelle DNA. These techniques include the selective inhibition of nuclear and organelle gene expression using different antibiotics, the analysis of transcription and protein synthesis in isolated organelles, the *in vitro* expression of organelle DNA and RNA, linkage mapping, restriction mapping, hybridization analysis, and sequencing. Organelle genomes differ fundamentally from nuclear genomes in terms of their structure, organization, stability and mechanisms of gene expression and regulation. Some of the unusual properties of organelle genomes are listed in Table 19.3.

**Table 19.1:** Four classes of mitochondrial dysfunction mutants in yeast

Original name	Mutant designation	Molecular basis
<b>Mitochondrial petite</b> Shows cytoplasmic inheritance	<i>mit</i>	Loss of function mutation in single mitochondrial gene due to point mutation
<b>Segregational petite</b> Shows normal Mendelian inheritance of slow growth phenotype	<i>pet</i>	Mutation in nuclear gene affecting mitochondrial function
<b>Suppressive petite</b> Generates all petite colonies when crossed to wild-type	<i>rho</i>	Loss of much of mtDNA leaving behind a small selfishly replicating circle which may outcompete wild-type mitochondria
<b>Neutral petite</b> Generates all wild-type colonies when crossed to wild-type	<i>rho</i> <sup>0</sup>	Loss of all mtDNA. Mitochondrial biogenesis is controlled mainly by the nuclear genome, so mitochondria are constructed even in the absence of their entire genome

Similar classes are seen in chloroplast mutants and animal mitochondrial mutants, reflecting two common themes: (1) dual control of organelle function by the organelle and nuclear genomes, and (2) the ability of mutant organelles to dominate wild-type organelles in the same cell.

**Table 19.2:** Some terms used to describe the behavior of organelle genomes

Term	Definition
Cytotype	A characteristic conferred by an agent in the cytoplasm rather than in the nucleus. Can refer not only to characteristics specified by cytoplasmic genes, but also to characteristics specified by the <i>products</i> of nuclear genes which are synthesized in and function in the cytoplasm (e.g. q.v. <i>P-elements</i> , <i>hybrid dysgenesis</i> )
Homoplasmic	From 'homozygous cytoplasm' — an individual with organelles of uniform genotype
Heteroplasmic	From 'heterozygous cytoplasm' — an individual with a mixture of organelles of different genotypes
Cytohets	Heteroplasmic cells. Higher eukaryotic cells tend to be homoplasmic unless a spontaneous mutation arises, although cytohets can also be produced by cell fusion. Cytohets arise naturally in lower eukaryotes with biparental cytoplasmic inheritance. In those species where the cytoplasmic genes of one parent are destroyed, cytohets can arise when this process fails. Such aberrant cells are termed <b>biparental zygotes</b>
Cytoplasmic segregation	The production of homoplasmic cells from heteroplasmic cells during mitotic division. Reflects the random partition of the organelle population between daughter cells, a stochastic process leading to bias which will be exaggerated through further rounds of division. Eventually, after several generations, all daughter cells are homoplasmic. Similar in principle to the segregation of multicopy plasmids (see Plasmids)
Heterokaryon, homokaryon, synkaryon	These terms concern nuclear rather than cytoplasmic genes but are often confused with cytoplasmic inheritance and are included here for clarity. A heterokaryon is a cell containing more than one nucleus and the nuclei are of different genotypes. A homokaryon is multinucleate but all nuclei have the same genotype. A synkaryon contains a diploid nucleus. A zygote may be called a synkaryon

**Table 19.3:** Some unusual properties of organelle genomes

---

The genome is generally circular and present as several to many copies
Most organelle genes are dedicated to gene expression functions (e.g. tRNA, rRNA, RNA polymerase). There are some genes encoding organelle functions (e.g. proteins for photosynthesis or oxidative phosphorylation), but most such proteins are encoded by nuclear genes and imported
The genome often exists as mixtures of sequence variants generated by recombination
Transcription and translation are regulated by prokaryote-like control sequences and <i>trans</i> -acting factors
Organelle gene expression is sensitive to antibiotics but not to inhibitors of the eukaryotic nuclear genes — this phenomenon is useful for the study of nuclear and organelle genes in isolation
Transcription is often complex, involving multiple initiation sites and polycistronic messages
Organelle genes are regulated primarily by posttranscriptional mechanisms including modulation of mRNA stability, RNA processing, protein synthesis and protein turnover
Gene expression is often characterized by complex RNA processing reactions including cleavage, <i>cis</i> - and <i>trans</i> -splicing, RNA editing and degradation
Organelles often use variants of the universal genetic code, reflecting a smaller repertoire of tRNA species with enhanced wobble interactions (see The Genetic Code)
Organelles may carry plasmids in addition to their genome, which often confer a phenotype on the host by mediating gene rearrangements (see Plasmids)

---

**Plastid genomes.** The **plastome** is the DNA isolated from a plastid (one of various plant organelles related to the chloroplast). The analysis of such DNA shows that each plastid contains the same DNA, which is the same as **chloroplast DNA (ctDNA)**. The genome is circular, with a narrow size range between 110 and 150 kbp. It encodes proteins and structural RNAs required for chloroplast gene expression, including tRNAs sufficient for all codons, rRNA, some ribosomal proteins and RNA polymerase. Additionally, the chloroplast encodes products with a direct function in photosynthesis, including components of photosystems I and II. Most polypeptides used in the chloroplast, however, are encoded by the nuclear genome and imported (q.v. *protein targeting*). The degree of nuclear/cytoplasmic integration is high; most chloroplast-encoded proteins associate with nuclear-encoded proteins at some time. In many plants, for instance, the large subunit of the dimeric enzyme ribulose 1,5-bisphosphate carboxylase-oxygenase (RuBisCO) is encoded by the chloroplast and the small subunit by the nuclear genome.

The chloroplast genome in some plant species is an almost totally unique sequence; others, however, are organized as two unique sequence regions separated by inverted repeats, the repeats containing several genes including the rRNA genes. Recombination occurs frequently between the repeats, generating inversion isomers and concatemers.

Chloroplast genes are controlled by *cis*-acting elements similar in structure to bacterial promoters and are able to function in *E. coli*. Transcription is often complex, generating *polycistronic mRNAs* (q.v.) of differing sizes which may be cleaved at specific processing sites. Gene regulation occurs primarily at the levels of RNA stability and protein synthesis, and a number of host-encoded proteins with roles in posttranscriptional regulation of chloroplast genes have been isolated. Many chloroplast genes contain *self-splicing introns* of the class II type (q.v.), and demonstrate other intriguing RNA processing reactions such as *trans-splicing* and *RNA editing* (q.v.). For protein synthesis, ribosome binding sites are similar to those found in *E. coli* and chloroplasts appear to use the unmodified universal genetic code.

**Mitochondrial DNA: Size, gene content and genome organization.** Mitochondrial genomes are extremely diverse and show striking characteristic differences between taxa in terms of size and structural organization. Like chloroplast genomes, however, mitochondrial genomes predominantly encode genes with gene expression functions (rRNA, tRNA, RNA splicing enzymes) as well as a few polypeptides concerned with mitochondrial function, although most mitochondrial proteins are synthesized in the cytoplasm and imported.

Yeast **mitochondrial DNA (mtDNA)** is approximately 80 kbp in length, which is about the



maximum size for fungal mtDNA. It contains some noncoding AT-rich DNA punctuated with GC-rich regions containing origins of replication. Many yeast mitochondrial genes have large introns, which may contain open reading frames encoding proteins which control intron splicing and transposition (q.v. *homing introns*). All three classes of *self-splicing intron* (q.v.) may be found in mitochondrial genes. tRNA genes are often found in functional clusters, whereas rRNA genes are dispersed. The mtDNA of other fungi contains less repetitive DNA and the smallest genomes are approximately 20 kbp.

Plant mtDNA shows an incredible diversity in size from 100 kbp to over 2.5 Mbp, the larger genomes containing a high proportion of repetitive DNA. In some species, the genome is of uniform size, but in many plants it has a complex organization involving linear and circular molecules of various sizes and structure generated by recombination. The full genome version is termed the **master circle**; smaller derivatives are **subgenomic circles**. Master circles contain repeats which are sites of recombination — generally a profile of recombination products is seen for each genome. Abnormally rearranged mtDNAs, **sublimons**, may become amplified, perhaps by selfish replication like *rho* mutants in yeast.

Animal mitochondrial genomes, by contrast, are typically small (less than 20 kbp) and organized parsimoniously. There are no introns, little space between genes and a single region of noncoding DNA which controls replication and gene expression. The space-saving features of animal mtDNA are remarkable, including *pangenomic transcription* (see below), overlapping genes and missing termination codons (these are added posttranscriptionally by polyadenylation of the RNA).

**Mitochondrial gene expression.** Animal and fungal mitochondrial transcription is *polycistronic* (q.v.), although there are many promoters in fungal mtDNA and only a single promoter for each strand of animal mtDNA, which is initially expressed as a **pangenomic** (full genome) transcript. In plants most mitochondrial mRNAs are monocistronic (the major exception being the rRNA genes), although they are transcribed from multiple initiation sites. All mitochondria use similar prokaryote-like promoter sequences and encode RNA polymerases related to prokaryotic enzymes, including the use of accessory factors similar to the  $\sigma$ -factor (q.v.). Yeast and animal mtDNA transcription requires a transcription factor related to the *HMG family* of proteins (q.v.).

RNA processing in mitochondria takes many forms. In plants, the rRNA transcripts are cleaved and matured, introns are spliced (although no *trans*-splicing — as seen in chloroplasts — is observed), and some genes may be subject to RNA editing. In animals, processing involves cleaving the polycistronic message into single gene fragments, a process possibly facilitated by the dispersed arrangement of tRNA genes which could provide secondary structure motifs for endonuclease activity. In animals both mRNA and rRNA from mitochondria may be polyadenylated. RNA processing in fungal mitochondria involves a 12-nucleotide 3' signal which controls maturation of the transcript.

Translation in mitochondria from plants, fungi and animals involves a modified genetic code. The modifications are taxon-specific and reflect smaller repertoires of tRNA genes, and in some cases unusual tRNA structures. The loss of tRNA species is compensated by a modification to wobble pairing, so more degeneracy is tolerated (see The Genetic Code). Plant mitochondria usually encode a complete set of tRNAs, although in certain species some are encoded by the nucleus and imported. Some algal and protozoan mitochondria possess a dramatically reduced set of tRNA genes or none at all. In these cases, many tRNAs are imported from the nucleus. Translation initiation involves prokaryote-like ribosome binding sites, and *N*-formylmethionine is the initiator amino acid (see Protein Synthesis). Translation is a predominant level of gene regulation in mitochondria.

**Replication of mitochondrial DNA.** Origins of replication have been identified in some plant and fungal mitochondria, but the replication mechanism is not understood: in some plant species, the system controlling replication must discriminate between the master circle and subgenomic circles (see Replication). In animals, leading strand (**heavy strand**, or **H-strand**) replication is initiated by a large RNA primer synthesized by RNA polymerase, which is cleaved by the ribonucleoprotein

endonuclease MRP (q.v. *ribozyme*). DNA synthesis is then initiated by DNA polymerase  $\gamma$  (q.v.), extending the primer and displacing the resident strand to generate a D-loop. The extension is often aborted at this point, probably due to motifs (**termination associated sequences**) within the D-loop. Readthrough of these sites, or reinitiation of a terminated strand, results in successful replication. Replication of the complementary strand (**light strand**, or **L-strand**) is initiated on the opposite side of the circle to the heavy strand and involves a specific mitochondrial primase. This unusual, continuous replication mechanism means that mtDNA is single stranded most of the time, which perhaps contributes to its relatively high *mutation rate* (q.v.).

**Kinetoplast DNA.** Protozoans contain a single, highly specialized mitochondrion termed a **kinetoplast** which, like other mitochondria, contains its own genome (**kinetoplast DNA**, **kDNA**). The organization of the kDNA genome is remarkable, with full genome-size **maxicircles** interlinked in a complex network with a multitude of shorter derivatives termed **minicircles**. Both maxicircles and minicircles appear to replicate using a rolling circle mechanism. Primary transcripts from kDNA demonstrate the most prolific *RNA editing* (q.v.), in extreme examples involving the deletion of more than 50% of nucleotides. This is facilitated by *guide RNAs* (q.v.), which appear to be the only products synthesized by the minicircles. Other unusual properties of protozoan kinetoplasts (often shared with algal mitochondria) are the lack of organelle-encoded tRNAs (see above) and the synthesis of rRNA in short fragments. The rRNA genes are broken up and dispersed throughout the genome. Functional RNAs are presumed to arise by intermolecular base pairing.

**Organelle plasmids.** Many plant mitochondria carry plasmids along with their genomes. These are diverse in structure (with genomes of linear or circular double-stranded DNA, or single- or double-stranded RNA), but many appear to be associated with the characteristic of **cytoplasmic male sterility** (**cms**), where cytoplasmic factors block the production of viable pollen. This promotes outcrossing and may therefore be of overall benefit to the plant. The cms phenotype can be suppressed by the expression of nuclear genes. Mitochondria of the cmsT line, for instance, produce a novel, plasmid-encoded protein which is responsible for the sterility phenotype in a poorly understood manner. The phenotype can be suppressed by increasing expression at two nuclear loci, *Rf1* and *Rf2* (for restoration of fertility), which act by preventing synthesis of the protein.

Plasmids in fungal mitochondria may be cryptic or they may confer the characteristics of **longevity** (extended life span) or **senescence** (cell death). The latter trait is inherited in a cytoplasmic fashion and appears to involve rearrangement of the mitochondrial genome, induced by plasmid integration. One example is the linear *kalilo* element of *Neurospora crassa*, although the same phenotype may be conferred by self-mobile introns, some of which can also exist as plasmids.

**Endosymbiont theory and promiscuous DNA.** Organelles probably evolved from prokaryotes which formed an endosymbiotic relationship with primitive eukaryotic cells. The evidence for this model includes similarity in genome structure, gene expression mechanisms, and antibiotic sensitivity between organelles and present day bacteria, and also reflects similarity between their gene and polypeptide sequences. Over an evolutionary time scale, both chloroplasts and mitochondria have lost genes to the nucleus and have thus become dependent on their hosts. The existence of **promiscuous DNA** (sequences of organelle origin inserted either into a different type of organelle or into the nuclear genome) supports this. Promiscuous DNA sequences may represent recent transposition events, and although the transposition mechanism is unknown, it is clear that rare gene transfer from the cytoplasm to the nucleus does occur.

**Other forms of cytoplasmic inheritance.** Mitochondrial DNA and chloroplast DNA are the most common and best-characterized forms of cytoplasmic inheritance. Other organelles, however, are also believed to contain DNA (e.g. centrioles), although less is known of their coding potential and functions. Cytoplasmic genetic information is also found in the form of endogenous microbes which exist in a parasitic or symbiotic relationship with their host. Several instances have been described

where viruses, bacteria and even protozoans are harbored in the cytoplasm of a larger eukaryotic cell, are transmitted maternally and confer a phenotype upon their host. The two major examples are found in *Drosophila*: one is a virus, sigma, conferring sensitivity to carbon dioxide, and the other a protozoan cell which produces a toxin that specifically kills male flies. This sex-ratio regulator is not transmitted to the (relatively few) males that survive.

## Reference

Clayton, D.A. (1992) *Int. Rev. Cytol.* **141**: 217–232.

## Further reading

- |  |  |
|--|--|
| <p>Larrson, N.G. and Clayton, D.A. (1995) Molecular-genetic aspects of human mitochondrial disorders. <i>Annu. Rev. Genet.</i> <b>29</b>: 151–178.</p> <p>Mattson, M.P. (1997) Mother's legacy: Mitochondrial DNA mutations and Alzheimer's disease. <i>Trends Neurosci.</i> <b>20</b>: 373–375.</p> | <p>Shadel, G.S. and Clayton, D.A. (1997) Mitochondrial DNA maintenance in vertebrates. <i>Annu. Rev. Biochem.</i> <b>66</b>: 409–435.</p> <p>Stern, D.B., Higgs, D.C. and Yang, J.J. (1997) Transcription and translation in chloroplasts. <i>Trends Plant Sci.</i> <b>2</b>: 308–315.</p> |
|--|--|

**This Page Intentionally Left Blank**



## Chapter 20

# Plasmids

### Fundamental concepts and definitions

- **Plasmids** are autonomous extrachromosomal replicons found commonly in prokaryotes, but also in eukaryotes and their organelles. They are generally covalently closed and supercoiled circles of double-stranded DNA with sizes ranging from 1 to 300 kbp. They contain from one to over 100 genes. Some plasmids integrate into the host genome and are described as episomes.
- Plasmids are parasitic structures (*selfish DNA* (q.v.)) which move between populations by conjugation, transformation, transduction or cell fusion (*see Gene Transfer in Bacteria*). Minimally, plasmids encode products required for their replication and maintenance in the host. Many also carry genes which promote self-transfer between cells (q.v. *conjugative plasmids*). Most bacterial plasmids also carry genes conferring a desirable but dispensable phenotype upon their host (e.g. antibiotic resistance). Such genes are often found on transposable elements which can jump between different plasmids in the same cell. Conversely, eukaryotic nuclear plasmids are generally **cryptic plasmids** (they confer no phenotype) and some appear to have originated from chromosomal DNA. Terms used in plasmid biology are defined in *Table 20.1*.
- The fundamental distinction between plasmids and viruses is that the former are maintained as stable extrachromosomal replicons which do not encode the coat proteins that enable viruses to form infectious particles (they can therefore only exist outside the cell as naked DNA). Some viruses can exist as plasmids, e.g. the P1 prophage is maintained as a plasmid during temperate infection, but the converse is never true.
- Plasmids are used widely as cloning vectors (*see Recombinant DNA*), but this chapter primarily concerns the behavior of natural plasmids.

### 20.1 Plasmid classification

**Classification by phenotype.** Plasmids can be classified in a number of ways. The simplest but least useful criterion is the phenotype they confer upon the host cell (*Table 20.2*). In bacterial plasmids, genes which are nonessential for replication, maintenance or transfer are often carried on mobile elements which can transpose to other plasmids and into the host chromosome. The phenotype conferred by a bacterial plasmid does not, therefore, reflect any intrinsic property of the plasmid molecule itself, and several distinct phenotypes can be conferred by the same plasmid. **Megaplasms** carry genes for resistance to many antibiotics and are of great concern to health authorities. Furthermore, because of the abundance of mobile elements and the tendency for plasmids to undergo recombination, plasmid structure itself is fluid. It is therefore difficult to devise a methodical system of nomenclature for plasmids based on phenotype alone (*Table 20.3*). Many eukaryotic plasmids lack a phenotype upon which to base such a classification system; for others, the phenotype arises from the behavior of the plasmid (e.g. integration and rearrangement of the chromosome), rather than from any encoded function.

**Classification by structure.** Although most plasmids exist as double-stranded closed circles of DNA, the definition of a plasmid does not exclude other structures. A number of single-stranded circular DNA plasmids have been identified in *Streptomyces* and *Clostridium* species; and linear double-stranded DNA plasmids have been isolated from several bacterial and eukaryotic sources, e.g. linear plasmids in *Borrelia hermsii* encode variant surface antigens and are responsible for relapsing fever. In eukaryotes with linear plasmids, it is not always clear which elements should be

**Table 20.1:** Definition of some terms used in plasmid biology

Term	Definition
<b>Basic replicon</b>	The minimal region of a plasmid able to replicate in the same manner as the full-sized plasmid — typically contains the origin of replication and <b>plasmid maintenance sequences</b> , elements that regulate replication and plasmid <i>maintenance</i> (q.v.)
<b>Conjugative and nonconjugative plasmids</b>	Plasmids which can, or cannot, promote their own transfer by <i>conjugation</i> (q.v.)
<b>Copy number</b>	The average number of plasmids per cell, usually measured by direct quantification of plasmid DNA or an encoded gene product
<b>cop mutant</b>	A plasmid carrying a mutation which affects copy number
<b>Cryptic plasmid</b>	A plasmid with no apparent phenotype
<b>Curing</b>	Spontaneous or induced plasmid loss. Spontaneous curing occurs at low frequency, but can be induced by intercalating agents and some antibiotics
<b>Dislodgment</b>	The rare displacement of a resident plasmid by a second, <i>compatible</i> plasmid, a phenomenon often interpreted as <i>incompatibility</i> (q.v.), but which may reflect, e.g., the activity of a restriction endonuclease encoded by the second plasmid
<b>Episome</b>	A plasmid or virus capable of both extrachromosomal replication and integration into the host genome, e.g. the F-plasmid (see Gene Transfer in Bacteria)
<b>Homoplasmid, heteroplasmid</b>	Describing cells containing one type of plasmid, or two distinct types of plasmid (q.v. <i>plasmid segregation</i> )
<b>Incompatibility</b>	The inability of two different types of plasmid to coexist in the same cell for more than a few generations in the absence of selection for both plasmids, reflecting common replication or partition mechanisms
<b>Invertron</b>	A linear plasmid in eukaryotes with long, perfect, inverted terminal repeats
<b>Killer system</b>	A maintenance system which ensures that cured daughter cells are destroyed (see Table 20.4)
<b>Maintenance system</b>	A system which ensures that plasmids are maintained in a population of dividing cells (see Table 20.4)
<b>Miniplasmid</b>	The basic replicon of a large plasmid such as the F-plasmid
<b>One-way incompatibility</b>	A phenomenon where the introduction of plasmid type 1 into a population of cells where plasmid type 2 is established results in incompatibility, but introduction of plasmid type 2 into a population where plasmid type 1 is established does not. This reflects similar but not identical replication mechanisms in the two plasmids
<b>Partition, partition system</b>	Partition is the distribution of plasmids into daughter cells during cell division. Partition systems are maintenance systems which ensure equal partition (see Table 20.4)
<b>Plasmid</b>	An autonomous, extrachromosomal replicon which is nonessential under normal growth conditions and not part of the cellular genome
<b>Plasmid origin</b>	The plasmid locus where DNA replication begins
<b>Plasmid segregation</b>	The separation of different types of plasmid into separate daughter cells at cell division (due to incompatibility)
<b>Promiscuous plasmid (broad host-range plasmid)</b>	A bacterial plasmid with a broad host range (usually including both gram-positive and gram-negative bacteria)
<b>Prime plasmid</b>	A plasmid episome which has excised aberrantly and carries part of the host chromosome. <b>Type I prime plasmids</b> have exchanged plasmid genes for host genes, <b>type II prime plasmids</b> carry all plasmid genes plus extra host genes (q.v. <i>F' plasmid</i> )
<b>Relaxed plasmid</b>	A plasmid whose replication does not require continued protein synthesis and whose copy number increases if protein synthesis is blocked, due to the removal of a negative regulator protein
<b>Stringent plasmid</b>	A plasmid whose replication requires continued protein synthesis and whose copy number thus falls if protein synthesis is inhibited

**Table 20.2:** Some phenotypes conferred by plasmids

Phenotype	Example of plasmid and host species
None (cryptic plasmid)	<i>Saccharomyces cerevisiae</i> , 2 $\mu$ plasmid
Antibiotic resistance	Enterobacteria, R6K
Antibiotic synthesis	Enterobacteria, ColE1
Antigen gain	<i>Yersinia pestis</i> , Vwa plasmid
Gas vacuole synthesis	<i>Halobacterium</i> spp. 'satellite DNA' plasmids
Heavy metal tolerance	<i>Pseudomonas</i> spp., FP2
Carries host genes	<i>Saccharomyces cerevisiae</i> , 3 $\mu$ plasmid
Longevity	<i>Podospira anserina</i> , pAL2-1
Metabolite utilization	<i>Pseudomonas</i> spp., OCT plasmid
Restriction modification system	Promiscuous plasmid N3
Senescence	<i>Neurospora crassa</i> (mitochondria), <i>kalilo</i>
Siderophore synthesis (iron transport)	Enterobacteria, certain ColV plasmids
Sterility	<i>Zea mays</i> (mitochondria), S1 and S2
Toxin synthesis	Enterobacteria, Ent P307
Tumor induction	<i>Agrobacterium tumorfaciens</i> , Ti plasmid
UV protection (SOS system)	Promiscuous plasmid R46

described as plasmids and which as chromosomes (i.e. which elements are extrachromosomal and which are part of the genome). In essence, there is no structural distinction between plasmids and chromosomes. The usual definition of a bacterial plasmid as nonessential under normal growth conditions is meaningless unless 'normal' growth conditions are defined unambiguously. For instance, a yeast artificial chromosome is nonessential, as is the mammalian Y-chromosome, but neither element is routinely defined as a plasmid.

There are also RNA plasmids. *Viroids* (q.v.) are specialized single-stranded circular RNA plasmids which carry no genes. Some bipartite, linear, double-stranded RNA elements found in yeast also conform to the definition of a plasmid. These are known as **killer factors** because they confer a killer phenotype upon the host (Table 20.3). They encode a coat protein which encapsulates the genome rather like that of an RNA virus but cannot infect other cells and are transmitted intracellularly. The killer factors thus occupy a middle ground between a plasmid and a virus, and can be classified as either (also q.v. *subviral agents*). Note that killer factors are not the same as *killer plasmids*; the latter are more conventional yeast DNA plasmids which also confer a killer phenotype upon the host cell.

**Classification on the basis of intrinsic properties.** A better plasmid classification system uses intrinsic properties such as transfer, replication and maintenance mechanisms.

In bacteria, plasmid transfer occurs by four routes — cell fusion, transformation, transduction and conjugation (see Gene Transfer in Bacteria). The first three processes are passive with respect to the plasmid, whereas conjugation is active and many plasmids carry genes which promote self-transfer between cells by this method. Bacterial plasmids may thus be classified into two major groups, **conjugative** and **nonconjugative**. Conjugative plasmids are subdivided into families based on their particular conjugation mechanism (Box 20.1).

In eukaryotes, plasmid transfer mechanism is a less useful criterion for classification. Horizontal plasmid transfer generally occurs only when cells fuse (e.g. syngamy, or the formation of a hyphal network in fungi) or occasionally by mechanical transfer (viroids spread in this manner). Occasionally, bacteria transfer plasmids to eukaryotes, as occurs in bacteria to yeast conjugation and in the specialized case of the *Agrobacterium tumorfaciens* Ti plasmid (q.v.).

## 20.2 Plasmid replication and maintenance

**Plasmid copy number.** Plasmid replication and maintenance is a totally intrinsic property applicable to all plasmids, allowing them to be classified according to replication and partition strategy and how their copy number is regulated.

**Table 20.3:** Classification and nomenclature of plasmids based on phenotype

Type of plasmid	Phenotype	Examples
Bacteriocinogenic	Encode bacteriocins (proteins which kill or inhibit growth of other bacteria, e.g. agrocins and colicins). Also encode immunity functions so that the host cell is not destroyed. <b>Killer plasmids</b> in yeast are analogous	AgK84, CloDF13, ColE1-K30, ColV, I-K94 Designation by encoded bacteriocin (e.g. agrocins 84, cloacin DF13, colicin E1, colicins V and Ia)
Cryptic	None	2 $\mu$ , <i>Mauriceville</i>
Degradative	Catabolic	Lac, TOL, Cit Designation by substrate (e.g. lactose, toluene, citrate)
F	Fertility (conjugal transfer)	The F-plasmid (see Gene Transfer in Bacteria)
R	Resistance (e.g. to antibiotics, heavy metals)	R1, R46, RK6, multiple resistance plasmids
Recombinant (constructed <i>in vitro</i> from parts of naturally occurring plasmids)	Usually antibiotic resistance — used as dominant selectable marker (q.v. <i>plasmid vectors</i> )	pBR322, pML31, pBluescriptII Recombinant plasmids usually designated 'p' followed by letters and numbers identifying originating laboratory. Commercial recombinant plasmids usually given trivial names indicative of uses
Virulence	Enables host to cause disease. Specifically refers to those plasmids encoding direct virulence functions (e.g. toxin synthesis, tumor induction), but also applies to indirect functions such as antibiotic resistance (which increases virulence by making host resistant to medical treatments)	Ent plasmids, Ti, Ri

This is an arbitrary system which is not based on any intrinsic property of the plasmid. The nomenclature is therefore not systematic, reflecting the abundance of trivial names and the fact that many plasmids encode several distinct functions.

Plasmid replication is autonomous, but may be coupled to the replication of the host genome and may be influenced by the host and its environment. Generally, plasmids require use of the host-encoded replication machinery (e.g. DNA polymerase, RNA polymerase, DNA primase), but encode other factors required for the initiation of replication and its regulation. Initiation occurs at a specific origin and is the rate-limiting step in plasmid replication. The frequency of successful initiation establishes the characteristic average number of plasmid molecules per cell, the **copy number**.

In bacteria, larger plasmids usually have a copy number of 1–3: their replication is coupled to that of the chromosome. These are termed **single-copy plasmids** and they are usually conjugative. Smaller plasmids have higher copy numbers, typically 10–30, and are termed **multicopy plasmids**. Replication occurs randomly and is self-regulated. The plasmids are usually nonconjugative. The replication control mechanisms which maintain the correct copy number do so by measuring the relative concentration of origins in the cell (see below). If a mutation disrupts the function of a key regulator, the copy number can change substantially. Severely repressed replication can lead to plasmid loss, and unregulated replication (**runaway replication**) can increase the copy number 10-fold.



Runaway replication is capped only when another component, e.g. an enzyme or substrate for DNA replication, becomes limiting. A plasmid carrying such a mutation is known as a *cop mutant*.

The behavior of eukaryotic plasmids has been studied in the yeast *S. cerevisiae*. Plasmids with the 2 $\mu$  origin are maintained at high copy number (<100), although replication occurs only in the *S-phase* (q.v.) and is coupled to genome replication. Conversely, plasmids which resemble chromosomes (i.e. those possessing ARS origins and a centromere) are maintained at a copy number of 1–2. The centromere is dominant to plasmid origins, so that artificial plasmids containing centromeres, ARSs and 2 $\mu$  origins behave like chromosomes (q.v. *yeast cloning vectors*).

**Plasmid maintenance.** As well as controlling self-transfer and replication, plasmids also encode **maintenance functions**, ensuring both daughter cells inherit plasmids following cell division. The distribution of plasmids to the daughters of a dividing cell is **partition**. Without maintenance, spontaneous *curing* (Table 20.1) can occur during partition. Cells without plasmids are more competitive than plasmid-containing cells under normal growth conditions (i.e. in the absence of selection for the phenotype conferred by the plasmid) because they do not divert their resources to plasmid functions — therefore, cured cells rapidly increase in the population at the expense of the plasmid-containing cells. Maintenance functions have been well characterized in bacteria and fall into several groups (Table 20.4). The existence of maintenance functions places plasmids, along with viruses and transposable elements, within the category of *selfish DNA* (q.v.).

**Plasmid incompatibility.** **Incompatibility** is the inability of two plasmids to coexist in the same host strain unless conditions are imposed which select for the phenotype conferred by both plasmids. Incompatible plasmids rapidly segregate in a growing unselected population to yield two homo-plasmid strains.

Incompatibility occurs between plasmids with similar strategies for the control of replication and/or partition. The control of plasmid copy number is predominantly by negative regulation; thus two distinct single-copy plasmids with the same regulatory mechanism in the same cell will

**Table 20.4:** Plasmid maintenance functions

Maintenance function	Description
<b>Partition (<i>par</i>) system</b>	A system found in low-copy number plasmids that accurately and equally distributes plasmid copies to either side of the cell prior to division. Partition systems require both <i>trans</i> -acting factors and a <i>cis</i> -acting element located on the plasmid, and may involve attachment of the plasmid to the cell membrane. The bacterial chromosome uses a similar partition mechanism (see The Cell Cycle). High copy number plasmids lack specific partition mechanisms and rely on the high probability that each daughter will receive at least one copy
<b>Killer (<i>kil</i>) systems (addiction systems)</b>	A system in which the plasmid encodes a stable <b>killer protein</b> (a protein which is lethal to the host cell) and an unstable 'antidote' molecule acting as an antagonist of the killer protein itself or an inhibitor of its synthesis. In cured cells, the killer protein outlasts the antidote and the cell is killed
<b>Cell division delay</b>	A system which delays cell division at low plasmid copy number. The F-plasmid is able to delay cell division by inducing the <i>SOS response</i> (q.v.).
<b>Recombination systems</b>	The plasmid encodes a <i>site-specific recombination system</i> (q.v.) which counteracts homologous recombination events leading to multimerization and ensures that monomers are available for partition. A similar function, encoded by the <i>xerC</i> and <i>xerD</i> genes of <i>E. coli</i> , ensures chromosome monomerization. The yeast 2 $\mu$ plasmid also encodes a site-specific recombination system which increases its copy number

Note that partition systems, killer systems and cell division delay systems ensure that plasmid segregation is better than random by preventing the birth of cured cells. Conversely, recombination systems do not prevent the birth of cured cells, and they can only achieve at best random plasmid segregation.

repress each other's replication until cell division, when they become segregated and repression is lifted. Mixed multicopy plasmids also demonstrate mutual repression until cell division. In the daughter cells, repression is lifted, and because multicopy plasmid replication is random, each type of plasmid has an equal chance to undergo replication. The type which succeeds will then achieve a higher copy number and will be at an advantage at the next round of division, eventually leading to the generation of homoplasmid cells. Multicopy plasmids thus take longer to segregate than single-copy plasmids. Single-copy plasmids with the same partition mechanism segregate due to competition for the same 'partition site', which is presumably represented only twice in the cell. Plasmid incompatibility provides a useful system of classification according to replication and partition strategy. Mutually incompatible plasmids (i.e. those with similar strategies) are placed into an **incompatibility group (Inc group)**, of which there are approximately 30 for *E. coli* and related enterobacteria (Box 20.1).

**Mechanisms of plasmid DNA replication.** Most of the typical closed circular double-stranded DNA plasmids of bacteria replicate similarly to the chromosome, with initiation characterized by the binding of specific initiation proteins called **Rep proteins** to repetitive elements at the plasmid origin (see Replication). For many plasmids (e.g. F, R1), Rep facilitates unwinding of the origin, allowing the loading of helicase and the establishment of either a single replication fork, or two replication forks. In other plasmids (e.g. pT181 of *Staphylococcus*), the Rep protein is a *nickase* (q.v.) which initiates *rolling circle replication* (q.v.). ColE1-related plasmids do not require Rep proteins because the host RNA polymerase is used to transcribe through the origin to generate a primer for the leading strand; such plasmids are thus under *relaxed control* (see below).

Linear bacterial plasmids are similar to linear virus genomes, e.g. they possess covalently sealed ends (e.g. *Borellia* plasmids) or terminal proteins and inverted repeats (e.g. *Streptomyces* plasmids) and may replicate using the same strategies employed by viruses (see Replication for overview and Viruses for specific examples). Linear plasmids of eukaryotic organelles often encode their own polymerases, which are presumably utilized for autonomous replication. The discussion below relates to strategies for the *regulation* of replication unique to plasmids.

**Multiple origins and iterons.** Many larger plasmids have multiple origins (often associated with different control mechanisms), making it difficult to assign such plasmids to single incompatibility groups. In some plasmids (e.g. the F plasmid), one particular origin is favored, and the role of the extra origins is uncertain. In others (e.g. R6K) the alternative origins are used with equal frequency. Many recombinant plasmid vectors have been deliberately designed to incorporate both prokaryotic and eukaryotic origins, allowing them to be maintained in both types of cells (q.v. *shuttle vector*).

Plasmid origins are often characterized by essential repetitive sequences termed **iterons**. These are common motifs in plasmids (e.g. plasmid RK2) and in some phages which can replicate like plasmids (e.g. bacteriophages P1 and  $\lambda$ ), as well as the bacterial chromosome itself (q.v. *origin of replication*). Plasmid iterons are binding sites for the Rep proteins, which may act alone or may associate with host initiator proteins (e.g. DnaA in *E. coli*). In plasmid RK6, for instance, regardless of which of the three origins is used, a sequence of seven direct tandem repeats located within the  $\gamma$ -origin is essential for replication.

**Control of plasmid replication by antisense RNA.** Plasmids with a ColE1 type origin of replication (including the vast majority of plasmid cloning vectors) initiate leading-strand DNA synthesis from a single RNA primer generated by the host RNA polymerase (q.v. *primer*, *priming strategy*). The origin is actually transcribed on both strands to yield two transcripts, RNA I and RNA II. RNA II is the primer, whereas RNA I, which is complementary to part of RNA II, acts as a repressor by sequestering RNA II into an inactive duplex. DNA synthesis begins at the origin, but transcription of RNA II begins 555 bp upstream of the origin and continues through it and beyond. The transcript must therefore be processed by cleavage at the origin to yield a functional primer. This processing is

carried out by RNaseH (q.v. *nucleases*) and is facilitated by the three-stem loop secondary structure adopted by RNA II.

Interaction between RNA I and RNA II culminates in their full *hybridization* (q.v.), which disrupts the secondary structure of RNA II and prevents it hybridizing to the DNA. This in turn prevents the cleavage reaction and results in a replication block. *In vitro*, RNA I is able to anneal to RNA II only when the latter is between 100 and 360 nt in length. This means that RNA I must hybridize to the nascent RNA II transcript during the early phase of its synthesis to repress replication. A plasmid-encoded protein called Rom enhances the rate of RNA I: RNA II hybrid formation once RNA II is greater than 200 nucleotides in length and is thus a key regulator of ColE1 replication.

The replication of the R1 plasmid is also regulated by antisense RNA. The plasmid-encoded initiator protein, RepA, is essential for the initiation of replication, but transcription of an adjacent gene generates an antisense RNA (*copA*) which binds to the leader region of the *repA* mRNA (designated *copT*) to prevent its translation (q.v. *antisense RNA*, *translational control*). The *copA* and *copT* RNAs both form secondary structures, and mutational analysis has shown that the interaction between them may involve these secondary structures rather than full duplex formation (a so-called **kissing complex**).

The prevalence of antisense RNA regulators in plasmids suggests that RNA offers a unique strategy for replication control for which protein is insufficient. In many diverse plasmids the same mechanism occurs: a short antisense inhibitory factor anneals to the 5' end of an essential functional RNA and blocks its activity. A possible explanation is that the maintenance of plasmid copy number requires an unstable regulator so that the plasmid population can respond rapidly to deviations from the standard copy number. RNA is a suitable candidate for such an unstable regulator (q.v. *antisense RNA*).

**Relaxed and stringent control.** Plasmid replication may be under **relaxed** or **stringent control**. If cellular protein synthesis is inhibited, for example by chloramphenicol, some plasmids continue to replicate and their copy number increases above normal vegetative levels. These are **relaxed plasmids**, and are typically small multicopy plasmids. Other plasmids cease to replicate along with the host chromosome when *de novo* protein synthesis is inhibited. These are described as **stringent plasmids**, and are typically large, single-copy, conjugative plasmids. Most general purpose plasmid cloning vectors are under relaxed control.

The ability of relaxed plasmids to continue to replicate in the absence of host protein synthesis reflects the stability of host proteins required for the initiation of replication, and the instability of plasmid-encoded regulators. For instance, plasmid ColE1 is under relaxed control because the initiation factor Rom has a negative role — it prevents initiation by facilitating the pairing of an antisense RNA with the leading strand primer. In the absence of Rom, plasmid replication is derepressed and copy number increases (very high number cloning vectors have no functional *rom* gene). Conversely, stringent plasmids such as F require a positively acting Rep protein for initiation. In the absence of Rep, plasmid replication is blocked regardless of the availability of host DNA replication proteins.

**Box 20.1:** Inc groups

**Problems associated with incompatibility.** Plasmids are assigned to **incompatibility groups (Inc groups)** based on their mutual incompatibility, a phenomenon in which a growing population of *heteroplasmid* cells (q.v.) segregates into two *homoplasmid* subpopulations (q.v.) because the plasmids have identical replication or partition control strategies. Incompatible plasmids can appear compatible, however, if they undergo recombination, or if there is a high copy number and segregation takes many generations. Conversely, compatible plasmids can appear incompatible if they undergo *dislodgment* (q.v.) or if the cell contains several undetected cryptic plasmids which express incompatibility. Finally, large plasmids often have multiple distinct origins (**multireplicon plasmids**) and cannot be assigned unambiguously to a unique Inc group.

**Inc group nomenclature.** In enterobacteria, incompatibility (Inc) groups are designated by a letter representing a particular plasmid conjugation system, followed by roman numerals and/or Greek letters to indicate incompatibility status (reflecting the replication and partition mechanism). Plasmids with F-type conjugation systems are divided into two incompatibility groups, IncFI and IncFII, with different replication control strategies, but other naturally occurring

conjugative plasmids have unique replication strategies. Plasmids from nonenteric bacterial species have also been assigned to incompatibility groups. In *Pseudomonas* spp., Inc groups are designated IncP1, P2, P3 etc., representing 15 or more different replication systems, the conjugation systems being less well characterized. There is overlap between the enterobacterial and *Pseudomonas* Inc groups representing the broad host-range plasmids; thus, e.g., groups IncP1, IncP3 and IncP4 are the same as enterobacterial groups IncP, IncC and IncQ, respectively.

**Basis of conjugation system classification.** Conjugation systems in enterobacteria are classified according to pilus morphology (which may be, for example, rigid or flexible, thick or thin) and the types of bacteriophage which adsorb to the pilus (see Viruses). F-type conjugation systems, for instance, are characterized by a thick but flexible pilus to which M13-like DNA phages and RNA phages such as Q $\beta$  and R17 adsorb. Incompatibility groups IncFI and IncFII represent two groups of plasmids with F-type conjugation systems and distinct replication strategies, the first exemplified by the F-plasmid itself, the second by plasmids R1 and R100.

**Further reading**

- Eguchi, Y., Itoh, T. and Tomizawa, J. (1991) Antisense RNA. *Annu. Rev. Biochem.* **60**: 631–652.
- Hiraga, S. (1992) Chromosome and plasmid partition in *E. coli*. *Annu. Rev. Biochem.* **61**: 283–306.
- Kornberg, A. and Baker, T.A. (1992) Plasmids and organelles. In: *DNA Replication* (2nd edn), pp. 637–689. W.H. Freeman, New York.
- Nordström, K. and Austin, S.J. (1989) Mechanisms that contribute to the stable segregation of plasmids. *Annu. Rev. Genet.* **23**: 37–69.
- Summers, D.K. (1996) *Plasmid Biology*. Blackwell Science, Oxford.



## Chapter 21

# The Polymerase Chain Reaction (PCR)

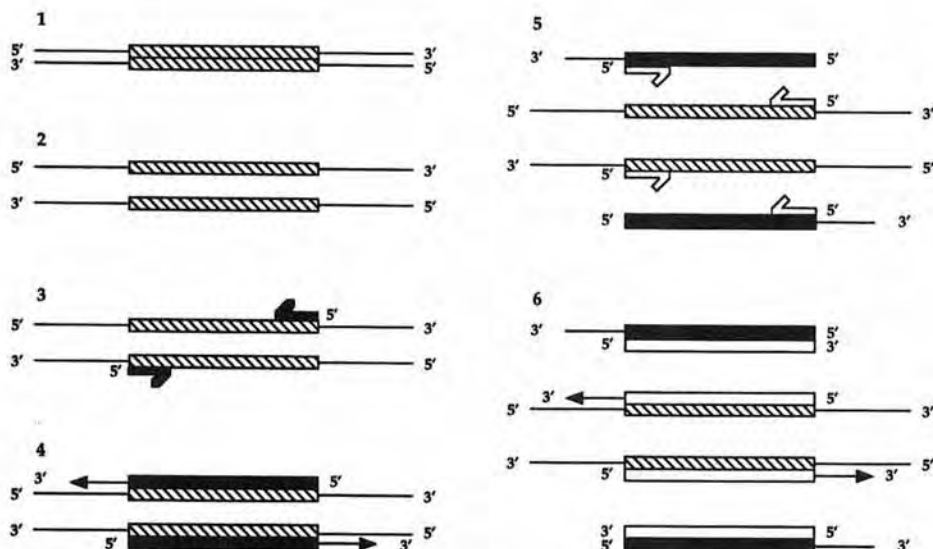
### Fundamental concepts and definitions

- The **polymerase chain reaction (PCR)** is a technique for amplifying DNA sequences *in vitro*, i.e. the DNA is not replicated in a cell, as is the case in *molecular cloning* (q.v.), but using purified enzymes.
- The basic PCR uses pairs of oligonucleotide primers designed to flank the target for amplification. The primers anneal to opposite DNA strands and face inwards, so that DNA synthesis proceeds across the central region. The reaction involves three stages carried out at different temperatures: **denaturation** of the double-stranded DNA (carried out at above 90°C), **annealing** of primers to the resulting single-stranded templates (carried out at ~50°C, the optimal temperature depending on the primer sequence), and **primer extension** to synthesize new DNA across the target region (carried out at ~70°C). Each group of three reactions is termed a **PCR cycle** and theoretically doubles the amount of the original target sequence (Figure 21.1).
- The PCR is performed in a small volume (20–100 µl) either in individual tubes or in multiwell plates, using a heating block with an automated thermal cycler for precise temperature control. The reaction contains the source DNA, the primers, the four deoxyribonucleoside triphosphates, a thermostable DNA polymerase and its reaction buffer, the most critical component of which is  $Mg^{2+}$ .
- The PCR is advantageous because it is quick and sensitive compared to transditional cloning methods, and remains efficient even when the source DNA is heavily degraded or must be isolated from difficult sources such as fixed tissue. However, the cycling tends to be error prone, the size of the products is limited and there is an absolute requirement for prior knowledge of target sequence.

### 21.1 Specificity of the PCR reaction

**Amplification of unique sequences.** Since all DNA polymerases require *primers* (q.v.) to initiate strand synthesis (see Replication), the target for PCR amplification can be specified by designing primers to anneal to particular unique DNA sequences. **PCR primers** are chemically synthesized *oligonucleotides* (q.v.) and their design is crucial for the success of the PCR. The most efficient primers for specific amplification reactions are 17–30 nucleotides in length, as this represents a sequence unlikely to be repeated by chance in the *unique sequence DNA* (q.v.) of higher eukaryotes. Primers usually have a GC content of approximately 50% and lack runs of the same nucleotide or significant secondary structures, both of which can cause looping out of primer residues and stabilization of the primer at an erroneous binding site. Additionally, primer pairs should not be self-complementary, as they then form **primer dimers** which act as templates for extension, resulting in spurious products.

The PCR is advantageous because it is quick and sensitive compared to transditional cloning methods, and remains efficient even when the source DNA is heavily degraded or must be isolated from difficult sources such as fixed tissue. However, the cycling tends to be error prone, the size of the products is limited and there is an absolute requirement for prior knowledge of target sequence.



**Figure 21.1:** The basic PCR reaction. In the first PCR cycle, the original target sequence (1), shown as hatched lines, is denatured (2) and primers anneal as shown (3). Primer extension across the central region (4) doubles the amount of target sequence, although the sizes of the first-cycle products (shown in black) are nonspecific. In the second cycle, the amplified target sequence is denatured and primers anneal as shown (5). Primer extension (6) doubles the amount of target sequence. The second-cycle products are shown in white. Half of them are nonspecific in size, whereas the other half are defined by the primer annealing sites and exactly correspond to the size of the target sequence. In subsequent cycles, these specific products accumulate exponentially, whilst the nonspecific products accumulate in a linear fashion and contribute little to the final product mix.

Some loss of specificity may occur if primers anneal at lower temperatures, as mismatches are tolerated and then become stabilized on the template by primer extension. This occurs when the components of the reaction are mixed at room temperature and can be avoided by **hot-start PCR**, where an essential component, usually the enzyme, is added to the reaction when it has reached the annealing temperature. Specificity can also be increased by the use of **nested primers**, where products from one amplification are subsequently amplified with a second set of primers which flank the same target site but internal to the original primers. Any spurious products amplified in the first reaction by mispriming are unlikely to also possess the correct sites for the internal primers, so only genuine products will be amplified in the second reaction (**nested PCR**).

**Applications for unique sequence amplification.** There are many applications for the amplification of unique sequences. As a diagnostic technique it can be used to confirm the presence of a given sequence in a complex source (e.g. to confirm the presence of a transgene or a plasmid insert), to detect polymorphisms (e.g. variation in minisatellite DNA; q.v. *DNA typing*) and to detect unknown mutations. Potentially unique sequences can be used to generate *sequence tagged sites* (q.v.) for applications in genome mapping and positional cloning (see *Genomes and Mapping, Recombinant DNA*). The inability of a primer to bind can also be diagnostic. **Allele-specific PCR** can identify, for example, point mutations responsible for diseases, if primers are designed to anneal at the site of the mutation and can anneal only to particular alleles. This type of detection is more sensitive than conventional *hybridization analysis* (q.v.) because it can detect minor sequence differences which might be undetectable by Southern hybridization. Additionally, since only small amounts of source material are required, it is useful where clinical samples are limited, e.g. chorionic villus sampling.

The PCR can also be used as a preparative technique to amplify specific fragments of DNA from

complex sources without recourse to molecular cloning. The main disadvantage of the PCR in this respect is its size limitation. PCR amplification is of little use for genomic cloning and tends to generate only partial cDNA fragments, equivalent to *expressed sequence tags* (ESTs) (q.v.). Also, there is an absolute requirement for some previous knowledge of target sequence, so although PCR is quick and economical, it is unlikely to fully replace conventional library-based methods of screening which provide a direct route to full-length clone isolation and allow diverse screening strategies such as immunological and complementation screening (q.v. *expression cloning*).

**Amplification of related sequences.** Not all PCR applications aim to generate specific products. It is possible for the PCR to identify families of related sequences or to amplify DNA from one species based on sequence information from another.

In such cases, specific primers may be designed around a highly conserved domain, allowing products with differing internal sequences to be identified. Another approach is to use **degenerate oligo-nucleotide primers** (DOP-PCR) a mixture of primers with alternative nucleotides at certain positions. This strategy, known as **homology screening**, involves the design of primers around a conserved domain but incorporating all known sequence variations in the family. Such an approach has successfully expanded several gene families, including the *POU domain* transcription factors (q.v.) and the *cyclins* and *cyclin-dependent kinases* (q.v.). A further use of degenerate primers is to amplify a specific target sequence corresponding to a known protein, when only the polypeptide sequence of the protein is known. In this case, the use of DOP-PCR reflects the degeneracy of the genetic code and is analogous to the use of degenerate oligonucleotide probes to isolate clones from cDNA libraries (see Recombinant DNA).

**Amplification of unrelated sequences.** A major application of PCR is the identification of microorganisms in, for example, infected tissue or contaminated water, allowing the correct treatment or cleansing strategy to be employed. Traditional tests based on morphological, metabolic and behavioral phenotypes are laborious and available for only a few species. The PCR can be used as a simple diagnostic test allowing the unambiguous identification of microorganisms based on their nucleotide sequence. An added advantage is that many different species can be assayed at once by using multiple sets of primers in the same reaction (**multiplex PCR**), each of which generates a specific diagnostic-sized product.

**Amplification of arbitrary sequences.** In certain cases, the aim of the PCR is not to generate one or more specific products or related products, but to produce a collection of purely arbitrary sequences which are used as the basis of further analysis. The two major applications of this strategy are cDNA cloning and gene mapping. The amplification is achieved using **arbitrary primers**, i.e. short primers (9–15 nt) which anneal to many sites and amplify a random collection of unrelated products. The arbitrary amplification of cDNA produces a library of partial cDNA fragments, ESTs, which can be used in the same way as sequence-tagged sites for genome mapping, but with the advantage of pinpointing actual genes. ESTs can also be used to characterize differentially expressed genes (q.v. *differential display PCR*, below). The arbitrary amplification of genomic DNA may also be useful in producing markers for genetic and physical mapping. **RAPDs** are **randomly amplified polymorphic DNA** markers used extensively to map plant genomes. Arbitrary genomic PCR products are separated by electrophoresis, and polymorphic products are easily identified as variable bands. Co-segregation of a trait with a particular polymorphic marker is evidence of linkage, and by characterizing the product and developing more specific primers, RAPDs can be converted into sequence-tagged sites for physical mapping. (See Genomes and Mapping).

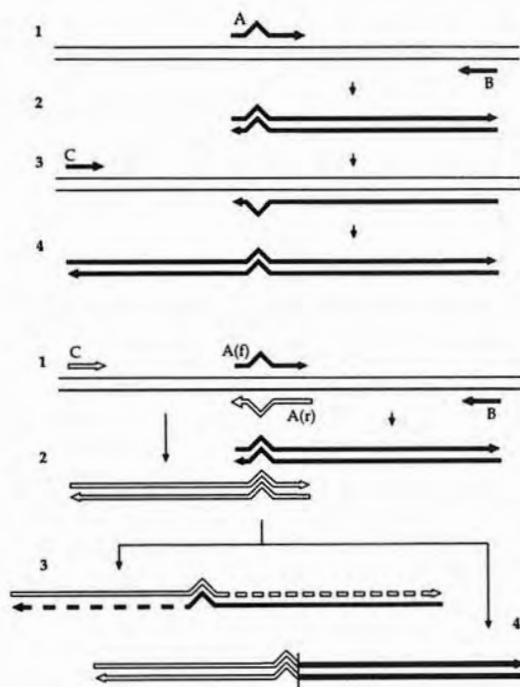
The ultimate form of arbitrary PCR is to amplify all target sequences. One way to achieve this is to add *linkers* (q.v.) to all source DNA molecules and amplify indiscriminately using primers that recognise the linkers (**ligation-adaptor, PCR, linker-primed PCR**). Alternatively, DOP-PCR can be carried out with *completely* degenerate oligonucleotides so *all* sequences are recognised.

**PCR mutagenesis and augmentation.** Primer design can be used not only to amplify sequences but also to alter them. This has two major applications, the introduction of point mutations (a form of *in vitro* mutagenesis, q.v.) and end-modification of PCR products to facilitate further manipulation.

In the first approach (**PCR mutagenesis**) primers are designed to mismatch with their target, and conditions are chosen so that annealing is still permitted. Amplification thus introduces the mutation into the amplified product, although only at the ends, which are specified by the primer. If a central replacement is required, the mutant product can itself be used as a large primer (a **megaprimer**) on the original template for extension, or an overlapping pair of PCR reactions can be carried out followed by combination of the two products, either by restriction digestion and ligation (if sites are available), or by **end-to-end recombination PCR**. These strategies are compared in Figure 21.2.

In the second approach, novel sequences are added to the 5' primer ends which do not pair with the template DNA, but allow extra sequences to be added on to each end of the PCR product. This has numerous applications: the addition of restriction endonuclease sites to facilitate subcloning, the addition of universal primer binding sites for sequencing, and the addition of bacteriophage promoters to facilitate *in vitro* transcription.

**Downstream applications.** Where a PCR product is required in large quantities and for numerous downstream applications, traditional *subcloning* techniques (q.v.) are still the most convenient form



**Figure 21.2:** PCR mutagenesis. The introduction of mutations into PCR products can be achieved using mismatching primers, but this only allows mutation at the ends of the PCR product. Two strategies allow any site in a given target fragment to be mutated. In the megaprimer strategy, amplification using mismatching primer A and primer B (1) generates a half-product (2) which can be used as a large primer, in combination with primer C (3), to generate the full-length mutated sequence (4). In the recombination PCR strategy, two amplifications are performed (1) to generate two half-products (2). Two mismatched primers are used with opposite polarities (A forward, A reverse) so the ends of the products overlap. The products can be mixed and joined by end-to-end recombination PCR (3) or, if there is a convenient restriction site, by ligation (4).



**Table 21.1:** Methods for cloning PCR products

Method	Advantages and disadvantages
Blunt-end cloning	Any PCR products may be cloned in this manner but, e.g. <i>Taq</i> products require polishing (removal of overhanging terminal bases to generate blunt ends). Subcloning is nondirectional
T-vectors	T-vectors have overhanging 5' thymidylate residues and allow efficient cloning of, e.g. <i>Taq</i> polymerase products which possess overhanging 3' adenylate residues. No processing of products is required, but subcloning is nondirectional
Linker primers	Inclusion of linker sequences in primers adds restriction sites to the end of PCR products. Products require restriction digestion prior to subcloning, but the process is efficient and directional if different 5' and 3' sites are used. The absence of internal restriction sites in the PCR product must be confirmed prior to subcloning
DISEC/TRISEC	Dinucleotide/trinucleotide sticky-end cloning. Products and vector require partial filling and exonuclease treatment, but subcloning is efficient and directional

of *in vitro* manipulation. Thus, PCR products must be inserted into cloning vectors and treated in the same manner as any other passenger DNA molecule (*see* Recombinant DNA). The error-prone thermostable DNA polymerases, such as *Taq* polymerase, tend to add a single nucleotide (usually dATP) to the 3' end of PCR products in a template-independent manner, whereas proofreading enzymes such as *Pfu* polymerase generate blunt-ended products. There are several choices of subcloning strategy for PCR products, which are listed in *Table 21.1*.

## 21.2 Advances and extensions to basic PCR strategy

**Reverse-transcriptase PCR (RT-PCR).** The PCR amplification of cDNA is termed RT-PCR as it involves an initial reverse transcription step prior to amplification. The first-strand cDNA synthesis reaction is carried out in a conventional manner (*q.v.* *cDNA synthesis*), but the second strand is synthesized in the first PCR cycle. Thermostable DNA polymerases with reverse transcriptase activity have recently been described, which may allow RT-PCR to be performed as a single reaction in the future.

RT-PCR is an extremely sensitive method for amplifying the sequences of RNA molecules, and can therefore be used to detect and isolate cDNA sequences from complex sources. Its advantages as a preparative technique over conventional library-based cDNA cloning methods (*see* Recombinant DNA) include speed, the requirement for only small amounts of target material, and tolerance of large amounts of contaminating rRNA and tRNA, allowing whole cellular RNA to be used as the source. Its disadvantages include the tendency to produce only part-length products (ESTs) and that, unlike a library, it is not a permanent resource. Arbitrary RT-PCR, using either random hexamers or longer arbitrary primers and oligo-dT primers which hybridize to the polyadenylate tails of mRNA molecules, can be used to generate representative pools of expressed sequence tags for further analysis. One application, as discussed above, is the assembly of collections of markers for physical genome mapping — expressed sequence tags are more likely to be unique sequences than those found in noncoding DNA. A second application is the identification of differentially expressed genes. In this technique, known as **differential display PCR** or **mRNA fingerprinting**, ESTs are amplified from different sources and compared side-by-side by electrophoresis. Differentially expressed genes are identified as extra or missing bands on the gel, and this can be a sensitive approach for detecting regulated gene products. A similar approach identifying differences in genomic DNA due to changes in the number or size of restriction fragments is termed **representational difference analysis (RDA)**.

As a diagnostic technique, RT-PCR can be adapted to both quantify and localize gene products. **Quantitative PCR** involves the amplification of a target cDNA in competition with known amounts of competing fragments added to the source prior to amplification. The amount of product generated by amplification of each competing fragment provides a linear scale which allows quantification of the true target. **In situ PCR** is a RT-PCR reaction which amplifies mRNA in its natural location in the cell, and allows the precise localization of even the most scarce transcripts. In each case, the PCR-based approach is more sensitive than the traditional hybridization-based methods (q.v. *northern blot*, *RNase protection*, *in situ hybridization*).

**Amplification of unknown sequences.** The basic PCR reaction is limited to the amplification of DNA lying between two defined primers, and therefore requires a previous knowledge of the sequence. However, it is often necessary to characterize the (unsequenced) DNA flanking a region for which the sequence is known, e.g. to clone a full length cDNA starting with an expressed sequence tag, or to examine the regulatory elements which lie upstream of an amplified gene segment.

**Inverse PCR** (or **inside out PCR**) allows the amplification of flanking sequences in genomic DNA. If DNA is digested with a restriction endonuclease, the target sequence for a given PCR will be embedded in a larger fragment of DNA containing both 5' and 3' flanking regions. This fragment can be circularized using DNA ligase and the flanking regions amplified using the same primer pair which generated the original (internal) product, but instead facing outwards so that they amplify the remainder of the circle. Other strategies with similar aims involve the addition of linkers to the ends of linear DNA molecules (to be used as primer binding sites), and the exploitation of intramolecular secondary structures such as stem-loops (**panhandle PCR**) and internal bubbles (**vectorette PCR**) for strand specific priming. Linkers are added to the ends of restriction fragments, and amplification is carried out with a known (gene-specific) primer and a linker primer. Products generated by amplification across the region between the two primers correspond to unknown flanking DNA. If the gene-specific primer is biotinylated, the products can be captured with streptavidin (**capture PCR**), and those nonspecific products generated by amplification between two linkers are discarded (q.v. *biotin streptavidin system*). A similar strategy is used to produce full-length cDNAs from expressed sequence tags. This technique, termed **RACE** (**rapid amplification of cDNA ends**) utilizes a gene-specific primer and an oligo-dT primer which hybridizes to the polyadenylate tail of the cDNA to specifically amplify the 3' end (**3' RACE**). The 5' end can be amplified in a similar fashion by adding an artificial tail to the second cDNA strand using *terminal transferase* (q.v.). Alternatively, a linker addition strategy similar to capture PCR can be employed (**5' RACE**).

**Asymmetric PCR.** In the PCRs discussed above, the primer pairs are added in equal amounts to generate equivalent numbers of copies of each target strand and thus produce a population of double-stranded amplification products. Another deviation from the standard PCR methodology is **asymmetric PCR** (**single-stranded PCR** (**ssPCR**)), where one of the primers is added in great excess, so that after a limited number of rounds of normal PCR amplification, one of the primers becomes depleted and the reaction switches to a linear accumulation of single strands. There are many applications for this technique, including the production of single-stranded DNA for conformational electrophoresis testing (q.v. *mutation detection*), for probe synthesis, or for sequencing. DNA sequencing can also be carried out in a variation of single-stranded PCR where only one primer is used and dideoxyribonucleotide triphosphates are added to the reaction (q.v. *cycle sequencing*).

### 21.3 Alternative methods for *in vitro* amplification

**Other amplification systems.** The PCR is the most widely applied method for *in vitro* enzymatic amplification, but it is not unique. Several other techniques have been developed for particular applications and these are listed in Table 21.2.

**Table 21.2:** Alternative *in vitro* amplification methods

System	Description and applications
Ligase chain reaction (LCR)	A technique used for the sensitive detection of a DNA sequence variation and discrimination between alleles. Two primers are used which anneal at adjacent sites on target DNA. If the target sequence is present, the primers can be ligated together and will act as sites for further annealing and ligation, leading to amplification of the primer ligation product. In the absence of the target, no ligation and no amplification occurs. Conditions can be chosen where only exactly complementary primers anneal, allowing sensitive discrimination between alleles differing by a point mutation ( <b>allele-specific ligation</b> ) (also q.v. <i>allele-specific hybridization</i> , <i>padlock probes</i> )
Transcription-based amplification; nucleic acid sequence-based amplification (NASBA)	A rapid amplification procedure based on transcription and reverse transcription. The target DNA or RNA is amplified by standard techniques using primers carrying a phage promoter. Further amplification is then carried out by transcription using phage RNA polymerase. The transcripts are reverse transcribed to generate cDNA for further amplification. The transcription reaction can generate a one thousand-fold amplification at each round compared to two-fold for PCR
Strand displacement reaction	A technique similar to PCR but not requiring a denaturation step in each cycle. After extension, primers are separated from their extension product by restriction endonuclease cleavage. Further extension then proceeds by strand displacement from the resulting nick. The amplified DNA must be protected from cleavage

## References

- Innis, M.A., Gelfand, D.H., Sninsky, J.J. and White, T.J. (eds) (1990) *PCR Protocols. A Guide to Methods and Applications*. Academic Press, San Diego, CA.
- McPherson, M.P. and Hames, B.D. (eds) (1995) *PCR 2: A Practical Approach*. IRL Press, Oxford.
- McPherson, M.P., Quirke, P. and Taylor, G.R. (eds) (1991) *PCR: A Practical Approach*. IRL Press, Oxford.
- Audic, S. and Beraud-Colomb, E. (1997) Ancient DNA is 13 years old. *Nature Biotech.* **9**: 855–858.
- Caetanoanollas, G. (1996) Scanning nucleic acids by *in vitro* amplification — new developments and applications. *Nature Biotech.* **14**: 1668–1674.
- Erlach, H.A. and Arnheim, N. (1992) Genetic analysis using the polymerase chain reaction. *Annu. Rev. Genet.* **26**: 479–506.
- Jansson, J.K. and Prosser, J.I. (1997) Quantification of the presence and activity of specific microorganisms in nature. *Mol. Biotech.* **7**: 103–120.
- McClelland, M., Mathieudauda, F. and Welsh, J. (1995) RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends Genet.* **11**: 242–246.
- Nanda, S.K. and Jain, S.K. (1995) *In vitro* nucleic acid amplification systems. *Curr. Sci.* **66**: 421–429.
- Ohan, N.W. and Heikkila, J.J. (1995) Reverse transcription polymerase chain reaction — an overview of the technique and its applications. *Biotech. Adv.* **11**: 13–29.
- Pena, S.D.J. and Chakraborty, R. (1994) Paternity testing in the DNA era. *Trends Genet.* **10**: 352–356.
- Powell, W., Machray, G.C. and Provan, J. (1996) Polymorphism revealed by simple sequence repeats. *Trends Plant Sci.* **7**: 215–222.
- Register, J.C. (1997) Approaches to evaluating the transgenic status of transformed plants. *Trends Biotech.* **15**: 141–146.
- Reischl, U. and Kochanowski, B. (1995) Quantitative PCR — a survey of the present technology. *Mol. Biotech.* **3**: 55–71.
- Sagerstrom, C.G., Sun, B.I. and Sive, H.L. (1997) Subtractive cloning: Past, present and future. *Annu. Rev. Biochem.* **66**: 751–783.
- Slatko, B.E. (1996) Thermal cycle dideoxy sequencing. *Mol. Biotech.* **6**: 311–322.
- Whelen, A.C. and Persing, D.H. (1996) The role of nucleic acid amplification and detection in the clinical microbiology laboratory. *Annu. Rev. Microbiol.* **50**: 349–373.

**This Page Intentionally Left Blank**



## Chapter 22

# Proteins: Structure, Function and Evolution

### Fundamental concepts and definitions

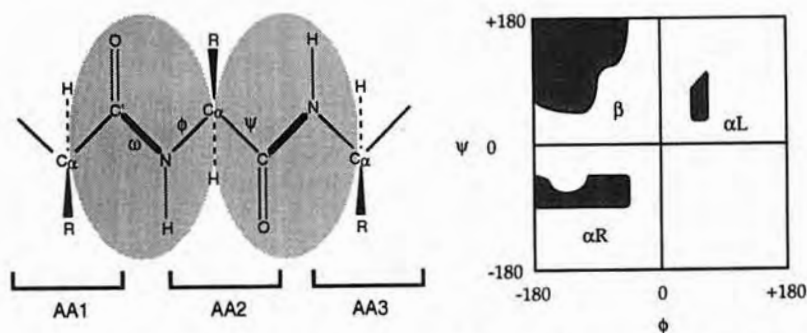
- Proteins are macromolecules composed of one or more polypeptide chains, each of which is a series of amino acid residues linked end-to-end by peptide bonds. The **primary structure** of a polypeptide is the linear sequence of residues, which is derived from a basic set of 21 amino acids specified by the genetic code (*see* The Genetic Code).
- Once incorporated into a polypeptide chain, individual amino acids may undergo **cotranslational** or **posttranslational modification**. This may involve simple chemical modification (e.g. methylation, hydroxylation), or the addition of large chemical groups (e.g. glycosylation, acylation). Minor modifications may be permanent and necessary for correct folding and protein function (e.g. hydroxylation of proline and lysine residues in collagen); more often, minor modifications are reversible, allowing the regulation of protein activity (e.g. phosphorylation of serine, threonine and tyrosine residues in eukaryotic signaling proteins; *see* Signal Transduction). The covalent addition of bulky chemical adducts is usually a permanent modification, and concerns either protein trafficking and processing in the cell (e.g. many secreted proteins are glycosylated), or the joining of a conjugated protein to its prosthetic group (e.g. conjugation of the heme group to cytochrome C).
- A newly synthesized polypeptide must fold into its **native conformation** (the conformation in which it is biologically active). Several levels of structural organization are observed in native proteins. **Secondary structure** refers to repeating local configurations of residues, e.g.  $\alpha$ -helices and  $\beta$ -sheets, generated mainly by the need for hydrogen bonds to form between peptide bond units. Secondary structures often group together to form larger and more complex arrangements (such as helix-turn-helix motifs), which may have particular roles interacting with other cellular components. **Tertiary structure** refers to the overall three-dimensional configuration of a polypeptide, reflecting many different types of chemical bonds, both covalent and noncovalent, which stabilize the most energetically favorable folding conformation. **Quaternary structure** refers to the arrangement of polypeptide subunits in a multimeric protein. It is unclear exactly how proteins adopt their native state given the very large number of alternative (**denatured**) conformations available. Folding is guided, either intrinsically (using a nucleation center or a series of favorable intermediate states), or extrinsically (using proteins termed **molecular chaperones**). Many proteins can exist in alternative stable conformations which show differential activity. Switching between these states can be regulated by covalent modification or noncovalent interactions with other molecules (**allostery**).
- Proteins can be grouped into families based on structural and functional similarities. In classical families, related proteins have evolved by **gene duplication and divergence**: an ancestral gene duplicates, and the nonallelic copies diverge by accumulating mutations. The mutations cause divergence of sequences, structures and expression patterns, allowing adaptation to novel functions. In some tandemly arranged gene families, the divergent pressure of mutation is countered by **sequence homogenization**, caused by nonallelic recombination events, i.e. unequal exchange and/or gene conversion. More complex patterns of evolution involve **chimeric proteins**, which have arisen by recombining **modules** (contiguous functional segments) of pre-existing proteins. The evolution of chimeric proteins has been facilitated by the intron/exon organization of eukaryotic genes: single exons may duplicate in tandem, resulting in a repetitive protein module within a single polypeptide (**exon repetition**); alternatively, exons may duplicate by dispersal, and colonize other genes to generate chimeric proteins (**exon shuffling**). Some protein modules, such as the zinc finger DNA-binding module, are very widely distributed.

- Proteins represent the predominant level of gene function, but whereas the resolution of protein structure is becoming routine, the determination of protein function is not always straightforward. In principle, the function of a protein can be determined by the analysis of structure, expression patterns, the effects of mutation and interactions with other cellular components. In some cases a single approach may be sufficient, but a combination is often required where one approach is uninformative (e.g. where mutation has no effect because of genetic redundancy). It is not yet possible to predict protein tertiary structure from primary sequence information, but thanks to the availability of an ever growing resource of primary sequence information, the structures and functions of many proteins can be inferred by homology to previously characterized molecules. The intensive analysis of single proteins is likely to be surpassed in the near future by genome-wide functional analysis (**functional genomics**), involving systematic investigation of the structure, expression, mutation and interactions of all the proteins synthesized in the cell (the **proteome**).

## 22.1 Protein primary structure

**The structure of polypeptide chains.** Polypeptides are linear chains of amino acid residues (Box 22.1) covalently joined by **peptide bonds** (bonds formed by condensation of the amino ( $-\text{NH}_2$ ) and carboxylic acid ( $-\text{COOH}$ ) groups of adjacent amino acids; Figure 22.1). The first residue in the polypeptide chain retains its amino group, and the last retains its carboxylic acid group. The ends of a polypeptide are thus termed the **N-terminus** and **C-terminus** respectively. The **primary structure** of a polypeptide is the **amino acid sequence**, conventionally read in the  $\text{N} \rightarrow \text{C}$  direction. This is analogous to the primary structure of nucleic acids — the nucleotide sequence — and the  **$\text{N} \rightarrow \text{C}$  polarity** of a polypeptide is colinear with the  $5' \rightarrow 3'$  polarity of its cognate mRNA (see Nucleic Acid Structure). All nascent polypeptides begin with methionine (see Protein Synthesis), although this is often cleaved off posttranslationally. In prokaryotes and eukaryotic organelles, the amino group of the N-terminal methionine is blocked by formylation after conjugation to the *initiator tRNA* (q.v.).

**Bond conformations in polypeptide chains.** As polypeptides are macromolecules built from amino acid monomers, it is tempting to think of them as beads on a string, with the amino acids representing rigid beads and the peptide bonds representing flexible linkers that allow folding. In fact, the converse is true. The peptide bonds adopt a rigid, planar conformation whereas the bonds within the amino acid residues, the  $\phi$ -bond ( $\text{C}\alpha\text{--N}$  bond) and the  $\psi$ -bond ( $\text{C}\alpha\text{--C}'$  bond), show different degrees of rotation (Figure 22.1). The peptide bonds ( $\omega$ -bonds) usually adopt a *trans*-configuration



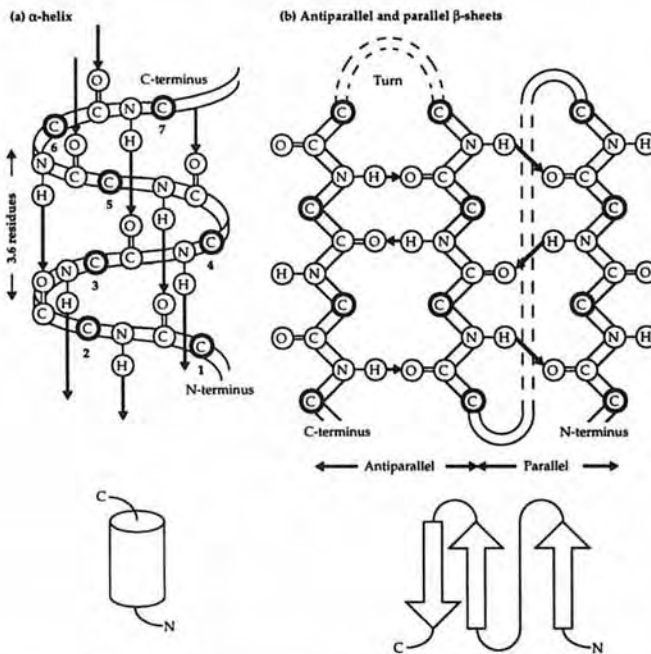
**Figure 22.1:** Bond conformation in the polypeptide backbone. Polypeptides consist of amino acyl residues (shown as AA1, AA2, AA3 in the left panel) linked by peptide bonds ( $\omega$ -bonds), shown as thick lines. The peptide bonds form planar, inflexible structures (defined by circles) usually in the *trans*-configuration. The  $\phi$  and  $\psi$  bonds of the amino acids rotate and allow protein folding. The permitted  $\phi$  and  $\psi$  bond combinations are shown by the Ramachandran plot (right panel) where  $\beta$  =  $\beta$ -sheet,  $\alpha\text{R}$  = conventional right hand  $\alpha$ -helix and  $\alpha\text{L}$  = left hand  $\alpha$ -helix.  $\text{C}\alpha$  is the central, tetravalent carbon,  $\text{C}'$  is the carbonyl carbon of each peptide bond.

(180° rotation), although rotation is prevented in the context of proline residues because of the ring structure of the secondary amino group. Peptide bonds flanking proline residues therefore adopt the *cis*-conformation.

A **Ramachandran plot** is a scatter graph with  $\phi$  and  $\psi$  angles plotted on alternative abscissa (Figure 22.1). It shows the combinations of  $\phi$  and  $\psi$  angles that are energetically favorable, i.e. not restricted by steric effects involving side chains and the polypeptide backbone. For most amino acids, permitted combinations fall into three main areas corresponding to the conformations of right-handed  $\alpha$ -helices,  $\beta$ -strands and left-handed  $\alpha$ -helices (see below). Glycine, which has a small side chain comprising a single hydrogen atom, can adopt a wider variety of conformations and thus plays a crucial role in the variability of protein structure. However, the overall folding of the polypeptide backbone can also force other amino acids to adopt unfavorable bond conformations.

## 22.2 Higher order protein structure

**Secondary structure.** Secondary structures in proteins are regular and repeating local configurations generated by intramolecular hydrogen bonds and other weak bonds. Although hydrogen bonds are formed by polar amino acid side chains (such as those of serine and threonine residues), the polypeptide backbone is itself polar because the amino nitrogen atom of each peptide bond unit can act as a hydrogen bond donor, and the carbonyl oxygen atom can act as a hydrogen bond acceptor. The periodic spacing of peptide bond units throughout the polypeptide allows regular ordered structures to form. Three broad classes of secondary structure are recognized, and are defined as



**Figure 22.2:** Major secondary structures in proteins. (a)  $\alpha$ -helix. Hydrogen bonds form between carbonyl oxygens and amino nitrogens four residues apart to generate a right handed helix with 3.6 residue periodicity. The arrows indicate hydrogen bonds and point to the positive pole of the dipole moment. (b)  $\beta$ -sheet. Extended  $\beta$ -strands can form sheets in antiparallel or parallel when interstrand hydrogen bonds are generated (arrows). Antiparallel strands are often separated by turns, whereas parallel strands are often separated by an  $\alpha$ -helix. Carbons with the thick outline are  $C_{\alpha}$  atoms.  $\alpha$ -helices and  $\beta$ -strands are often represented by cylinders (or coils) and arrows respectively in protein topology diagrams.

**helix, sheet and turn.** Of these,  $\alpha$ -helices and  $\beta$ -sheets are the most common, and occur when consecutive residues have the same  $\phi$  and  $\psi$  bond angles (Figure 22.2). Secondary structures not fitting any of the three categories are described as **coil**.

**$\alpha$ -helices** are usually right handed in polypeptides made from L-amino acids and occur where consecutive residues have average  $\phi$  and  $\psi$  angles of approximately  $-60^\circ$  and  $-45^\circ$ , respectively, corresponding to a block of values in the lower left quadrant of the Ramachandran plot (Figure 22.1). The curvature of the backbone allows hydrogen bonds to form between peptide bond units four residues apart. This aligns the peptide bonds throughout the  $\alpha$ -helix in the same orientation, which amplifies the slightly polarized charge distribution of each peptide unit, so giving the helix a significant dipole moment. Interactions involving helices thus reflect not only the chemical groups of individual residues, but the cumulative charge distribution over the entire structure. Helices vary in length from four to over 40 residues (1–12 turns of helix). There is some bias to the amino acids appearing in helices, with hydrophobic residues such as alanine and leucine found commonly and some other residues such as serine found only rarely. An **amphipathic  $\alpha$ -helix** has hydrophobic residues on one surface and charged or polar residues on the opposite surface. This generates a structure that can interact with both polar and nonpolar chemical environments (q.v. *leucine zipper*, *ion channel*).

**$\beta$ -sheets** form from regions of the polypeptide chain termed  **$\beta$ -strands**, where bond angles are almost fully extended, corresponding to a broad range of values in the upper left quadrant of the Ramachandran plot (Figure 22.1). Several  $\beta$ -strands align in parallel, antiparallel or mixed arrays, allowing hydrogen bonds to form between peptide bond units in different strands. These structures are often termed **pleated sheets** because the  $C\alpha$  atoms within each strand are displaced alternately above and below the plane of the sheet.  $\beta$ -sheets are usually twisted in a right handed direction, sometimes to such an extent that they form propeller-like structures. In some proteins,  $\beta$ -strands are arranged in perpendicular arrays to generate a lattice type sheet arrangement.

$\alpha$ -helices and  $\beta$ -sheets are joined together by linker residues which may adopt their own secondary structures by hydrogen bonding, in which case they are described as **turns**. The simplest turns are very short — a  $\beta$ -turn, for instance, is generated when hydrogen bonds form between peptide bond units located three residues apart, resulting in a hairpin turn in the polypeptide backbone. Alternatively, linker residues may have no secondary structure, in which case they are described as **loops**. Lacking intramolecular hydrogen bonds, loops often form hydrogen bonds with water and are therefore usually found at the surface of proteins.

**Supersecondary structures.** Almost all proteins contain some regions which adopt either  $\alpha$ -helical or  $\beta$ -sheet structure, connected by turns or loops. Two or more secondary structures often combine to form a more complex structural unit which may be termed a **supersecondary structure** or **motif**<sup>1</sup>. Some combinations are particularly common and form the basis of structural *domains* (see below), or facilitate specific interactions with other molecules in the cell. A supersecondary structure observed in three or more otherwise unrelated proteins is termed a **superfold**. Numerous supersecondary structures have been described, and some common examples are listed in Table 22.1. They form when side chains from adjacent  $\alpha$ -helices and  $\beta$ -sheets pack against each other.

**Tertiary, quaternary and interactive structure.** Protein **tertiary structure** refers to the overall three-dimensional conformation of a polypeptide and reflects the packing together of secondary and supersecondary structures to form compact globular *domains*. A **domain** is the smallest unit of protein tertiary structure (c.f. *module*) and may be considered a unit of independent or quasi-independent function. Small proteins may comprise a single domain, but large proteins often assemble from

<sup>1</sup>The term *motif* is used in an alternative way to describe a short region of a polypeptide sequence recognized because it is conserved in two or more proteins. A motif in this sense is not necessarily a structural motif as discussed above (see section on motifs, modules and domains later in the chapter).



**Table 22.1:** A selection of commonly occurring supersecondary structures in proteins. Many of these can be viewed on the SCOP database (see end of chapter)

Structure	Components
<i>Simple supersecondary structures: <math>\alpha</math>-helices</i>	
Coiled coil	Two or more $\alpha$ -helices coiled round each other. This is a particularly stable structure often found in fibrous structural proteins such as $\alpha$ -keratin, collagen and fibrinogen. A two helix coiled coil is formed by dimerization of <i>leucine zippers</i> (q.v.). The helical repeat of a normal $\alpha$ -helix is 3.6 residues, but in a two helix coiled coil, distortion reduces the periodicity to 3.5 residues allowing interaction between leucine residues at every seventh position (see Nucleic Acid-binding Proteins)
Helix-turn-helix	Two $\alpha$ -helices joined by a turn. This structure is a common prokaryotic and eukaryotic DNA-binding motif (see Nucleic Acid-binding Proteins)
Helix-loop-helix	Two $\alpha$ -helices joined by a linking region of loop. A common motif, found in the dimerization modules of some DNA-binding proteins (e.g. Achaete-scute, Sisterless-a, MyoD) and in their inhibitors (e.g. Id, Deadpan) (q.v. <i>basic helix-loop-helix</i> , <i>sex determination</i> ) and in calcium-binding proteins, where the helices designated E and F adopt an orthogonal helix-loop-helix structure termed an <b>EF-hand</b>
<i>Simple supersecondary structures: <math>\beta</math>-sheets</i>	
$\beta$ -hairpin	Two antiparallel $\beta$ -strands joined by a loop. Found in many proteins either alone or as part of a larger $\beta$ -sheet. There is no specific function associated with this structure
$\beta$ - $\alpha$ - $\beta$ motif	Two parallel $\beta$ -strands separated by an $\alpha$ -helix. Most $\beta$ -sheets containing parallel $\beta$ -strands are formed using this motif, as the two ends of the $\beta$ -strands which are joined lie at opposite sides of the sheet
$\beta$ -arch	A structure formed by two adjacent $\beta$ -strands linked by a loop, but where the strands lie in different sheets
$\beta$ -bulge	Extra residues in a $\beta$ -strand which cause a distortion in a $\beta$ -sheet
<i>Complex supersecondary structures: <math>\alpha</math>-helices</i>	
Four helix bundle	Four antiparallel $\alpha$ -helices packed together to form a hydrophobic core with hydrophilic residual groups exposed. Found in many predominantly $\alpha$ -helical proteins including Rop and ferritin
Globin fold	Eight $\alpha$ -helices arranged in a complex manner so that helices adjacent in the polypeptide primary structure are not necessarily adjacent in the secondary structure. The helices define a pocket which forms the active site of the protein. The globin fold is highly conserved in evolution: in the hemoglobins and myoglobins it binds the heme group
<i>Complex supersecondary structures: <math>\beta</math>-sheets</i>	
$\beta$ -barrel	A structure formed by a large antiparallel $\beta$ -sheet when it rolls up so that the first $\beta$ -strand is joined to the last by hydrogen bonds and a closed cylinder is formed. Each $\beta$ -strand is joined to the next by a hairpin turn or loop. These structures often form pockets for binding small molecules, e.g. the retinol binding protein contains a $\beta$ -barrel which accommodates the retinol molecule
$\beta$ -propeller	A $\beta$ -sheet which is twisted so that strands adopt a radial arrangement. The influenza virus neuraminidase protein comprises six $\beta$ -propeller motifs, each comprising four antiparallel $\beta$ -strands. The six motifs are arranged to form a symmetrical barrel-like domain
$\beta$ -sandwich	Two or more $\beta$ -sheets which pack on top of each other
Greek key	A four-strand antiparallel $\beta$ -sheet where the four $\beta$ -strands in the primary sequence (1-2-3-4) are arranged with the topology 4-1-2-3. Found in many proteins with antiparallel $\beta$ -sheets

Continued

Jelly roll	Four Greek key motifs adopting a barrel structure, so-called because of the way the polypeptide backbone wraps around the barrel. Found in e.g. influenza virus hemagglutinin
<b>Complex supersecondary structures: <math>\alpha</math>-helices and <math>\beta</math>-sheets</b>	
$\alpha\beta$ -barrel	A structure formed by sequential $\beta$ - $\alpha$ - $\beta$ motifs which roll up into a cylinder in which $\beta$ -strands are parallel and enclosed by $\alpha$ -helices. Common motif in enzymes, e.g. pyruvate kinase, aldolase, enolase, RuBisCO, glucose isomerase and Triosephosphate IsoMerase (often termed a <b>TIM barrel</b> for this reason)
$\alpha\beta$ -open sheets	Structures formed by sequential $\beta$ - $\alpha$ - $\beta$ motifs which do not roll up and thus form sheets in which parallel $\beta$ -strands are flanked by $\alpha$ -helices on each side. A highly variable motif found in many proteins including hexokinase and phosphoglycerate mutase
Rossmann fold	Alternating $\beta$ - $\alpha$ structures which fold to form a motif comprising a central $\beta$ sheet surrounded by $\alpha$ -helices. Common in nucleotide binding proteins and ribosomal proteins
<b>Other structures</b>	
Omega loop	A long loop whose termini lie close together. This is not strictly a secondary structure because there are no hydrogen bonding interactions within the loop

multiple domains which, theoretically, can fold and function in isolation or as part of a chimeric protein (q.v. *foldon*, *chimeric protein*, *fusion protein*, *domain swap*).

Whereas many polypeptides function in isolation, others assemble into oligomeric or multimeric complexes containing either copies of the same polypeptide (**homomultimers**, e.g.  $\beta$ -galactosidase) or different polypeptides (**heteromultimers**, e.g. hemoglobin). An additional level of **quaternary structure** can be recognized in multimeric complexes, reflecting the spatial organization of the individual components (**protomers**) and the chemical bonds formed between them. Individual protomers may function either independently or interdependently within the complex. In the latter case, quaternary structure is important in the regulation of protein activity (q.v. *cooperative binding*).

A further level of structural organization is seen when proteins interact with nonprotein molecules: cofactors, ligands and substrates. Proteins which function only in the presence of a non-polypeptide **cofactor** are termed **conjugated proteins**. They are usually noncovalently associated with their cofactors (e.g. coenzymes, nucleotides, metal ions), and may be termed **apoproteins** in the absence of cofactor and **holoproteins** in the presence of cofactor. Some cofactors, however, are covalently joined to their cognate proteins, in which case they are termed **prosthetic groups** (e.g. the heme group in hemoglobin). In each case, the activity of the protein is altered by its association with the cofactor, both by the influence of the cofactor on protein structure and by any unique chemical properties possessed by the cofactor. Protein interaction with ligands and substrates may also involve covalent and noncovalent bonds, although the former are usually reversible (c.f. *suicide enzyme*). The change in conformation caused by ligand or substrate binding may be important for the protein to function (q.v. *allostery*, *induced fit*).

Secondary structure forms predominantly under the influence of hydrogen bonds, but many different types of chemical bond (covalent and noncovalent) contribute to tertiary, quaternary and interactive structure. Owing to the complexity of these bonds, it is currently impossible to predict higher order structures from primary amino acid sequence with accuracy. Tertiary and quaternary structure, and the structure of proteins interacting with cofactors, ligands and substrates, must therefore be characterized directly using biophysical methods (Box 22.2).

**Covalent bonds in higher order protein structure.** In some proteins, the only covalent bonds are those of the polypeptide backbone and amino acid residual groups. Many proteins, however,

achieve their native states through additional covalent interactions in the form of **disulfide bonds** which form between the **sulfhydryl (-SH) groups** of cysteine residues. Intramolecular disulfide bonds are necessary for stable tertiary structure in many proteins, but intermolecular disulfide bonds can also form to stabilize quaternary structure. This may involve the formation of bonds between copies of the same polypeptide or between polypeptides encoded by different genes (e.g. during antibody synthesis, disulfide bonds join the two immunoglobulin heavy chains, but also join each heavy chain to a light chain). Insulin provides an interesting variation. There are three disulfide bonds in the mature protein, one intramolecular bond within the A-chain, and two intermolecular bonds joining the A- and B-chains. However, in the nascent polypeptide, all three bonds are intramolecular because the A- and B-chains are encoded by the same gene and are initially joined by an interstitial C-peptide which is removed by cleavage after disulfide bonds have formed.

Covalent bonds are also used to join prosthetic groups to proteins, e.g. the heme group of cytochrome C is attached covalently to several residues. Coordinate bonds are often responsible for metal ion binding, as in *zinc finger* (q.v.) modules and bacterial cupric and mercuric ion binding proteins, but the iron atom of the heme group is covalently joined to all its cognate proteins, including cytochrome C and hemoglobin.

**Noncovalent bonds in higher order protein structure.** Amino acids can be roughly divided into those with hydrophobic (nonpolar) side chains and those with hydrophilic (polar or charged) side chains (*Box 22.1*). In a protein, it is thermodynamically unacceptable to expose predominantly hydrophobic residues to water, therefore soluble proteins possess a **hydrophobic core** where the nonpolar residues are sequestered, and a polar or charged surface exposed to the solvent.

There are few charged residues **buried** in the interior of a protein, but polar residues are not excluded and the backbone, which is itself polar, must run through the core. Polar atoms generally make hydrogen bonds with water, but where this is not possible, as in the hydrophobic core of a protein, the hydrogen bonding potential must be taken up by secondary structure. Otherwise, the protein can be denatured — the free energy of stabilization is typically equivalent to that generated by one or two hydrogen bonds. The formation of secondary structures, predominantly  $\alpha$ -helices and  $\beta$ -sheets, neutralizes the polar atoms of the backbone, and polar side chains can also be neutralized by hydrogen bonding with the backbone and with each other. The hydrophobic core is further stabilized by van der Waals' interactions (hydrophobic attraction), which increase as neutral atoms approach each other until the point of contact. This is reflected by close packing of atoms in the protein core, through surface complementarity of the various secondary structural elements.

Protein surfaces are generally rich in polar and charged residues although nonpolar residues are also exposed. In this manner, proteins can make noncovalent contacts with other molecules using electrostatic forces, hydrogen bonds and van der Waals' interactions. Electrostatic forces are important where protein and target have opposite charges, as in the 'salt-bridge' interaction between highly basic histones and the negatively charged phosphate backbone of DNA. Van der Waals' forces are particularly important for interactions between complementary surfaces, where water is excluded and binding brings the appropriate chemical groups into close proximity. Conversely, water molecules can play an active bridging role in protein interactions involving hydrogen bonds, particularly in the interaction of proteins with DNA (*see Nucleic Acid-binding Proteins*).

**Protein folding.** Polypeptides are synthesized linearly (*see Protein Synthesis*) and must fold to adopt their correct secondary and tertiary conformations. The 'correct' structure is the native state of the molecule, i.e. the structure it adopts when biologically active. There is an infinite number of alternative denatured conformations, and the **Levinthal paradox** states that a random search through all these conformations would take an infinite length of time. Thus, **protein folding** must follow a defined pathway directed by energetically favorable interactions. Three models have been devised to explain the **protein folding problem**, i.e. the sum of processes which allow proteins adopt their native state.

Two models propose the existence of stable intermediate states through which folding could proceed. In the **framework model**, adjacent residues throughout the polypeptide interact to form secondary structures without forming a defined tertiary structure. The number of possible arrangements of the secondary structures is limited, and random diffusion of these preformed structures is sufficient to identify and adopt the native conformation. The **hydrophobic collapse model** predicts that the energetically favorable process of excluding water molecules from nonpolar side chains would drive an initial folding reaction to form the hydrophobic protein core, and that the remainder of the protein could then reorganize in the more limited spectrum of conformational arrangements available. In each case there is a compact late intermediate, a **molten globule**, which is rich in secondary structure and is arranged in approximately the right conformation but loosely packed. The consolidation stage of protein folding would thus involve a reorganization of this intermediate to generate the native state.

Alternatively, the **nucleation model** suggests that protein folding is initiated by specific residues in the polypeptide which form a nucleus around which other structures build. This global folding needs no intermediates and could occur using a well-defined nucleus or a weakly defined nucleus which condenses as structure forms around it (**nucleation-condensation**). Most studies of protein folding have identified intermediate folding states which support the framework or collapse models. More recently, however, several small proteins have been shown to fold in a simple two-state manner with no intermediates, and a single transition state. Such globally folding units are termed **foldons**. Therefore, while the formation of a native-like molten globule from a series of productive but dynamic intermediate states is an attractive folding mechanism for most proteins, others may fold in a single step using a nucleation mechanism.

**Molecular chaperones.** *In vivo*, many newly synthesized proteins are unable to fold into their native conformations spontaneously (**self-assembly**), either because the folding process requires the transition of an energetically unfavorable intermediate state or because there are several, equally stable alternative folding pathways available, only one of which is native. In these cases, correct folding is directed by proteins termed **molecular chaperones**, which recognize denatured states by binding to exposed residues normally buried in the native protein (e.g. hydrophobic residues exposed to the solvent). Correct folding reflects a cycle of chaperone-substrate binding and release which is often dependent on ATP hydrolysis and the activities of chaperone accessory proteins (**cochaperones**). Molecular chaperones are necessary to prevent illegitimate interactions between denatured proteins in the cell, which could result, for example, in protein aggregation (q.v. *inclusion body*, *prion*).

Chaperones direct protein folding, unfolding, refolding and assembly at many levels, including: (i) initial folding following protein synthesis; (ii) refolding following, for example, heat induced denaturation (many heat shock and other stress-induced proteins are chaperones); (iii) unfolding and refolding to allow translocation across membranes; (iv) interconverting alternative conformations of allosteric proteins; (v) refolding e.g. enzymes which become denatured as part of their activity; (vi) preparing proteins for degradation; and (vii) controlling the formation of multimeric and protein-ligand complexes.

There are many chaperone families, differing in their specificities and folding mechanisms. The **nucleoplasmins** are nuclear chaperones which control nucleosome assembly (see Chromatin). Other important chaperone families include the **Hsp70** class, which stabilize nascent polypeptides and facilitate membrane translocation, and the **chaperonins**, which control initial folding. The mechanisms of chaperone activity have become clearer recently following the solution of the structure of *E. coli* DnaK (a member of the Hsp70 family which works in concert with two cochaperones, DnaJ and GrpE, required for substrate binding and release), and the determination of structural changes in both the chaperone and its substrate in the folding cycle of *E. coli* GroEL (the archetypal chaperonin). GroEL forms a 14-mer structure comprising two inverted heptameric rings, capped at one end by a heptamer of the cochaperone protein GroES. Each GroEL subunit binds ATP, which is used to



provide energy for substrate folding. GroES facilitates the release of correctly folded product. Several rounds of ATP hydrolysis are required for full folding.

Other, less well-characterized chaperone families include Hsp90 and Hsp100, which act predominantly after protein synthesis and may function to prevent and/or reverse protein misfolding and aggregation under stress. Furthermore, while some chaperones are predominantly cytosolic, others, such as Grp78 in eukaryotes and proteins encoded by the *E. coli sec* genes, function specifically in the secretory system (Box 22.3).

**Allostery and cooperativity.** The interconversion of alternative **conformers** (conformational isomers) or alternative quaternary states can be used to regulate protein activity. Protein conformation is controlled by interactions with other molecules, and may involve either reversible posttranslational covalent modifications, or noncovalent associations. In either case, the result is a reversible switch in protein function.

The regulation of protein activity by covalent modification is common in eukaryotic signaling pathways and in the control of gene expression. Many proteins are regulated by phosphorylation, e.g. intracellular signaling proteins such as MAP kinase, and transcription factors such as the retinoblastoma protein RB1 (see Signal Transduction, The Cell Cycle). Furthermore, the reversible acetylation of histones is important in the control of chromatin structure (see Chromatin). Covalent modifications control protein activity in two ways. Firstly, they can alter protein tertiary structure and affect quaternary interactions and interactions with other molecules, e.g. by exposing or sequestering a particular domain, such as a catalytic site. Secondly, the modification may play a direct role in interactions with other molecules. Tyrosine phosphorylation, for example, adds negative charge to the protein causing compensatory reorganization of tertiary structure. However, phosphorylated tyrosines also act as direct binding sites for certain ligands, such as proteins containing *SH2 domains* (q.v.).

The regulation of protein conformation by noncovalent interactions involves the induction of a conformational change at one site by ligand binding at another. Proteins may possess several active centers that communicate with each other by conformational changes, a phenomenon termed **allostery**. The ligand may be a small effector molecule (e.g. in the control of transcription — q.v. *transcriptional regulation, lac operon, nuclear receptor family*), another protein (e.g. in the spread of prion diseases, see below) or an intercellular signaling protein (e.g. in the activation of a receptor tyrosine kinase by a growth factor, or the opening of a ligand-gated ion channel; see Signal Transduction). Prions may represent a unique example of an allosteric control *chain reaction*: prions are pathological conformational isomers of a normal cellular protein called PrP. Contact between a prion and normal PrP may induce a conformational change in the latter, resulting in its conversion into a prion (q.v. *refolding model*). This causes the accumulation of prions which, being misfolded conformers, aggregate in the cell to form pathological plaques (q.v. *transmissible spongiform encephalopathies*).

A further example of the conformational control of protein activity is **cooperative binding**. In this case, regulation occurs at the quaternary level: the binding of a ligand to one protomer changes its conformation and through quaternary interactions increases the affinity of other protomers. Hemoglobin binds oxygen cooperatively: oxygen binding to one globin causes a conformational change in the other globins which increases their oxygen-binding efficiency. Some proteins bind each other cooperatively. The binding of single-stranded DNA-binding protein (SSB) to DNA causes a conformational change in the protein which allows other SSB molecules to bind with greater efficiency, resulting in a filament of protein surrounding the DNA.

## 22.3 Protein modification

**Classes of protein modification.** During or following synthesis, all polypeptides undergo some form of covalent modification before they form functional proteins (Table 22.2). Structurally, such

**Table 22.2:** A summary of programmed enzymatic protein modifications with roles in protein structure and function, protein targeting or processing and the flow of genetic information

Covalent modification	Examples
<i>Substitutions (minor side chain modifications)</i>	
Minor side chain modification — permanent and associated with protein function	Hydroxylation of proline residues in collagen stabilizes triple helical coiled coil tertiary structure Sulfation of tyrosine residues in certain hormones Iodination of thyroglobulin $\gamma$ -carboxylation of glutamine residues in prothrombin
Formation of intra- and intermolecular bonds	Formation of disulfide bonds in many extracellular proteins, e.g. insulin, immunoglobulins
Minor side chain modification — reversible and associated with regulation of activity	Phosphorylation of tyrosine, serine and threonine residues regulates enzyme activity, e.g. <i>receptor tyrosine kinases</i> , <i>cyclin dependent kinases</i> (q.v.) Many side chains are also methylated although the function of this modification is unknown Acetylation of lysyl residues of <i>histones</i> (q.v.) regulates their ability to form higher-order chromatin structure and plays an important role in the establishment of <i>chromatin domains</i> (q.v.)
<i>Augmentations (major side or main chain modifications)</i>	
Addition of chemical groups to side chains — associated with protein function	Addition of nucleotides required for enzyme activity (e.g. adenyl groups added to glutamine synthase in <i>E. coli</i> ) Addition of <i>N</i> -acetylglucosamine to serine or threonine residues of some eukaryotic cytoplasmic proteins Addition of cholesterol to <i>Hedgehog family</i> (q.v.) signaling proteins controls their diffusion Addition of prosthetic groups to conjugated proteins, e.g. heme group to cytochrome C or globins
Addition of chemical groups to side chains — associated with protein targeting or trafficking	Acylation of cysteine residue targets protein to cell membrane Addition of GPI membrane anchor targets protein to cell membrane <i>N</i> -glycosylation of asparagine residues in the sequence Asn-Xaa-Se/Thr is a common modification in proteins entering the secretory pathway ( <i>Box 22.3</i> ) <i>O</i> -glycosylation of Ser/Thr occurs in Golgi ( <i>Box 22.3</i> ) Ubiquitination of proteins targeted for degradation ( <i>Box 22.3</i> )
End group modification	Acetylation of N-terminal amino acid of many cytoplasmic proteins appears to relate to rate of protein turnover Acylation of N-terminal residue targets proteins to cell membrane, e.g. myristylation of <i>Ras</i> (q.v.)
<i>Cleavage (removal of residues)</i>	
Cleavage of peptide bonds	Co- or posttranslational cleavage of initiator methionine occurs in most cytoplasmic proteins Cotranslational cleavage of <i>signal peptide</i> (q.v.) occurs during translocation across endoplasmic reticulum membrane ( <i>Box 22.3</i> ) for secreted proteins Maturation of immature proteins ( <b>proproteins</b> ) by cleavage: e.g. activation of <b>zymogens</b> (inactive enzyme precursors) by proteolysis, removal of internal C-peptide of proinsulin, cleavage of <i>Hedgehog</i> proteins into N-terminal and C-terminal fragments Processing of genetic information: e.g. cleavage of <i>polyproteins</i> (q.v.) synthesized from poliovirus genome and mammalian tachykinin genes, splicing out of <i>intons</i> (q.v.)

modifications can be divided into three groups: (i) minor substitutions — minor changes to amino acid side chains; (ii) augmentations — the addition of bulky chemical groups to particular amino acid residues; and (iii) cleavage — the removal of residues from the nascent polypeptide. Functionally, modifications may be classed as neutral or essential for protein function, or required for protein folding, trafficking and processing. Most augmentations are required for protein sorting (some are also essential for protein function). Several types of modification play a role in the rate of protein turnover and degradation.

Modifications may also be classed as permanent or reversible. Cleavages and most augmentations are permanent. Substitutions may be permanent with an essential structural role (e.g. formation of disulfide bonds between cysteine residues), or reversible, fulfilling a regulatory function (e.g. phosphorylation of tyrosine residues). While these are enzymatic modifications programmed by the cell, other modifications occur in a nonenzymatic manner and are usually associated with protein aging. Such modifications include oxidation, deamidation and, for blood proteins, reaction with glucose.

## 22.4 Protein families

**Conventional gene and protein families: molecular taxonomy.** A **gene family** is a group of genes with a significant level of sequence identity, and a **protein family** is a similarly related group of gene products. Members of gene and protein families which bear particular resemblance to each other are grouped into **subfamilies**, and more distant relationships require **superfamilies** and **megafamilies** representing higher orders of molecular taxonomy.

Conventional gene families are conserved in sequence throughout the entire length of the coding region. The evolution of such families can be explained in two ways — by divergence from a common ancestor or by convergence from unrelated ancestors. The average protein is approximately 300–350 amino acid residues in length, corresponding to a coding region size of approximately 1 kbp. The chance of random independent convergence upon the same nucleotide sequence is therefore  $1:4^{1000}$ , and the chance that two sequences will independently evolve to 50% identity is  $1:4^{500}$ . Sequence conservation over the moderate to large physical distances spanned by genes is therefore taken as *prime facie* evidence for **homology**<sup>1</sup>, a term which means relationship through common ancestry. Conversely, there is a significant chance that small sequence motifs, such as transcriptional regulatory elements and splice sites, can arise independently by stochastic processes.

Sequence divergence reflects the accumulation of mutations over time, and as sequences continue to diverge, eventually there comes a time when homology can no longer be detected through sequence conservation. Thirty percent sequence identity can be detected by alignment and the sequences thus related are clearly homologous. Distant homology, where sequence identity is < 30%, requires computer analysis. Homology may be inferred from conserved protein folds (i.e. conserved tertiary structure), but at this point it is often likely that the same tertiary structure

<sup>1</sup>Homology is strictly an absolute term, i.e. two sequences are either homologous or not — there is no degree of homology. To quantify the relatedness between sequences the terms **percent identity** (the frequency of exactly conserved bases/amino acids) or **percentage similarity** (the frequency of exactly conserved amino acid and conservative changes) are used (however, chimeric proteins may be described as **partially homologous**, see main text). The degree of relatedness between homologous sequences reflects the time since divergence (q.v. *molecular clock*). The comparison of orthologous sequences therefore allows the creation of **molecular phylogenies**, and the comparison of paralogous sequences provides information concerning the evolution of gene families. A problem with sequence alignments is the introduction of gaps. Sequences may diverge by the deletion or insertion of nucleotides or amino acids, as well as by substitution. The problem is that by introducing gaps at will, any pair of sequences can be made to match, so there must be a **gap penalty** in sequence alignments which reduces the level of identity between sequences as more gaps are introduced. This is usually an arbitrary penalty, which is higher for gap introduction than for extending a preexisting gap.

could have arisen by convergence, but using entirely different sequences. The 'protein structure code' which converts amino acid sequences into tertiary structure is highly degenerate; thus many different sequences can adopt the same structure. Homology inferred through structure is therefore uncertain unless backed up by intermediate sequence relationships.

Gene family relationships may be described as orthologous or paralogous. **Orthologous** relationships refer to genes which have diverged by speciation, i.e. family members represent genes performing the same functions in different species (e.g. human  $\beta$ -globin and mouse  $\beta$ -globin). **Paralogous** relationships refer to genes which have duplicated and diverged within a genome, i.e. paralogous genes are members of *multigene families* (q.v.) (e.g. human  $\beta$ -globin and human  $\alpha$ -globin). In large multigene families, particularly closely related genes may be placed into **paralogous subgroups**, e.g. the mouse *Hox* genes are paralogous, but those representing equivalent positions within each cluster are more closely related to each other than any of them are to the remaining *Hox* genes (see Development: Molecular Aspects).

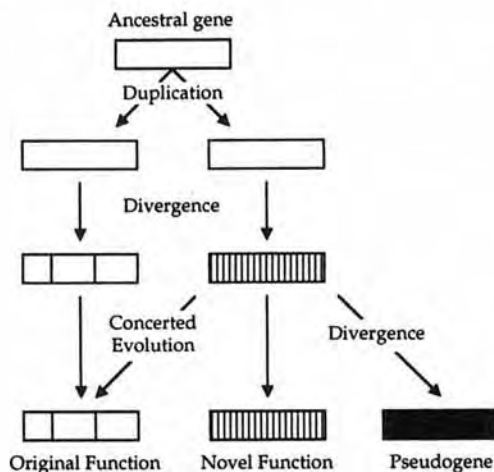
**Gene duplication and divergence: the origin of new functions.** The evolution of complex organisms from simpler ones presents a problem: from where do new functions arise? Increasing biological complexity (e.g. the transition from unicellular to metazoan organisation) requires increasing numbers of functions, ultimately specified by genes. Thus, the origin of new functions must reflect the creation of new genes, or diversification of the use of pre-existing genes. Both processes seem to have occurred during evolution: total gene number has increased with biological complexity (see Gene Structure and Mapping) as has the production of multiple products from single genes, particularly through the use of *alternative splicing* (q.v.).

New genes can in principle arise from three sources: (i) through import from other genomes; (ii) through intrinsic duplication of preexisting genes; and (iii) spontaneously from random, noncoding DNA by the accumulation of point mutations. The third source can effectively be discounted because it is very unlikely (see earlier probability calculation). The import of genes (**horizontal gene transfer**) is a comparatively rare event (see Gene Transfer in Bacteria; q.v. *Ti* plasmid, *promiscuous DNA*, *acute transforming retrovirus*) and does not strictly generate new functions, although they may be new to the importer. Most new genes thus arise by intrinsic **gene duplication** followed by **divergence**. The frequency of such events explains the predominance of multigene families in cellular genomes.

Once gene duplication takes place, the cell has two copies of essentially the same locus, and the genes will initially behave in a manner which appears superficially allelic (q.v. *pseudoalleles*, *genetic redundancy*). If the gene product is essential, selection will act on only one locus allowing the other to mutate (Figure 22.3). In most cases, the accumulation of mutations results in loss of function and the production of a *pseudogene* (q.v.) which is eventually lost altogether. Occasionally, mutations alter the structure of the gene product in such a way that it can adopt a new function. More often, however, novel functions arise initially from the acquisition of new expression patterns, followed later by structural diversification. As an example, many enzymes in metazoans are found as multiple isoforms (**isoenzymes**, **isozymes**) encoded by distinct differentially expressed genes. In mammals, there are three isoforms of the glycolytic enzyme enolase, one distributed ubiquitously, the others expressed specifically in mature neurons and muscles, respectively. This presumably represents an early stage of diversification, as each protein carries out the same catalytic reaction, but may have adapted to suit each particular intracellular environment, e.g. chloride tolerance in the case of neuron-specific enolase. Even significantly diverged proteins, such as the *Drosophila* proteins Paired and Gooseberry and the mouse Pax-3 protein, can substitute for each other if the genes are driven by heterologous regulatory elements (e.g. *gooseberry*, expressed under the *paired* promoter, can rescue *paired*<sup>-</sup> mutant flies). Evolution may therefore be driven as much by mutations in *cis*-acting DNA elements as by those affecting protein structure.

Another, more complex route to independent function is combinatorial action, where the initial-





**Figure 22.3:** The evolution of protein families. A single ancestral gene may duplicate, in which case selection pressure applies to only one copy and the other can accumulate mutations (lines). The divergence may lead to a novel function or to loss of function (creating a pseudogene). In tandem gene clusters, sequence homogenization through nonallelic recombination events can result in concerted evolution.

ly redundant gene copies diverge but maintain a related function and can cooperate to generate further functional diversity. For example, extensive duplication of an ancestral G-protein-coupled receptor gene has provided the multitude of alternative odorant receptors in humans and other mammals. These receptors act alone and in complex combinatorial signaling networks to identify distinct smells.

With the adoption of new functions, duplicated genes become nonallelic. The genetic consequences of gene duplication and divergence therefore begin with full redundancy, progress to partial redundancy and eventually to full functional independence. However, it is also possible for unrelated gene products to converge upon a single function, especially in the complex signaling pathways of eukaryotes which diverge, converge and engage in cross-talk (*see Signal Transduction*). An example is the vertebrate organizer which secretes at least three unrelated proteins — Noggin, Follistatin and Chordin — to block signaling by bone morphogenetic proteins (*see Development: Molecular Aspects*).

**Mechanisms of gene duplication.** Gene duplication, the creation of paralogous genes, can occur at three levels: (i) an isolated gene duplication event; (ii) whole genome duplication; and (iii) a large scale duplication involving a whole chromosome or chromosome segment.

**Selective gene duplication** events can occur by a number of different mechanisms depending upon the existing copy number of the gene. If there is already more than one copy of the gene in the genome, *unequal crossing over*, *unequal sister chromatid exchange* and *replication slipping* (q.v.) can all generate duplications (as well as deletions) in repetitive DNA. How do the copies arise in the first place? The duplication of a single copy gene involves more complex mechanisms including: (i) replicative transposition encompassing genomic DNA; (ii) chromosome breakage followed by out of register end joining of sister chromatids; and (iii) unequal exchange at short sequence motifs which have arisen by chance. More complex rearrangements involving large sections of unlinked DNA may also contribute to isolated gene duplications (q.v. *gene amplification*).

Many paralogous gene relationships, particularly in larger genomes, may reflect ancient **whole genome duplication** events. Comparative analysis of genome size and gene number suggests that the human genome arose through at least two rounds of whole genome duplication. Many genes

represented once in the *Drosophila* and *C. elegans* genomes are present as four paralogous copies in mammals, the primary example being the four *Hox* gene clusters (q.v.). Whole genome duplication is frequent in plants and, although deleterious in mammals, is tolerated by many invertebrates (q.v. *polyploidy*). There is a single *Hox* cluster in the genome of *Amphioxus*, which is regarded as the closest living relative of the vertebrates. It is likely, therefore, that an ancestral species predating the vertebrate lineage underwent a series of genome duplication events followed by divergence at the chromosome level to restore diploidy. The tetraploid state was presumably transient in the mammalian lineage, as large scale chromosome rearrangements would restrict homologous chromosome pairing to specific partners during meiosis (many fish, however, are tetraploid). Divergence occurring at the gene level at the same time, would involve the loss of some genes and the rearrangement of others. Eventually, only a few traces of the ancient duplication event would remain, as **paralogous chromosome segments**. For example, human chromosomes 12 and 17 each contain one of the four *HOX* clusters and paralogous members of several other gene families (collagen, enolase, retinoic acid receptor, keratin, integrin, *WNT* and aldehyde dehydrogenase), with broadly conserved linkage.

Large scale chromosome rearrangements have also contributed to the evolution of multigene families. The *comparative mapping* (q.v.) of mammalian genomes has shown that such rearrangements have occurred frequently, so that *syntenic regions* (q.v.) are restricted to small chromosome segments. As well as obscuring the evidence of tetraploidisation, chromosome rearrangements have resulted in large scale **subgenomic duplication** events. In particular, the two arms of human chromosome 1 appear to be paralogous, each containing genes for glutamic-oxaloacetic acid transaminase, blood coagulation factors, two different types of tRNA, and a ferritin heavy chain. This suggests that chromosome 1, the largest human chromosome, may have arisen through an ancient *Robertsonian translocation* event (q.v.), involving two smaller chromosomes.

**Sequence divergence and homogenization.** As discussed above, gene duplication is typically followed by sequence divergence which may result in the acquisition of new functions or, alternatively, loss of function for one of the copies. If duplicated genes do evolve new functions, these tend to be conserved in later speciation events with the result that the *orthologous genes are more highly conserved than paralogous genes*. This situation is observed for most dispersed multigene families, e.g. the enzyme isoform families such as enolase, where mammalian orthologs are more highly conserved than the paralogs within each species.

In tandemly arranged gene families, however, the high frequency of nonallelic recombination events, such as unequal crossing over, unequal sister chromatid exchange and gene conversion, results in **sequence homogenization**. In such cases, paralogs are more closely related than orthologs, because independently arising mutations are likely to be fixed in each species. This phenomenon, defined as **concerted evolution**, is common for tandemly clustered genes, e.g. the histone genes, but does not apply to all clusters. The globin clusters, for instance, show evolution more typical of dispersed genes (i.e. orthologs are more highly conserved than paralogs), probably because the genes have individual roles in development and sequence homogenization would be deleterious. The *rRNA genes* (q.v.) provide an unusual example where concerted evolution also occurs in dispersed repeats (in this case, each dispersed repeat comprises a series of tandemly arranged rRNA genes). This reflects the close association of rDNA clusters in the nucleolus, allowing frequent non-homologous recombination events *in trans* to maintain sequence homogeneity (q.v. *satellite association*). *Trans*-interactions between dispersed genes are otherwise rare events (q.v. *homology-dependent silencing*, *co-suppression*, *trans-sensing*).

**Protein chimeras: motifs, modules and domains.** Conventional gene families are homologous over their entire lengths, and their evolution can be discussed in terms of simple duplication and divergence. However, a family relationship can also involve a particular conserved region, which is

present in two otherwise entirely dissimilar proteins. Any unit of conserved sequence in a gene or protein may be termed a **box** or **motif**<sup>1</sup> — motifs range from short sequence elements such as the DEAD box RNA helicase motif, to larger functional units such as zinc fingers, which may properly be designated *modules*. A **module** is a contiguous segment of a protein which performs a specific function. Modules are typically made up of several motifs, although the zinc finger is an exception. A module differs from a protein *domain* in that the latter is a unit of tertiary structure and does not necessarily form from contiguous elements of primary sequence. The importance of this distinction is that modules can be inserted into or deleted from proteins by moving a contiguous segment of DNA from one gene to another; this may not be possible for all protein domains. Unfortunately, the terms motif, module and domain are often used loosely and interchangeably.

Genes and proteins which are partially homologous (i.e. homologous over a particular conserved segment) may arise due to intense selection for a specific module, while the remainder of the protein is free to mutate and diverge. In other cases, however, the same motif or module may be found in proteins which are very obviously derived from totally distinct lineages. Through a variety of gene rearrangement processes, modules can become combined in new ways to generate novel **chimeric proteins**. These complicate molecular taxonomy because the proteins can contain modules representing several different families at the same time, and the source of each module is often impossible to determine.

**Evolution of modular structure.** The evolution of repeated modules within single proteins and related modules in different proteins involves the same mechanisms which cause whole gene duplication (tandem or dispersed), but on a subgenic scale. Modular evolution is prolific in eukaryotes, because the presence of introns allows recombination of segments of coding DNA without the need for precise recombination junctions.

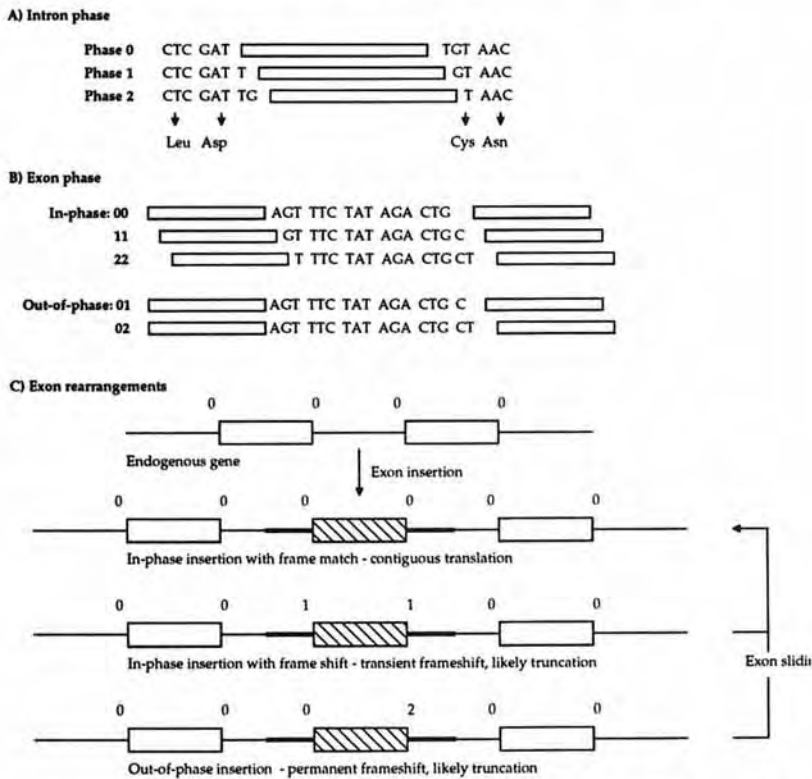
Repetitive modules within genes can arise by **exon duplication (exon repetition)** where a single exon undergoes tandem duplication. This can offer two immediate evolutionary advantages: (i) for structural proteins, exon duplication can extend a specific structural domain (e.g. the triple helical domain of collagen, which is encoded by multiple similar exons); (ii) for other proteins, exon duplication increases the dosage of a specific functional domain (**dosage repetition**), which may increase protein activity, or in exceptional cases, allow posttranslational cleavage to generate multiple functional units (e.g. tachykinin and ubiquitin genes). Like whole gene duplication and divergence, exon duplication is usually followed by **exon divergence**, leading to structurally conserved modules with alternative functions. Recombination may cause **exon deletion** as well as exon repetition, resulting in the removal of a particular module from a protein.

The dispersive duplication of exons can introduce single exons into the intron of a preexisting gene. If the exon corresponds to a protein module, a new function would be conferred upon the recipient protein. This type of event could occur by recombination between nonallelic genes, perhaps involving short repetitive elements or *transposons* (q.v.) within nonallelic introns. However, this would tend to cause gene fusion rather than single exon integration, unless there was a rare double crossover event. A more likely explanation for exon insertion is transposable element activity, e.g. aberrant excision (where some flanking host DNA is transposed at the expense of part of the element) or by cooperative transposition (where two elements cooperate to transpose an interstitial segment of DNA) (see Mobile Genetic Elements). The generation of chimeric proteins by mixing preexisting modules is termed **exon shuffling**. It provides a rapid route for the production of novel proteins from modules which have been functionally honed by conventional evolution.

<sup>1</sup>The term *motif* is used in an alternative way to describe a unit of supersecondary structure in a protein, i.e. a specific configuration of two or more secondary structural elements. A conserved sequence motif may correspond to a structural motif, but this is not necessarily the case. Note that the term *box* is used not only to describe motifs in genes and proteins (e.g. DEAD box), but also motifs in noncoding DNA (e.g. TATA box).

A further form of chimeric protein evolution is **whole gene fusion**, exemplified by eukaryotic **multienzyme proteins** whose prokaryotic orthologs are encoded by separate genes. For example, vertebrates possess a multienzyme protein with three nucleotide biosynthesis activities: glycineamide ribonucleotide synthetase (GARS), aminoimidazole ribonucleotide synthetase (AIRS) and glycineamide ribonucleotide transformylase (GART). In vertebrates, each activity is carried out by a different module of one protein, encoded by the single *GART* gene. In bacteria, three enzymes are encoded by three *separate* genes, whereas in yeast GARS and AIRS are carried by one enzyme and GART is encoded separately. The situation is more complex in *Drosophila*, where the multienzyme protein has four domains because the AIRS module has been duplicated. Originally, it was thought that such whole gene fusion events involved recombination between the last intron of one gene and the first of another. However, the discovery of a polyadenylation site within the intron separating GARS and AIRS modules in the vertebrate protein suggests the fusion may have occurred in intergenic DNA, followed by the modification of regulatory elements and splice signals to allow cotranscription and processing.

An important aspect of all forms of exon rearrangement — repetition, shuffling, deletion and whole gene fusion — is **intron phase** (Figure 22.4). This refers to how the coding region is



**Figure 22.4:** Intron phase and consequences for exon rearrangements. (a) Intron phases are defined by the position within the codon where the coding region is interrupted. (b) Exon phase is defined by the phase of the flanking introns. There are three types of in-phase exon and six types of out-of-phase exon (two examples are shown). (c) The endogenous gene has two exons in phase 0,0. Insertion of a further 0,0 phase exon results in contiguous translation of the new exon. Insertion of a 1,1 phase exon results in a transient frameshift so that the inserted exon is translated out of frame. This is usually deleterious unless the exon is small. Insertion of an out-of-phase exon results in an uncorrected frameshift. Both transient and permanent frameshifts can be corrected by small insertions or deletions, or by moving the exon–intron boundary through splice site mutation (exon sliding).



interrupted and therefore applies only to coding region introns. The most common introns are **phase 0**; these interrupt the coding sequence between codons (i.e. between the third base of one codon and the first base of the next). **Phase 1** and **phase 2** introns interrupt the coding sequence between bases one and two, or two and three of the codon, respectively. While the phase of a single intron is irrelevant, the phase of a **fusion intron**, generated by exon rearrangement is vitally important in the translation of the new gene. Exons can be defined according to the phase of the two flanking introns and those which begin and end at the same position with respect to codon boundaries (i.e. the flanking introns have the same phase) are termed **in-phase exons**: These can undergo unlimited duplications and can participate in exon shuffling events because they do not disturb the overall reading frame. Conversely, **out-of-phase exons** begin and end at different positions within the codon (i.e. the flanking introns have different phases) and the insertion of such an exon generates a frameshift. Even in-phase exons can cause an internal transient frameshift if they are inserted into a gene by fusion to introns with different phases. The interpretation of the new exon is unpredictable in these cases and may cause truncation by uncovering an adventitious termination codon. Such alterations to gene function may be corrected by point mutations which result in **exon sliding**, the shifting of the exon start and stop positions with respect to the surrounding introns. This can be caused by small insertions or deletions and by mutations affecting the position of the splice sites. Exon sliding can also 'correct' shuffling events in which the participating exon is out-of-phase.

**Evolutionary origins of introns.** In higher eukaryotes, most genes are interrupted by introns, whereas the genes of bacteria and many lower eukaryotes generally lack introns (q.v. *gene architecture*). Since their discovery in 1977, there has been intense debate about the origin of introns and their function. One theory is that spliceosomal introns evolved from self-splicing introns by acquisition of the ability to splice *in trans* (the selfish DNA model of intron origin). The similar splicing mechanisms of nuclear introns and group II self-splicing introns supports this model (see RNA Processing), however, the splice sites recognized by the two types of intron are distinct. It is possible that nuclear spliceosomal introns and group II introns evolved from a common lineage. Other models suggest that nuclear introns evolved separately as byproducts of the intergenic DNA that separated ancient genes, or as coding segments that were discarded during the evolution of alternative splicing. The last model is particularly provocative, because it suggests that splicing evolved before introns, a hypothesis supported by some modern-day genes which can undergo differential exon splicing in the absence of introns.

A further contentious issue is the age of introns. **Early intron theories** propose that introns were present in the earliest genes, and have been selectively lost from the prokaryotic lineage. The early introns may have arisen from self-splicing introns, or may have evolved from the noncoding DNA separating the ancient genes and spread by **reverse splicing**. In the latter model, the evolution of splicing was driven by the advantage of increasing gene size. Early intron theories are supported by the positional conservation of introns in ancient gene families, such as the globin gene family, and by the correspondence between exons and functional protein modules (presumably the relics of the original genes). **Late intron theories** propose that introns have inserted into genes relatively recently. They are supported by evidence for random intron insertion, i.e. insertions which are neither conserved in gene families nor dividing proteins into neat modules, such as the introns found in the collagen gene family.

Whatever their origins, introns have proliferated rapidly during eukaryote radiation and have facilitated both modular evolution (exon repetition and shuffling) and alternative splicing as advantageous mechanisms to increase the functional repertoire of the eukaryotic genome. The size of introns has also increased, so that many vertebrate genes are not only rich in introns, but are predominantly represented by intron material (q.v. *gene architecture*; c.f. *Puffer fish genome*).

**Selective expansion of protein families.** The analysis of information from genome and EST projects (see Gene Structure and Mapping) has allowed the distribution and abundance of particular protein

modules to be documented. Data from the complete genome sequences and the many genome sequencing projects underway has shown that certain protein families are highly successful, but that success is often lineage-dependent. Many of the protein modules which are ubiquitous, including those making up metabolic enzymes and core components of the protein synthesis machinery, show minimal proliferation. Conversely, other modules, such as zinc fingers, immunoglobulin modules, G-protein-coupled receptors and protein tyrosine kinases, have multiplied disproportionately, but only in certain groups of organisms. One of the largest protein families in yeast is the GAL4 transcription factor family (q.v. *zinc binuclear cluster*), but this appears to be entirely restricted to fungi. Similarly, immunoglobulin modules are found only in animals, and are particularly abundant as tandemly repeated modules within proteins. In some cases, the proliferation of distinct protein families with related functions in different kingdoms suggests a stochastic influence. For example, protein kinases are widely used as signaling molecules by all living organisms, but while serine/threonine and tyrosine kinases represent two of the largest protein families in eukaryotes, they are scarce in bacteria. Conversely, histidine kinases are well represented in bacteria but not in eukaryotes. In other cases, the expansion of protein families may reflect a functional innovation during evolution: the proliferation of EGF and immunoglobulin modules in animals probably reflects their ability to mediate cell-cell contacts, which is advantageous for the complex cell interactions which occur during development (*see* Development: Molecular Aspects).

## 22.5 Global analysis of protein function

**The proteome.** The *genome* is the full complement of genetic information in a cell (*see* Gene Structure and Mapping), and is the store of the global program required to manage and reproduce the cell (or multicellular organism). The **proteome** represents the entire collection of proteins which are encoded by the genome and hence the global functional spectrum of the genome. The proteome is a complex system, even though it represents only a fraction of the genome (i.e. the coding sequences). This is because there are numerous ways to use single genes to generate multiple products (*see* Gene Expression and Regulation), so the proteome contains many overlapping, structurally similar products. The proteome interacts with the environment to generate the **phenome**, a representation of the total sum of characters displayed by an organism.

**Functional genomics.** In recent years, there has been an explosion in the amount of sequence information due to the success of genome mapping and sequencing projects and concerted efforts to characterize cDNA sequences (*see* Gene Structure and Mapping). Functions can be tentatively assigned to new gene sequences by comparison with previously cloned genes whose products have already been functionally characterized. However, many novel genes show no relationship to previously characterized families, and *de novo* functional analysis is usually carried out on an individual basis, by looking at expression patterns, the effects of mutation, and interaction with other cellular components.

To make sense of the large amount of data arising from sequencing projects, the intensive functional analysis of single genes must be replaced, or at least complemented by genome-wide approaches to functional analysis (**functional genomics**). This involves the systematic analysis of the expression, mutation and interactions of the proteome.

Several techniques have been developed recently for the simultaneous analysis of the expression of all genes in the genome, providing an accurate characterization of cell type-specific gene expression profiles and the response of a cell to the environment. **DNA microarrays** and **oligonucleotide chips** allow cDNAs or oligonucleotides to be precisely gridded and hybridization analysis facilitates the qualitative and quantitative detection of RNA expression. The **serial analysis of gene expression (SAGE)** technique is a PCR-based method where short sequence tags corresponding to specific RNA molecules are amplified, concatemerized and cloned, providing a linear array of markers to identify expressed genes. These techniques are useful to probe cell responses, and complement

approaches such as high throughput *in situ* hybridization (q.v.) to assemble gene expression databases and to help with the construction of **genetic networks** (see below).

Mutation approaches to the determination of gene function are widely used on a single gene basis in all model organisms. In the most genetically amenable organisms, the systematic disruption of all genes is feasible and such a project is already underway for the fully sequenced *S. cerevisiae* genome, with an international consortium of laboratories working to disrupt each of the 6000 genes using *targeting vectors* (q.v.) containing oligonucleotides cassettes which can be identified by PCR. *Gene targeting* (q.v.) in yeast is very efficient, but this is not the case for other eukaryotes. Additionally, for the large genomes of model vertebrates (e.g. mice), an enormous input of resources would be required for a systematic knockout project, and this approach appears unlikely to be used in the near future. However, it may be possible to generate a library of *embryonic stem cells* (q.v.) with each gene disrupted. Such a library could be used in concert with a complete genome sequence to provide a resource for the production of mice lacking any particular gene. This approach obviates the need for housing and breeding large collections of mutants (q.v. *gene trapping*).

At the protein level, techniques for the simultaneous monitoring of the abundance and state of modification of all proteins in the cell are being developed. One approach is based on 2-D electrophoresis (Box 22.2) and another is chip-based, but using antibodies instead of oligonucleotides. The analysis of protein interactions is a powerful approach to the determination of protein function. A number of different techniques can reveal protein-protein interactions and the interaction between proteins and nucleic acids, but *phage display* (q.v.) and the *yeast two hybrid system* (q.v.) are particularly applicable to high throughput studies, and a project to systematically catalog the interaction of all yeast proteins is currently underway. One problem with biochemical approaches to functional interactions is that many of the interactions discovered are not physiologically relevant. This may be because the products are not usually expressed at the same time or in the same cell type, i.e. the interaction is adventitious. Here, a combination of approaches — interactions, expression patterns and mutant phenotypes, are needed for the full elucidation of protein function.

### Box 22.1: Amino acids

**Standard amino acids.** Amino acids are amphoteric molecules (they can function as acids or bases) with the general structure  $H_2N-CHR-COOH$ . They are abundant in living organisms, occurring either as free dipolar ions (zwitterions), short peptides (e.g. hormones) or in proteins. Each has a central tetravalent carbon atom ( $C\alpha$ ) with four coordinated groups, three of which (the amino and carboxylic acid groups and the hydrogen atom) are invariant. The fourth group, which is known as the **residual group (R)** or **side chain** is variable and determines the physical and chemical properties of the molecule. There are hundreds of different types of amino acid, each with a different side chain, but proteins are formed from a basic repertoire of 21 standard types specified by the *genetic code* (q.v.). One of these, selenocysteine, is very rare (q.v. *selenoproteins*). The amino acids can be placed into categories based on the chemical properties of their side chains: some are composed entirely of hydrocarbon groups and are hydrophobic; others are polar,

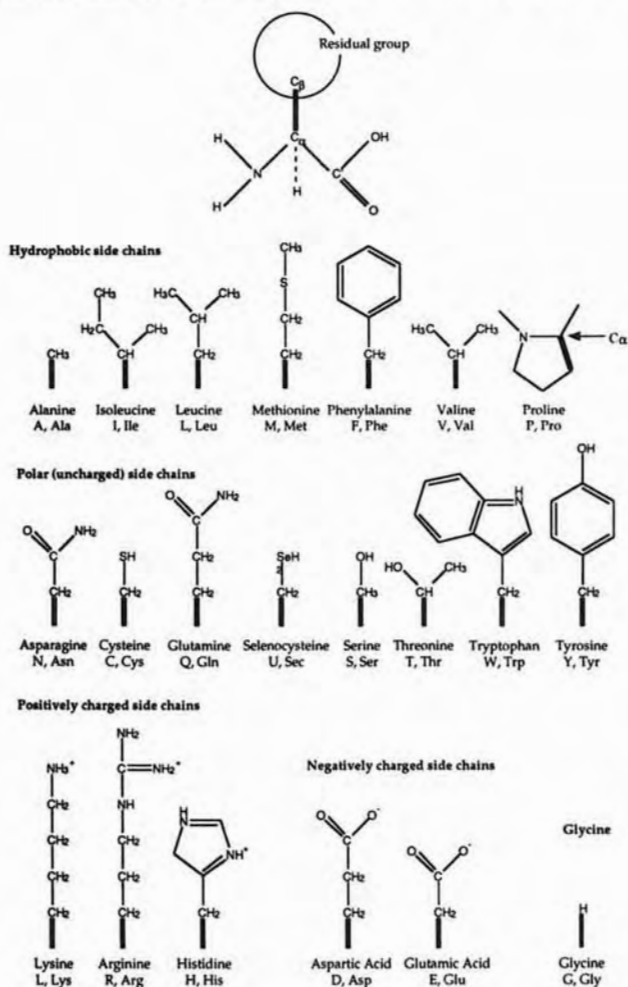
containing either amide or hydroxyl groups that can form hydrogen bonds; others contain charged residues which can form salt bridges. The figure below shows a generic amino acid and the side chains of the 21 standard types grouped according to their chemical properties. Substitutions within groups are often *conservative* (q.v.) whereas those between groups are usually *nonconservative* (q.v.). The first carbon atom in the side chain is  $C\beta$  and the thick bar represents the  $C\alpha-C\beta$  bond. Amino acids are abbreviated using either a **three letter** or **one letter assignment** and the specific assignments are shown for each amino acid. There are also three assignments for ambiguous residues: Glx, Z specifies glutamine or glutamic acid and Asx, B specifies aspartame or aspartic acid (these pairs are difficult to distinguish in some types of chemical analysis). Xaa or X is used where the nature of a residue is unknown or, in a consensus sequence, unimportant.

All amino acids except glycine exist as stereoisomers because the tetravalent  $C\alpha$  carbon atom is a

chiral centre. However, almost all amino acids found in proteins are L-isomers, reflecting the specificity of the protein synthesis enzymes (this specificity is thought to have evolved by chance, a 'frozen evolutionary accident'; conversely, most sugars in biological systems are D-isomers). Proline is classed as one of the standard amino acids but, because it contains a secondary amino group, it is actually an *imino* acid.

**Nonstandard amino acids.** Individual amino acids may be subject to cotranslational or posttranslational chemical modification, increasing the repertoire of bond forming capabilities of proteins. Such modifications may be permanent or reversible and may involve minor changes or the addition of large chemical groups (see discussion in the main text). Importantly, all these modifications occur *after* the standard amino acid has been incorporated into the polypeptide chain. There are also a number of situations where unusual amino acids are used as substrates for incorporation. Methionine and serine may

be modified before incorporation, to generate *N*-formylmethionine and selenocysteine respectively. In each case, the modification takes place on a charged tRNA, which is different from the tRNA used to incorporate the standard amino acid (the initiator tRNA rather than tRNA<sup>Met</sup> for methionine, and tRNA<sup>Sec</sup> rather than tRNA<sup>Ser</sup> for serine). While *N*-formylmethionine and unformylated methionine are functionally equivalent (blocking the formylation reaction has no effect), selenocysteine performs a specialized role completely different to that of serine, and it is thus regarded independently as a standard amino acid, even though the route to its synthesis is unorthodox. Another example of the incorporation of nonstandard amino acids in proteins is the D-alanine residues in the peptidoglycan of bacterial cell walls. Peptidoglycan is composed of glycosylated tetra-D-alanine units cross-linked by penta-L-glycine bridges. It is this unusual cross-linking reaction which is prevented by the antibiotic penicillin.





**Box 22.2: Methods for studying protein structure**

**Methods for studying gross properties of proteins.** Proteins can be separated by *electrophoresis* (q.v.) on the basis of their mass or charge. Size-fractionation is achieved by **SDS-PAGE** (polyacrylamide gel electrophoresis in the presence of the detergent sodium dodecylsulfate). SDS has a large negative charge and binds to protein main chains. The amount of SDS bound to a protein is therefore roughly proportional to its size, and the excess negative charge of many SDS molecules effectively cancels any charge carried by the protein itself. This allows separation on the basis of size alone by sieving through pores in the gel. Separation by charge is achieved by **isoelectric focussing**, where proteins move to their isoelectric point (pH equilibrium point) in a pH gradient. The two techniques can be combined in **2-D electrophoresis** for high resolution protein separation. Electrophoresis can be used to fractionate, identify and isolate proteins and to determine their approximate mass by using markers. Masses can be determined more accurately by ultracentrifugation (q.v. *Svedberg unit*) and, more recently, by an innovation in mass spectrometry where proteins are sprayed into the spectrometer in a volatile solvent which evaporates rapidly.

**Antibodies in protein analysis.** Antibodies are secreted *immunoglobulins* (q.v.) which have great specificity for their cognate antigens. These molecules can therefore be exploited to detect and purify proteins. Antibodies can be produced by injecting antigen into rabbits and purifying antibodies from the blood. Such antibodies are **polyclonal** (produced by different B-lymphocytes) and usually recognize more than one feature or **epitope** of the antigen. **Monoclonal antibodies** are generated by fusing B-lymphocytes to myeloma cells to produce clones of immortalized lymphocytes termed **hybridomas**. Hybridomas produce a single type of antibody and are immortal. More recently, it has been possible to produce antibodies using recombinant DNA techniques. As well as producing antibodies for detecting and purifying proteins, recombinant antibodies can be engineered with novel properties, such as catalytic activity (**abzymes**) and as tools for gene therapy (q.v. *intrabodies*).

Antibodies can be conjugated to radioactive, fluorescent or enzymatic labels for detection of specific proteins immobilised on solid supports following electrophoresis (**western blot**, **immunoblot**) or as part of a library screen (**immunoscreening**, q.v. *expression library*). The same principle can be used to detect proteins *in situ*, to determine the temporal and spatial distribution of proteins (*in situ* immuno-

histochemistry). Antibodies can be used to purify proteins, by immunoprecipitation or affinity chromatography, and labeled antibodies can be used as a sensitive quantitative assay for antigens, e.g. in the enzyme-linked immunosorbent assay (ELISA).

**Methods for studying primary polypeptide structure.** The quickest route to the determination of primary polypeptide structure is not by direct analysis of the polypeptide, but by cloning and sequencing the corresponding cDNA. However, this is not always possible, and cDNA sequences provide only the sequence of the nascent polypeptide, whereas the functional protein may be cleaved and modified. The direct analysis of polypeptide sequence is facilitated by automated **Edman degradation**, where the terminal amino acid residue is labeled and then specifically cleaved and identified. This process has been automated using machines termed **sequencers**, and the complete sequence of polypeptides up to 50 residues in length can be determined in a single run. Larger proteins can be characterized by first cleaving the protein into short peptide fragments using specific proteases or dipeptide-specific chemicals (e.g. hydroxylamine specifically cleaves asparagine-glycine bonds). **Protein microsequencing** is carried out by gas phase automated Edman degradation followed by high pressure liquid chromatography. This allows direct sequencing of picomole amounts of protein, such as bands eluted from polyacrylamide gels, and provides a direct route to the design of degenerate oligonucleotides which can be used to isolate the corresponding cDNA from a suitable *library* (q.v.).

**Methods for studying higher order protein structure.** It is not yet possible to determine higher order protein structure from primary sequence data and therefore a collection of biophysical methods are used to probe tertiary and quaternary structures directly.

**Circular dichroism (CD)** describes the optical activity of asymmetric molecules characterized by differing absorption spectra in left and right circularly polarized light. CD spectrophotometry between 160 and 240 nm allows rapid characterization of protein secondary structure because  $\alpha$ -helix,  $\beta$ -sheet and coil generate distinct CD spectra.  $\alpha$ -helices, for instance, generate a characteristic absorbance spectrum with a peak of positive differential absorption ( $\Delta A$ ) at about 190 nm and twin peaks of negative differential absorption at about 210 and 220 nm.

Two techniques, X-ray crystallography and nuclear magnetic resonance spectroscopy, can be

used to study protein atomic structure. **X-ray crystallography** uses precisely oriented protein crystals to scatter X-rays onto a detector. X-rays are scattered by electrons and the amplitude of the scattering effect is proportionate to the number of electrons in the atom. The waves can reinforce or cancel each other, and the manner in which this occurs depends upon the spatial orientation of different atoms in the protein. The resulting image, a series of spots or **reflections** of differing intensities, can be used to reconstruct an image of the protein using a mathematical function termed a **Fourier transform**. The data are used to determine electron densities and phases to create an electron density map. The interpretation of the map depends on the amount of data used in the Fourier synthesis, and this governs the resolution of the final structure. The maximum resolution of X-ray crystallography is approximately 0.1–0.2 nm.

**Neutron scattering** can be used to enhance X-ray crystallographic images because neutrons generate strong diffraction patterns from small atoms, including hydrogen. **Fiber diffraction**, using X-rays or neutrons, can be used to analyze the structure of elongated fibers such as nucleic acids and long fibrous proteins (e.g. collagen, keratin). Unlike crystallographic analysis, which generates precise three-dimensional images, fiber diffraction patterns represent a two-dimensional average of the cylindrical cross-section of the fiber (q.v. *double helix*).

**Nuclear magnetic resonance (NMR) spectroscopy** is used to analyze the structure of proteins in solution. The resolution is similar to that of X-ray crystallography, but the technique is

applicable only to molecules of low molecular weight. The basis of NMR is that some nuclei, including hydrogen, nitrogen and phosphorus (as well as rare isotopes of carbon and oxygen) possess intrinsic magnetism and can switch between magnetic spin states in an applied magnetic field by absorbing electromagnetic energy (a similar technique, **electron spin resonance (ESR) spectroscopy**, is applied to paramagnetic materials, i.e. those containing unpaired electrons). Absorbance can be recorded as a resonance frequency, which is specific for each type of nucleus. Additionally, the resonance frequency is influenced by surrounding electron density, so that atoms in different chemical environments undergo a **chemical shift** and absorb energy at different resonance frequencies. This can be used to discriminate between different chemical groups (methyl, aromatic, etc.). The manner in which nuclear magnetic resonance decays after a magnetic pulse is also highly informative, because it depends on molecular structure and spatial configuration. The **nuclear Overhauser effect (NOE)** is the result of the transfer of magnetic energy through space and occurs only if interacting nuclei are less than 0.5 nm apart. Two-dimensional **NOE spectroscopy (NOESY)** identifies atoms which are close together as symmetrical peaks superimposed over the typical one dimensional NMR chemical shift spectrum which lies along the diagonal of the plot. **Spin-spin coupling** is the transfer of magnetic energy through chemical bonds to neighboring nuclei. This can also be investigated by other forms of two-dimensional NMR, termed COSY and TOCSY.

### Box 22.3: Protein targeting, sorting and processing

**Protein traffic in the cell.** **Protein targeting** describes the processes directing proteins to particular destinations in the cell (c.f. *gene targeting*). In bacteria, the choice of destination is between the cytoplasm, the inner and outer cell membranes, and the periplasmic space between them; proteins can also be secreted. In eukaryotes, proteins can also be targeted to any one of several intracellular organelles, as well as to the nucleus. Since all bacterial proteins and most eukaryotic proteins are synthesized in the cytoplasm, targeted proteins must carry recognizable sequences or structures which allow them to be transported to the appropriate

cellular compartment. This process is termed **protein sorting** or **protein trafficking**.

**The nature of sorting information.** Many proteins destined for particular cellular compartments contain either a specific amino acid sequence or an arrangement of residues with particular chemical characteristics which interact with the cell's sorting machinery. These conserved residues are termed **signal sequences**. They are often found at the termini of polypeptides (**signal peptides**) and may be cleaved off when the protein reaches its destination. A protein with a signal sequence that is later discarded is termed a **preprotein**, and the signal

peptide a **presequence**. This differs from a protein activated by proteolysis (Table 22.2) which is termed a **proprotein** (proteins with presequences which are also activated by proteolysis are termed **preproproteins**, e.g. preproinsulin). Where a protein must cross several interfaces within the cell, a number of signal sequences may be arranged in series and may act sequentially. N-terminal signal peptides are the first structures of the nascent protein to emerge from the ribosome, allowing protein sorting to occur cotranslationally. C-terminal and internal signals, e.g. those for nuclear or peroxisomal localization, and **signal patches** — signals formed on the surface of the protein by folding, but whose components are not contiguous in the primary sequence — function posttranslationally. Proteins have to be maintained in an unfolded state or must be unfolded before being transported across membranes, a process which is facilitated by *molecular chaperones* (see main text). Whereas signal peptides act directly to target proteins, signal patches tend to act indirectly, as sites for glycosylation, the sugar residues being recognized by the sorting apparatus of the cell. Sorting signals can be disguised by interaction with other proteins and this can be used as a regulatory mechanism, e.g. the transcription factor NF- $\kappa$ B has a nuclear import signal which is masked by the inhibitory factor I- $\kappa$ B, thus regulatory systems which destroy I- $\kappa$ B allow nuclear uptake of NF- $\kappa$ B and activation of the genes it controls (see Signal Transduction, Transcription).

**The secretory pathway.** In eukaryotes, proteins destined for secretion are initially targeted to the endoplasmic reticulum (ER). This requires an N-terminal hydrophobic signal peptide which is recognized by a cytoplasmic ribonucleoprotein complex termed the **signal recognition particle (SRP)**. The SRP comprises six proteins and a small cytoplasmic RNA (**7S RNA**), which is homologous to the *Alu element* (q.v.). It binds to the signal sequence as it emerges from the ribosome and stalls protein synthesis until the complex reaches the ER. Here, the SRP binds to its receptor, a dimeric **docking protein** located on the endoplasmic reticulum membrane (the SRP has three functional domains which sponsor signal binding, receptor binding and elongation arrest respectively). The ribosome then associates with its own receptor, a trimeric transmembrane protein called Sec61, and protein synthesis recommences. The polypeptide is fed into the ER lumen, a process termed **vectorial discharge** or **cotranslational import**, through the Sec61 complex (some proteins require an additional component called TRAM). After synthesis, GTP exchange releases the ribosome from the SRP, and the SRP itself is released from its receptor back into the cytoplasm.

The signal sequence is cleaved as the polypeptide enters the ER lumen by a pentameric signal peptidase. Proteins with asparagine residues in the tripeptide motif Asn-Xaa-Ser/Thr (known as a **sequon**) are N-glycosylated with preformed oligosaccharide units. Not all sequons are glycosylated so other residues may also be involved in the determination of a glycosylation site or the structure of the protein itself may inhibit the glycosylation of certain polypeptides.

Transmembrane proteins possess an internal hydrophobic **stop transfer sequence** which lodges the protein in the ER membrane. The remainder are packaged into vesicles and pass to the Golgi apparatus, although those which are to be retained in the ER lumen are recognized and selectively returned to the ER. Such proteins often have a **retention signal** with the consensus sequence KDEL.

In the Golgi apparatus, N-linked glycan chains can be further modified and *de novo* O-glycosylation of serine, threonine and hydroxylysine residues may occur. In some cases, the glycosylation is required for correct protein folding and function (q.v. *expression cloning*) whereas in others the glycosylation itself acts as a targeting signal. Proteins with a signal patch are modified by addition of mannose-6-phosphate, which targets them to the lysosomes. The remainder of proteins are secreted, although some may possess sequences which cause them to be retained in the membrane of one of the organelles in the secretory pathway (see below).

In bacteria, proteins destined to be secreted are synthesized as preproteins with N-terminal signal sequences sometimes termed **leader peptides**. Bacterial signal sequences are short ( $\leq 25$  residues) and comprise a hydrophobic central core which can adopt an  $\alpha$ -helical structure, flanked by regions containing several charged residues. A number of chaperone proteins can bind to the nascent protein as the leader peptide emerges from the ribosome, to prevent misfolding. One such protein, SecB, plays a predominant role in protein export because it binds another component of the secretory system, SecA, a chaperone associated with the translocation apparatus. SecA mediates translocation by feeding the substrate protein through the translocation apparatus (comprising transmembrane proteins SecE and SecY) in a manner which is dependent on ATP hydrolysis. The leader peptide is then cleaved by an enzyme called **leader peptidase**. In the mature protein, amino acids with small residual groups are often found adjacent to the cleavage site (the -1 position) and the next but one upstream residue (the -3 position). This phenomenon, is known as the **-1 and -3 rule**.



**Targeting proteins to membranes.** In eukaryotes, many proteins entering the secretory pathway are destined to be retained in a particular membrane. The stop transfer sequence allows a protein to be retained in the ER membrane itself, but by combining various similar signals (**membrane anchor signals**) with particular retention signals, a protein can also be targeted to the Golgi or lysosomal membrane. Any membrane-associated proteins lacking retention signals will eventually arrive at the cytoplasmic membrane, but this is only one of several routes to this destination. Proteins in the Golgi which carry a **GPI sequence** become conjugated to a GPI (glycosyl-phosphatidylinositol) anchor which attaches the protein to the inner membrane surface. Proteins in the cytoplasm which become modified by fatty acylation (e.g. myristoylation) are also targeted to the membrane.

**Protein transport into the nucleus and intracellular organelles.** Small proteins can diffuse readily through nuclear pores but larger proteins (> 50kDa) destined for the nucleus must be imported, a process which involves the recognition of a **nuclear localization sequence (NLS)**. NLSs are generally short, and rich in proline and basic residues, but lacking in hydrophobic residues. Some NLSs are contiguous, e.g. the SV40 T antigen NLS (**Pro-Pro-Lys-Lys-Lys-Arg-Lys**), while some are bipartite, with two proline/basic-rich regions interrupted by a sequence of polar residues (e.g. the p53 NLS is **Lys-Arg-Ala-Leu-Pro-Asn-Asn-Thr-Ser-Ser-Ser-Pro-Gln-Pro-Lys-Lys-Lys**). A dimeric cytoplasmic protein, **importin**, is required for the docking stage of nuclear import. Importin  $\alpha$  binds the NLS of the substrate and importin  $\beta$  recognizes and binds to a component of the nuclear pore, **nucleoporin**. A GTPase called Ran then causes the importins to dissociate. It is thought that nuclear import may involve a series of such reactions, with the substrate being docked at nucleoporins running through the pore.

Proteins destined for import into mitochondria or chloroplasts can be targeted to various compartments: the outer or inner membranes, the intermembrane space, the internal matrix/stroma, and in chloroplasts, the thylakoid membrane or lumen. Proteins for import into organelles usually contain an N-terminal **transit sequence**. This is a 15–70 residue sequence, rich in polar and basic residues (chloroplast transit sequences tend to be richer in polar residues than those of mitochondria), which is recognized by a receptor on the organelle surface. The transit sequence by default targets proteins to the mitochondrial matrix or chloroplast stroma, wherein it is cleaved by a specific peptidase. The protein is translocated across outer and inner membranes at the same time, where they are brought into contact

by the translocation apparatus. Further signals are required for targeting to alternative compartments. Organelle localization signals are arranged in a hierarchical manner, so alternative signals become active when the N-terminal transit sequence is cleaved. The peptidase which removes the transit sequence is located in the matrix/stroma, so organelle proteins must transiently enter the matrix/stroma even if their final destination is elsewhere. Proteins which are inserted into the membrane contain hydrophobic stop transfer sequences.

**Protein degradation pathways.** Proteins vary widely in stability, with half lives ranging from a few minutes (e.g. many regulatory proteins) to several weeks or longer (e.g. many structural and storage proteins: collagen, hemoglobin). There are several distinct degradation pathways in the cell whose activities can vary according to nutritional status and signaling from other cells, and which may involve protein modification. Many characteristically unstable proteins contain **PEST sites** (i.e. rich in proline, glutamine, serine and threonine) and target them for rapid proteolysis (q.v. *cyclins*). Other proteins may be targeted for degradation by posttranslational modification, either involving chemical modification of specific residues (e.g. phosphorylation) or conjugation to the small protein **ubiquitin**. The N-terminal residue of a protein and its state of modification can also play a role in the regulation of protein turnover by specifying the protein as a potential target for **ubiquitination**. This is known as the **N-end rule**.

The lysosomal pathway is a ubiquitin-independent pathway which generally involves relatively stable proteins. Proteins are internalized by the budding off of vesicles, a process termed **microautophagy**. During starvation some cytoplasmic proteins may be taken into the lysosome directly, by recognition of a signal peptide.

Ubiquitin-dependent degradation is the principle degradation pathway for cytosolic proteins, involving the activation of the small protein ubiquitin, which is transferred to its substrates by a carrier protein and targets them for degradation in a large multiprotein complex termed the **proteasome**. Ubiquitin is activated by conjugation to a protein called E1. It is then transferred to protein E2, which in turn transfers ubiquitin to its target protein, identified by binding to protein E3 (**ubiquitin ligase**). Substrate specificity is controlled by both E2 and E3, of which there are several isoforms which may target different types of protein. E2 transfers ubiquitin to the  $\epsilon$  amino group of lysine residues on the target protein. Ubiquitin itself then becomes a target for ubiquitination on Lys 46, resulting in a polyubiquitinated substrate, which is the signal for proteasome-mediated degradation.



## References

- Branden, C. and Tooze, J. (1991) *Introduction to Protein Structure*. Garland Press, New York.
- Doolittle, R.F. (1985) Proteins. *Sci. Amer.* 253: 88–99.
- Lewin, B. (1996) *Genes VI*. Oxford University Press, Oxford, pp. 663–711.
- Richardson, J.S., Richardson, D.C., Tweedy, N.B., Gernet, K.M., Quinn, T.P., Hecht, M.H., Erickson, B.W., Yan, Y., McClain, R.D., Donlan, M.E. and Suries, M.C. (1992) Looking at proteins: representations, folding, packing and design. *Biophys. J.* 63: 1186–1220.
- Stryer, L. (1995) *Biochemistry (4th edn)*. W.H. Freeman, New York.
- Allen, J.B., Wallberg, M.W., Edwards, M.C. and Elledge, S.J. (1995) Finding prospective partners in the library — the 2-hybrid system and phage display find a match. *Trends Biochem. Sci.* 20: 511–516.
- Britten, R.J. (1996) DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl Acad. Sci. USA* 93: 9374–9377.
- Cooke, J., Nowak, M.A., Boerlijst, M., and Maynard-Smith, J. (1997) Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends Genet.* 13: 360–363.
- Cooper, A.A. and Stevens, T.H. (1995) Protein splicing — self-splicing of genetically mobile elements at the protein level. *Trends Biochem. Sci.* 20: 351–356.
- Davidson, J.N. and Peterson, M.L. (1997) Origin of genes encoding multi-enzyme proteins in eukaryotes. *Trends Genet.* 13: 281–285.
- Dibb, N.J. (1993) Why do genes have introns? *FEBS Lett.* 325: 135–139.
- Doolittle, R.F. (1995) The origins and evolution of eukaryotic proteins. *Phil. Trans. R. Soc. Lond.* 349: 235–240.
- Evans, M.J., Carlton, M.B.L. and Russ, A.P. (1997) Gene trapping and functional genomics. *Trends Genet.* 13: 370–375.
- Fersht, A.R. (1997) Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* 7: 3–9.
- Gabor Miklos, G.L. and Rubin, G.M. (1996) The role of the genome projects in determining gene function: Insights from model organisms. *Cell* 86: 521–529.
- Goffeau, A., Barrell, B.G., Bussey, et al. (1996) Life with 6000 genes. *Science* 274: 546–567.
- Gorlich, D. and Mattaj, I.W. (1996) Nucleocytoplasmic transport. *Science* 271: 1513–1518.
- Govindarajan, S. and Goldstein, R.A. (1996) Why are some protein structures so common? *Proc. Natl Acad. Sci. USA* 98: 3341–3345.
- Heinkoff, S., Greene, E.A., Pietrovski, S., Bork, P., Attwood, T.K. and Hood, L. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science* 278: 609–614.
- Hieter, P. and Boguski, M. (1997) Functional genomics: it's all how you read it. *Science* 278: 601–602.
- Hochstrasser, M. (1996) Ubiquitin-dependent protein degradation. *Annu. Rev. Genet.* 30: 405–439.
- Holland, P.W.H., Garcia-Fernandez, J., Williams, N.A. and Sidow, A. (1994) Gene duplications and the origins of vertebrate development. *Development (Suppl.)*: 125–133.
- Lander, E.S. (1996) The new genomics: Global views of biology. *Science* 274: 536–539.
- Levitt, M., Gerstein, M., Huang, E., Subbiah, S. and Ysai, J. (1997) Protein folding — the endgame. *Annu. Rev. Biochem.* 66: 549–579.
- Loomis, W.F. and Sternberg, P.W. (1995) Genetic Networks. *Science* 269: 649.
- Martin, J. and Hartl, F.U. (1997) Chaperone-assisted protein folding. *Curr. Opin. Struct. Biol.* 7: 41–52.
- Neupert, W. (1997) Protein import into mitochondria. *Annu. Rev. Biochem.* 66: 863–917.
- Rapoport, T.A., Jungnickel, B. and Kutay, U. (1996) Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes. *Annu. Rev. Biochem.* 65: 801–848.
- Rechsteiner, M. and Rogers, S.W. (1996) PEST sequences and regulation by proteolysis. *Trends Biochem. Sci.* 21: 267–271.
- Roder, H. and Colon, W. (1997) Kinetic role of early intermediates in protein folding. *Curr. Opin. Struct. Biol.* 7: 15–28.
- Rothman, J.E. (1994) Mechanisms of intracellular transport. *Nature* 372: 55–68.
- Schatz, G. and Dobberstein, B. (1996) Common principles of protein translocation across membranes. *Science* 271: 1519–1526.
- Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. *Science* 278: 631–637.
- Walter, P. and Johnson, A.E. (1994) Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. *Annu. Rev. Cell Biol.* 10: 87–119.
- Xue, L. and Noll, M. (1996) The functional conservation of proteins in evolutionary alleles and the dominant role of enhancers in evolution. *EMBO J.* 15: 3722–3731.

## Websites

- The yeast protein database — <http://www.proteome.com/YPDhome.html>
- Structural classification of proteins (SCOP) database — <http://scop.mrc-lmb.cam.ac.uk/scop>

**This Page Intentionally Left Blank**

## Chapter 23

# Protein Synthesis

### Fundamental concepts and definitions

- Protein synthesis is the level of gene expression where genetic information carried as a messenger RNA (mRNA) molecule is translated into a polypeptide.
- The components of the protein synthesis machinery are mRNA, the template which contains the code to be translated, ribosomes, large ribonucleoprotein particles which are the sites of protein synthesis, transfer RNA (tRNA), versatile **adaptor molecules** which carry amino acids to the ribosome and facilitate the process of translation, and accessory factors associating transiently with the ribosome, required for ribosome assembly and disassembly, and its activity during elongation.
- Like other polymerization reactions, protein synthesis has stages of **initiation**, **elongation** and **termination**, each of which may be regulated. The protein synthesis mechanism is similar in prokaryotes and eukaryotes, but there are subtle differences in the nature of the components and their order of assembly. There are major differences, however, concerning the context in which protein synthesis occurs. In bacteria, transcription and protein synthesis occur simultaneously in the cytoplasm (allowing cross-regulation between different levels of gene expression) and mRNA may be polycistronic. In eukaryotes, transcription is confined to the nucleus and the mRNA is exported to the cytoplasm for translation. Nascent mRNA is extensively processed before export and is usually monocistronic (*see RNA Processing*).
- Following synthesis, a polypeptide undergoes further processing before becoming active. It must fold correctly, a process sometimes requiring the assistance of a *molecular chaperone* (q.v.). It may be cleaved, and specific residues may be chemically modified or conjugated to small molecules. Such modifications are often associated with the targeting of proteins to specific compartments in the cell or for secretion. Proteins may need to associate noncovalently with other proteins or with nonpolypeptide cofactors for their full activity. For a discussion of these processes, *see Proteins: Structure, function and evolution*.

### 23.1 The components of protein synthesis

**Messenger RNA.** Messenger RNA (mRNA) is the template for protein synthesis. It has two essential features: an **open reading frame** (a sequence of translatable codons) and a **ribosome binding site** (where the small ribosome subunit binds and the rest of the ribosome assembles). There are important distinctions between prokaryotes and eukaryotes with respect to the organization of these sites, and also concerning other aspects of the life of a typical mRNA molecule. These differences and their consequences are summarized below.

(1) In bacteria, transcription and translation occur simultaneously in the same cellular compartment, whereas in eukaryotes, transcription is restricted to the nucleus and RNA must be exported into the cytoplasm for translation.

(2) Bacterial mRNA has a limited half-life (several minutes for the most stable transcripts). Some eukaryote mRNAs are also unstable, but most are stable for hours or even days (e.g. in eggs).

(3) Bacterial transcripts are used directly for translation, whereas eukaryotic transcripts are extensively processed and modified beforehand. Processing reactions include the splicing of introns and 3' end polyadenylation; both may regulate the efficiency of translation, either directly or by modulating mRNA stability. A further modification is the synthesis of a 5' end 7-methylguanosine cap, which has a direct role in ribosome binding. Some transcripts are also edited (q.v. *RNA editing*).

(4) Ribosome binding in bacteria depends on a conserved motif in the mRNA which complements part of the ribosomal 16S rRNA. Binding occurs wherever this sequence is found, including internally, allowing bacterial transcripts to be polycistronic. Eukaryotic ribosomes are docked onto the mRNA by a protein which recognizes the modified 5' cap. Therefore binding cannot occur internally, and eukaryotic mRNAs are almost universally monocistronic (some RNA viruses, however, have managed to circumvent this restriction in order to express genes located on their own — polycistronic — genomes; q.v. *internal ribosome entry site*). The structures of typical bacterial and eukaryotic mRNAs are shown in Figure 8.2 (see The Gene).

The broad consequences of these differences are reflected in the regulatory strategies used by bacteria and eukaryotes. Bacterial mRNA is transcribed, translated and degraded in quick succession, and the close association of these three processes allows a significant amount of cross-regulation between different levels of gene expression (q.v. *attenuation*, *retroregulation*). Conversely, in eukaryotes there is a considerable delay between transcription and translation while the mRNA is first processed and then exported from the nucleus. Both of these events are potential targets for regulation.

**Ribosomes.** Ribosomes are large, abundant ribonucleoprotein complexes upon which protein synthesis occurs. They are found free in the cytoplasm and, in eukaryotes, associated with the membrane of the rough endoplasmic reticulum. Ribosomes may function alone (**monosomes**) although it is common to see them in clusters concurrently translating the same mRNA (**polysomes**). Polysomes can be extracted from cells and used to purify mRNA (q.v. *poly(A)<sup>+</sup> RNA*).

All ribosomes comprise two dissimilar sized subunits, the **large** and **small subunits**. Each subunit consists of several **ribosomal RNAs (rRNAs)** and numerous **ribosomal proteins (r-proteins)**. Their relative sizes are often expressed in Svedberg units (Box 23.1). In *E. coli*, the 70S ribosome is composed of a small 30S subunit and a large 50S subunit. The small subunit contains 21 different proteins (named S1–S21), and the 16S rRNA. The large subunit comprises 34 proteins (named L1–L34) and the 23S and 5S rRNAs. Some proteins are common to both subunits (e.g. L6 = S20). Eukaryote ribosomes are larger (80S) and contain more components. The small (40S) subunit comprises 33 proteins and the 18S rRNA whilst the large (60S) subunit contains 50 proteins and three rRNAs of 28S, 5.8S and 5S. The 5.8S rRNA is homologous to the 5' portion of the bacterial 23S rRNA and forms complementary pairs with the equivalent eukaryotic 28S rRNA. Archaeal ribosomes resemble those of bacteria, but some genera contain extra subunits similar to those of eukaryotes.

The spatial organization of the ribosome is complex. rRNA makes up 60–65% of the total mass and is essential for structural integrity and function, adopting complex tertiary and quaternary conformations by intra- and intermolecular base pairing. The manner in which proteins interact with the RNA and each other within this network has been studied in great detail for the *E. coli* ribosome using cross-linking and footprinting studies, and structural analysis by neutron scattering and diffraction. Although the primary nucleic acid sequences of rRNAs from different species vary considerably, it appears that most of the secondary structures are conserved, suggesting that all ribosomes share a common organization.

Several domains of the ribosome have particularly important functions during protein synthesis. The small subunit contains a binding site for mRNA and two major binding sites for tRNA. The **A-site (amino acyl-tRNA site)** binds incoming charged tRNAs during elongation, whilst the **P-site (peptidyl-tRNA site)** binds the tRNA carrying the nascent polypeptide chain. Bacterial ribosomes possess a third **E-site (exit site)** to which spent tRNAs are dispatched prior to ejection. The large subunit possesses a **peptidyltransferase domain**, which provides the catalytic activity for peptide bond formation, and a GTPase domain, whose activities are required for translocation of the ribosome along the mRNA. The roles of these sites during protein synthesis are discussed in more detail below.

**Transfer RNA.** Before the *genetic code* (q.v.) was understood, Francis Crick proposed the **adaptor hypothesis** to explain how the nucleotide sequence in mRNA could be translated into protein. The



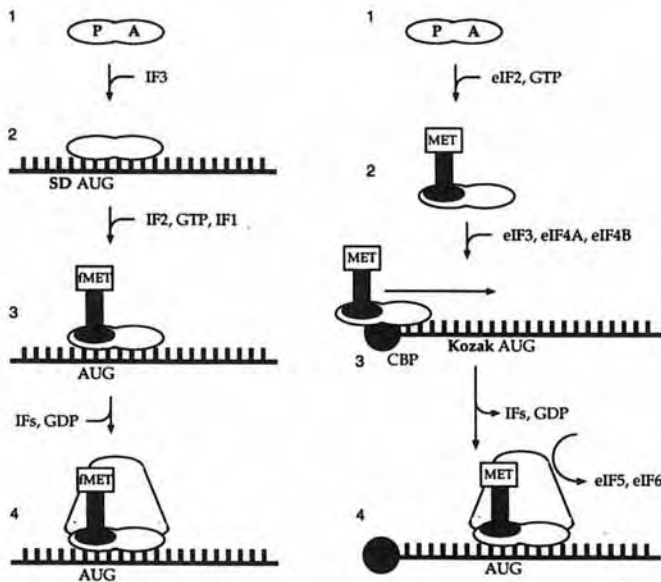
model predicted the existence of an adaptor molecule which would recognize both the nucleic acid sequence of the message and the appropriate amino acid, bringing the two together at the ribosome.

The adaptor molecule is **transfer RNA (tRNA)**. The tRNAs are a relatively homogeneous family of RNA molecules, usually 75–100 nucleotides in length, which are extensively processed during their production (see RNA Processing). They possess a characteristic secondary and tertiary structure (Box 23.2), most importantly the **acceptor stem** (to which the amino acid binds) and the **anticodon loop** (which carries the three nucleotide **anticodon** that forms complementary base pairs with codons in the mRNA). Bacterial cells contain up to 35 different tRNAs, and eukaryotic cells up to 50. This number is lower than the number of possible codons in the genetic code, but greater than the number of amino acids specified by the code. This indicates that individual tRNAs can recognize more than one codon (reflecting *wobble pairing*, q.v.), but that different tRNAs may be charged with the same amino acids (these are *isoaccepting tRNAs*, q.v.).

The tRNAs are **charged** (conjugated to their corresponding amino acids) by enzymes termed **amino acyl tRNA synthetases**. There is one enzyme for each amino acid, and therefore each synthetase recognizes all its cognate isoaccepting tRNAs. The charging mechanism and its regulation are considered in detail elsewhere (see The Genetic Code).

### 23.2 The mechanism of protein synthesis

**Overview.** The initiation stage of protein synthesis includes the assembly of the protein synthesis components, the positioning of the ribosome at the translation start site, and the placement of the first amino acid in the polypeptide (Figure 23.1). Initiation ends when the first peptide bond is formed. In both prokaryotes and eukaryotes, the small ribosome subunit binds to the mRNA before



**Figure 23.1:** Initiation of protein synthesis in bacteria and eukaryotes. In bacteria (left) the small subunit (1) binds to the Shine-Dalgarno (SD) sequence (2) and recruits the initiator tRNA (3), whereupon the large subunit binds (4). In eukaryotes (right), the small subunit (1) associates with eIF-2-GTP and the initiator tRNA to form a preinitiation complex (2) which recognizes the cap-binding protein (CBP) at the 5' cap of the mRNA, binds, and scans along until it finds an initiator codon in the correct context (3) before the large subunit binds (4). For clarity, the small ribosome subunit is shown to span just six bases, although in reality it spans 30–40 bases, allowing it to interact with the bacterial SD sequence (or the eukaryotic Kozak sequence) and the initiation codon simultaneously. The A-site and P-site of the ribosome are indicated.

the large subunit, and binding requires a number of **initiation factors (IFs)**, which are released when the large ribosome subunit is recruited. It is thus the small subunit that recognizes the **ribosome binding site** on the mRNA. The first amino acid is attached to a special **initiator tRNA**, which has the unique property of being able to enter the partial P-site on the small ribosome subunit. The **initiator codon** is usually AUG; it marks the beginning of the open reading frame and sets the reading frame for the rest of the polypeptide. All other **elongator tRNAs** enter the *complete* A-site in the fully assembled ribosome during the elongation stage. Elongation involves a cycle of three reactions — recruitment, transpeptidation and translocation. Each step requires **elongation factors (EFs)** and the hydrolysis of GTP provides energy. The net effect of the three reactions is to transfer amino acids from the cytoplasm onto the nascent polypeptide chain and to shunt the ribosome along the mRNA in units of three nucleotide residues. Protein synthesis ceases when the ribosome encounters a termination codon, where release factors attach to the ribosome and cause the translation machinery to disperse.

**Initiation in bacteria — detailed mechanism.** In eubacteria, the 30S ribosomal subunit binds to the mRNA only when associated with initiation factor IF3, which stabilizes the structure of the small subunit and prevents premature interaction with the large subunit. Although IF3 controls the ability of the small subunit to bind to the mRNA, the specificity of the recognition is provided by the ribosome itself. The ribosome binding site on the mRNA is the **Shine–Dalgarno sequence** (consensus UAAGGAGG) and forms base pairs with a complementary sequence in the 16S rRNA of the ribosome. The initiator codon (usually AUG but sometimes GUG or UUG) lies ~10 nt downstream of the Shine–Dalgarno sequence and correct positioning of the ribosome aligns the initiator codon with the P-site of the ribosome subunit.

The initiator codon specifies the amino acid methionine. There are two tRNAs carrying methionine in *E. coli*. One ( $\text{tRNA}_f^{\text{Met}}$ ) specifically recognizes initiation codons and the other ( $\text{tRNA}_m^{\text{Met}}$ ) recognizes internal AUG codons (internal GUG and UUG codons are recognized by  $\text{tRNA}^{\text{Val}}$  and  $\text{tRNA}^{\text{Leu}}$ , respectively). Once the initiator tRNA has been charged, the methionine is modified by formylation of the amino group. The signal for this may be unpaired residues in the terminal position of the double-stranded acceptor stem, which are paired in elongator tRNAs (methionyl- $\text{tRNA}_m^{\text{Met}}$  cannot be formylated). This signature also prevents the initiator tRNA entering the ribosomal A-site during elongation, and the modification, since it blocks the amino group, prevents N-formylmethionine forming peptide bonds with upstream residues. The initiator tRNA is recruited to the P-site of the small ribosome subunit during initiation. The P-site provides a special molecular environment which allows access only to the initiator tRNA, and not to the elongator tRNAs. The specificity of the initiator tRNA is conferred by three G:C base pairs in the anticodon stem, which are absent from elongator tRNAs. N-formylation is not required for initiator specificity — initiator tRNAs containing ordinary methionine are as efficient as correctly modified initiators. A degree of *wobble* (q.v.) in the third anticodon position allows AUG, GUG and UUG to act as initiators, but  $\text{tRNA}_f^{\text{Met}}$  recognizes these codons with diminishing efficacy; thus the choice of start codon can be used to establish a constitutive control over the efficiency of the initiation of protein synthesis.

The initiator tRNA is carried to the P-site by IF-2 in the presence of GTP. IF-1 then binds to the complete initiation complex to stabilize it. All initiation factors are released when the 50S subunit is recruited. This process requires the hydrolysis of GTP to provide energy.

**Initiation in eukaryotes — detailed mechanism.** Eukaryotic and bacterial initiation mechanisms are similar, the major differences reflecting the order of assembly of the components, the larger number of eukaryotic initiation factors, and that normal methionine is used both for initiation and elongation in eukaryotes. Thus the initiator and elongator methionyl-tRNAs ( $\text{tRNA}_i^{\text{met}}$  and  $\text{tRNA}_m^{\text{met}}$ , respectively) carry identical amino acids and differ only in the structure of the tRNA itself.

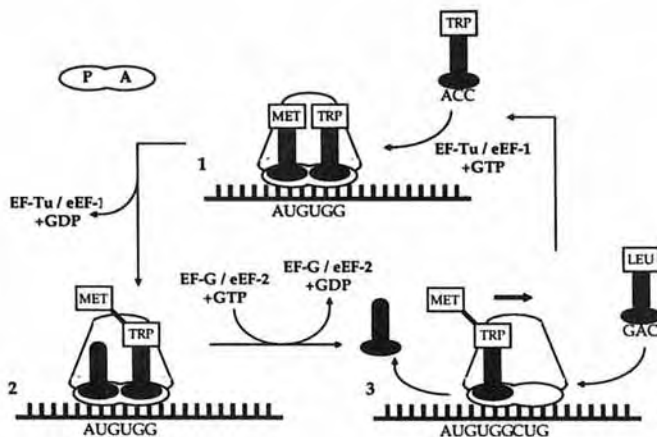
In eukaryotes, the small ribosome subunit is unable to bind to the mRNA without first becoming associated with the initiator tRNA. A ternary complex of  $\text{tRNA}_i^{\text{met}}$ , eIF-2 and GTP bind the small

subunit to form the **preinitiation complex**, which then associates with the mRNA. The ribosome binding site is defined by the initiation codon, although the sequence context surrounding this may be important and is known as the **Kozak consensus** (ACCAUGG). Initially, the preinitiation complex binds at the 5' 7meG cap structure and scans along the mRNA to find the first available initiation codon. Several initiation factors are required for this process, including a cap binding protein (eIF-4F), eIF-3, eIF-4A and eIF-4B. The cap binding protein facilitates initial binding of the 40S subunit. eIF-4A and 4B cooperate to remove secondary structure from the 5' untranslated region (UTR), then these and eIF-3 assist 40S subunit binding to form the **initiation complex**. The large subunit is associated with a further initiation factor, eIF-6. Binding of the large subunit to the initiation complex involves the loss of eIF-2 and eIF-3 from the preinitiation complex and the loss of eIF-6 from the large subunit, a process mediated by eIF-5. Initiation factors eIF-2 and eIF-6 are both **antiasociation factors**: they prevent ribosome subunits interacting in the cytoplasm. The other initiation factors dissociate from the ribosome as the large subunit binds, a process requiring GTP hydrolysis.

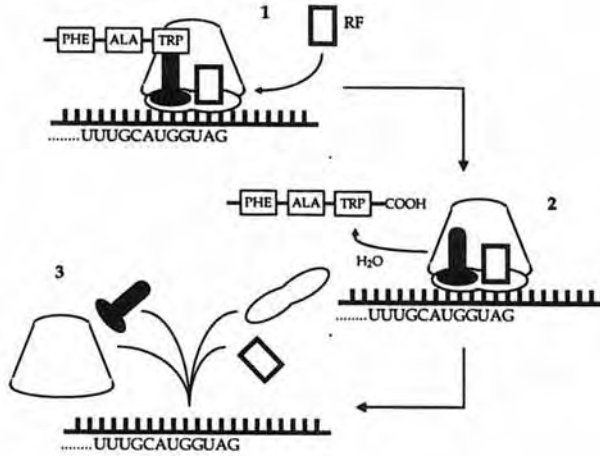
**The elongation cycle.** The elongation cycle, essentially identical in bacteria and eukaryotes, occurs in three stages — recruitment, transpeptidation and translocation (*Figure 23.2*).

Following initiation, the initiator aminoacyl-tRNA occupies the ribosomal P-site and the next codon is aligned with the A-site. The first elongator aminoacyl-tRNA enters the A site and its anticodon pairs with the codon, a recognition mediated by an **elongation factor** (EF-Tu in bacteria, eEF-1 in eukaryotes, each associated with GTP).

The peptide bond joining the initiator tRNA in the P-site to methionine is then broken and a new peptide bond is formed between methionine and the amino acid occupying the A-site. The initiator tRNA is **discharged**. This **transpeptidation** reaction is mediated by the peptidyltransferase activity of the large ribosomal subunit. The GTP associated with EF-Tu (or eEF-1) is hydrolyzed and the elongation factor is ejected. EF-Tu/GDP is inactive, and the nucleotide is displaced by a second elongation factor EF-Ts, which is itself displaced by GTP. EF-Tu can now be reused. There appears to be no counterpart to EF-Ts in eukaryotes, although eEF-1 is an aggregate of several polypeptides, one of which may perform the function of EF-Ts.



**Figure 23.2:** The elongation cycle of protein synthesis. (1) Recruitment: a charged tRNA enters the ribosomal A-site. (2) Transpeptidation: the peptide bond joining the amino acyl residue to the P-site tRNA is transferred to the amino acyl residue occupying the A-site. (3) Translocation: the spent tRNA is ejected from the P site and the ribosome translocates three bases downstream on the mRNA, shifting the A-site tRNA to the P-site and aligning the next codon with the A-site, ready for recruitment of the next charged tRNA. The A-site and P-site of the ribosome are indicated.



**Figure 23.3:** Termination of protein synthesis. When the ribosome encounters a termination codon, a release factor binds in the A-site. This releases the nascent polypeptide from the peptidyl-tRNA in the P-site and disassembles the ribosome, allowing the components of protein synthesis to be recycled.

The next stage is **translocation**. The initiator tRNA occupying the P-site is ejected (in bacteria it is moved to a further ribosomal domain termed the E-site, whereas in eukaryotes the tRNA dissociates from the ribosome). The methionyl residue is thus tethered to the amino acid occupying the A-site by a peptide bond, and the tRNA in the A-site is now referred to as the **peptidyl-tRNA**. The ribosome then translocates three bases along the mRNA while the peptidyl-tRNA remains associated with its codon. The peptidyl-tRNA is thus transferred to the P-site and the mRNA moves through the ribosome so that the next codon is aligned with the A-site. The geometry of the translocation step depends upon the mRNA-tRNA interaction: mutant tRNAs with four base anticodons cause four base translocations, resulting in *frameshifting* or *frameshift suppression* (q.v.). Translocation requires a second elongation factor (EF-G in bacteria, eEF-2 in eukaryotes, also associated with GTP) and involves GTP hydrolysis. In *E. coli* and some other bacteria, translocation also requires a free rRNA species, 4.5S rRNA, whose role is unclear.

**Termination of protein synthesis.** Termination occurs when the ribosome encounters one of three *termination codons* (q.v.). These usually have no cognate tRNAs, and translation ceases (however, c.f. *selenocysteine insertion*, *nonsense suppressors*). The completed polypeptide is released and the tRNA, mRNA and ribosomal subunits disperse and are recycled for further translation events (Figure 23.3). Termination requires a number of protein **release factors** (RFs) which bind in the A-site and recognize the nonsense codons. These facilitate the release of the polypeptide from the last peptidyl-tRNA, the ejection of the discharged tRNA from the ribosome, and the disassembly of the ribosome. In *E. coli* there are two release factors recognizing different pairs of termination codons: RF1 recognizes UAA and UAG, whereas RF2 recognizes UAA and UGA. Each is assisted by a third factor, RF3. In eukaryotes, a single GTP-dependent factor (eRF) recognizes all three termination codons, but not with the same efficiency. Readthrough may therefore occur, especially at weak termini, reflecting competition between release factors and tRNA. The misreading of termination codons is an important aspect of translational regulation (discussed below) and may be influenced by secondary structure in the mRNA.

### 23.3 The regulation of protein synthesis

**Constitutive control of protein synthesis.** The efficiency of protein synthesis is under constitutive control, but may also be regulated either globally or at the level of individual transcripts.



Constitutive levels of protein synthesis are dependent upon mRNA structure and reflect such factors as mRNA stability, the sequence and context of the ribosome binding site, the presence of secondary structure, the choice of initiation codon, and codon bias throughout the open reading frame.

Secondary structure and the choice of termination codon also permit a group of regulatory phenomena collectively described as **programmed misreading**, where the normal interpretation of the sequence of codons is suppressed. Examples of programmed misreading include readthrough, selenocysteine insertion, and frameshifting. **Readthrough** occurs at weak termination codons (i.e. termination codons where release factors are limiting), and there is competition between release factors and tRNAs with tolerable anticodon sequences. Selenocysteine insertion occurs at the termination codon UGA, and requires a secondary structure, the *selenocysteine insertion sequence* (q.v.). **Frameshifting** or **recoding** occurs when the ribosome pauses at a secondary structure or rare codon, shifts forwards or backwards by a single nucleotide, and continues translation in a different reading frame. This can be used to change the reading frame midway through translation (e.g. in the retroviral *pol* gene) to induce early truncation (e.g. the *MS2 lysis* gene) or to prevent truncation and extend a gene product (e.g. in the *E. coli dnaX* gene).

In bacteria, the operon environment allows a novel form of translational regulation where the translation of downstream genes is dependent on the translation of upstream genes in a polycistronic message. The spacing of the individual open reading frames is important. Where the space between successive open reading frames is greater than approximately 30 nucleotides, translation begins with discrete initiation events and each locus is independent. If there is a shorter gap, the ribosome can reinitiate by bridging the open reading frames, i.e. without first dissociating from the template. The initiation signal is different from the normal SD-sequence, and spontaneous assembly of ribosome subunits in the typical manner occurs with low efficiency. A mutation which blocks or interrupts translation of the upstream gene therefore also effects translation of the downstream gene; this is termed a **polar mutation** (e.g. q.v. *lac operon*).

The translation of eukaryotic mRNA is usually cap-dependent and facilitated by scanning for the first initiator codon. The picornavirus family provides an exception, in that the genomes contain **internal ribosome entry sites** (IRES), i.e. motifs which form secondary structures allowing internal initiation of protein synthesis. Internal initiation is not dependent upon the cap-binding protein eIF-4F, and occurs when this protein is inactive (the *picornaviridae* block host protein synthesis by inhibiting eIF-4F as part of their infection strategy). However, no viral proteins are required for IRES function, i.e. initiation is dependent upon other host initiation factors. Internal initiation may also occur for certain cellular transcripts, including *Drosophila Antennapedia* and mouse *fgf2*. The abundance of IRES motifs and their significance in the control of gene expression is unknown; they have been defined in functional terms and appear to share no conserved structural features.

**Global regulation of protein synthesis.** The components of the 'basal apparatus' of protein synthesis (the initiation, elongation and termination factors, and the components of the ribosome) may be regulated and may exercise a global control over protein synthesis. For example, several eukaryotic viruses shut down host protein synthesis by phosphorylating the initiation factor eIF-2, and this strategy may also be used by the cell itself where global repression of protein synthesis is required (e.g. when cells are subjected to heat shock). Phosphorylation of eIF-2 in *S. cerevisiae* allows leaky scanning, where the ribosome can skip weak initiators and bind to strong ones. This lifts translational repression of genes such as *GCN4*, where there are several unproductive AUG codons between the cap and the definitive start codon on the mRNA. When eIF-2 is unphosphorylated, the ribosome attempts to initiate protein synthesis at the first AUG, which is followed by an in-frame termination codon.

**Narrow domain regulation of protein synthesis.** The translation of specific mRNAs can also be regulated individually. Well-characterized examples include the mammalian ferritin mRNA and ribosomal protein L5 mRNA in *E. coli*.

In the ferritin system, translation of ferritin mRNA is blocked when the concentration of intracellular iron is low. The inhibition depends upon **iron-response elements (IREs)** in the transcript, and is mediated by IRE-binding protein, which is inactivated in the presence of iron. The ferritin IRE is present in the 5' untranslated region of the transcript and IRE-BP binding inhibits protein synthesis by preventing ribosome scanning. In the ferritin mRNA, IREs are not associated with AU-rich mRNA instability sites and do not affect mRNA turnover, as is the case for transferrin receptor mRNA (see RNA Processing).

The L5 system represents a feedback control for the synthesis of ribosomal components. Ribosomal protein S8 binds to a stem loop structure formed by residues 588–651 of the 16S rRNA. A similar structure is formed by the 5' untranslated region and the first 30 bases of the coding region of the L5 ribosomal protein mRNA. Protein S8 binds to L5 mRNA (albeit with lower efficiency than does 16S rRNA) and inhibits protein synthesis. Excess 16S rRNA sequesters all the S8 protein and thus derepresses L5 translation, whereas if 16S rRNA is limiting (and there is therefore no need for L5 translation), the excess S8 protein prevents L5 translation.

A final example of specific translational regulation is provided by *antisense RNA* (q.v.). This is found mainly in prokaryote systems (e.g. in the control of Q protein synthesis in bacteriophage  $\lambda$ , in the control of plasmid replication genes, and in the control of transposase synthesis in various transposons, see Viruses, Plasmids, Mobile Genetic Elements), but also in eukaryotes (e.g. in the control of Fgf-2 synthesis in the chick limb bud). More recently, antisense suppression of translation has been used as a strategy to block gene function without *in vitro* or targeted mutagenesis allowing the effects of null phenotypes to be determined without gene disruption (q.v. *transgenic organisms*, *gene therapy*, *cosuppression*).

#### Box 23.1: Measuring the mass of macromolecules

**Relative molecular mass.** The mass of a molecule is often represented as a **relative molecular mass** ( $M_r$ ), which is its mass relative to that of one twelfth of a carbon atom (one **atomic mass unit** or one **dalton**). Thus the mass of macromolecules such as proteins and nucleic acids is often stated in **kilodaltons** (kD, kDa), which are equivalent to 1000 atomic mass units. The  $M_r$  of a molecule can be determined theoretically by adding together the  $M_r$  of each atom, or empirically by sizing against markers of known mass (e.g. by gel electrophoresis).

**Nucleic acids.** Nucleic acids, which are often larger than proteins, are usually measured in terms of the number of nucleotide subunits they contain. The size of single stranded nucleic acids is expressed as **bases (b)** or **nucleotides (nt)** (which are the same), or for larger molecules, **kilobases (kb)**. The size of double-stranded molecules is expressed as **base pairs (bp)**, **kilobase pairs (kbp)** or, for large DNA molecules such as chromosomes, **megabase pairs (Mbp)**. The monomers of nucleic acids are more evenly distributed and more homogenous in weight compared with those of proteins, and thus the relative molecular mass can be derived from length in bases or base pairs with reasonable accuracy.

**Svedberg units.** The mass of macromolecules and cell components can also be estimated from their

**sedimentation coefficient** determined by ultracentrifugation. This is measured in second(s) and reflects the rate of sedimentation in a unit field. Sedimentation coefficients for macromolecules tend to lie within the range  $1 \times 10^{13}$  to  $2 \times 10^{15}$ , with  $1 \times 10^{13}$  sec equivalent to one **Svedberg unit (S)**. Svedberg units are often used to express the mass of ribonucleoprotein particles and RNA molecules such as ribosomal RNAs and ribosome subunits (also q.v. *spliceosome*, *signal recognition particle*).

**Conversions.** Using the sedimentation coefficient,  $M_r$  can be determined as follows:

$$M_r = \frac{RTs}{D(1 - v\rho)}$$

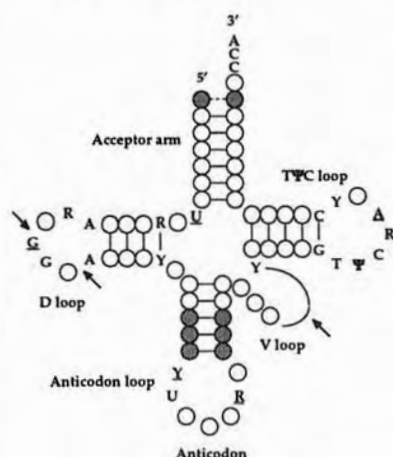
where  $R$  is the gas constant,  $T$  is the temperature,  $s$  is the sedimentation coefficient,  $D$  is the diffusion constant,  $v$  is the partial specific volume and  $\rho$  is the density of solvent. As a rule of thumb, the average  $M_r$  of one base pair of DNA is taken to be 650 Da, and that of one base of RNA 335 Da. These are the average values for nucleotide subunits and include one phosphate group per nucleotide residue. They can be used to estimate lengths in bases or base pairs from  $M_r$  or Svedberg units and convert length measurements into  $M_r$ .

**Box 23.2:** The structure and function of tRNA (also see Nucleic Acid Structure)

**Primary structure.** All tRNAs are of similar length, 75–100 nt, and possess up to 10% modified and hypermodified bases. About 20% of the bases in tRNA are **invariant** or **semi-invariant** (i.e. the type of base — purine or pyrimidine — is conserved) and play an essential role in the adoption of tertiary structure or interact with other components of the translation machinery. More than half of the bases are variable but form intramolecular complementary base pairs. For these bases, the ability to form pairs, rather than the specific base itself is an important determinant of tRNA structure and function. Other variable bases are found in unpaired loops, including the anticodon loop. Some variable bases give each tRNA a specific signature recognized by the *amino acyl-tRNA synthetases* (q.v.). In some tRNAs, the anticodon loop plays a role in this specificity, but in others external residues alone are involved and substitution can switch the identity of the tRNA, causing it to be charged incorrectly and to introduce the incorrect amino acid every time its particular codon arises. In some cases, this identity can be resolved to a single base pair (e.g. tRNA<sup>Ala</sup>). The anticodon alone, however, is responsible for the specificity of translation: if the amino acid on a charged amino acyl-tRNA is chemically modified, the modified amino acid is inserted in the nascent polypeptide. Thus, like immunoglobulins, tRNAs have common effector functions mediated by constant residues and idiosyncratic functions (interactions with amino acyl-tRNA synthetases and interaction with mRNA) mediated by variable residues. The remainder of the bases in tRNA appear to be expendable: some of these occur in the D-loop, but most in the extra arm (see below).

**Secondary structure — the cloverleaf model.** Several regions of hyphenated dyad symmetry in the tRNA primary sequence allow the formation of four double-stranded stems. The ends of the tRNA pair to form the **acceptor stem**, which contains an invariant 3' terminal CCA motif. During charging, amino acids are attached to the terminal adenosine residue of this motif, either to the 2' or 3' hydroxyl group (q.v. *amino acyl-tRNA synthetase*). The internal residues of the tRNA fold to form three major loops: the **TΨC loop**, so-called because of its invariant sequence GTΨCRANYC (Ψ is the modified residue pseudouridine); the **D-loop**, so-called because it often contains the modified residue dihydrouridine; the **anticodon loop**, which carries the three-nucleotide anticodon sequence that pairs with codons in the mRNA. There are 4–5 residues

between the TΨC arm and the anticodon arm in **class 1 tRNAs**. This may expand to form an **extra arm** in **class 2 tRNAs**, the largest secondary structure in the molecule, containing up to 25 bases. The function of the extra arm is unknown.



**Consensus structure of tRNA.** Invariant and semi-invariant bases are identified, variant bases are shown as circles. Underlining indicates a base is usually modified, although the type of modification varies. Ψ is pseudouracil. Arrows indicate the positions where extra bases may be inserted, which in the D-loop include the modified base dihydrouracil. The shaded bases are important in initiator tRNAs, where the anticodon stem carries three consecutive G:C base pairs and the acceptor stem terminal bases are unpaired.

**Tertiary structure.** The tertiary structure of several transfer RNAs have been solved by X-ray crystallography. The molecules adopt a compact, L-shaped conformation with the TΨC and acceptor stems forming one continuous double helix and the D loop and anticodon loop forming another. These helices are held perpendicular to each other, with the TΨC loop and D-loop forming a rigid core and the acceptor stem and anticodon loop forming flexible arms at opposite ends of the molecule.

**Nomenclature.** Each tRNA molecule is named according to its cognate amino acid using the three-letter designation, e.g. tRNA<sup>Ile</sup> is the tRNA carrying isoleucine. Isoaccepting tRNAs are given numbers, e.g. tRNA<sup>Ile</sup><sub>1</sub>, tRNA<sup>Ile</sup><sub>2</sub>. The nomenclature of the initiator tRNAs is different, and is discussed in the main text. Amino acyl-tRNA synthetases are named

according to their substrates, e.g. isoleucyl-tRNA synthetase charges tRNA<sup>Ile</sup> with isoleucine. The name is usually abbreviated to the three-letter designation, i.e. Ile-tRNA synthetase. A charged tRNA is named according to the same convention, i.e. following the example above, isoleucyl-tRNA<sup>Ile</sup> or Ile-tRNA<sup>Ile</sup>.

**Supplementary functions.** As well as their essential

role in translation, tRNAs perform other functions in the cell. Amino acyl-tRNAs act as donors of amino acids in other cellular reactions, and individual tRNAs are known to act as primers for reverse transcription in the replication cycle of certain *retroviruses* (q.v.). tRNA or similar structures have been shown to regulate translation in the *his* operon of *E. coli* and to play an important role in *attenuation* (q.v.).

### Further reading

- Farabaugh, P.J. (1996) Programmed translational frameshifting. *Annu. Rev. Genet.* **30**: 507–528.
- Green, R. and Noller, H.F. (1997) Ribosomes and translation. *Annu. Rev. Biochem.* **66**: 679–716.
- Jacobson, A. and Peltz, S.W. (1996) Interrelationships of the pathways of messenger-RNA decay and translation in eukaryotic cells. *Annu. Rev. Biochem.* **65**: 693–739.
- Mountford, P.S. and Smith, A.G. (1995) Internal ribosome entry sites and dicistronic RNAs in mammalian transgenesis. *Trends Genet.* **11**: 179–184.
- Ruizechevarria, M.J., Czaplinski, K. and Peltz, S.W. (1996) Making sense of nonsense in yeast. *Trends Biochem. Sci.* **21**: 433–438.



## Chapter 24

# Recombinant DNA and Molecular Cloning

### Fundamental concepts and definitions

- **Recombinant DNA** is generated *in vitro* by covalently joining DNA molecules from different sources. The technology associated with the construction and application of recombinant DNA is referred to as **genetic engineering**, **gene splicing** or **gene manipulation**.
- The basis of recombinant DNA technology is a key set of enzymes and techniques which allow DNA to be manipulated and modified precisely. The major enzymes used are listed in *Table 24.1*. The most fundamental techniques include: (i) cutting DNA with sequence-specific bacterial endonucleases (**restriction endonucleases**) to generate defined DNA fragments, and using the enzyme **DNA ligase** to join them; (ii) separating nucleic acids on the basis of size by **gel electrophoresis**; (iii) detecting specific sequences in complex mixtures by **nucleic acid hybridization**; (iv) introducing DNA into cells; and (v) amplification of specific DNA molecules, either by molecular cloning or using the polymerase chain reaction.

- In cell-based **molecular cloning**, the DNA fragment of interest is amplified *in vivo* in a population of proliferating cells. A DNA molecule is unable to replicate if it lacks an *origin of replication* (q.v.) so this is provided by a **cloning vector**, a genetic element derived from a plasmid or virus which is exploited to carry extra DNA (**donor**, **foreign**, **insert** or **passenger DNA**). The donor and vector DNAs are covalently joined, producing a **recombinant vector**. This is introduced into host cells, and replicates as the cells proliferate, producing many copies (clones) of the donor DNA. The recombinant vector is episomal and can therefore be separated from chromosomal DNA on the basis of its unique physico-chemical properties, and this facilitates the recovery of large amounts of cloned donor DNA from bulk cultures.

The alternative technique of *in vitro* DNA amplification using the polymerase chain reaction is quicker and easier than molecular cloning and is extremely sensitive and robust. However, only relatively short sequences can be amplified by this method and the enzymes used tend to be less accurate than those found in cloning hosts, leading to some heterogeneity in PCR products. Additionally, no previous knowledge of the donor DNA sequence is required for molecular cloning, whereas the PCR requires primers designed to anneal at sites flanking a specific DNA target. The PCR is not discussed further in this chapter (see The Polymerase Chain Reaction (PCR)).

- The DNA inserted into a vector for cloning arises from either a primary or secondary source. A **secondary source** is a previously isolated DNA clone, the donor DNA being isolated from within that clone for further manipulation, in a procedure known as **subcloning**. On the other hand, **primary cloning** is the isolation of donor DNA from its original or **primary source**, which is either whole genomic DNA or a population of cDNAs. Although it is occasionally possible to isolate a particular cDNA or genomic DNA fragment directly from a source of low complexity, primary cloning usually requires a **DNA library**, a representative collection of all the DNA fragments from a given source cloned in vectors. The desired clone must be isolated from this library by exploiting some unique property of the donor DNA, e.g. its sequence, or a structural or functional property of the protein it encodes. This process is known as **screening**.
- Once a particular DNA clone has been isolated, it may be exploited in a great number of ways. However, research applications fall into one or more of the following broad categories, which are discussed in more detail below: (i) characterization of gene and genome structure, and of gene expression; (ii) physical gene mapping and positional cloning (see Gene Structure and Mapping); (iii) expression cloning; (iv) functional analysis of genes and their products, and of regulatory elements; (v) *in vitro* mutagenesis; and (vi) gene transfer and transgenesis.

**Table 24.1:** Principle enzymes used to manipulate DNA *in vitro*, with their activities and major applications

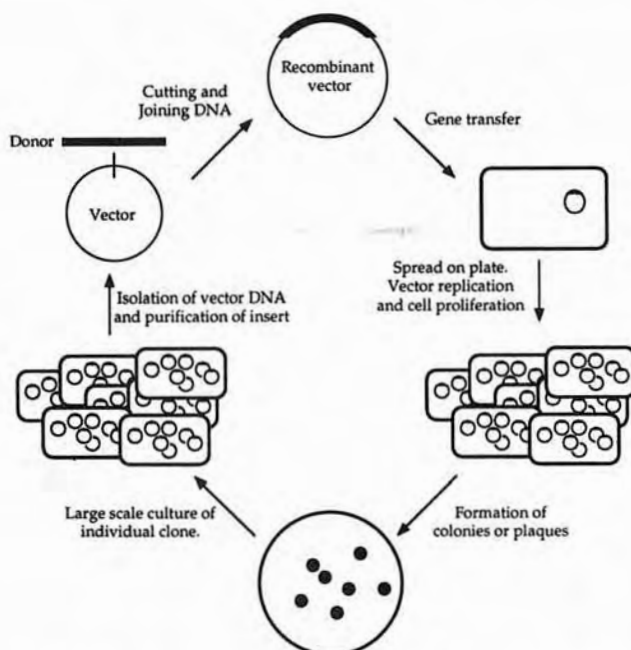
Enzyme	Principle properties and applications
Klenow Fragment (part of <i>E. coli</i> DNA polymerase I; see Replication)	Lacks 5'→3' exonuclease activity Used for general DNA synthesis purposes, e.g. labeling by random priming, end-filling, second-strand cDNA synthesis, primer extension
T4 DNA polymerase	Active 3'→5' exonuclease activity Used for replacement labeling, generating blunt-ended DNA
T7 DNA polymerase	Rapid and highly processive. Modified version, <b>Sequenase</b> , lacks 5'→3' exonuclease activity and is used for DNA sequencing
Thermostable DNA polymerases (e.g. <i>Taq</i> DNA polymerase)	Function at elevated temperatures Used for PCR, and applications where DNA secondary structure is problematical
Reverse transcriptase	RNA-dependent DNA polymerase activity Used for first strand cDNA synthesis, RNA sequencing
RNA polymerase (T7, T3, SP6)	DNA-dependent RNA polymerase with strong promoter-specificity Used for <i>in vitro</i> transcription, RNA labeling
Restriction endonucleases	Sequence specific DNA endonucleases, over 1000 commercially available Used for clone mapping and preparing DNA fragments
S1 nuclease, mung bean nuclease	Single-strand-specific DNA endonuclease. Used for mapping DNA:RNA hybrids, cleaving single-stranded DNA
RNaseA	General RNA endonuclease Used to remove contaminating RNA from DNA
DNase I	General DNA endonuclease Used to introduce nicks into DNA and remove contaminating DNA from RNA
T4 DNA ligase	Generates phosphodiester bonds Used for ligation of cohesive and blunt termini
Calf intestinal alkaline phosphatase	Removes 5' phosphate groups from DNA and RNA Used to block self-ligation and for 5' end-labeling
T4 polynucleotide kinase	Adds 5' phosphate groups to DNA and RNA Used for 5' end-labeling
Terminal deoxynucleotidyl transferase	Template-independent DNA polymerase Used for 3' end-labeling, tailing

## 24.1 Molecular cloning

**Principles of molecular cloning.** Molecular cloning is an *in vivo* technique for producing large quantities of a particular DNA fragment. Essentially, there are four steps in the procedure (Figure 24.1):

- (i) construction of a recombinant vector, which involves cutting, modifying and joining donor and vector DNA *in vitro*;
- (ii) introduction of the recombinant vector into a suitable host cell;
- (iii) selective propagation of cells containing the vector (this is the cloning step);
- (iv) extraction and purification of the cloned DNA.

**Cutting, modifying and joining DNA molecules.** DNA can be cleaved randomly by a number of mechanical, chemical or enzymatic methods, the extent of the treatment governing the average size of fragment produced (Table 24.2). While such methods may be useful for generating random, overlapping fragments of genomic DNA, the only way to generate precise and defined fragments is to use **restriction enzymes** (Box 24.1), bacterial endonucleases which recognize short specific nucleotide sequences termed **restriction sites**. Most class II restriction enzymes recognize restriction sites with dyad symmetry, and cleave phosphodiester bonds at the same position in the half recognition sequence on each DNA strand. A few enzymes cleave at the axis of symmetry and produce



**Figure 24.1:** The principle of molecular cloning. The donor DNA to be cloned is ligated into a vector, generating a recombinant molecule. This is introduced into host cells and the cells are spread on a solid medium. The recombinant vector replicates as the cells proliferate, and clones can be identified as colonies or plaques. Individual clones are picked and placed into liquid culture, from which large quantities of homogenous recombinant DNA can be isolated.

**Table 24.2:** Sources of DNA for cloning and strategies for cutting, modifying and joining DNA molecules to generate recombinant vectors

#### Sources of DNA for cloning:

- Genomic DNA
- cDNA
- Previously isolated clone (subcloning)
- PCR product
- Chemically synthesized oligonucleotide

#### Mechanisms of DNA cleavage:

- Random cleavage by mechanical shearing (e.g. sonication, vortexing), chemical treatment (e.g. acid/alkali hydrolysis), or nonspecific endonucleases → ragged fragments
- Precise cleavage using sequence-specific restriction endonuclease → blunt or sticky fragments

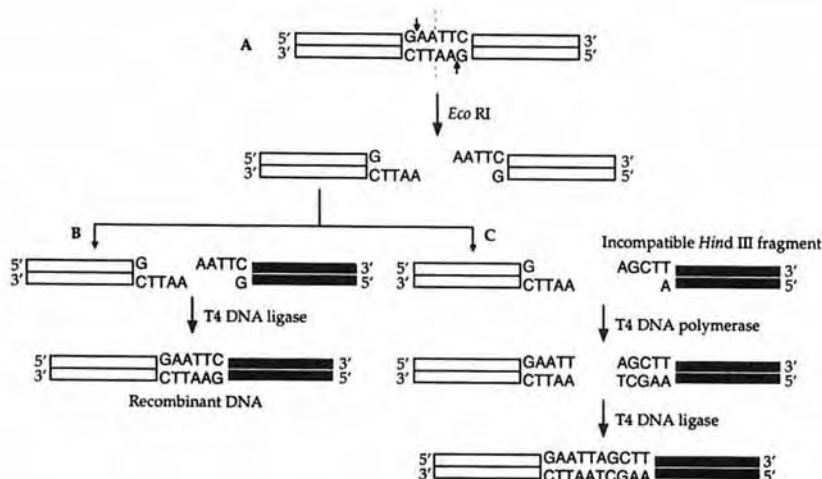
#### End modification strategies:

- End-filling and trimming with T4 DNA polymerase → blunt fragments
- Partial end-filling with Klenow DNA polymerase → alternative sticky fragments
- Addition of linkers followed by restriction enzyme digest → sticky fragments
- Addition of adaptors → sticky fragments

#### Mechanisms of DNA joining:

- Ligation of cohesive termini
- Ligation of blunt termini
- Homopolymer tailing

See also *Table 21.1* for PCR subcloning strategies.



**Figure 24.2:** Cutting, modifying and joining DNA. (a) The sequence GAATTC is the recognition site for the restriction endonuclease *Eco*RI. It has dyad symmetry (i.e. the sequences on each strand read the same in the 5'→3' direction) but the cleavage sites (small arrows) are displaced from the axis of symmetry (dashed line) so that four-nucleotide 5' overhangs are generated, i.e. cohesive termini. (b) These can associate with foreign DNA fragments (black strands) bearing the same termini, facilitating covalent joining of the strands by DNA ligase to generate a recombinant molecule. (c) If the foreign DNA has incompatible ends, one strategy is to blunt both fragments with T4 DNA polymerase and then join them. Other strategies for joining incompatible molecules are discussed in the text.

blunt-ended fragments, but most cleavage sites are displaced from the axis of symmetry, so that the fragments produced have complementary single-stranded overhangs (**cohesive** or **sticky ends**) (Figure 24.2). By hydrogen bonding, these can associate with the ends of other DNA fragments generated by the same enzyme (or one with the same specificity). The base pairing between the overhanging ends holds the terminal residues of each fragment in adjacent positions allowing them to be efficiently joined by **DNA ligase** (Figure 24.2).

The simplest strategy for cloning is thus to cut the donor and vector DNA with the same restriction enzyme and join them with DNA ligase. However, this allows the vector to reclose without an insert, generating a high background of nonrecombinant vectors. There are two procedures which prevent vector self-ligation: (i) the open vector can be dephosphorylated using the enzyme **alkaline phosphatase** — the 5' phosphate groups at each end of the linearized vector are required for ligation and if these are removed, the vector cannot reclose unless 5' phosphate groups are provided by a bridge of donor DNA; (ii) the vector and donor can be prepared using a pair of restriction enzymes which generate incompatible ends — the vector cannot reclose unless the gap is bridged by an insert prepared using the same enzymes. A further advantage to the second strategy is that the orientation of the insert can be predicted (**directional cloning**).

In many cases, compatible ends are not available for ligation (e.g. where donor and vector must be cut with incompatible enzymes, where the donor DNA is randomly sheared and thus has ragged ends, or where the donor is cDNA). Under these circumstances, there are alternative joining strategies which involve DNA end-modification (Table 24.2). One common method is to make all the fragments blunt-ended by filling or trimming overhanging termini. Bacteriophage T4 DNA polymerase performs both functions, and is routinely used to generate blunt fragments. Blunt end ligation is less efficient than sticky end ligation and is nondirectional, but the efficiency is high enough for most subcloning applications. Another strategy is to add linkers or adaptors to the ends of the donor and/or vector DNA. **Linkers** are double-stranded oligonucleotides which, in this case, contain restriction



sites. These can be ligated onto blunt donor DNA, which is then digested with a restriction enzyme to generate cohesive fragments suitable for ligation. **Adaptors** are predigested linkers. Ligation of adaptors to a blunt donor DNA generates a fragment ready for ligation without prior restriction endonuclease treatment; this is useful where the donor DNA has an internal site for the restriction endonuclease being used. It is also possible to join donor and vector noncovalently by **homopolymer tailing**. Blunt or ragged donor DNA is extended (tailed) by adding nucleotides to the 3' end of each strand using the enzyme *terminal deoxynucleotidyl transferase* (q.v.). If only one type of nucleotide is used in the reaction, *homopolymer* tails are formed. A complementary homopolymer tail can be added to the 3' end of each strand of the linearized vector. When the two are mixed, base pairing between the tails forms a relaxed circle. This can be transformed into bacteria and is repaired *in vivo*.

**Cloning vectors.** Molecular cloning involves the amplification of donor DNA by replication in a host cell. However, since donor DNA generally lacks an origin of replication, it must be joined to a suitable replicon to facilitate cloning. Such a replicon is termed a **cloning vector**, and is a derivative of a plasmid, virus or chromosome. Ideal cloning vectors display the following properties:

- (i) they are episomal, i.e. they are not integrated into the host genome, so they can be separated easily from bulk chromosomal DNA;
- (ii) they replicate autonomously, and to a high copy number, thus facilitating the amplification of any foreign DNA they carry;
- (iii) they allow vector-carrying cells to be selected over those lacking vectors;
- (iv) they allow cells carrying recombinant vectors to be selected over those carrying nonrecombinant vectors;
- (v) they are versatile, with conveniently placed unique cloning sites, and ancillary functions enabling them to be used for downstream manipulative applications after cloning.

Naturally occurring plasmids and bacteriophages make poor vectors because they lack cloning versatility. An important initial step in vector development was the construction of plasmids with unique restriction sites (**cloning sites**), allowing the insertion of donor DNA at a specific position without loss of vector sequence. Ideally, these sites are placed within a selectable or visible marker gene to facilitate recombinant selection by insertional inactivation. Most vectors in current use have a **multiple cloning site** or **polylinker**, a cluster of unique restriction sites which provide a convenient position for the introduction of donor DNA prepared using a variety of enzymes.

Alternative vector systems are used for cloning different sized DNA fragments for different purposes (Table 24.3). General purpose plasmid vectors are used for subcloning and downstream manipulations because they are small, easy to handle and extremely versatile. Plasmid vectors carry markers allowing the selection of transformed cells, which form **colonies** on selective media. Bacteriophage  $\lambda$  vectors are used for DNA library construction because they have a greater capacity than plasmids, they are more stable in long-term storage, and plaques are easier to screen than colonies. Cells carrying phage are not selected as such, it is the phage themselves which are selected on the basis of their ability to lyse bacterial cells and thus form **plaques** on bacterial lawns. A number of vectors also exploit features of both phage and plasmids. **Cosmids** are plasmids carrying a  $\lambda$  *cos* site, allowing them to be packaged into phage heads. The basic cosmid is very small, so these vectors can accommodate large donor DNA fragments and are used for genomic library construction and *contig mapping* (q.v.). **Phagemids** are plasmids carrying the bacteriophage M13 (or similar) origin of replication, which allows the plasmid to replicate as a single-stranded DNA phage when appropriate phage functions are supplied *in trans*. These vectors are used to produce single-stranded DNA for applications such as sequencing, *in vitro* mutagenesis and probe synthesis. Finally **phasmids** are composite  $\lambda$ -plasmid vectors, basically  $\lambda$  insertion vectors containing an entire plasmid. Such vectors, e.g.  $\lambda$ ZAP, are extremely versatile, allowing cDNA libraries to be constructed in phage vectors but excised as plasmids for easy downstream manipulations, without subcloning.

**Table 24.3:** Principle features and applications of different cloning vector systems**General purpose cloning vectors****Plasmid vectors**

**Basis:** naturally occurring multicopy relaxed plasmids. Most contemporary plasmid vectors are based on the ColE1 replicon (see Plasmids)

**Introduction into host:** Transformation of naked plasmid DNA, either by chemically assisted transformation, electroporation or lipofection

**Vector selection:** vectors carry dominant selectable marker genes, e.g. antibiotic resistance markers, allowing vector-containing cells to form colonies on selective media

**Recombinant selection:** usually by insertional inactivation of selectable or visible marker

**Size of donor DNA:** 0–20 kbp, although inserts above 5–10 kbp are unstable

**Major applications:** subcloning and downstream manipulation, cDNA cloning and expression

**Comments:** most plasmid vectors are high copy number vectors, which facilitates amplification of the donor DNA. Occasionally, cloned DNA may affect the host cell, e.g. by titrating out host-encoded transcription factors etc., and a number of **low copy number vectors** have been designed to clone sequences which cause dosage sensitivity (**poison sequences**). Some low copy number vectors provide inducible control of replication, so that a burst of plasmid synthesis can occur when cell numbers have increased sufficiently (**runaway vectors**). Many improvements have been made to plasmid vectors to increase their versatility in downstream applications. **Transcription vectors** incorporate bacteriophage promoters flanking the multiple cloning site and allow the production of sense and antisense RNA by *in vitro* transcription. *E. coli* promoters can be included to facilitate protein overexpression, and insertion in-frame with an upstream gene fragment allows the production of fusion proteins and proteins targeted for secretion (q.v. *expression cloning*). **Reporter genes** (q.v.) can be incorporated to allow analysis of regulatory elements. Origins of replication from other species allow replication and maintenance in *E. coli* and a foreign host (q.v. *shuttle vector*). Also q.v. *phagemids*, below

**Phagemid vectors**

**Basis:** plasmid replicon with additional bacteriophage M13 (or similar) origin of replication

**Introduction into the host:** as for plasmids

**Vector selection:** as for plasmids

**Recombinant selection:** as for plasmids

**Size of donor DNA:** as for plasmids

**Major applications:** production of single-stranded DNA

**Comments:** Originally, single-stranded DNA viruses such as M13 and f1 were used to generate single-stranded recombinant DNA for sequencing, probe synthesis, etc. These phages were poor vectors, however, lacking dispensable genes and unstable with even moderately sized inserts. The phage origin allows the plasmid to replicate as a single-stranded DNA phage genome if phage functions are supplied *in trans* by a helper phage. Phagemids thus combine the versatility of plasmids with the ability of the phage to generate ssDNA

**Bacteriophage  $\lambda$  vectors**

**Basis:** Bacteriophage  $\lambda$

**Introduction into host:** transfection of naked phage DNA or *in vitro* packaging of recombinant phage into  $\lambda$  capsids followed by transduction of donor DNA through natural infection route

**Vector selection:** lytic phage form plaques on bacterial lawns

**Recombinant selection:** insertion vectors — by insertional inactivation of visible marker (interruption of *cl* gene prevents lysogeny and therefore phage form clear rather than turbid plaques; contemporary vectors carry  $\Delta lacZ$  gene allowing blue-white selection), Replacement vectors —  $Spi^-$  phenotype (loss of sensitivity to P2 infection) caused by loss of *gam* and *red* loci on central 'stuffer fragment'. Positive selection for large insert size ( $\lambda$  cannot form plaques if genome falls below 75% wild-type size)

**Size of donor DNA:** insertion vectors — 0–10 kbp; replacement vectors 9–23 kbp. Insert size range defined by efficient plaque formation between 75–105% wild-type genome size

**Major applications:** insertion vectors — cDNA cloning and expression libraries; phage display. Replacement vectors — genomic DNA cloning

**Comments:** there are two types of  $\lambda$  vector. **Insertion vectors** have unique restriction sites which allow the cloning of small DNA fragments in addition to the  $\lambda$  genome. **Phasmids** (q.v.) are insertion vectors containing a plasmid. **Replacement vectors** have paired sites defining a central **stuffer fragment** which contains genes

Continued

for lysogeny and recombination, not essential for the lytic cycle. The stuffer can be removed and donor DNA inserted between the arms

### **Cosmid vectors**

*Basis:* plasmid containing bacteriophage  $\lambda$  cos site

*Introduction into host:* *in vitro* packaging of recombinant cosmid into  $\lambda$  capsids followed by transduction through normal infection route

*Vector selection:* as for plasmids

*Recombinant selection:* insertional inactivation of visible marker and positive selection for minimum insert size

*Size of donor DNA:* 30–45 kbp. Insert size range defined by efficient plaque formation between 75–105% wild-type  $\lambda$  genome size

*Major applications:* genomic library construction

*Comments:* cosmids exploit the  $\lambda$  packaging site both for efficient delivery to the host cell and selection for large insert sizes. Cosmids are widely used for genome mapping and contig assembly

### **High capacity cloning vectors**

#### **Bacterial artificial chromosomes (BACs, fosmids)**

*Basis:* *E. coli* F plasmid

*Introduction into the host:* electroporation

*Vector selection:* dominant selectable marker

*Recombinant selection:* size of insert

*Size of donor DNA:* > 300 kbp

*Major applications:* analysis of large genomes

*Comments:* low frequency of rearrangement and chimerism. Vector maintained at one or two copies per cell and thus generates a low yield of donor DNA

#### **P1 vectors and P1 artificial chromosomes (PACs)**

*Basis:* bacteriophage P1

*Introduction into the host:* *in vitro* packaging and transduction

*Vector selection:* dominant selectable marker gene

*Recombinant selection:* various — in one system, positive selection for interruption of a lethal marker is used

*Size of donor DNA:* ~100 kbp

*Major applications:* analysis of large genomes

*Comments:* low frequency of rearrangement and chimerism. Vector maintained at low copy number but can be amplified by inducing bacteriophage P1 lytic cycle

#### **Yeast artificial chromosomes (YACs)**

*Basis:* *S. cerevisiae* centromere, telomeres and autonomously replicating sequences (chromosome origins of replication)

*Introduction into the host:* Transfection of yeast spheroplasts

*Vector selection:* Dominant selectable marker (rescue of auxotrophy)

*Recombinant selection:* Size of insert

*Size of donor DNA:* > 2000 kbp

*Major applications:* Analysis of large genomes, YAC transgenic mice (q.v.)

*Comments:* YACs are the highest capacity cloning vector but suffer from several disadvantages including high frequency spontaneous deletions and clone chimerism. The size of the recombinant vector requires specialized electrophoresis systems for resolution and it is sometimes difficult to separate YACs from endogenous yeast chromosomes. Maintained at low copy number (usually one per cell)

---

Finally, the analysis of large eukaryotic genomes has demanded the development of high capacity vectors — **artificial chromosomes** — for genomic mapping and the structural and functional analysis of large genes and gene complexes. The yeast artificial chromosome is the most developed of these vectors and has the greatest capacity, but shows a high frequency of **clone chimerism** (coligation and maintenance of unlinked donor DNA fragments). More recently, a number of artificial chromosome vectors based on bacterial plasmids have gained popularity. They have a smaller capacity than YACs, but chimeric inserts are much less common.

**DNA transfer to cloning host.** Once a recombinant vector has been constructed *in vitro*, it must be introduced into host cells for cloning. *E. coli* is the major host for general cloning purposes, but this bacterium is not naturally competent to take up DNA from the surrounding medium. An artificial state of competence can be brought about by chemical treatment, such as incubation in the presence of divalent cations. A brief heat shock stimulates DNA uptake, allowing the generation of  $10^7$ – $10^9$  bacterial colonies or  $10^4$ – $10^5$   $\lambda$  plaques per  $\mu\text{g}$  of vector DNA under optimal conditions. The uptake of plasmid DNA by bacteria is termed **transformation** and that of naked phage DNA **transfection**, although the mechanism is in each case identical (transformation and transfection have different meanings when applied to eukaryotic cells, see Table 24.11 below). **Electroporation (electrotransformation)** is an alternative technique where DNA enters cells through pores created by transient high voltage. This is also a highly efficient method for introducing DNA into cells and can generate up to  $10^9$  colonies per  $\mu\text{g}$  vector DNA. It is especially useful for low copy number plasmid vectors such as BACs.

Although these techniques are adequate for subcloning, a higher efficiency of DNA transfer is required for the construction of representative  $\lambda$  libraries. Phage and cosmid vectors can be transferred with high efficiency by **transduction**, which involves first packaging the vector in bacteriophage  $\lambda$  heads (*in vitro* packaging). This is accomplished by mixing recombinant vector DNA with phage head precursors, tails and packaging proteins, and then infecting bacterial cultures with the packaged clones. DNA enters the cells through the normal phage infection route and up to  $10^6$  colonies (cosmids) or plaques ( $\lambda$  vectors) can be generated per  $\mu\text{g}$  vector DNA.

The introduction of DNA into yeast cells is discussed in Box 24.4 and DNA transfer to animal and plant cells is discussed in Table 24.10.

**Vector and recombinant selection.** Neither DNA manipulation nor gene transfer procedures are 100% efficient. Thus, at the start of any cloning experiment, there will be a large population of cells lacking the vector, and of those containing the vector, a moderate proportion will contain nonrecombinant vectors. Both the empty and nonrecombinant cells may proliferate at the expense of the recombinant population, so it is desirable to identify and preferably eliminate such cells.

**Vector selection** is selection for cells carrying a vector — this is usually positive and direct selection, i.e. the vector possesses or confers a property which can be selected. Bacterial cells transformed with plasmid vectors are positively selected for dominant antibiotic resistance markers carried by the plasmid, effectively maintaining a population of plasmid-containing cells. Alternative markers are used in eukaryote systems, e.g. rescue of *auxotrophy* (q.v.) is used to select YACs in yeast, although this requires special auxotrophic host strains. For phage vectors, the phage itself is selected by its ability to form plaques representing areas of lysed cells on a bacterial lawn.

**Recombinant selection** is the selection of cells carrying recombinant vectors over those carrying nonrecombinant vectors. A number of different selection systems are employed depending on the vector type. Recombinant plasmids are usually identified by insertional inactivation of a second marker, either a second antibiotic resistance marker (a process requiring a replica plating selection step) or a visible marker. A current popular strategy is **blue–white selection**. Plasmids carry a non-functional, truncated allele of the *lacZ* gene, which encodes a small N-terminal fragment of the  $\beta$ -galactosidase protein termed the  $\alpha$ -peptide. This can be complemented by an allele encoding the remainder of the polypeptide, which is found in specially modified host strains such as JM101 ( **$\alpha$ -complementation**). Functional  $\beta$ -galactosidase converts the colorless chromogenic substrate X-gal (see Table 24.7) into a blue precipitate. The  $\Delta lacZ$  gene contains an integral polylinker allowing insertional inactivation by donor DNA. Therefore recombinant cells form white colonies and nonrecombinants form blue colonies on the appropriate detection media. A number of direct negative selection plasmids have also been designed where the second marker gene is a conditional lethal, allowing cells containing nonrecombinant vectors to be counterselected under restrictive conditions on the basis of their loss of sensitivity to the marker. However, many of these vectors require specialized host strains and selection systems which are not widely available.



Recombinant  $\lambda$  insertion vectors are usually selected visibly (either by disruption of the *cl* gene — which prevents lysogeny and thus generates clear rather than turbid plaques — or by disruption of an integrated  $\Delta lacZ$  gene as discussed above). Recombinant  $\lambda$  replacement vectors are subject to dual positive selection for their ability to infect bacteria lysogenic for phage P2, and for the size of donor DNA. Wild-type phage have an  $Spi^+$  phenotype because they are sensitive to P2 infection, i.e. they will not superinfect cells already infected with phage P2. This sensitivity is conferred by the *gam* and *red* loci on the stuffer fragment, which is removed and replaced by donor DNA. Hence only recombinant vectors form plaques on P2 lysogens. Additionally,  $\lambda$  only forms infectious particles if the recombinant genome is 75–105% of the wild-type genome size. The upper limit of 105% dictates the maximum insert size in both insertion and replacement vectors. However, while the genome size of nonrecombinant insertion vectors is approximately 100% wild-type size, nonrecombinant replacement vectors lacking the stuffer fragment fall below the 75% lower limit and do not form plaques. The same principle applies to cosmid vectors because they are packaged in phage heads: vector selection is dependent on antibiotic resistance like conventional plasmid vectors, but recombinant selection depends on size of insert-like standard  $\lambda$  vectors.

It is not always necessary to select for recombinant vectors. In simple subcloning experiments, the ligation reaction can be controlled so that a very low background of nonrecombinants is generated and there is a high probability that random colony picking will identify the desired clone. Also, where colonies or plaques are to be assayed by hybridization, nonrecombinants simply fail to hybridize to the probe and can be eliminated from further analysis (q.v. *colony screening, plaque lift*).

**Recovery of cloned DNA.** After transfer of recombinant DNA to host cells, the cells are cultured for a short time to allow recovery and then **plated out** (spread on solid medium) to form colonies or plaques under the appropriate selective regime. Usually, each colony or plaque represents a clone of identical cells or phage and can be **picked into** (removed from the plate and transferred to) liquid medium for a second round of cloning (this time in isolation from other clones). Plating out is the process which *fractionates* the heterogeneous population of recombinant vectors into isolated homogeneous clones, and indicates that on average, each host cell takes up a single recombinant molecule. Plating is a vital step in DNA library screening, where each colony or plaque represents a different part of the genome, or a different cDNA.

The final step in molecular cloning is the recovery of the cloned DNA. Traditional methods involve cell lysis followed by a series of selective precipitation, sedimentation and dialysis steps which are laborious, expensive and time consuming. More recent innovations include the purification of DNA by adsorption to glass beads or to resin in spin columns. A number of convenient kits are available commercially to obtain high yields of pure plasmid or phage DNA from cells and lysates within a few hours.

## 24.2 Strategies for gene isolation

**Isolating DNA fragments from simple and complex sources.** The techniques discussed above allow any DNA sequence to be inserted into a vector and cloned to facilitate further analysis and manipulation. Under circumstances where the source DNA is not complex, or where it is highly enriched for a particular sequence, it may be possible to isolate the desired donor DNA fragment directly, and insert it into a vector for cloning. This is applicable to genomic DNA fragments from small genomes (i.e. those of plasmids, some viruses and animal mitochondria), previously obtained clones, PCR products and cDNAs representing superabundant messages in particular tissues (e.g. the globins, ovalbumin).

In most cases, however, the source of a particular target sequence is complex (e.g. the average human gene is diluted one millionfold by the DNA of the human genome). It is therefore necessary to construct a **DNA library**, a representative collection of all DNA fragments from a particular

**Table 24.4:** Screening strategies to isolate specific genes from cDNA or genomic libraries, depending on the source and the information available

Information available	Screening strategy
<b>Functional cloning — no expression required</b>	
Transcript is superabundant in a particular source tissue	Enrichment cloning — clones isolated randomly from cDNA library and sequenced to confirm product
Partial nucleotide sequence known	Screen library with oligonucleotide probe <sup>a</sup>
Partial clone available (e.g. cDNA used to screen genomic library), or screen based on homology to related cloned sequence	Screen library with cloned fragment or with homologous gene at lower stringency <sup>a</sup>
Partial polypeptide sequence known	Screen library with degenerate oligonucleotides ( <b>guessmers</b> ) <sup>a</sup>
Differential expression between two tissues	Plus and minus screening Enrich library for differentially expressed clones by subtractive hybridization <sup>a</sup> (q.v. <i>difference cloning</i> )
<b>Functional cloning involving cDNA expression</b>	
Mutant available	Screen by complementation of mutant phenotype ( <b>phenotypic rescue</b> )
Antibody available	Screen expression library by immunological detection
Specific properties of product	Screen expression library by specialized technique e.g. southwestern blotting for DNA-binding protein, interaction with other proteins using yeast <i>two hybrid system</i> (q.v.), substrate conversion by enzymes, etc.
<b>Positional cloning</b>	
Structural mutant available	If mutant caused by deletion, clone from genomic subtraction library (q.v. <i>difference cloning</i> )
Mutant caused by transposon insertion	If mutant caused by insertion of transposable element, screen library generated from mutant using transposon sequence as probe and isolate clone by plasmid rescue (q.v. <i>transposon tagging</i> , <i>plasmid rescue</i> )
Position of gene on chromosome	Positional cloning by <i>chromosome walking</i> (q.v.) from linked marker, or chromosome breakpoint. Marker may be a transposable element for <i>enhancer trap vectors</i> (q.v.)
<b>Nonscreening methods</b>	
Genome mapping and sequencing	Systematic analysis of clones spanning entire genome (see Gene Structure and Mapping)
Expressed sequence tags	Industrial scale cloning and characterization of random cDNA clones (q.v. <i>expressed sequence tags</i> )

<sup>a</sup>A PCR-based approach can be used as an alternative (see Polymerase Chain Reaction (PCR)).

source cloned in vectors. There are two major types of library: **genomic libraries**, prepared from total genomic DNA and **cDNA libraries**, prepared by reverse transcription of a population of mRNA molecules. The challenge is then to identify the sequence of interest in the (usually large) background of unwanted sequences by a process termed **screening**. The screening strategy chosen depends upon the information available (Table 24.4).

**Genomic libraries.** The size of a genomic library (i.e. the number of individual clones required to represent the whole genome) depends not only on the average size of donor DNA fragments and the size of the genome, but also on the desired probability that a given region of the genome will be represented. It is not sufficient simply to generate a library upon the principle that every sequence is represented once. Due to differential cloning efficiency and sampling errors during vector

construction and gene transfer to the host, there will be some sequences represented more than once and some not represented at all. It is also desirable to have overlapping fragments, as this facilitates the assembly of clone contigs to generate complete physical maps and gene sequences (see Gene Structure and Mapping). Overlapping fragments of the desired average size can be generated by random shearing of total genomic DNA or by a minimal restriction digest using pairs of 4-cutter restriction enzymes (these have abundant restriction sites whose distribution is essentially random, thus by cutting these sites infrequently, random and overlapping fragments are produced). The latter strategy is convenient because the fragment ends are cohesive and donor DNA can be directly ligated into the vector.

The following formula is used to estimate the total number of clones,  $N$ , required to achieve inclusion of all sequences with probability  $p$ , given that  $n$  is the number of clones theoretically required to span the genome once, i.e. a **genome equivalent**:

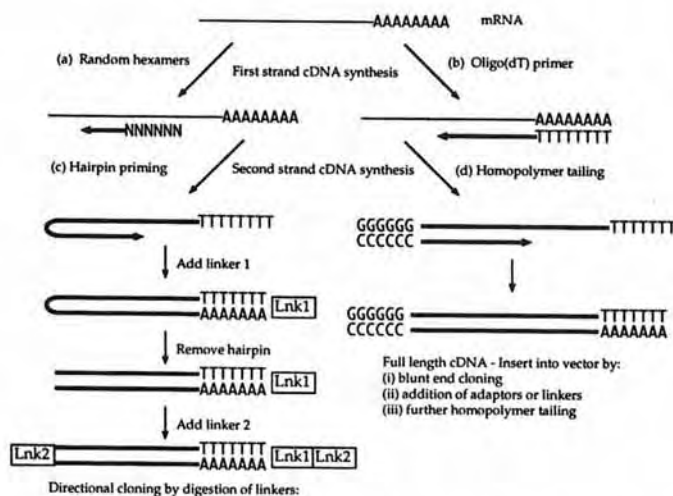
$$N = \frac{\ln(1-p)}{\ln(1-1/n)}$$

This predicts that to achieve a 95% probability of including a given sequence in a library,  $3-4n$  clones must be prepared, and to achieve a 99% probability (the usual gold standard) the library must contain  $4-5n$  clones. An *E. coli* genomic library prepared in  $\lambda$  vectors (average insert size 20 kbp) would thus require 800–1000 clones, whereas the equivalent human library would require more than half a million clones to achieve the same probability of inclusion.

Library size can be reduced by using larger insert sizes in higher capacity vectors for initial screening (i.e. cosmids or artificial chromosomes). Additionally, if the chromosome locus of the desired gene is known, it is possible to create **chromosome-specific genomic libraries** by isolating individual chromosomes and cloning from them. Chromosome separation may be achieved by **fluorescence-activated chromosome sorting (FACS)**, where chromosomes are separated according to their differential ability to bind certain dyes, or by the use of *monochromosomal somatic cell hybrids* (q.v.). It is also possible to generate libraries from specific regions of chromosomes, either by using chromosomes carrying deletions as the source material for library construction, or by **chromosome microdissection**.

**cDNA libraries.** cDNA is **complementary DNA**, i.e. DNA which is complementary to mRNA. cDNA libraries are prepared by reverse transcribing a population of mRNAs and preparing **double-stranded cDNA** clones. cDNA libraries thus differ from genomic libraries in sequence representation, gene sequence architecture and application of screening methods:

- (i) cDNA libraries represent a source of mRNA where particular transcripts will be abundant and others rare. Thus, unlike genomic libraries which theoretically represent all gene sequences equally, cDNA libraries will be comparatively enriched for some sequences and depleted for others. This can be exploited to isolate abundantly represented cDNAs, but on the other hand cDNAs representing rare transcripts can be difficult to isolate. cDNA libraries prepared from different cell types (or different developmental stages, or cells exposed to different treatments) will contain some common sequences and some unique sequences. This can be exploited to isolate differentially expressed genes (q.v. *difference cloning*).
- (ii) cDNA libraries represent only expressed DNA, thus they lack introns, regulatory elements and intergenic DNA. cDNA libraries are therefore of little use for investigating gene structure or regulation, but the clones are generally smaller than genomic clones, and eukaryotic cDNAs can be expressed in bacteria (which cannot splice introns). Splice variants will generate different but partially overlapping cDNA clones.
- (iii) Genomic libraries are screened by hybridization. However, because cDNA can be expressed in bacteria, *expression libraries* (q.v.) can be used for diverse screening strategies, such as immunological screening and screening by complementation (Table 24.4).



**Figure 24.3:** cDNA cloning. First strand cDNA is prepared by reverse transcribing poly(A)<sup>+</sup> RNA using either (a) random hexanucleotide or (b) oligo(dT) primers. Second-strand synthesis may be (c) self-primed or (d) the first strand may be tailed, and a complementary primer used for second-strand synthesis. Self-priming involves some loss of 5' sequence due to nuclease treatment, but directional cloning is possible. DNA is shown as thick lines and RNA as thin lines.

The preparation of a cDNA library (Figure 24.3) typically involves the isolation of poly(A)<sup>+</sup> RNA, which represents the majority of mRNA in the cell (rRNA and tRNA can be isolated directly by fractionation because of their abundance and size homogeneity). **Poly(A)<sup>+</sup> RNA selection** can be achieved by the extraction of polysomes or, more usually, by passing total cellular RNA through a column containing polyuridine or polydeoxythymidine immobilized to the matrix (affinity selection). Purified poly(A)<sup>+</sup> RNA is then reverse transcribed. A common strategy is to prime **first strand cDNA synthesis** with an oligo(dT) primer, although this tends to generate a 3' end bias in the final library because long mRNAs are not fully reverse transcribed. An alternative strategy is to use random hexanucleotides which anneal at arbitrary positions along all RNAs and produce overlapping cDNA fragments which can be assembled into full-length cDNAs. **Second strand cDNA synthesis** may be self-primed, as the first strand has the tendency to loop back to form a hairpin. This can be exploited for directional cDNA cloning (Figure 24.3) but involves some loss of 5' sequence. A more efficient method is to tail the first cDNA strand (e.g. with polycytidine) and then use an oligo(dG) primer to initiate second strand synthesis. Double-stranded cDNAs prepared in this manner may be inserted into vectors either by addition of linkers/adaptors or by homopolymer tailing.

**Screening strategies.** Classical screening strategies require some knowledge of the biochemistry of the gene, either some sequence information, or an exploitable property of the product which will allow the gene to be isolated from an expression library. These approaches are listed in Table 24.4 under the heading **functional cloning** and can be employed in the absence of any idea of the *position* of the gene. The opposite approach is **positional cloning**, the isolation of a gene without any biochemical or functional information but a knowledge of its approximate location on a chromosome map.

Although both functional and positional cloning strategies are useful, if laborious, approaches to gene isolation, they are being superseded by recent developments in factory-style **nonscreening strategies** involving the systematic cloning and characterization of genomic DNA and cDNA. This is demonstrated by the wealth of knowledge emerging from the genome projects and the analysis of human *expressed sequence tags* (q.v.) (see Gene Structure and Mapping).



**Functional cloning.** Functional cloning strategies can be divided into three groups involving: (i) exploitation of *sequence information* — this does not require expression of the cloned gene; (ii) exploitation of *information about gene expression* — this does not require expression of the cloned gene; and (iii) exploitation of *protein function* — this does require expression of the cloned gene, and is hence specific to cDNA libraries.

Traditional sequence-based approaches involve the use of partial sequence information to isolate clones by hybridization (see Box 24.3). A simple example is the use of a cDNA clone to isolate the cognate genomic clone or *vice versa*. Alternatively, clones may be isolated by homology to related but nonidentical sequences which have already been cloned (e.g. another member of the same multigene family, or a similar gene from a different species, or a highly conserved module such as the homeobox). If a polypeptide sequence is known, oligonucleotide probes may be designed to the corresponding nucleotide sequence, although in this case some *degeneracy* (q.v.) must be built into probe design to incorporate all possible codes for the same peptide fragment.

Approaches which use information about gene expression but do not actually require the clone to be expressed exploit differences in the representation of particular cDNA clones, reflecting the abundances of the corresponding RNAs in the source material. Such approaches include the isolation of a cDNA clone because it is superabundant in a particular library (**enrichment cloning**) and the exploitation of differences between libraries or RNA sources to isolate differentially expressed genes (q.v. *difference cloning*).

Screening strategies which rely on cDNA expression exploit the structure or function of the encoded polypeptide. The most straightforward of these procedures is the immunological screening of an expression library, where a labeled antibody is used to identify clones expressing a particular polypeptide. Similarly, a variety of other molecules can be used as probes, including DNA and RNA to detect nucleic acid-binding proteins, and proteins to detect protein-protein interactions. The activity of a protein can also be screened, e.g. by testing clones encoding enzymes for their ability to carry out a specific reaction, or by testing for rescue of a mutant phenotype.

**Positional cloning.** If the approximate position of a gene is known, it may be cloned on that basis with no knowledge of its sequence or biochemical function. Positional cloning starts with the identification of a nearby marker which has already been cloned. The proximity of the marker to the target gene (e.g. an unknown gene responsible for an inherited human disease) may be determined by linkage analysis or low resolution physical mapping (see Gene Structure and Mapping).

In humans and other animals, the first stage in positional cloning is a **chromosome walk**. This involves using the proximal marker as a probe and screening a genomic library for overlapping clones. Positive clones can then be used as probes for subsequent steps until the region to which the gene was mapped has been covered, and a set of contiguous clones has been prepared spanning the region of interest. The direction and progress of the walk can be assessed by *in situ* hybridization to metaphase chromosomes. This technique is simple in principle, but long-range chromosome walks are technically challenging owing to the presence of repetitive DNA or unclonable sequences (which will not be represented in the library). Such obstacles can effectively be bridged by **chromosome jumping**: large fragments of genomic DNA are cloned in cosmid vectors so that DNA sequences originally located many kbp apart in the genome are brought together as the vector closes. The DNA from the closure sites can be subcloned to generate a **jumping library** which allows rapid progress along the chromosome. The effects of repetitive DNA can be controlled to some degree by competitive hybridization to **Cot-1 DNA**, the DNA which anneals quickly in reassociation experiments (q.v. *Cot analysis*) and corresponds to satellite DNA and superabundant transposable elements.

Once the mapped locus has been cloned, there are numerous techniques which can be used to identify **candidate genes** within the region (see Table 12.10). However, as the number of mapped and cloned genes rises, it is becoming increasingly possible to select a group of candidate genes by scanning genome databases for genes already mapped to the region covered by the walk. Once such a

group of candidate genes has been identified, it is then necessary to confirm which one of them is the target gene by showing that affected individuals have mutations and unaffected individuals do not. Some methods for **mutation screening** are listed in Table 24.5.

In plants, the abundance of repetitive DNA makes chromosome walking impossible, but the availability of high-density physical maps in many species based on markers such as *RAPDs* (q.v.) means that markers are often found close enough to genes to be included on the same genomic clone. Since this procedure allows a suitable marker to be used directly as a probe to screen a genomic library, the procedure has been termed **chromosome landing** and is analogous to *transposon tagging* (q.v.).

**Difference cloning.** **Difference cloning** is a functional cloning approach which exploits differences in representation between DNA sources to isolate differentially expressed genes. The majority of genes expressed in any tissue are housekeeping genes and are present in most cDNA libraries. To specifically target the unique genes, an approach termed **differential screening** or **plus and minus screening** involves screening duplicate cDNA libraries with labeled mRNA or cDNA from two different tissues. The labeled mRNA from the library source hybridizes to most of the clones, whereas the labeled mRNA from the second source hybridizes to only a proportion of the clones, those representing the common housekeeping genes. The clones identified by one probe but not the other are differentially expressed and can be isolated for further characterization.

A more efficient route to identifying differentially expressed genes is to generate a library enriched for unique clones. This can be achieved by **subtractive hybridization**, where cDNA from one source is used to remove common RNA from a second source, leaving behind the unique RNA to be used for library construction. The use of biotin-labeled cDNA and streptavidin affinity capture (Box 24.4) has been invaluable in the development of this technique. cDNAs from one source are prepared by incorporating biotin-labeled nucleotides. The cDNAs are denatured, mixed with mRNA from a second source and allowed to hybridize. If an excess of labeled cDNA is used, most of the common RNA will form cDNA:RNA heteroduplexes which can be isolated by affinity to immobilized streptavidin beads along with any cDNA homoduplexes and single strands. If several rounds of subtraction are used (i.e. by adding fresh cDNA), the remaining population of mRNAs is highly enriched for unique transcripts and can be used to construct a **subtracted cDNA library**.

The same general strategy can be used in genomic libraries for the positional cloning of genes localized by the analysis of chromosomal deletions. If genomic DNA can be prepared from normal and mutant (deleted) sources, subtractive hybridization can enrich for genomic clones spanning the deletion. This technique led to the successful cloning of the Duchenne muscular dystrophy (DMD) gene.

Difference cloning can also be carried out using PCR-based techniques (q.v. *representational difference analysis, differential display, suppression PCR*).

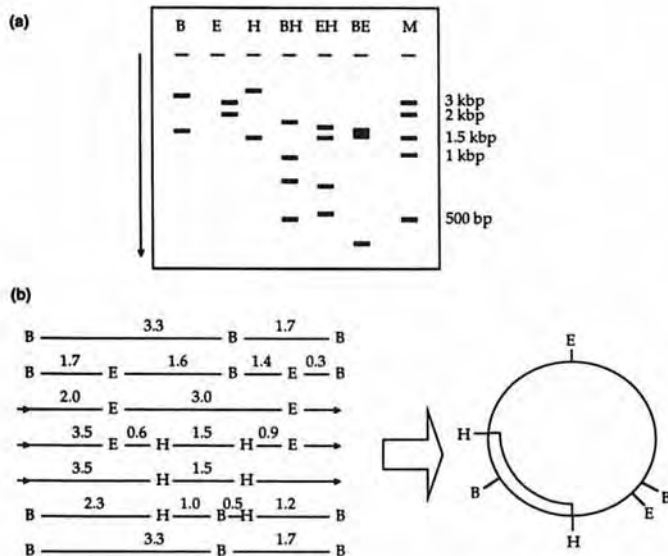
### 24.3 Characterization of cloned DNA

**Restriction mapping.** Once a novel clone has been isolated, the first stage of analysis is the creation of a **restriction map**, a physical clone map showing the relative positions of the restriction sites for a number of different restriction enzymes. The vector is digested with a panel of restriction enzymes and the restriction fragments separated by *gel electrophoresis* (Box 24.2). The position of each restriction site is resolved by determining the sizes of fragments produced by single and double restriction digests, and partial digests (an example is shown in Figure 24.4). As well as simple analysis by electrophoresis, end-labeling the donor DNA can help to identify the borders of the map and individual restriction fragments can be labeled and used to identify overlapping fragments by hybridization to a Southern blot (Box 24.3). The restriction map shows the size of the donor DNA, and identifies sites

**Table 24.5:** Mutation screening for identifying mutations in cloned DNA

Technique	Basis of detection
<i>Techniques for detecting various types of mutation in the whole genome</i>	
Cytogenetic mapping	Identification of chromosome aberrations by cytogenetic analysis. May show correspondence between phenotype and particular rearrangement
Comparative genome hybridization	A form of competitive <i>in situ</i> hybridization which allows the identification of differentially amplified genes. Very useful for detecting amplification mutants causing cancer (see Oncogenes and Cancer)
<i>High resolution techniques for the localization of point mutations and indels to particular DNA fragments</i>	
RFLP analysis	Analysis of cloned DNA for <i>restriction fragment length polymorphisms</i> (q.v.) caused by creation/obliteration of restriction sites (point mutations) or repeat expansion/contraction (indels)
Denaturing gradient gel electrophoresis (DGGE)	Duplex DNA migrates through a gel in which there is an increasing gradient of denaturant until the strands separate. Homoduplexes of different alleles, and heteroduplexes denature under slightly different conditions generating individual bands
Constant denaturant capillary electrophoresis (CDCE)	Homoduplexes of different alleles and heteroduplexes demonstrate different mobilities in a gel when partially denatured under constant denaturing conditions
Single-strand conformational polymorphism (SSCP) analysis	Single-stranded fragments of DNA are amplified by PCR and separated by electrophoresis in a nondenaturing gel. Point mutations affect the tertiary structure adopted by single strands and influences their mobilities. This technique does not precisely localize a mutation but since it can only be applied to short fragments (<300 nt) it can scan for mutations within genes
Restriction endonuclease fingerprinting	More refined SSCP analysis where amplified DNA from different alleles is digested with restriction enzymes to generate different sized fragments containing the mutation, at least one of which is likely to show conformational polymorphism. The position of the mutation can then be localized by creating a <i>restriction map</i> (q.v.)
Protein truncation test	Coupled <i>in vitro</i> transcription and translation of cDNA clone followed by analysis of proteins by electrophoresis. Detects protein truncations caused by deletions/rearrangements and point mutations
<i>Techniques to detect the exact position of point mutations within DNA fragments</i>	
Chemical cleavage of mismatch	Heteroduplex DNA is specifically cleaved at mispaired cytosines (by piperidine following modification by hydroxylamine) and thymines (by osmium tetroxide), and fragments separated by electrophoresis on a sequencing gel to determine the position of the mismatch
Enzyme mismatch cleavage	Cleavage of heteroduplex DNA by T4 endonuclease VII followed by electrophoretic separation and analysis as described above
Mismatch repair enzyme cleavage	Cleavage of heteroduplex DNA by mismatch repair enzymes. For example <i>E. coli</i> MutY detects A:G and A:C mismatches. Other enzymes can be used to detect different sets of mismatched bases. Followed by electrophoretic separation and analysis as described above
Sequence comparison	Sequencing, to directly identify positions of indels and mismatches

Note that while *mutation screening* is used for identifying *unknown* mutations, *mutation detection* (q.v.) is used for identifying known mutations. Many of the high resolution screening techniques exploit heteroduplex DNA, either to locate mutations as mismatches or through their differing mobilities in electrophoretic gels compared to homoduplexes. The origin of heteroduplex DNA reflects the common strategy of amplifying pools of DNA from heterozygotes, by PCR resulting in the random association of wild-type and mutant strands during the thermal cycling.



**Figure 24.4:** The principle of restriction mapping. By digesting e.g. a plasmid with a panel of restriction enzymes, a physical map can be constructed based on restriction sites. By analyzing the size of electrophoretic bands generated by single and double digests (a), the relative positions of different sites can be deduced (b) forming a complete plasmid map. Restriction endonucleases are identified by their initial letters (B = *Bam*HI, E = *Eco*RI and H = *Hind*III; BH, BE and EH are double digests; M is the marker lane, where restriction fragments of known size are run for comparison). Fragment sizes are in kb pairs. Note that where different fragments of similar size are generated, they comigrate on the gel. This can be seen in the *Bam*HI/*Eco*RI double digest.

which may be used for subcloning. The pattern of restriction fragments generated may also be used to identify overlapping clones for contig mapping (q.v. *restriction fragment fingerprinting*).

Standard 6-cutter enzymes such as *Eco*RI and *Hind*III are useful for mapping plasmid and phage clones because the restriction sites occur on average once in every 4 kbp. For larger vectors such as cosmids and artificial chromosomes, *rare cutters* are useful (Box 24.1). This is termed **long range restriction mapping** and requires specialized electrophoresis methods to separate the large fragments generated (Box 24.2).

The ultimate characterization of a cloned DNA is to determine its nucleotide sequence. This is not possible for large clones, which must be subcloned to provide fragments small enough for direct sequencing. Techniques for DNA sequencing and its role in genome analysis projects are considered in Box 12.2).

**Transcript analysis.** Transcript analysis is the investigation of gene structure and expression at the RNA level. It is important to characterize both gene architecture and expression parameters, and the methods used are discussed in Table 24.6. **Northern blotting** is a useful technique which can determine whether or not a gene is expressed in a given tissue, and if so at what level. It also shows the size of a transcript and reveals any alternative products (e.g. splice variants). An important step in gene characterization is to map the transcriptional start and termination (or polyadenylation) sites and intron/exon boundaries. Such precise structural details can be obtained by **nuclease mapping** and **primer extension**, techniques which exploit the structural differences between mRNA and the corresponding genomic DNA (Figure 24.5). One of the most powerful transcript analysis techniques is *in situ* hybridization, which allows the cellular and subcellular characterization of RNA expression patterns (see also Box 24.3).



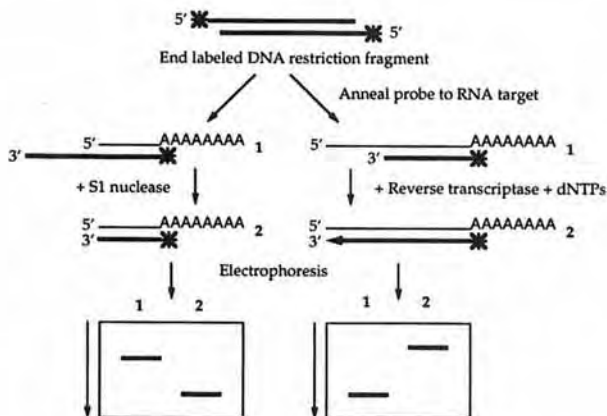
**Table 24.6:** Techniques for the analysis of gene structure and gene expression at the RNA level

Method	Principle, and role in the study of RNA structure and expression
<b>Northern blotting</b>	<p><i>Principle:</i> total RNA or poly(A)<sup>+</sup> RNA is transferred to a filter following electrophoretic separation and is hybridized to a complementary probe</p> <p><i>Structure:</i> shows transcript size and presence of splice variants, but only useful for relatively abundant RNAs</p> <p><i>Expression:</i> provides rough comparison of levels of RNA in different cell lines or tissues</p>
<b>Nuclease mapping</b> (Figure 24.4)	<p><i>Principle:</i> exploits enzymes which cleave single-stranded nucleic acids but not double-stranded nucleic acids (nuclease S1 from <i>Aspergillus oryzae</i>, RNase A). A genomic restriction fragment is labeled and hybridized to RNA. Duplex regions (where the genomic DNA and RNA are colinear) are protected from nuclease activity, but single-stranded tails not represented in the RNA are not</p> <p><i>Structure:</i> comparison of denatured fragments from nuclease-treated and untreated samples reveals size difference which can locate transcribed/nontranscribed sequence boundaries (i.e. transcriptional start site) and intron/exon boundaries</p> <p><i>Expression:</i> RNase protection assay can also be used as a sensitive method to quantitate RNA levels</p>
<b>Primer extension</b> (Figure 24.4)	<p><i>Principle:</i> a primer (an oligonucleotide primer or a restriction fragment of a clone) is annealed near the 5' end of mRNA and extended by reverse transcription</p> <p><i>Structure:</i> comparison of denatured fragments from extended and unextended samples reveals size difference which can locate transcriptional start site. Particularly useful for addressing 5' heterogeneity and start site usage (q.v. <i>multiple start site, alternative promoter usage</i>)</p>
<b>In situ hybridization</b>	<p><i>Principle:</i> a labeled cRNA probe is hybridized to endogenous RNA <i>in situ</i>, i.e. in its normal cellular location</p> <p><i>Expression:</i> invaluable method for the determination of detailed expression patterns at the cellular and subcellular levels. The technique can be applied to tissue sections, isolated cells and, for smaller specimens, in wholemount (requiring nonisotopic probes). <i>In situ</i> hybridization to tissue sections or isolated cells can be used to quantitate expression levels</p>
<b>RT-PCR</b>	<p><i>Principle:</i> reverse transcription of mRNA followed by PCR on resulting cDNA</p> <p><i>Structure:</i> <i>inverse PCR</i> (q.v.) can be used to map intron/exon architecture by amplifying from an exon-specific primer into a neighboring intron. The most important application of PCR to transcript analysis, however, is the isolation of full length rare cDNA clones by <i>RACE</i> (q.v.)</p> <p><i>Expression:</i> quantitative RT-PCR can be used to determine the levels of particular RNA molecules even if they are very rare. Expression patterns can be determined by <i>in situ</i> PCR For further discussion of all these techniques, see The Polymerase Chain Reaction (PCR)</p>

See Figure 24.4 for examples of nuclease mapping and primer extension analysis.

## 24.4 Expression of cloned DNA

**Rationale for the expression of cloned DNA.** While the vectors discussed above have been designed simply to clone DNA fragments and permit their isolation, many plasmid and  $\lambda$ -based vectors are equipped with functional regulatory elements which allow the donor DNA to be expressed. There are several reasons for attempting to express cloned genes, including (i) the production of large quantities of labeled cRNA to use as probes; (ii) the construction of **expression libraries** (cDNA libraries where the donor DNA of each clone is expressed, allowing screening for structural or functional properties of the encoded polypeptide); (iii) the analysis or exploitation of gene function at the protein level; (iv) the commercial production of proteins; (v) the production of antibodies; and (vi) the entrapment of interacting molecules.



**Figure 24.5:** Transcript analysis to position the transcriptional start site of a gene. An end-labeled restriction fragment is denatured and the antisense DNA strand hybridized to RNA. In nuclease protection, the part of the restriction fragment projecting beyond the start site remains single stranded and is degraded by a single-strand-specific nuclease such as S1 nuclease. In primer extension, any single-stranded RNA projecting beyond the restriction fragment 'primer' can be copied by reverse transcription extension of the primer. In each case, comparative electrophoresis of treated and untreated samples reveals a size difference which can approximately locate the transcriptional start site. If the fragments are run next to a sequencing ladder, the exact position of the start site can be mapped. The same techniques can be used to determine any structural differences between RNA and DNA, e.g. intron/exon boundaries. DNA is shown as thick lines and RNA as thin lines.

It is convenient to carry out **expression cloning** in *E. coli* because it is the major cloning host, and by adding appropriate transcription and translation control sequences to a basic cloning vector, it becomes an **expression vector** which facilitates the **overexpression** of donor DNA. *E. coli* plasmid expression vectors carry strong and usually inducible promoters such as the *lac* or *trp* operon promoters, or the T7 late promoter (with the T7 RNA polymerase gene expressed from an inducible promoter). Expression vectors also contain a transcriptional terminator and a ribosome-binding site containing a consensus Shine-Dalgarno sequence. The arrangement of these elements is optimized to produce stable RNA and high yields of protein.

A problem with any expression strategy is the stability of the expression vector. Cells expressing large amounts of a foreign protein (or even an endogenous protein) grow more slowly than wild-type cells, hence there is selection for mutants which have either lost the expression vector altogether or have reduced expression levels. The use of inducible promoters can help to avoid this problem, as large quantities of cells containing the vector can be grown, then expression can be induced and the cells harvested quickly. Other important aspects of expression strategy include the control of vector segregation and the design of expression vectors to limit the likelihood of spontaneously occurring structural mutations. The addition of a *leader peptide* (q.v.) sequence to the vector allows the expressed protein to be secreted from the cell. This may increase stability by removing overexpressed protein from the intracellular environment and it may be necessary for cell survival if the expressed protein is toxic. In any case, secretion facilitates the purification of the protein from cell culture. A vector containing a leader peptide upstream of the expression cloning site is a **secretion vector**.

**Native and fusion proteins.** Depending upon the particular application, cloned genes can be expressed to produce **native proteins** (in their natural state) or **fusion proteins**, where the foreign polypeptide is fused to a vector-encoded polypeptide, e.g. a bacterial leader peptide (as discussed above) or a fragment of a larger protein such as  $\beta$ -galactosidase. Native proteins are

preferred for therapeutic use because fusion proteins can be immunogenic in humans. However, fusion proteins are usually more stable in *E. coli* because they resemble endogenous proteins, whereas native proteins can be targeted for degradation. Fusion proteins also offer other advantages: they can be easily purified (e.g. by affinity to an antibody which recognizes the vector-derived polypeptide) and the function of the vector polypeptide is often retained in the fusion product allowing it to be used as a reporter. It is sometimes possible to cleave the vector-derived polypeptide from the fusion protein using specific proteases, to yield native protein.

For native protein synthesis, donor cDNA is inserted into the expression vector downstream of the transcriptional and translational regulatory elements, and often carries its own initiation codon. For fusion protein synthesis, the open reading frame of the donor DNA must be inserted in-frame with that of vector gene fragment so that they are read contiguously. This can be accomplished by careful choice of subcloning strategy, but it is more convenient to use a shotgun approach where the correct reading frame is achieved by chance. Suitable strategies include the use of three vectors which have cloning sites in different reading frames, using a downstream fusion protein and selecting for readthrough of the insert, or by inserting the donor DNA into the vector using homopolymer tailing, which generates homopolymer joints of random length.

**Disadvantages of expression cloning in *E. coli*.** Although the overexpression of cloned genes in *E. coli* has facilitated the industrial-scale synthesis of many prokaryotic and eukaryotic proteins, there are a number of problems associated with this system. As discussed above, native eukaryotic proteins can be unstable in *E. coli*, and this reduces the protein yield, sometimes to minimal levels. Another problem is that overexpressed foreign proteins can form insoluble **inclusion bodies** which must be broken up by harsh chemical treatments, although the size and number of inclusion bodies can be reduced by growth at lower temperatures, and by overexpression of the *E. coli* *molecular chaperones* (q.v.) GroEL and GroES.

Even if a foreign protein is expressed successfully, however, it may not function in the same manner as its natural, endogenously synthesized counterpart. Whereas some overexpressed eukaryotic proteins are produced in an active form (e.g. granulocyte-colony stimulating factor), many others are not (e.g. epidermal growth factor, EGF), reflecting the absence of correct posttranslational modification. Once a polypeptide is expressed, it must be correctly folded and processed in order to function (see Proteins: Structure, Function and Evolution). However, *E. coli* often fails to fold and process eukaryotic proteins — it does not cleave or glycosylate proteins, nor does it form correct disulfide bonds, probably because it lacks the *molecular chaperones* (q.v.) present in eukaryotic cells. Failure at any of these stages can result in the absence of functional protein even if the polypeptide is efficiently expressed and stable. Active proteins expressed in *E. coli* may reflect their limited structural constraints, e.g. some glycosylated proteins, such as interleukin-3, appear to function equally well in the presence or absence of glycosylation.

**Eukaryote expression hosts.** Where bacterial cells fail to process expressed proteins correctly, eukaryote cells may be used as alternative expression hosts. Expressing proteins in eukaryotic cells also allows the analysis of protein function in a eukaryotic environment, and the use of intracellular signal sequences allows protein targeting to particular organelles (see Proteins: Structure, Function and Evolution). Three types of eukaryotic host are used for protein overexpression: yeast, insect cells and mammalian cells, each with tailored expression vectors and harvesting strategies.

Yeast are often the first recourse for expressing eukaryotic genes which yield poor results in bacteria: as microorganisms, they are genetically amenable, easy to handle and can be used for large-scale cultures. Yeast are useful not only for their ability to express foreign eukaryotic proteins, they also provide a suitable environment for studying eukaryotic protein function (**surrogate genetics**), they provide a system for studying protein-protein interactions (q.v. *two hybrid system*) and they provide large capacity cloning vectors for genome analysis (q.v. *yeast artificial chromosomes*).

There has therefore been a considerable interest in the development of yeast cloning and expression vectors (Box 24.5). Yeast expression vectors are usually based on the  $2\mu$  plasmid (YEps) although YIps may be used for stable integration of DNA into the genome. The expression vectors contain strong, constitutive or inducible yeast transcriptional control sequences (e.g. the phosphoglycerate kinase gene promoter) usually in combination with a  $2\mu$  plasmid terminator and a yeast selectable marker for reversion to *prototrophy* (q.v.). Few yeast proteins are secreted, but most posttranslational modification takes place in the secretory pathway, so yeast secretion vectors are required for the expression of glycosylated mammalian proteins (as well as for toxic proteins). Such vectors carry a yeast 5' leader peptide signal sequence, often that of the secreted mating factor pheromone (q.v. *mating type switching*).

Yeast cells do not perform all the correct posttranslational modifications required for the function of mammalian proteins. Many mammalian proteins function correctly with different glycosylation patterns, but those expressed for therapeutic use are often immunogenic if incorrectly modified. The use of yeast glycosylation mutants can sometimes address this problem. The **baculovirus expression system** provides an alternative where foreign genes are expressed from the strong promoter of the nonessential polyhedrin gene in cultured insect cells. The yield of protein in this system is impressive — up to 100  $\mu$ g of recombinant protein from  $10^9$  infected cells — and the range of posttranslational modifications carried out is well documented. Other advantages of the baculovirus system include its capacity — inserts of up to 20 kbp have been successfully cloned — and more than one foreign gene can be expressed simultaneously allowing the expression and study of protein complexes.

Mammalian cells provide the ultimate expression system for human therapeutic proteins but are the least efficient. Few mammalian expression vectors are maintained episomally, so long-term expression requires constructs that integrate into the genome (q.v. *transient transfection*, *stable transfection*). Mammalian expression constructs may be based on bacterial plasmids, which are transfected into cells, or mammalian viruses, which transduce DNA into the cells. Most viral vectors are recombinant integrating viruses (e.g. retroviruses, adenoviruses) although vectors containing a herpesvirus origin of replication are maintained episomally. Efficient mammalian expression vectors carry a strong, constitutive viral promoter (e.g. the SV40 early promoter and enhancer), or an inducible promoter (e.g. the heat shock promoter, or a modified *E. coli lac* promoter). Heterologous inducible promoters are advantageous because only the foreign gene is induced, and not other endogenous mammalian genes. Expression vectors also carry a strong translation initiation sequence (q.v. *Kozak consensus*), a polyadenylation site, and usually an intron, which may be essential for the efficient expression of some genes. Some of the most efficient mammalian expression vectors exploit *gene amplification* (q.v.). Certain genes are maintained at a high copy number by drug selection, e.g. the dihydrofolate reductase (*DHFR*) gene in cells are exposed to methotrexate (see Box 15.3). Analysis of the amplified regions (**amplicons**) has shown that they contain much more DNA than the amplified gene itself, i.e. *nonselected* DNA is co-amplified. Expression vectors which contain the *DHFR* gene adjacent to the cloning site for the foreign gene are integrated into the genome and selected for amplification. The foreign gene is thus co-amplified with *DHFR* and is strongly expressed.

## 24.5 Analysis of gene regulation

**Reporter genes.** Analogous to the principle of expressing cloned genes by subcloning them in vectors with suitable regulatory elements, cloned regulatory elements can be 'expressed' by subcloning them in vectors providing a suitable gene to regulate, facilitating an assay for gene regulation. In principle any gene will suffice, but it is convenient to use a gene whose expression can be determined easily and quantitatively using a simple assay, and whose product is absent from the host cell. Examples of such **reporter genes** are listed in Table 24.7. Alternative **reporter vectors** are



**Table 24.7:** A comparison of reporter genes

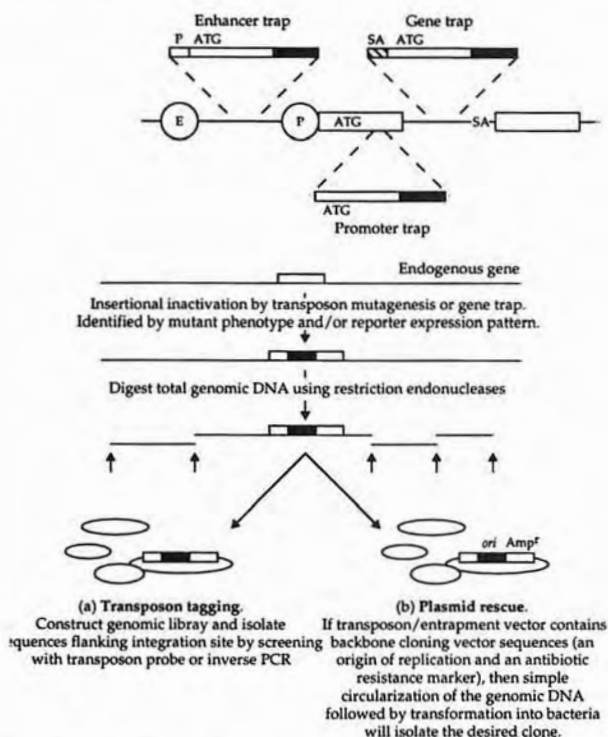
Reporter genes (and products)	Uses <sup>a</sup>
<i>lacZ</i> ( $\beta$ -galactosidase) from <i>E. coli</i>	Widely used reporter system. The enzyme converts chromogenic substrate X-gal into blue precipitate for localization of gene expression. Converts ONPG into soluble yellow product for quantification. Inducible by IPTG (q.v. <i>lac operon</i> )
<i>luc</i> (luciferase) from fireflies	Highly sensitive reporter which generates a bioluminescent product when exposed to substrate luciferin
<i>cat</i> (chloramphenicol acetyltransferase (CAT)) from <i>E. coli</i> Tn9	A useful reporter for <i>in vitro</i> assays but protein gives poor resolution <i>in situ</i> . CAT acetylates chloramphenicol and the extent to acetylation can be determined by thin layer chromatography in a <b>CAT assay</b>
<i>GUS</i> ( $\beta$ -glucuronidase) from <i>E. coli</i>	Generally used reporter in plant systems, converts chromogenic substrate X-gluc into blue precipitate for localization of gene expression
<i>gfp</i> (green fluorescent protein) from jellyfish	A reporter which, because it is autoluminescent, has the distinct advantage that it can be used in living systems

<sup>a</sup>X-gal = 5-bromo-4-chloro-3-indolyl- $\beta$ -D-galactopyranoside; X-gluc = 5-bromo-4-chloro-3-indolyl- $\beta$ -D-glucuronide; ONPG = O-nitrophenyl- $\beta$ -D-galactopyranoside; IPTG = isopropyl- $\beta$ -D-thiogalactopyranoside.

available for the analysis of different regulatory elements: **promoter probe** vectors contain a reporter gene downstream of a polylinker, for the insertion and testing of putative promoter elements; **enhancer probe** vectors contain a reporter gene driven by a minimal promoter and a polylinker for the insertion of putative enhancer elements. There are also **terminator probe** vectors for the analysis of bacterial transcriptional terminators.

The analysis of gene regulation may involve comparing the activity of a series of **reporter constructs** (reporter genes with upstream cloned regulatory elements), in which the regulatory elements have been modified by *in vitro* mutagenesis (q.v.). Such analysis can be carried out by *in vitro* transcription using different cell lysates, by transient transfection of reporter constructs into cells, or for multicellular organisms, by introducing the construct into the germline (a **reporter transgene**, q.v. *transgenic* animals and plants). The *in vitro* transcription and cell line approaches are simple experiments but restricted in the information they provide. A particular disadvantage of the cell line approach is that the transiently episomal vectors do not accurately represent *in vivo* regulatory conditions, lacking the level of control afforded by chromatin structure and distant *cis*-regulatory elements. Additionally, the high copy number of the vector may titrate out transcription factors (q.v. *sequestration*). **Reporter transgenics** provide a more accurate representation of endogenous gene regulation and allow the spatial and temporal effects of modulating regulatory elements to be addressed. However, the reporter transgene may be subject to position and dosage effects, resulting in ectopic or restricted expression patterns and variable expression levels.

**DNA-protein interaction.** By comparing reporter constructs with different mutations, it is possible to localize putative regulatory elements quite accurately. However, to define those elements precisely, the reporter construct approach must be complemented by (i) sequence analysis to identify functional motifs, and (ii) experiments to demonstrate protein-DNA interactions. There are several *in vitro* techniques for the analysis of protein-DNA binding, including the *electrophoretic mobility shift assay*, *DNase I footprinting* and *methylation interference* (for discussion of these techniques, see Nucleic Acid-Binding Proteins). The identification of protein-binding motifs in DNA also provides a direct route to the isolation of clones encoding novel transcription factors. Oligonucleotides corresponding to the putative regulatory elements can be labeled and used to screen an expression library for interacting proteins in a procedure known as a *southwestern screen* (see Nucleic Acid-Binding Proteins).



**Figure 24.6:** Entrapment vectors. (a) There are three types of entrapment vector: the **enhancer trap**, which carries a minimal promoter and responds to endogenous enhancer elements; the **gene trap**, which carries a splice acceptor site and responds to endogenous splice donors following intron insertion; and the **promoter trap**, which carries a simple initiation codon and is activated by insertion into the first exon. (b) Strategies for cloning trapped genes include isolating **tagged** DNA from a genomic library (by hybridization or inverse PCR) or **plasmid rescue**. Enhancer trap vectors may integrate some distance from the endogenous gene influenced by the enhancer, and cloning the endogenous gene involves chromosome walking.

**Entrapment vectors.** Reporter genes are widely used to assay cloned regulatory elements, but they can also be used to characterize the expression patterns of unknown endogenous genes. **Entrapment vectors** are constructs containing a reporter gene which integrate into the genome at random positions and respond to nearby *cis*-acting regulatory elements by generating a reporter expression pattern. There are three types of entrapment vector (Figure 24.6) whose properties are listed in Table 24.8. The technique was pioneered in *Drosophila* using recombinant P-element vectors, but is now applied to other organisms including plants (*Ac-Ds* elements) and mice (recombinant retroviruses or simply randomly integrated DNA; q.v. *random insertion transgenesis*). Advances in trapping technology have facilitated the development of modified gene trap vectors which can target specific classes of gene, e.g. those encoding secreted proteins.

With each type of vector, it is usually possible to directly clone the endogenous gene which activates the trap because it is **tagged** with the vector sequence (Figure 24.6). Transposon tagging (q.v.) is a widely used technique to isolate genes mutated by transposon insertion, and entrapment vectors (many of which are recombinant transposable elements) can be exploited in the same way. The vector is used as a probe to screen a genomic library, and isolated genomic clones containing the vector sequence also carry genomic DNA flanking the site of insertion. This can be used to identify overlapping genomic clones and cognate cDNAs. If the vector also contains backbone sequences from a plasmid cloning vector, simple circularization of digested genomic DNA will yield a plasmid containing flanking genomic DNA. This technique, known as **plasmid rescue**, allows rapid molecular characterization of interesting mutants and reporter expression patterns.

Table 24.8: Entrapment vectors

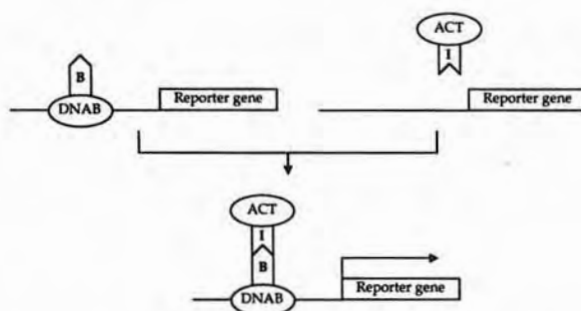
Entrapment vector	Structure	Properties
Enhancer trap	Minimal promoter upstream of reporter gene	Activated by insertion adjacent to endogenous enhancer. Orientation independent and does not necessarily depend on the expression of endogenous target gene(s). Occasionally mutagenic. Often possible to clone endogenous genes from genomic library by chromosome walking from vector insert site (q.v. <i>positional cloning</i> )
Gene trap	Splice acceptor site upstream of reporter gene	Activated by insertion into intron. Dependent on orientation, reading frame and expression of surrounding gene. Usually mutagenic due to exon insertion. Gene may be cloned from genomic or cDNA library by <i>transposon tagging</i> (q.v.)
Promoter trap	Translational initiation sequence upstream of reporter gene	Activated by insertion into an exon. Dependent on orientation, reading frame and expression of surrounding gene. Usually mutagenic due to exon interruption. Gene may be cloned from genomic or cDNA library by <i>transposon tagging</i> (q.v.)

Gene and promoter trap vectors generate fusion proteins and are reading frame dependent as well as orientation dependent. The use of *internal ribosome entry sites* (q.v.) upstream of the reporter gene has alleviated this restriction.

## 24.6 Analysis of proteins and protein-protein interactions

**Antibodies to detect and purify proteins.** The interaction between antibodies and their cognate antigens is highly sensitive and specific. Antibodies can therefore be exploited as probes to study gene expression at the protein level in much the same way that nucleic acid probes are used at the DNA and RNA levels. In principle, antibodies can be used in three ways: (i) to isolate and purify proteins (e.g. affinity chromatography and immunoprecipitation); (ii) to detect proteins on membranes or *in situ* (e.g. western blot, immunohistochemistry and immunological screening of expression libraries) and (iii) to interfere with or modify protein function (q.v. *intrabodies*, *abzymes*). For further discussion of these techniques, see Proteins: Structure, Function and Evolution.

**The yeast two hybrid system.** Traditional methods for analyzing protein-protein interactions include immunoprecipitation (the precipitation of an antigen following antibody binding, and the elution and analysis of any interacting proteins), and screens for *suppressor* and *enhancer* mutations (which respectively ameliorate or augment the phenotype associated with a given mutation, and often identify interacting proteins; see Box 15.4). The yeast two hybrid system is an expression library-based system used to characterize protein-protein interactions. It takes advantage of the modular nature of eukaryotic proteins, specifically that the DNA-binding and activation domains of transcription factors can function independently (q.v. *domain swap*). Furthermore, several transcription factors have been identified whose DNA-binding and activation domains are encoded by separate genes, so that noncovalent association facilitates transcriptional activation (e.g. Oct-1 and the herpes simplex virus VP16 transactivator; see Transcription). The yeast two hybrid system detects proteins which interact with the product of a cloned gene and facilitates the isolation of cDNAs encoding these unknown interacting proteins (Figure 24.7). The known gene is expressed as a fusion protein with a classic DNA-binding domain such as that of GAL4 — this is the **bait** for the **interaction trap**. An expression library is constructed where cDNAs are expressed as fusion proteins with a transcriptional activation domain, and any **interactors**, proteins which interact with the bait, recruit the activation domain to the DNA-binding domain. The final component of the system is a reporter gene which is activated by the hybrid transcription factor. Interactors are thus identified by reporter gene expression, and the corresponding cDNA can be isolated and characterized.



**Figure 24.7:** The principle of the yeast two hybrid system. The aim of the experiment is to identify proteins interacting with the bait protein (B). This is expressed as a fusion protein with a DNA-binding domain (DNAB). An expression library is constructed where clones are expressed as fusion proteins with a transcriptional activation domain (ACT). Cells expressing the bait fusion protein are transformed with expression vectors from the library and reporter constructs containing a binding site for the bait DNA-binding domain. Neither the bait protein nor any putative interactor protein (I) can activate the reporter gene alone. However, any library proteins which interact with the bait activate transcription of the reporter gene allowing positive clones to be identified.

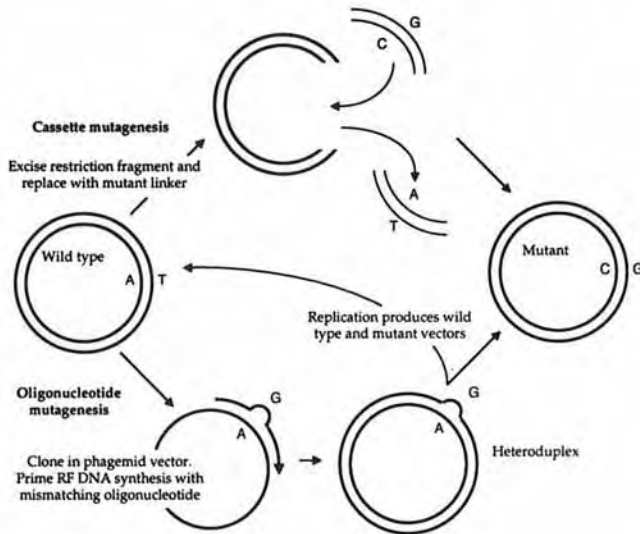
**Phage display.** This technique involves the expression of foreign peptides on the surface of bacteriophage by cloning oligonucleotide cassettes in-frame into phage coat protein genes. The resulting **fusion phage** retain their infectivity and can form plaques in the usual manner. A **phage display library** can be screened in the same way as a conventional expression library, using antibodies or other proteins as probes. The major application of phage display technology is the screening of peptides generated by random mutagenesis for improved affinity or binding specificity. This is useful in protein engineering, for improving the performance of commercially important enzymes, etc., and for the construction of recombinant antibodies. In the latter case, phage display equates to artificial affinity maturation (q.v. *somatic hypermutation*), and circumvents the need for hybridoma cells lines. Interacting phage particles can be purified from a background of up to  $10^8$  noninteracting phage.

## 24.7 *In vitro* mutagenesis

**Altering cloned genes.** Once a DNA molecule has been cloned, *in vitro* mutagenesis techniques can be used to introduce sequence changes. These can be (i) specific mutations which allow functional comparison between mutant and wild-type clones, e.g. to identify critical amino acid residues or regulatory elements, or (ii) random mutations at a defined region which allow the screening of many variants, e.g. to identify those with improved performance (e.g. q.v. *phage display*). Cloned DNA subjected to *in vitro* mutagenesis can be used to replace the homologous sequence in the genome of a cell, or of an entire animal or plant, for functional testing *in vivo* (q.v. *transplacement*, *gene targeting*). This is a reverse approach to classical genetic analysis, in which random mutations are generated and screened to select mutants for the gene or system of interest (q.v. *genetic screen*).

**Introducing specific point mutations.** The introduction of specific base substitutions or small indels at defined sites in a cloned DNA molecule is **site-directed mutagenesis**, and several approaches can be used. **Cassette mutagenesis** involves the excision of a fragment of donor DNA using restriction endonucleases and its replacement by a synthesized oligonucleotide carrying the desired mutation. Alternatively, **oligonucleotide mutagenesis** requires the donor DNA to be cloned in a phagemid vector so that single-stranded DNA is produced. The desired mutation is contained in an oligonucleotide which anneals to the single-stranded vector leaving a mismatch. The oligonucleotide acts





**Figure 24.8:** Site-directed mutagenesis. To introduce a specific point mutation *in vitro*, a small restriction fragment may be removed and replaced by a synthetic linker containing the desired mutation (cassette mutagenesis). If this is not possible, the insert may be cloned in a phagemid vector which produces single-stranded DNA, and replicative form synthesis may be primed with a mismatching primer. This generates a heteroduplex replicative form which produces wild-type and mutant replicative forms in the subsequent round of replication.

as a primer for DNA synthesis so that the *replicative form* (q.v.) of the vector is a heteroduplex, containing one mutant and one wild-type strand (*PCR mutagenesis* (q.v.) employs a similar primer mismatch strategy). Subsequent replication produces homoduplex wild-type and mutant vectors, the latter being identified by hybridization analysis. This is an efficient mechanism for generating mutations, but requires disablement of the host mismatch repair system and the selection of mutant vectors over the wild-type. Numerous **strand selection strategies** facilitate this process, e.g. protection of the mutant strand with restriction enzyme-resistant thionucleotides, allowing the wild-type strand to be digested.

**Systematic mutational analysis.** The functional analysis of for example a regulatory sequence in DNA often begins with the generation of large deletions and substitutions (using restriction endonucleases and exonucleases) to determine the position of essential regions. **Deletion mutagenesis** involves the use of restriction enzymes to remove segments of donor DNA from a clone. **Unidirectional deletions** can be generated with exonucleases such as *E. coli* exonuclease VII, which act upon a particular type of DNA end substrate. Unidirectional deletions allow the creation of a **nested set** of deletions (where one end is common and the other variable). These are useful for analyzing regulatory elements and also for mapping point mutations. **Scanning mutagenesis** is the systematic replacement of each part of a clone to determine its function. In the analysis of regulatory elements, **linker scanning mutagenesis** allows the deletion of small blocks of sequence and replacement by oligonucleotide *linkers* (q.v.) at each position along the clone, thus preserving the spatial relationship of the remaining DNA motifs. In the functional analysis of proteins, **homolog-scanning mutagenesis** involves replacing each segment of the protein with a homologous region from a related protein to identify functionally specific residues. This technique can be extended to systematically replace each amino acid with a different residue, which is achieved by site-directed mutagenesis (q.v. *domain swap*).

**Random mutagenesis.** Where site-directed mutagenesis allows the accurate introduction of specific point mutations, **random mutagenesis** is the most rapid way to analyze large numbers of different mutations for their effects. Random point mutations can be introduced by using an error-prone DNA polymerase (e.g. reverse transcriptase, *Taq* DNA polymerase) under conditions where inaccuracy is favored and in a host strain where all DNA repair systems have been disabled. Random mutations can also be targeted to a specific region of a clone using *degenerate* (q.v.) oligonucleotides for cassette mutagenesis, oligonucleotide mutagenesis or PCR mutagenesis. More recently, techniques have been developed involving the PCR-mediated repair of randomly fragmented genes by *in vitro* homologous recombination. Although not widely used, this technique has been very successful for generating variants of the enzyme  $\beta$ -lactamase with higher activity than the wild-type enzyme.

## 24.8 Transgenesis: gene transfer to animals and plants

**Mechanisms of gene transfer into higher eukaryotic cells.** One of the most important basic techniques of molecular cloning is the introduction of DNA into cells. As discussed above, there are several highly efficient procedures used routinely to introduce DNA into bacterial and yeast cells, facilitating the cloning and expression of animal and plant genes. However, the functional analysis of animal or plant DNA often requires the reintroduction of cloned DNA into the species of origin.

A range of techniques allows the introduction of DNA into eukaryotic cells (Table 24.9). Most involve forcing cells to take up naked DNA (transfection), but some gene transfer techniques are based on the transduction of DNA packaged in viral capsids, and uniquely in plants, there is a procedure based on the conjugal transfer of a bacterial plasmid (also see Gene Transfer in Bacteria).

Most bacterial and yeast vectors are episomal, i.e. maintained outside the genome as autonomous replicons. In contrast, higher eukaryotes tend to lack nuclear plasmids and latent episomally maintained DNA viruses (herpesvirus-based vectors in mammals are an exception). The consequence of this is that DNA cannot be maintained episomally in higher eukaryotes as conveniently as it can in bacteria and yeast. In some cases, this is not important because it is unnecessary for DNA to be stably maintained: many experiments, such as reporter gene assays, can be carried out relatively quickly and **transient transfection**, (where DNA introduced into the cell but is eventually lost by dilution and degradation) is sufficient. **Stable transfection** is required for techniques such as protein overexpression, and in the absence of episomal vectors, this is generally achieved by the stable integration of DNA into the genome. DNA transfected directly into higher eukaryotic cells frequently integrates randomly into the genome, providing a relatively easy mechanism for the genetic transformation of animal and plant cells in culture.

**Transgenic animals and plants.** In multicellular organisms, application of the gene transfer technology discussed above to *totipotent* (q.v.) cells allows the generation of animals and plants where every somatic cell has the same modified genotype. Such organisms are described as **transgenic**<sup>1</sup>, and transmit their newly acquired genetic determinants through the germline as simple Mendelian traits.

The route to transgenic plants is relatively simple because of the naturally occurring and highly efficient Ti plasmid-based gene delivery system and because differentiated plant cells are totipotent and, at least in some species, can be persuaded to regenerate into whole adult plants under the appropriate conditions (Box 24.6). Differentiated animal cells, by contrast, are restricted in their potency, thus transgenic animals must be generated by the manipulation of eggs or cells derived from early

<sup>1</sup>Transgenesis was originally defined as the introduction of an *alien* gene (i.e. one from a different species) into the germline of an animal or plant, but it is convenient to use the term to cover all forms of germline manipulation including the introduction of extra copies of an endogenous gene, targeted disruptions, and the introduction of antisense genes. Some researchers prefer to use the term *targeted* rather than transgenic to describe animals and plants if the germline alteration is subtle (as in the double replacement strategy to generate point mutations) or if it involves genes solely derived from that species.

**Table 24.9:** Methods for introducing DNA into eukaryotic cells

Method	Comments
<b>Cell transfection methods (uptake of naked DNA)</b>	
Polyethylene glycol	Protoplast fusion or direct uptake of DNA in presence of $\text{Ca}^{2+}$ ions and polyethylene glycol. Efficient but labor intensive and, in yeast and plants, requires regeneration of cells from spheroplasts/protoplasts
Chemical transfection	Many methods, e.g. lithium acetate transfection of yeast, calcium phosphate or DEAE-dextran transfection of animal cells. DNA is internalized by endocytosis. These methods are generally efficient for both transient and stable transfection, except DEAE-dextran transfection of animal cells, which is inefficient for stable transfection
Lipofection	DNA complexed with cationic liposomes and taken up by endocytosis. A highly efficient method for transfecting animal cells and yeast and plant spheroplasts/protoplasts. Often works on mammalian cells difficult to transfect using other methods
Electroporation	Naked DNA taken into cells through transient pores created by brief pulses of high voltage. A very efficient method for the transfection of yeast, plant and animal cells
Direct injection	100% efficient though labor-intensive method for introducing DNA into cells. Routinely applied to animal oocytes, eggs and zygotes and to cells which are difficult to transfect by other methods
Microballistics (biolistics)	The use of <b>microprojectiles</b> , tungsten or gold particles coated with DNA, which are fired into cells at high velocity using a <b>gene gun</b> (originally a modified shotgun, but recently more refined apparatus has been developed using high pressure blasts of air or electric discharges). Gives efficient transfection of plant cells without removing cell walls. Can also be used to transfect whole plant and animal tissues
<b>Transduction methods (DNA transferred in viral capsid)</b>	
Recombinant viruses	Viral vectors are used predominantly in mammals. Several viruses are exploited as vectors — herpesviruses are maintained episomally, retroviruses and adenoviruses integrate into genome (also q.v. <i>baculovirus expression system</i> ). Plant viruses have been developed as vectors but not widely exploited due to success of Ti plasmid vectors. Viral promoters are often exploited in expression vectors to drive high level constitutive gene expression
<b>Conjugation methods (DNA transferred to eukaryotic cell by bacterial conjugation)</b>	
Ti vectors	Living plants and plant cells in culture are transformed by T-DNA excised from the Ti plasmid of <i>Agrobacterium tumefaciens</i> . Recombinant T-DNA is a highly efficient gene transfer vector, although only dicotyledonous plants are transformed (Box 24.6)

**Transfection**, when applied to eukaryotic cells, means the uptake of any naked DNA (not just phage DNA as in bacteria). Conversely, **transformation** refers to a change in genotype brought about by the integration of DNA into the genome (in bacteria, transformation is the process of introducing naked, nonviral DNA into the cell and is equivalent to eukaryotic transfection). The terms transformation and transfection tend to be used synonymously in yeast.

embryos. There are three ways to produce transgenic animals (Box 24.7). The most widely used is the simple injection of DNA into the nucleus of an egg, leading to random integration of DNA into the genome. In mice, transgenic technology is more refined and *gene targeting* (q.v.) in embryonic stem (ES) cells allows specific genetic modifications. As well as germline transformation, techniques have also been developed for the transfer of genes to somatic cells in living organisms. The most ambitious application of **somatic transgenesis** is in **gene therapy**, the use of nucleic acid to treat or even correct human diseases (Box 24.8).

Transgenic animals and plants provide an unparalleled resource for the study of gene function. Unlike similar investigations *in vitro* or in cultured cells, the inserted genetic material in a transgenic organism, the **transgene**, can be studied in the context of the whole organism (i.e. in terms of spatial and developmental expression). This can be exploited in the analysis of loss and gain of gene function effects, and the testing of regulatory elements. Particularly important in this context is the use of transgenic mice as models of human disease and cancer. Another important application of transgenic technology is the overexpression of foreign proteins. **Animal pharming** is a euphemistic term to describe the production of commercially valuable proteins — especially drugs — in animal milk, by driving transgenes with promoters from endogenous milk-protein genes. This approach provides an abundant and renewable source of recombinant proteins in a form that is readily purified. Similarly, transgenic plants provide a resource for the production of chemicals, fuels, drugs and novel foods. Transgenic technology can also be used to increase the performance of commercially important animals and plants by adding new traits (e.g. herbicide- and pest-resistance) and improving on existing ones (e.g. the yield or quality of fruit, grain, meat and milk).

**The fate of DNA transferred to eukaryotic cells.** As discussed above, DNA transferred to higher eukaryotic cells is rarely maintained episomally, but is integrated into the genome. Exogenous DNA can interact with the genome in three ways: (i) it can integrate randomly by illegitimate recombination; (ii) it can integrate at a specific site by single cross-over homologous recombination (cointegration or fusion); and (iii) it can replace a fragment of the genome by double cross-over homologous recombination or gene conversion (**transplacement**).

In yeast, homologous recombination is extremely efficient and most genome integration events occur either by single cross-over integration or transplacement. In mammals, illegitimate recombination occurs  $10^5$  times more frequently than homologous recombination because of the highly active end-joining repair system. For this reason, most DNA introduced into animal and plant cells is randomly integrated at sites of where adventitious DNA strand breaks have occurred, and careful strategies are required to select for rare homologous recombination events (Figure 24.9).

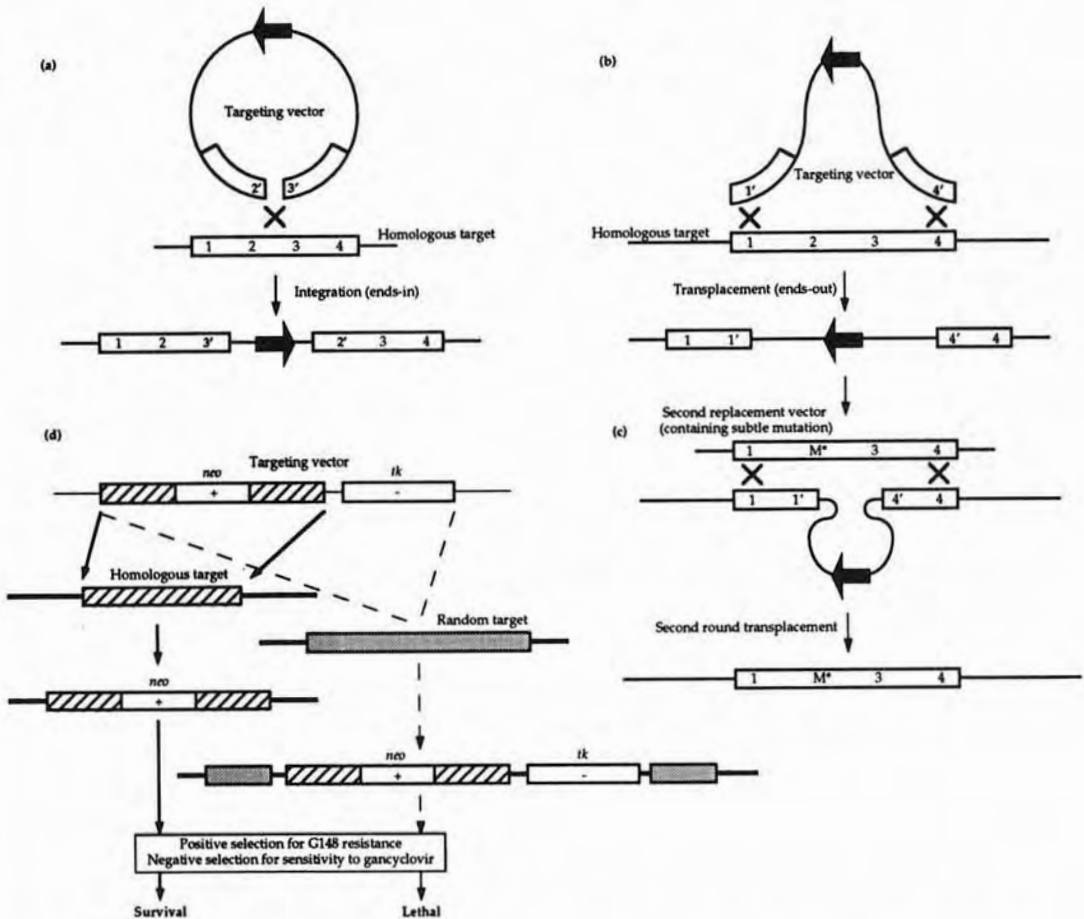
A vector designed specifically to introduce DNA into the genome is described as a **gene delivery vector** or **suicide vector** (the latter because it is meant neither to be maintained nor recovered). Homologous interactions between the host genome and vector occur only if there is a shared region of homology, and such interactions are stimulated by vector linearization because linear DNA initiates recombination by invading the homologous duplex (see Recombination).

**Random integration transgenesis — gain of function effects.** The simplest strategy for introducing germline changes into an animal or plant genome is to allow exogenous DNA to integrate randomly. This is the only available approach for most organisms and is sufficient to study or obtain both the gain of function and dominant loss of function effects of transgene integration.

Gain of function effects result not only from introducing a foreign gene into the genome, but also from increasing the level or scope of expression of an endogenous gene (e.g. by introducing extra copies under the control of a strong viral promoter — **overexpression studies** — or an alternative cell type-specific promoter — **ectopic expression studies**). The expression of foreign genes is often a primary goal in the biotechnology industry whereas the research community tends to focus on the analysis of gene function and regulation. The analysis of reporter gene expression under the control of a normal and modified regulatory elements has been extensively employed in the study gene regulation.

Several problems are associated with random integration approaches to transgenesis: (i) transgenes are often subject to position effects which may cause silencing, varying ectopic or restricted expression patterns and variable expression levels; (ii) transgenes may be subject to dosage effects — the number of integrating copies cannot be controlled; (iii) particularly in plants, but also in mammals, the integration of multiple copies of a DNA sequence into the genome results in epigenetic silencing phenomena (see DNA Methylation and Epigenetic Regulation); (iv) there may be unspecified effects upon endogenous gene expression.





**Figure 24.9:** Gene targeting. Targeted DNA integration may be achieved using one of two classes of targeting vector: (a) an insertion vector (single cross-over site, with ends in) or (b) a transplacement vector (two cross-over sites, with ends out). The insertion vector integrates completely into the genome whereas the transplacement vector replaces part of the genome with the homologous vector sequence. In both cases, large segments of the vector remain in the genome because of the need to use dominant selectable markers (arrow). To achieve subtle targeted mutations, such as a point mutation (shown as M\*), a second round of replacement is therefore necessary (c). Homologous recombination is a rare occurrence in mammalian genomes while random integration is very common. Dual positive-negative selection is therefore employed (d) e.g. the using the *E. coli neo* and herpesvirus *tk* genes. The *neo* gene allows positive selection for resistance to the antibiotic G418, while the *tk* gene confers sensitivity to the thymidine analogue gancyclovir. The *tk* gene is placed outside the homology domain of the targeting vector so that it is only introduced into the genome by random insertion. Therefore only those cells having undergone homologous recombination will be resistant to both gancyclovir and G418.

Recent refinements in transgenic technology have helped to alleviate integration position effects. These reflect (i) the influence of heterologous regulatory elements and chromatin domain structure at the site of integration, and (ii) the fact that transgenes are often small and lack the distant regulatory elements that normally confer position independence upon them. In both animals and plants, it has been found that by flanking the transgene with *boundary elements* (q.v.) position effects can be reduced, perhaps by specifying the transgene as an independent chromatin domain. The more recent development of **YAC transgenics**, mice carrying yeast artificial chromosome transgenes, has allowed

large segments of DNA to be integrated into the mouse genome so that genes stand a good chance of being influenced by all their endogenous regulatory elements. YAC transgenics are invaluable for the study of large genes and long range regulatory phenomena such as the activity of locus control regions and enhancers, chromatin domain effects, parental imprinting and somatic hypermutation.

**Random integration transgenesis — loss of function effects.** The study of loss of function effects often requires targeted disruption of a particular gene followed by breeding to homozygosity (see next section). However, randomly integrated transgenes can also be used to study loss of gene function, although usually only if they are dominant to wild-type (because introducing a recessive mutant allele into the genome will have no effect). Occasionally, a randomly integrating transgene will happen to disrupt an endogenous gene (**insertional inactivation**), in which case a phenotype may be produced in the heterozygote (dominant mutations, usually due to haploinsufficiency) or the homozygote (recessive mutations). This is a crude and accidental form of mutagenesis and is untargeted. However, the principle of random insertional mutagenesis by integration of a transgene can be exploited in large scale genetic screens (q.v. *transposon tagging*, *gene trap*)

Dominant loss of function effects can be generated in several ways: (i) if a mutant allele acts in a dominant negative manner, a randomly integrating transgene will disrupt the function of the wild-type alleles; (ii) selective **cell ablation** can be achieved by expressing a toxic protein such as ricin under the control of a tissue-specific promoter; this can be used to investigate the effects of killing all cells in which a particular gene is expressed; (iii) dominant or partially dominant **gene knock-down** effects can be achieved by expressing antisense RNA or a ribozyme construct targeted to a specific gene — these may inhibit gene function by degrading or inactivating the mRNA (*Box 24.8*); (vi) similarly, gene knockdown at the protein level can be achieved by expressing a recombinant antibody, which binds to and inhibits the activity of a specific protein (see *Box 24.8*).

**Gene targeting by homologous recombination.** **Gene targeting** is a form of *in vivo* site-directed mutagenesis involving homologous recombination between a **targeting vector** containing one allele and an endogenous gene represented by a different allele. Two types of targeting vector are used: **integration vectors (ends-in vectors)** where cleavage within the homology domain stimulates a single cross-over resulting in integration of the entire vector; and **transplacement vectors (ends-out vectors)**, where linearization occurs outside the homology domain and a double cross-over or gene conversion event within the homology domain replaces part of the genome with the homologous region of the vector (*Figure 24.9*).

There are many applications of gene targeting:

- (i) **Gene knockout (targeted disruption)** which can be achieved by inserting a cassette anywhere in the integration vector, or within the homology domain of a transplacement vector (shown as a black arrow in *Figure 24.9*). This cassette is usually a dominant selectable marker, such as the bacterial *neo* gene, which allows selection of targeted cells.
- (ii) **Allele replacement.** One allele is replaced by another, e.g. to investigate the effects of a subtle mutation. This requires two rounds of replacement because the need for selection means that both integration and transplacement vectors leave vector sequence in the genome (*Figure 24.9*).
- (iii) **Gene knock-in**, a novel application where one gene is replaced by another (nonallelic) gene. This is achieved by inserting the incoming gene as a cassette within the homology domain, and is most readily achieved when swapping alternative members of multigene families.
- (iv) **Gene therapy.** In this case, a mutant nonfunctional allele is replaced by a normal allele (see *Box 24.8*).

Gene targeting is an efficient process in yeast and is being actively applied in the systematic project to knockout of all 6300 genes. In mice, gene targeting is carried out by transfection of ES cells and is a very inefficient process compared to random integration. The positive-negative strategy required to select the rare targeted cells is shown in *Figure 24.9*. Notwithstanding these limitations,

the technique has been invaluable in the analysis of gene function, including many genes with important roles in development. However, one unexpected finding from such experiments is the high level of genetic redundancy for developmental genes, with the consequence that many null mutant mice show surprisingly mild phenotypes (q.v. *redundancy*).

**Inducible transgene activity.** An extra level of control can be engineered into transgenic organisms by placing the transgene under inducible control. Two forms of control are commonly used: (i) inducible promoters to switch gene expression on and off; and (ii) inducible site-specific recombination systems which facilitate not only the control of gene expression, but also cell type-specific gene deletions and chromosome rearrangements.

Inducible transgenes have been widely used for overexpression and ectopic expression studies. Heat shock induction is often used in *Drosophila* and plants. In mice, a number of different systems have been tried with varying results. Heterologous regulation systems have been most successful because there is little residual activity and induction is specific to the transgene rather than coactivating endogenous genes. Examples include the *Drosophila* ecdysone promoter, which responds to the *Drosophila* moulting hormone, and the **Tet system**, which responds to tetracycline induction.

**Site-specific recombination** (q.v.) is a form of recombination involving short conserved sequences (recombinators) and proteins which recognize them and catalyze recombination between them (recombinases). The particular arrangement of pairs of recombinator elements can stimulate deletion, inversion or translocation (cointegration) events (see Box 25.4). If a recombinase gene and the recombinator elements recognized by the encoded enzyme are inserted into a transgenic organism, targeted DNA arrangements occur. Targeted deletions can be used for gene knockout (e.g. by deleting the entire gene) or gene reactivation (e.g. by deleting an insert which separates a gene from its promoter). Targeted chromosome rearrangements can also be produced. The power of this technique derives from control of the recombinase. The recombinase gene can be activated in a cell type specific manner or under inductive control. In the first case, this allows cell type specific gene knockouts to be generated, and in the second case, gene knockouts can be generated at any stage in the life cycle of the organism, which is useful e.g. if the gene to be knocked has *pleiotropic* effects (q.v.) but is embryonic lethal. The Cre-lox recombinase system has been widely exploited particularly in transgenic mice and the *S. cerevisiae* 2 $\mu$  plasmid FLP-FRP system has been well-developed in *Drosophila*. The endogenous functions of these systems are discussed in Box 25.4.

#### Box 24.1: Essential tools and techniques I: Restriction endonucleases

Enzyme class	Features
Class I	Three subunit complex with individual recognition, endonuclease and methylase activities Mg <sup>2+</sup> , ATP and S-adenosylmethionine (SAM) required for activity Recognition site is bipartite and cleavage occurs at <i>random</i> site > 1 kb away
Class II	Endonuclease and methylase are separate single-subunit enzymes recognizing the same target sequence Mg <sup>2+</sup> required for activity Recognition site usually shows dyad symmetry — there are several subclasses based on recognition site structure. Cleavage occurs at <i>precise</i> site within or near to recognition site on both strands
Class III	Endonuclease and methylase are separate two-subunit complexes with one subunit in common Mg <sup>2+</sup> and ATP required for activity. SAM stimulatory but not essential Recognition site is unipartite. Cleavage site is <i>variable</i> , about 25 bp downstream of recognition site. Cleavage occurs on one strand only

*Continued*

**Classes of restriction endonucleases.** Restriction endonucleases (restriction enzymes) are bacterial endonucleases which recognize specific nucleotide sequences (restriction sites) typically 4–8 base pairs in length. Their physiological role is *host controlled restriction and modification* (q.v.) hence each endonuclease is associated with a cognate DNA methylase to protect host DNA from **autorestriction**. There are at least three restriction enzyme classes (see table below) but only the class II enzymes are useful for constructing recombinant DNA molecules: they always cleave DNA at precisely the same phosphodiester bond relative to the restriction site and generate defined products — restriction fragments.

**Nomenclature.** Restriction endonucleases are designated by a three letter species identifier in *italic* (e.g. *Eco* = *E. coli*, *Hin* = *H. influenzae*) followed, if necessary, by further letters and/or numbers in roman type to indicate strain type or vector if the restriction phenotype is conferred by a plasmid or a phage (e.g. *EcoRI*, *Hind*, *BamHI*). Finally, if more than one restriction system exists in the same cell it is designated by a roman numeral, e.g. *HindIII*. Where necessary, the endonuclease and cognate methylase of a restriction-modification system can be specified by the prefixes R. and M. respectively, e.g. R.*BamHI*, M.*BamHI*.

**Distribution of class II restriction sites and frequency of cleavage.** The frequency with which a class II restriction endonuclease cleaves DNA is dependent upon the size of its restriction site (the enzymes may be described as 4-cutters, 6-cutters, etc.). The frequency of any motif in random sequence DNA is  $1/4^n$ , where  $n$  is the size of the motif. Hence, 4-cutter enzymes such as *Sau3AI* (GATC) tend to cleave DNA once every ~250 bp, whereas 6-cutters such as *EcoRI* (GAATTC) generate fragments with an average size of 4 kbp and 8-cutters such as *Not I* (GCGGCCGC) generate fragments with an average size of 65 kbp. Fragment sizes also depend on the base composition of the substrate. **Rare cutters** have large recognition sites and/or recognize sequences which are underrepresented in a particular genome. *Not I* is an 8-cutter whose restriction site is GC-rich (and thus slightly underrepresented in mammalian genomes — 40%GC) and contains two CpG motifs (which are heavily depleted in mammalian DNA). The estimated average fragment size for a *Not I* digest of mammalian DNA is thus ~95 kbp. Rare cutters are useful for preparing *cosmid libraries* and *long-range restriction maps* (q.v.) (also q.v. *intron-encoded endonucleases*, *HO endonuclease*).

**Properties of class II restriction sites and restriction fragments.** Class II restriction sites generally show dyad symmetry. If cleavage occurs at the axis of symmetry, **blunt** or **flush ends** are generated. However, if the cleavage positions are not directly opposite each other, a staggered break is generated, producing either 5' or 3' overhanging termini (**sticky** or **cohesive ends**). Generally, restriction fragments produced by the same enzyme are compatible and those produced by different enzymes are incompatible. Some exceptions are discussed below.

*The same restriction endonuclease does not always generate compatible fragments.* Restriction sites may be **specific**, in which case the nucleotide sequence is invariable and all ends generated by the endonuclease are **compatible** (e.g. *HindIII* always cuts at the sequence AAGCTT). Other sites contain one or more **ambiguous** nucleotides, which increases the frequency of the sequence in random DNA but means that ends generated by the enzyme are not always compatible (e.g. *HindII* cuts at the sequence GTYRAC, and produces four different types of sticky ends). Restriction sites are **unipartite** if the recognition sequence is continuous or **bipartite** if it shows hyphenated dyad symmetry (e.g. *EcoNI* cuts at the sequence CCTNNNNNAGG where N is any nucleotide). Cleavage at a bipartite site does not generate universally compatible fragments because of the arbitrary nature of the central residues. Under suboptimal conditions, the specificity of some restriction endonucleases can be reduced so that only part of the normal recognition site is recognized. This is known as **star activity** (e.g. at suboptimal pH, *EcoRI*, which usually recognizes the site GAATTC will recognize only the internal AATT sequence). The enzyme *BcgI* is unique in that it cleaves the DNA twice on each strand, generating a tiny fragment containing the restriction site. *TaqII* is unique in that it recognizes two unrelated sites.

*Different restriction endonucleases may generate compatible fragments.* Restriction endonucleases which recognize different sites can sometimes generate compatible sticky ends. This occurs if one enzyme recognizes a site which is embedded in the larger site of another, a **nested site**. *BamHI* recognizes the sequence GGATCC and *Sau3AI* recognizes the internal tetranucleotide GATC; both generate GATC 5' overhangs which are compatible. Joining, however, generates a **hybrid site** which may be cleaved by only one of the original enzymes or both, or in some cases neither (in the example *Sau3AI* cleaves the *BamHI/Sau3AI* hybrid site, but *BamHI* cleavage depends on the flanking residues). Restriction enzymes from different sources may recognise the same restriction site. Such enzymes



are termed **isoschizomers** if they cleave at the same position and **neoschizomers** (or **heteroschizomers**) if they cleave at a different positions. *SmaI* and *XmaI* both recognize the hexanucleotide site CCCGGG. However, whereas *SmaI* cleaves at the axis of symmetry and generates blunt fragments, *XmaI* cleaves between the first and second cytosine residues and generates CCGG 5' overhangs.

**Methylation sensitivity.** Every restriction endonuclease has a **cognate methylase** which modifies restriction sites in the host genome by methylation and prevents autorestriction (q.v. *host restriction and modification*). Thus, all restriction endonucleases are to some degree methylation sensitive. Some

restriction enzymes however, due to the nature of their restriction sites, are also sensitive to genome-wide methylation such as Dam and Dcm methylation in the *E. coli* genome, and the methylation of CG or CNG motifs in eukaryote genomes (see DNA Methylation and Epigenetic Regulation). The availability of isoschizomers differing in methylation sensitivity (**heterohypermomers**) is useful for mapping methylated DNA. For instance, both *HpaII* and *MspI* recognize the sequence CCGG but only the former is sensitive to methylation of the internal cytosine. These enzymes can thus be used to determine the positions of methylated CpG motifs in higher eukaryotic genomes (q.v. *HTF island*).

#### Box 24.2: Essential tools and techniques II: Gel electrophoresis

**Gel electrophoresis.** **Electrophoresis** is the separation of molecules in an electric field on the basis of their charge and size. **Gel electrophoresis** is the standard method used to resolve mixtures of large molecules (i.e. proteins and nucleic acids) because there is no convection in gels, allowing individual fractions to form sharply defined bands. Samples are loaded in a narrow zone at one end of the gel, defined by wells formed during gel casting. An electric field is then applied across the gel and the samples move out of the wells at different velocities according to size and charge. Since all nucleic acids have the same negative charge on the phosphate backbone, their mobility is determined only by size and shape. Proteins have different charges and are separated according to both charge and size, but by denaturing proteins in the presence of the detergent sodium dodecyl sulfate, the charges are equalized allowing separation by molecular weight (q.v. *western blot*).

**Standard gel electrophoresis for nucleic acids.** DNA and RNA molecules ranging from oligonucleotides to 20 kbp restriction fragments can be resolved in standard electrophoresis gels. Two types are used: horizontal **agarose gels** for the analysis and preparation of fragments between 100 bp and 20 kbp in size with moderate resolution, and vertical **polyacrylamide gels** for the analysis and preparation of small molecules with single nucleotide reso-

lution (required e.g. for DNA sequencing). In each case the average pore size of the gel can be altered by changing its concentration, and different concentrations can be used to resolve different size ranges of nucleic acids. Nucleic acids in agarose gels are usually detected by staining with the intercalating dye ethidium bromide which fluoresces under UV light. Bands in polyacrylamide gels are usually detected by autoradiography, although silver staining can also be used.

**Adaptations for large DNA molecules.** Nucleic acids change conformation as they move through gels, alternating between extended and compact forms. Their velocity depends upon the relationship between the pore size of the gel and the globular size of the nucleic acids in their compact form, with larger molecules moving more slowly. Once a critical size has been reached, however, the compact molecule is too large to fit through any of the pores and can move only as an extended molecule, a process termed **reptation**. At this point, the mobility of DNA becomes independent of size, resulting in the comigration of all large molecules. To fractionate large DNA molecules such as YACs and long-range restriction fragments, electrophoresis is carried out with a pulsed electric field. The periodic field causes the DNA molecule to reorient; longer molecules take longer to realign than shorter ones, so delaying their progress through the gel and allowing them to

be resolved. DNA molecules up to 200 Mbp in size have been separated by various pulsed field-based methods (summarized below).

**2-D electrophoresis.** Electrophoresis in two dimensions exploits different properties of molecules in each dimension and allows finer resolution. 2-D protein electrophoresis involves isoelectric focusing in the first dimension (separation on the basis of charge in a pH gradient) followed by addition of SDS

and separation in the second dimension primarily on the basis of molecular weight. 2-D electrophoresis of DNA allows separation of molecules with the same size but different conformations (e.g. topoisomers, conformational isomers, replication intermediates). 2-D DNA electrophoresis involves separation in the first dimension on the basis of size, followed by the addition of ethidium bromide to induce conformational changes allowing resolution of structural isomers in the second dimension.

Method (GE = gel electrophoresis)	Brief description
<i>Constant field orientation methods</i>	
Pulsed field (PFGE)	Field applied in short pulses; resolution of molecules < 400 kbp
Field inversion (FIGE)	Field pulsed and alternates in polarity; resolution of < 800 kbp
<i>Variable field orientation methods</i>	
Pulsed field gradient (PFGGE)	Field pulsed and alternates orthogonally; resolution of < 2 Mbp
Orthogonal field alternation (OFAGE)	As above, but alternate fields at 45° instead of 90°, which improves interpretation of band mobilities
Transverse alternating field (TAFGE)	As PFGGE but orthogonal field runs transversely through gel
Contour clamped homogenous electric fields (CHEF) and programmed autonomously controlled electrodes (PACE)	Gel surrounded by multiple electrodes arranged in a polygonal pathway; resolution < 7Mbp

#### Box 24.3: Essential tools and techniques III: Nucleic acid hybridization

**Nucleic acid hybridization.** Complementary base pairing between single-stranded nucleic acids underlies some of the most important biological processes: replication, transcription, protein synthesis and its regulation, RNA splicing, recombination and DNA repair. **Nucleic acid hybridization** describes a range of techniques which exploit the ability of double-stranded nucleic acids to undergo **denaturation** or **melting** (separation into single strands) and for complementary single strands to spontaneously **anneal** (form a duplex). Duplex DNA can be denatured and the same strands can then **reanneal** or **renature** to form **homoduplexes**. Alternatively, single strands can anneal to alternative complementary partners, such as a labeled nucleic acid probe to form a **hybrid duplex** or **heteroduplex**. The power of the technique is that a labeled nucleic acid probe can detect a complementary molecule in a complex mixture, with great specificity and sensitivity. Hybridization can occur between DNA and DNA, DNA and RNA or RNA and RNA, and

may be intramolecular or intermolecular. Hybridization can occur between nucleic acids in solution, or where one is in solution and the other immobilized (either on a solid support or fixed *in situ* in a cell).

**Hybridization parameters.** The stability of a duplex nucleic acid is dependent upon both intrinsic and extrinsic factors. Intrinsic properties influencing duplex stability reflect the number of hydrogen bonds holding two single strands together, and include the length of the duplex, its GC content and the degree of mismatch between the complementary partners. The shorter the duplex, the lower the GC content and the more mismatches there are, the fewer hydrogen bonds hold the two strands together, and the easier they are to denature. Extrinsic properties influencing duplex stability reflect the presence of environmental factors which interfere with hydrogen bonds. Increasing temperature causes the disruption of hydrogen bonds, thus duplexes being to melt as temperature increases

*Continued*

(**thermal melting**). The chemical environment is also important:  $\text{Na}^+$  ions increase the stability of the duplex whereas destabilizing agents such as formamide disrupt hydrogen bonds.

The intrinsic properties of a given duplex are constant, so the ability of a given duplex to be maintained can be controlled by modulating the extrinsic conditions, which are collectively defined as **stringency**. The intrinsic stability of a duplex can be measured by determining its **melting temperature** ( $T_m$ ) in a constant chemical environment. The denaturation of double-stranded nucleic acids causes a shift in the absorbency of UV light at 260 nm wavelength, a **hypochromic effect** which can be assayed by measuring optical density ( $\text{OD}_{260}$ ).  $T_m$  is defined as the temperature corresponding to 50% denaturation, i.e. where the  $\text{OD}_{260}$  is midway between the value expected for double-stranded DNA and single-stranded DNA. The  $T_m$  of perfectly complementary duplexes of various compositions can be calculated as shown below. The  $T_m$  falls by  $1^\circ\text{C}$  for each 1% of mismatch, and  $0.6^\circ\text{C}$  for each 1% of formamide in the hybridization solution.

Nucleic acid hybridization experiments can therefore be used to determine the complementarity between two nucleic acids by establishing the  $T_m$ . Conversely, the  $T_m$  can be used to direct hybridization at precise stringency, allowing the hybridization of some molecules and not others. Under some circumstances, it may be desirable to detect only fully complementary sequences, in which case high stringency conditions are used. In other cases, it may be desirable to detect fully complementary sequences and related sequences, in which case lower stringency conditions can be chosen to detect a particular degree of complementarity.

DNA:	$T_m = 81.5 + 16.6(\log_{10}[\text{Na}^+]) + 0.41(\% \text{GC}) - 500/\text{length}$
RNA; RNA-DNA:	$T_m = 79.8 + 18.5(\log_{10}[\text{Na}^+]) + 0.58(\% \text{GC}) + 11.8 (\% \text{GC})^2 - 820/\text{length}$
Oligonucleotides:	$T_m = 2(\text{no. of AT pairs}) + 4(\text{no. of GC pairs})$

**Solution hybridization.** The hybridization of two nucleic acids mixed in solution allows the investigation of sequence complexity, genome organization and gene structure. In the past, Cot analysis of genome complexity and gene distribution, and Rot analysis of transcript abundance and expression parameters were major applications of solution hybridization (see Box 12.1), but the advent of genome mapping and sequencing projects has rendered this type of experiment obsolete. However, any molecular reaction which involves the annealing

of single-stranded nucleic acids in solution (including primers for PCR, *in vitro* mutagenesis, primer extension, cDNA synthesis and random priming; and in techniques such as subtractive hybridization, nuclease protection and homopolymer tailing) is taking advantage of solution hybridization.

**Simple filter hybridization.** Filter or membrane hybridization improves the detection of hybridized molecules by immobilizing the denatured target nucleic acid on a solid support. The transfer of nucleic acids onto such a support, which is often a nitrocellulose filter or a nylon membrane, is termed **blotting**. The simplest form of blotting is when the denatured sample is placed directly onto the membrane (a **dot blot**). Alternatively, the target can be applied through a slot, which allows the area of the filter covered by the target to be defined (**slot blot**). Once transferred, the nucleic acid is immobilized on the membrane. This is often achieved by baking or cross-linking under UV light, although contemporary charged nylon filters bind nucleic acids spontaneously. The membrane is then incubated in a hybridization solution containing the probe and hybridisation is carried out for several hours. The filter is then washed and the probe detected (q.v. *nucleic acid probes* for discussion of probe synthesis and detection). This is a rapid diagnostic technique which allows the presence or absence of particular sequences to be confirmed and quantification of the target sequence.

**Southern and northern hybridization.** A more sophisticated approach involves the separation of DNA fragments or RNA by electrophoresis in a gel before blotting. The capillary transfer of electrophoretically fractionated DNA from a gel to a solid support was first carried out by Edward Southern and is called a **Southern blot**. By extension, a similar technique for the immobilization of electrophoretically fractionated RNA is a **northern blot**\*. The principle behind these techniques is that the position of DNA fragments or RNA molecules on the filter represents their positions in the gel which reflect size fractionation.

Southern blots have many applications, and these are divided into two groups: (i) simple Southern hybridization and (ii) genomic Southern hybridization. Simple Southern blots are used to complement restriction mapping studies of cloned DNA, to identify overlapping fragments and assemble clone contigs. Genomic Southern blots involve the digestion of whole genomic DNA and its fractionation by electrophoresis. For most genomes, digestion with standard six-cutter enzymes generates millions of fragments of varying lengths which produce an

unresolvable smear on an electrophoretic gel, but hybridization can identify and characterize individual fragments. One major application of genomic Southern is to identify structural differences between genomes, through the alteration of restriction fragment sizes (restriction fragment length polymorphisms, RFLPs). Many pathogenic mutations in humans can be identified by Southern blotting in this way. Point mutations can create or abolish restriction sites and therefore alter the pattern of bands observed. Large deletions can remove two consecutive restriction sites and hence delete an entire restriction fragment (q.v. *loss of heterozygosity*). Increases in restriction fragment sizes are also seen when DNA has integrated into the genome, e.g. through the insertion of a transposable element or of foreign DNA in transgenic mice. The analysis of RFLPs in hypervariable DNA allows the characterization of microsatellite polymorphism (q.v. *DNA typing*). A second major application of genomic Southern is to study families of related DNA sequences. A probe may identify not only its cognate target, but also other targets which are unknown, and the number of identified targets may increase as stringency is reduced. The same stringency conditions can be used to screen DNA libraries in an attempt to isolate the related clones representing novel members of a multigene family. The same technique may be used to identify related sequences between species (q.v. *zoo blot*).

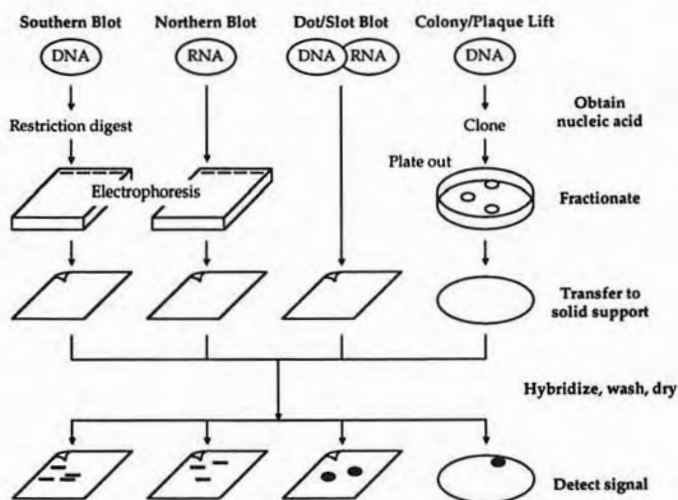
**Allele-specific hybridization.** A major application for DNA blots (Southern blots and dot blots) is **allele-specific hybridization** in the analysis of human disease loci. Oligonucleotide probes are exquisitely sensitive to base mismatches, and hybridization conditions can be controlled so that a

single mismatch results in hybridization failure. This can be exploited to detect specific alleles generated by point mutations (**mutation detection**, c.f. *mutation screening*). Similar PCR-based techniques involve primers which likewise fail to hybridize at single base mismatches (q.v. *allele-specific PCR*, *oligonucleotide ligation assay*, *ligase chain reaction*).

**Reverse hybridization.** Classic Southern and northern hybridization involves using a simple homogeneous probe to screen a complex mixture of immobilized target molecules. **Reverse hybridization (reverse Southern, cDNA Southern)** involves the opposite approach of immobilizing the cloned DNA and hybridizing to it a complex probe mixture such as labeled whole RNA or cDNA. This technique is useful for the rapid, high throughput expression studies where multiple clones are tested simultaneously, e.g. to confirm that each cloned gene is expressed in a given tissue without performing many individual hybridizations.

\*Southern blotting is named after its inventor and should always be used with an initial capital letter, but northern blotting (of RNA) and western blotting (of protein) were named by analogy and should not. There have been several attempts to popularize the *eastern blot*, most recently as a technique for separating and immobilizing lipids, but the term is not widely used. There are **southwestern** and **northwestern blots**, which are modified western blots in which the probe is a labeled nucleic acid, used to detect nucleic acid-binding proteins (q.v. *western blot*).

**Colony blots and plaque lifts.** Another example of nucleic acid blotting being used to precisely reproduce a pattern is **colony blotting** or **plaque lifting**.



Continued



In these techniques, which are used to screen DNA libraries or analyze the results of cloning experiments, a nitrocellulose or nylon membrane is laid over bacterial colonies or phage plaques on a plate in order to transfer some cells/phage onto the support. The DNA is then denatured and immobilized, and screened for a particular donor DNA by hybridization. Positive signals can be referred back to the original plate where the corresponding colony or plaque can be removed, cultured and large amounts of donor DNA isolated.

**In situ hybridization.** *In situ* hybridization is the hybridization of a nucleic acid probe to a target which remains in its normal cellular location. This

technique has three major applications — the cytogenetic mapping of cloned genes onto chromosomes (also useful for the characterization of chromosome aberrations; q.v. *FISH*), the detection of virus genomes, and the localization of mRNA expression. The latter is a powerful and rapid technique for determining gene expression patterns, complementing the use of antibodies to detect protein expression. Cells or tissues are fixed, permeabilized and incubated with antisense RNA or oligonucleotide probes, either as tissue sections or in wholemount. Advances in nonradioactive probe technology allow the expression of several genes to be analyzed simultaneously using different colorimetric assays.

#### Box 24.4: Essential tools and techniques IV: Nucleic acid probes

**Probe structure.** A **probe** is a nucleic acid which has been **labeled**, i.e. chemically modified in some way which allows it, and hence anything it hybridizes to, to be detected. There are three major types of probe: **oligonucleotide probes**, which are synthesized chemically and end-labeled, **DNA probes**, which are cloned DNAs or PCR products and may either be end-labeled or internally labeled during *in vitro* replication, and **cRNA probes (complementary RNA probes, riboprobes)** which are internally labeled during *in vitro* transcription of cloned DNA. RNA probes and oligonucleotide probes are gener-

ally labeled as single-stranded molecules. DNA may be labeled as a double-stranded or single-stranded molecule, but it is only useful as a probe when single stranded and must be denatured before use.

**Probe labeling.** The different ways of generating probes are shown in the table below. Probes of the highest **specific activity** (proportion of incorporated label per mass of probe) are generated by **internal labeling**, where many labeled nucleotides are incorporated during DNA or RNA synthesis. **End-labeling** involves either adding labeled nucleotides

Labeling method	Probe	Comments
5' end-labeling	Oligo/DNA	Replacement of 5' phosphate group with labeled $\gamma$ -phosphate group of free nucleotide catalyzed by <b>T4 polynucleotide kinase</b>
3' end labeling	Oligo/DNA	Tailing with labeled nucleotides using terminal transferase
Nick translation	DNA	Nicks introduced into dsDNA by DNase I and free 3' ends extended by DNA polymerase I using labeled nucleotides.
Random priming	DNA	Short random primers annealed to denatured DNA and extended by DNA polymerase using labeled nucleotides. Higher specific activity probes than nick translation
Primer extension	DNA	As above but using a specific primer. Used e.g. to label PCR products during thermal cycling
Single strand synthesis	DNA	ssDNA produced by M13/phagemid vectors or by <i>asymmetric</i> PCR (q.v.) using labeled nucleotide. Strand-specific
<i>In vitro</i> transcription	RNA	ssRNA produced by <i>in vitro</i> transcription using labeled nucleotide. Strand-specific

*Continued*

to the 3' end of a DNA strand or exchanging the 5' phosphate group for a labeled moiety. DNA probes may also be end-labeled for specific applications where identifying one end of the molecule is important (e.g. q.v. *restriction mapping*, *DNase footprinting*, *transcript analysis*).

**Isotopic and nonisotopic labeling systems.** Traditionally, nucleic acids have been labeled with radioisotopes such as  $^{32}\text{P}$  and  $^{35}\text{S}$  (and more recently  $^{33}\text{P}$ ) which are detected by autoradiography. These **radiolabeled probes** are very sensitive, but their handling is subject to stringent safety precautions and the signal decays relatively quickly. More recently, a series of **nonisotopic labeling systems** have been developed which generate colorimetric or chemiluminescent signals. A widely used label is **digoxigenin**, a plant steroid isolated from digitalis. This can be conjugated to nucleotides and incorporated into DNA, RNA or oligonucleotide probes and then detected using an antibody. Another system uses **biotin**, a vitamin, and the bacterial protein **streptavidin** which binds to biotin with extraordinary affinity. Biotin-conjugated nucleotides

are incorporated as a label and detected using streptavidin. The detecting molecule can be conjugated to fluorescent dyes or enzymes which facilitate signal detection. The advantage of such systems is that they can also be used to extract nucleic acids from complex mixtures (**affinity capture**). The biotin/streptavidin system is widely exploited to capture and extract specific DNA fragments from complex mixtures (e.g. q.v. *subtractive hybridization*, *capture PCR*).

**Padlock probes.** A recent development in probe technology is a probe structure consisting of two segments complementary to the target joined by a nonspecific link. The two ends of the probe hybridize to adjacent segments of the target (ends in) and can be joined by DNA ligase to form a topologically closed loop wound around the target DNA. Such **padlock probes** are thus extremely sensitive, because a locked probe will remain in contact with its target even under superstringent washing conditions. Point mutations which prevent ligation are therefore simple to detect by this method, compared to *allele-specific hybridization* (q.v.) which requires precise control of stringency conditions.

#### Box 24.5: Yeast cloning vectors

**Classification and applications of basic yeast vectors.** The first yeast vectors, **yeast integrative plasmids (YIps)**, were based on *E. coli* plasmid vectors and transformed cells at low frequencies due to infrequent random integration into the yeast genome; they are not maintained episomally. All yeast plasmid vectors have an *E. coli* origin of replication (q.v. *shuttle vector*), one or more selectable markers which function in both yeast and *E. coli* (usually markers which select for reversion of auxotrophy) and one or more cloning sites to insert donor DNA. Other yeast plasmid vectors differ from YIps in possessing a yeast origin of replication (in addition to the *E. coli* origin) allowing maintenance outside the yeast genome. **Yeast episomal plasmids (YEps)** carry the origin of replication from the *S. cerevisiae* 2 $\mu$  plasmid. Such vectors are maintained episomally at a high copy number and transform yeast cells at high frequency. **Yeast replicating plasmids (YRps)** contain an *ARS* element (q.v.), the yeast chromosomal origin of replication: these are unstable and usually remain in the mother

cell during budding; occasionally they may integrate like YIps. **Yeast centromere plasmids (YCps)** contain a centromere. They are maintained at a low copy number and are transmitted as Mendelian traits. **Yeast artificial chromosomes (YACs)** contain a centromere, an *ARS* element and telomeres. These high capacity vectors are used to clone large DNA fragments (Table 24.3) and more recently for the generation of transgenic mice. Yeast vectors based on the retrotransposon Ty amplify in the genome by transposition and allow high level expression of integrated genes. The principle features of the yeast vectors are summarized in the table below.

**Yeast targeting vectors.** YIps containing homology regions with endogenous genes can be used for gene targeting. Two vector types are used. Standard YIps possess a unique restriction site within the homology region which favors a single cross-over and hence insertional interruption at the target locus. **Multicopy integration vectors** have been developed using this

principle, by targeting the reiterative rRNA genes. The second type of vector (**yeast transplacement plasmid**) stimulates a double cross-over and replaces a segment of chromosomal DNA, producing a stable, single-copy integration at a defined site. Similar principles are used in mammalian targeting vectors (q.v. *gene targeting*) although homologous recombination is rare compared to random integration events in mammalian genomes.

**Transformation of yeast.** DNA can be introduced

into yeast cells in much the same way as bacteria. An efficient method using intact cells is to suspend cells in 0.1 M lithium acetate and add DNA and polyethylene glycol (PEG), followed by a brief heat shock. Yeast cells can also be transformed by electroporation. A highly efficient but laborious method is to remove the cell wall and generate **spheroplasts**, which then readily take up DNA in the presence of calcium ions and PEG. The intact cell methods are suitable for most applications, but spheroplasts are required for transformation of YAC vectors.

Vector	Components	Properties	Applications
Ylp/YTp	<i>E. coli</i> plasmid origin	Low transformation frequency. Not maintained episomally, only as integrated element (cannot be recovered).	Stable transformation Useful for surrogate genetics
YEp	2 $\mu$ plasmid origin	High transformation frequency, stable episomal maintenance at approx. 100 copies per cell	Functional analysis by complementation
YRp	Yeast chromosomal <i>ARS</i> element	High transformation frequency but unstable maintenance due to association with mother cell. Is maintained as integrated element (see Ylp)	Functional analysis by complementation, or stable integration
YCp	Yeast chromosomal <i>ARS</i> element and centromere	High transformation frequency, stable episomal maintenance with low copy number	Functional analysis especially if gene dosage effects deleterious
YAC	Yeast centromere, <i>ARS</i> element and telomeres	Stable maintenance as chromosome if length above 50 kb, 1–2 copies per cell	Highest capacity cloning vector available, useful for generating libraries of large eukaryote genomes
Ty	Ty retrotransposon	Disarmed Ty vectors in YEpS are strong expression constructs. Autonomous Ty vectors integrate into host chromosome	High yield expression

**Box 24.6:** Genetic manipulation of plants

**Crown gall disease.** *Agrobacterium tumefaciens* is a Gram-negative soil bacterium which causes **crown gall disease** in many dicotyledonous plants. The bacterium infects wounded cells at soil level (the *crown* of the plant) causing them to proliferate to form a tumor or *gall*. Crown gall cells demonstrate two novel properties: they proliferate in the absence of plant growth hormones and they synthesize one or more modified amino acids termed **opines**.

**The Ti plasmid.** The oncogenic potential of *A. tumefaciens* is conferred by a plasmid, the **Ti (tumour-inducing) plasmid**, which ranges in size from 5–450 kbp depending on the strain. There are two types of Ti plasmid each inducing the synthesis of a different class of opine, either **octapines** or **nopalines**. Generally, all Ti plasmids share four conserved regions. The most important of these is the **T-DNA** which carries genes encoding plant hormones and opine synthesis enzymes. Crown gall disease results from transformation of the plant genome with this part of the plasmid in a process analogous to bacterial *conjugation* (q.v.). This requires a second region of the plasmid, the **vir region** containing the **virulence genes**. A third region of the plasmid carries conjugation genes concerned with whole plasmid transfer between bacteria. A final conserved region encodes functions concerned with opine utilization. Thus, by infecting and transforming a wounded plant cell, the bacterium converts the crown of the plant into a safe refuge and factory producing specialized nutrients which it is able to utilize.

**Structure of the T-DNA.** The T-DNA region of the Ti plasmid is defined by 25 bp imperfect direct repeats. Its structure differs between octopine and nopaline plasmids. In the former, the T-DNA is divided into two segments with  $T_L$  carrying the plant hormone genes and the gene encoding octopine synthase, and  $T_R$  carrying genes for the synthesis of other opines, e.g. agropines. The nopaline plasmid T-DNA is a single segment with plant hormone genes and nopaline synthase genes on the right and genes for the synthesis of other opines on the left.

**Transfer of T-DNA to the plant genome.** The transfer process (see figure) is controlled by the virulence genes which are located in the *vir* region of the plasmid. *virA* and *virG* encode regulators which respond to the phenolic compounds released by wounded plant cells. *VirA* is a receptor which becomes autophosphorylated when stimulated by its ligand and transfers the phosphate group to the *VirG* protein, a transcriptional activator of the

remaining *vir* loci, each of which is an operon containing multiple open reading frames. The functions of several of the downstream *vir* genes are known. *VirD1* and *VirD2* form the endonuclease which nicks the T-DNA and initiates transfer. T-DNA is transferred as a single strand (the **T-strand**) coated with a single-stranded DNA-binding protein encoded by *virE2*. The *VirE2* protein has a nuclear localization sequence which is responsible for the transfer of the T-DNA to the plant nucleus. The T-DNA integrated into the plant genome has a precise right-hand border located within 1–2 bp of the 25 bp repeat but a variable left border which can be located at the left 25 bp repeat or up to 100 bp inside the T-DNA. Adjacent to the right-hand border is a sequence called **overdrive** which enhances the transfer process and binds proteins encoded by the *virC* operon.

**Ti plasmids as plant gene transfer vectors.** Their ability to introduce foreign DNA into the plant genome with high efficiency makes Ti plasmids attractive vectors for gene transfer into plants. The natural plasmids are unsuitable as vectors for two reasons: firstly they are oncogenic and secondly they are large and are thus unsuitable for *in vitro* manipulation. The oncogenesis problem has been solved by the use of **disarmed Ti plasmids** where the oncogenes have been disrupted or deleted. The genes carried by the T-DNA play no part in its ability to transform the host genome, so the internal region can be manipulated at will. The strong opine synthase promoters are exploited to express both novel transgenes and dominant selectable markers for recombinant selection. The size problem has been addressed in two ways. In the first strategy, the T-DNA region is cloned into a small **intermediate vector** for manipulation and then reintroduced into *A. tumefaciens* where it may recombine with the endogenous Ti plasmid. The targeted Ti plasmid can then be used for transformation. In the second strategy, a **binary vector system** is used where the T-DNA is cloned and manipulated in a small plasmid and the *vir* genes are supplied *in trans* on a second plasmid. In both cases, the manipulated T-DNA is transferred to the plant genome efficiently.

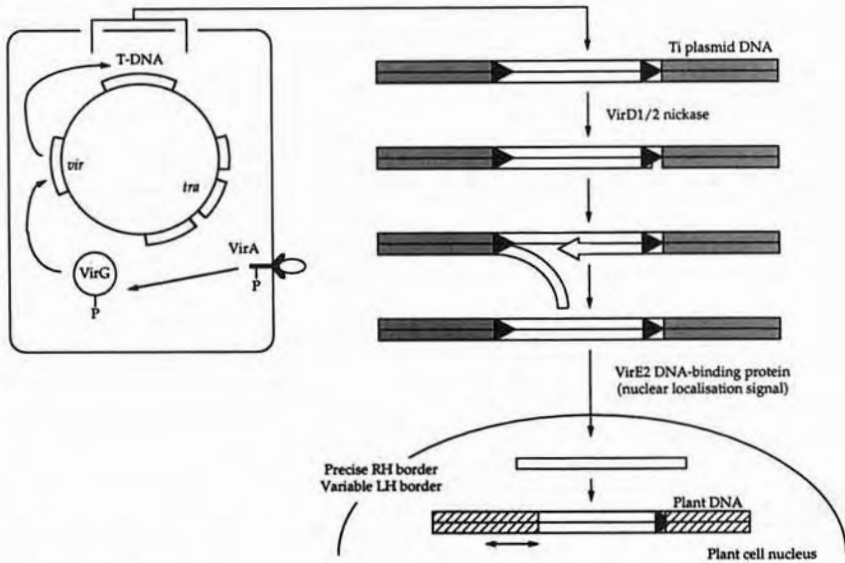
**Transformation of dicots and monocots.** The general strategy for transforming dicotyledonous plants is to cut leaf discs (causing cell injury) and then incubate with *Agrobacterium* carrying recombinant disarmed Ti vectors. The discs can then be transferred to shoot-inducing medium and simultaneously selected for markers carried on the T-DNA.

Continued



After a few weeks, shoot discs are transferred to root-inducing medium, and after another few weeks plantlets can be transferred to soil. Although the Ti plasmid-based transformation system is an efficient and widely used mechanism of gene transfer, it is restricted primarily to dicotyledonous plants which are susceptible to infection. For the manipulation of other plants (including the major cereal species:

rice, wheat, corn and maize), alternative techniques have been developed. Electroporation of protoplasts is suitable for some of these species, although not all can be regenerated from single cells. The most widely applicable technique is microballistic transfection, which has been used to transform a wide variety of both dicot and monocot plants, including rice, wheat and maize.



T-DNA transfer to the plant genome

#### Box 24.7: Transgenic animals

**Routes to the germline transformation of animals.** Unlike plant cells, differentiated animal cells are unable to regenerate into entire organisms. Transgenic animals must thus be generated by the manipulation of totipotent cells such as eggs and the cells of very early embryos. Generally, three forms of genetic manipulation have been used to generate transgenic animals (see figure): (i) microinjection of DNA; (ii) infection with recombinant retroviruses; and (iii) transfection of embryonic stem cells. All three routes have been used to generate transgenic rodents, but microinjection is the predominant route to other transgenic animals.

**Microinjection of DNA.** To generate transgenic mice by this approach, DNA is microinjected directly into the male pronucleus of a recently fertilized egg (the male pronucleus is chosen because it is larger than the female pronucleus). Fertilized eggs are obtained from the dissected oviducts of superovulated females recently mated to fertile males. Only the transgene need be injected (i.e. no vector is required). The surviving embryos are cultured until the blastocyst stage and then implanted into pseudopregnant foster females. DNA integration often occurs as a tandem array (**transgenome**) resulting from end-to-end joining of the injected DNA fragments by illegitimate recombination. There is usually only one integration

*Continued*

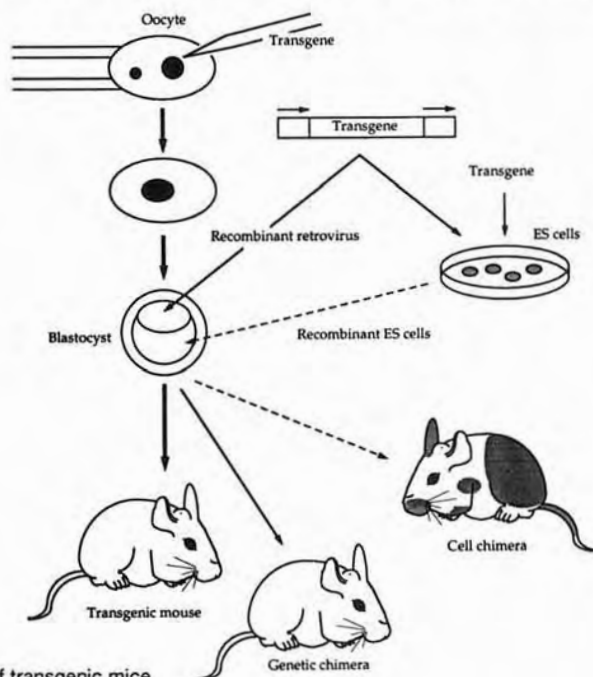
site, and the transgenome is transmitted in a Mendelian fashion, thus up to 50% of offspring may be transgenic. Occasionally, integration may not take place until after one or two early cell divisions, in which case the mouse is a genetic chimera. The random integration position and copy number means that transgenes are often subject to position and gene dosage effects. They may also occasionally disrupt endogenous genes.

The microinjection of eggs or early embryos is used to generate transgenic cows, goats, sheep, pigs, frogs (*Xenopus*), fish (e.g. the zebrafish), worms (*Caenorhabditis*) and flies (*Drosophila*). In *Drosophila*, germline transformation is mediated by recombinant defective *P-elements* (q.v.) which are cloned in plasmids and injected into embryos along with a second plasmid containing a *cis*-defective (wings-clipped) *P-element*, i.e. one which supplies transposase in *trans* but is unable to mobilize itself. If the recombinant *P-element* is introduced into the germline of an M-cytype strain (which lacks endogenous *P-elements* and thus has no endogenous transposase), the transgene will become stably integrated into the germline (q.v. *hybrid dysgenesis*).

**Retroviral transduction.** Preimplantation mouse embryos which are exposed to retroviruses often integrate single copies of the provirus into the genome of one or more cells. Recombinant defective retroviruses carrying transgenes can therefore

be used to integrate foreign DNA into the mouse genome. Retroviruses have several disadvantages compared with pronuclear microinjection. The transgene DNA has to be cloned in a suitable vector which has a limited capacity of 8–9 kbp, a helper virus is needed for integration (and despite precautions, this can result in the proliferation of the virus after transduction, causing arbitrary gene disruption), and the transgene may not be expressed due to *de novo* methylation of the integrated provirus (see DNA Methylation and Epigenetic Regulation). Retroviruses are therefore useful for some applications but not versatile for the production of transgenic mice.

**Embryonic stem (ES) cells.** A third way to generate transgenic mice involves the use of totipotent **embryonic stem cells**, immortalized cells derived from the inner cell mass of an early embryo. Such cells can be transfected with naked DNA or transduced with a recombinant retrovirus, and transformed cells can be injected into blastocysts where they may become incorporated into the embryo and contribute to the developing mouse. The ES cell method is advantageous for a number of reasons: successfully transformed cells can be selected prior to transfer to the embryo, so the success rate is higher, and ES cells can therefore be used to detect rare homologous recombination events and are thus suitable for generating targeted mutant mice (see main text). The



Routes for the production of transgenic mice

injection of ES cells into blastocysts generates cell chimeric mice, so the animals born to the foster mothers must be mated before germline transmission of the transgene can be established. ES cells are advantageous in this respect because sex and coat color can be used as cell and genetic markers. If male ES cells from mouse strain 129 (with the dominant agouti coat color phenotype) are injected into female blastocysts of C57B10/J mice (with the recessive

black phenotype), the first generation animals will be cell chimeric, i.e. they will have patches of each color depending on the clonal extent of each cell line. Germline transmission of the transgene can be established by the agouti phenotype in the second generation, and the success of transmission can be increased by mating chimeric *males* with black females, because the male ES cells are likely to have produced the chimera's reproductive tissue.

#### Box 24.8: Gene therapy

**The scope of gene therapy.** The recombinant DNA revolution has furnished the medical community with many new approaches in the fight against human disease. As well as providing new diagnostic tools and animal disease models, a number of novel therapeutic strategies have resulted directly from the ability to clone and manipulate human genes. These include the expression cloning of human gene products, the development of novel vaccines, and the engineering of therapeutic antibodies. A completely different approach involves the therapeutic use of DNA to alleviate disease. Such **gene therapy** can involve genetic modification of cells in a living patient (*in vivo* **gene therapy**) or the genetic modification of cultured cells which are then returned to the patient (*ex vivo* **gene therapy**). *In vivo* gene therapy encompasses both genetic modification of target cells (somatic transgenesis) and the therapeutic use of DNA as an epigenetic treatment (i.e. without changing the nucleotide sequence). The therapy can be used to ameliorate or correct conditions caused by human gene mutation, or to prevent infectious disease (e.g. by interfering with the life cycle of a virus).

**Gene augmentation therapy.** The traditional approach to gene therapy is where DNA is added to the genome to replace a lost function, and can be described as **gene augmentation therapy (GAT)**. The aim of gene augmentation therapy is to correct a loss of function effect, e.g. caused by a deletion, by adding the functional allele. Transferred genes may be stably integrated into the genome (in which case there is the potential to permanently correct the defect, especially if stem cells are transformed) or may be maintained episomally (in which case there is an inevitable decay in the maintenance of gene

expression and treatment may need to be repeated). Several GAT clinical trials are currently underway including cystic fibrosis, adenosine deaminase deficiency and familial hypercholesterolemia.

**Gene inhibition therapy at the nucleic acid level.** For the treatment of dominant negative loss of function mutations, and gain of function mutations, gene augmentation therapy is less powerful: additional functional copies of a dominantly malfunctioning gene are unlikely to affect the phenotype. In principle, a valid approach to the treatment of such diseases would be **targeted correction** (i.e. allele replacement) or gene knockout to remove the mutant allele. However, this is a very inefficient process in practice, even in cultured cells, and its application to the correction of genetic defects in many somatic cells, especially *in vivo*, awaits further technical improvements. A novel approach which may play a role in gene therapy in the future is targeted correction at the RNA level by using ribozymes or RNA editing enzymes to correct pathogenic mRNAs.

An alternative strategy, which is presently undergoing clinical trials for the treatment of several types of cancer, is the use of nucleic acids to inhibit gene expression. The advantage of this approach is that the inhibitor can be tailored to inhibit a specific allele, so expression of any normal functioning alleles is not affected. The introduction of antisense genes allows the stable and permanent expression of antisense RNA, which binds to (mutant) mRNA and prevents translation (it may also target the mutant mRNA for degradation). Furthermore, increasing use is being made of antisense constructs containing ribozymes which degrade the mRNAs to which they bind. Oligonucleotides can be used for epigenetic gene therapy (i.e. therapy which

does not involve changes to the genome) and can act in two ways. Antisense oligonucleotides act in the same way as antisense RNA, by binding to mRNA and causing inhibition and degradation (in this case, probably by recruiting RNaseH, which digests the RNA strand of DNA:RNA hybrids). Secondly, pyrimidine-rich oligonucleotides can potentially form triple helix structures with purine-rich strands of DNA (by *Hoogsteen base pairing*, q.v.), which blocks transcription (**triple helix therapy**). **Peptide nucleic acid (PNA)** is an analog of DNA where the phosphate backbone is replaced by a neutral peptide backbone. PNA oligonucleotides can be used as probes, and as agents for gene therapy. They form very stable triple-helix structures.

**Gene inhibition therapy at the protein level.** Targeted inhibition of gene expression can also take effect at the protein level, by the expression within a cell of genetically engineered antibodies (**intra-bodies**) which bind to and inactivate mutant proteins. Antibodies are not the only molecules being developed for protein-level gene therapy. Any protein acting as a multimer is a potential target for a dominant negative inhibitory proteins, and this strategy is being explored to prevent viral coat protein assembly, including that of the HIV. A novel approach is to use degenerate oligonucleotides to identify specific oligonucleotide sequences which interact with proteins. These oligonucleotides, or **aptamers**, can then be used to inactivate specific mutant proteins.

**Methods of gene transfer and expression.** The most efficient gene transfer vectors are based on mammalian viruses, specifically retroviruses, adenoviruses and the adeno-associated viruses (which integrate into the genome and may therefore mediate permanent correction of genetic defects) and the herpesviruses (which are neurotrophic and remain episomal). More direct approaches to gene transfer include injection (e.g. into muscle) and microballistic techniques which are used successfully to generate transgenic animals and plants. A transfer procedure based on the packaging of DNA in liposomes is also popular in cancer gene therapy, although the transfer efficiency is low.

The specificity of therapeutic gene expression may be an important factor in disease control or correction. Transferred genes can be linked to cell-type specific promoters, but this regulation is leaky due to the influence of endogenous regulatory elements on integrated genes. Specificity can also be achieved if particular cells can be targeted in some manner. Direct targeting to specific cell types is possible by exploiting the tropic properties of viruses, by conjugating DNA to ligands for cell-specific receptors (allowing the DNA to be internalized by receptor-mediated endocytosis) or by controlling the site of delivery mechanically (injection and microballistic approaches). Aerosols are used to introduce recombinant viral vectors into the lungs (e.g. for the treatment of cystic fibrosis).

## References

- Brown, T.A. (ed.) *DNA Cloning: a Practical Approach*. 2nd Edn. (4 volumes). IRL Press, Oxford.
- Brown, T.A. (ed.) *Essential Molecular Biology: a Practical Approach* (2 volumes). IRL Press, Oxford.
- Old, R.W. and Primrose, S.B. (1996) *Principles of Gene Manipulation: an Introduction to Genetic Engineering*. Blackwell Science, Oxford.
- Allen, J.B., Wallberg, M.W., Edwards, M.C. and Elledge, S.J. (1995) Finding prospective partners in the library — the 2-hybrid system and phage display find a match. *Trends Biochem. Sci.* **20**: 511–516.
- Anderson, W.F. (1992) Human gene therapy. *Science* **256**: 808–813.
- Beddington, R. (1992) Transgenic mutagenesis in the mouse. *Trends Genet.* **8**: 345–347.
- Coutre, L.A. and Stinchcomb, D.T. (1996) Anti-gene therapy: the use of ribozymes to inhibit gene function. *Trends Genet.* **12**: 510–515.
- Evans, M.J., Carlton, M.B.L. and Russ, A.P. (1997) Gene trapping and functional genomics. *Trends Genet.* **13**: 370–375.
- Jain, S.M. (1993) Recent advances in plant genetic engineering. *Curr. Sci.* **64**: 715–724.
- Monaco, A.P. and Larin, Z. (1994) YACs, BACs, PACs and MACs — Artificial chromosomes as research tools. *Trends Biotech.* **12**: 280–286.
- Mountford, P.S. and Smith, A.G. (1995) Internal ribosome entry sites and dicistronic RNAs in mammalian transgenesis. *Trends Genet.* **11**: 179–184.
- Nielsen, P.E. (1996) Peptide nucleic acids — a new dimension to peptide libraries and aptamers. *Methods Enzymol.* **267**: 426–433.
- Parimoo, S., Patanjali, S.R., Kolluri, R., Xu, H.X., Wei, S.M. (1989) *Molecular Cloning: A Laboratory Manual* (3 volumes). Cold Spring Harbor Press, Cold Spring Harbor, NY.
- Watson, J.D., Gilman, M., Witkowski, J. and Zoller, M. (1992) *Recombinant DNA*. 2nd Edn. Scientific American Books, New York.

## Further reading



- H. and Weissman, S.M. (1995) cDNA selection and other approaches in positional cloning. *Analytical Biochem.* **228**: 1–17.
- Peterson, K.R., Clegg, C.H., Li, Q.L. and Stamatoyannopoulos, G. (1997) Production of transgenic mice with yeast artificial chromosomes. *Trends Genet.* **13**: 61–66.
- Sauer, B. (1993) Manipulation of transgenes by site-specific recombination — use of Cre recombinase. *Methods Enzymol.* **225**: 890–900.
- Sokol, D.L. and Murray, J.D. (1996) Antisense and ribozyme constructs in transgenic animals. *Transgenic Res.* **5**: 363–371.
- Subtractive cloning: Past, present, and future. *Annu. Rev. Biochem.* **66**: 751–749.
- Tanksley, S.D., Ganai, M.W. and Martin, G.B. (1995) Chromosome landing — a paradigm for map-based gene cloning in plants with large genomes. *Trends Genet.* **11**: 63–68.
- Zhang, J.Z. and Chai, J.H. (1995) Methods for finding new genes in positional cloning. *Prog. Biochem. Biophys.* **22**: 126–132.

## Websites

- Transgenic mice and targeted mutations database <http://www.gdb.ors/dan/tbase/tbase.html>
- Restriction enzymes database <http://www.neb.com/rebase/>

**This Page Intentionally Left Blank**

## Chapter 25

# Recombination

### Fundamental concepts and definitions

- **Recombination** is any process generating new combinations of preexisting genetic material. **Intermolecular** or **interchromosomal recombination** generates new combinations by mixing discrete chromosomes (independent assortment of eukaryotic chromosomes at meiosis, and reassortment of segmented viral genomes), whereas **intramolecular** or **intrachromosomal recombination** is an enzyme-dependent process where new combinations of genetic material arise by cutting and joining DNA. There are five types of intramolecular recombination (Table 25.1), although the molecular reactions — cleaving DNA, exchange of strands between duplexes, DNA repair and resolution — are similar in each case.
- Recombination and mutation are regarded as two discrete mechanisms of genetic change: recombination rearranges information already present, and mutation introduces new information to the genome. Although these are useful definitions, mutation and recombination are intertwined at the molecular level. Many recombination events (especially transposition and illegitimate recombination) cause gene disruption and are described as mutations. Conversely, recombination is required to repair potentially mutagenic or lethal DNA lesions.
- Recombination is exploited in two ways: to map genetic loci and to manipulate genes and genomes. Genetic mapping works on the principle that the further apart two loci lie on the chromosome, the more likely it is that a homologous recombination event will occur between them. Therefore, the proportion of recombinant products of a given cross provides an estimate of physical distance (see Gene Structure and Mapping). Homologous recombination is also exploited for gene targeting, whereas site-specific recombination facilitates inducible deletion and chromosome rearrangement. Transposition and illegitimate recombination are also exploited, and can be used for gene transfer and integration (see Recombinant DNA and Molecular Cloning, Mobile Genetic Elements).

### 25.1 Homologous recombination

**Homologous recombination — roles and mechanisms.** Homologous recombination is homology-dependent, but not sequence-dependent, so any two DNA molecules of related sequence can undergo recombination by this process. There are two primary roles for homologous recombination, genetic mixing and DNA repair, although it is not clear which role the process first evolved to fulfill. Other processes also rely on homologous recombination: the passive movement of transposable elements (see Mobile Genetic Elements), the replication of certain bacteriophage (see Viruses), mating-type switching in yeast and related phenomena (see below), and chromosome segregation during meiosis (see below). Homologous recombination is exploited for *genetic mapping* (q.v.) of eukaryotic genomes (see Gene Structure and Mapping) and *gene targeting* (q.v.) — the deliberate replacement of one allele with another following artificial introduction of DNA into the cell (see Recombinant DNA and Molecular Cloning).

Homologous recombination models fall into three classes. In **copy-choice models**, recombination occurs during DNA replication: the nascent strand switches onto a new template as it elongates. In **breakage and reunion models**, recombination occurs in the absence of DNA replication: DNA strands are broken, exchanged between duplexes and religated. There are also hybrid models combining features from both of the above. All three types of recombination occur under different circumstances. Homologous recombination during meiosis is primarily by strand breakage and reunion, whereas DNA repair by recombination involves copy choice or hybrid mechanisms.

**Table 25.1:** Different types of intramolecular recombination**Homologous recombination**

Requires homology between the recombining partners. The proteins mediating this process (e.g. RecA in *E. coli*) are not sequence-specific but are homology-dependent. Long regions of homology are often involved (e.g. during meiosis)

**Site-specific recombination**

Does not require homology between recombining partners. The proteins mediating this process (**site-specific recombinases**) recognize short, specific DNA sequences in the donor and recipient molecules, and interaction *between the proteins* facilitates recombination. Homology often exists between the donor and recipient sites because the same recombinase protein binds to both recognition sites

**Transposition**

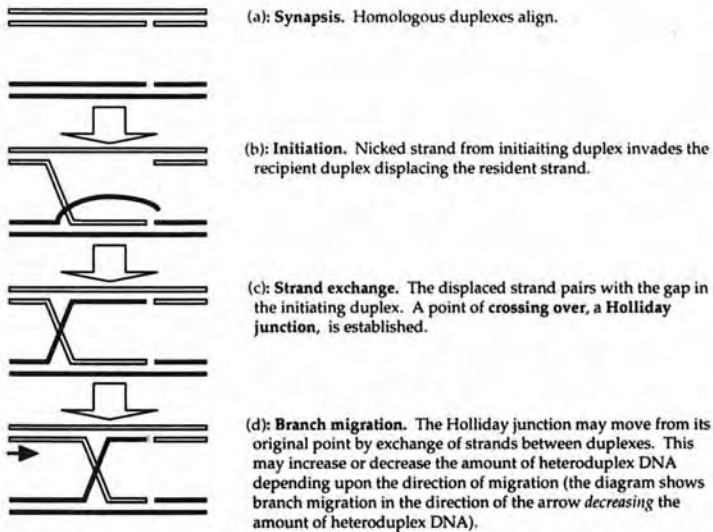
Does not require homology between recombining partners. The proteins mediating this process (**transposases, integrases**) recognize short, specific DNA sequences in one of the recombining partners only, which is a transposable element (the site of recognition is usually at the junction between the transposable element and the host DNA). The recipient site is usually relatively nonspecific in sequence, and recombination integrates the transposable element into the host DNA (see Mobile Genetic Elements)

**Illegitimate recombination**

Requires little or no homology between recombining partners and results from aberrant cellular processes. Includes illegitimate end-joining, and strand-slipping or looping during replication. *Unequal crossing over* (q.v.) is often described as illegitimate, although the mechanism is normal (albeit misplaced) homologous recombination

**Artificial recombination**

Recombination carried out by DNA ligation (q.v.) *in vitro* using purified enzymes and substrates (see Recombinant DNA)



**Figure 25.1:** The **Holliday model** of homologous recombination between intact duplexes, showing synapsis and strand-transfer stages. The classic model involves duplexes with nicks in equivalent positions. However, this is an unusual situation — normally, only one of the duplexes is nicked and the invading strand displaces the resident strand as a topologically sealed unit — a **D-loop (displacement loop)**. Recombination can also occur between two intact duplexes, in which case *topoisomerase* (q.v.) is required to allow the exchanged strands to wind round their complementary partners (q.v. *paranemic joint*, *plectonemic joint*).



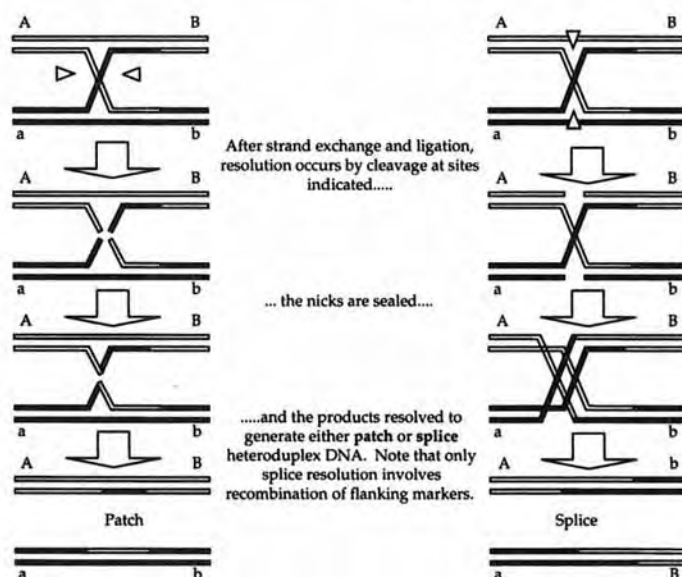
**Steps in homologous recombination.** Homologous recombination is divided into four stages: synapsis, strand transfer, repair and resolution. Synapsis and strand transfer are shown in Figure 25.1.

(1) In **synapsis**, homologous duplexes are aligned.

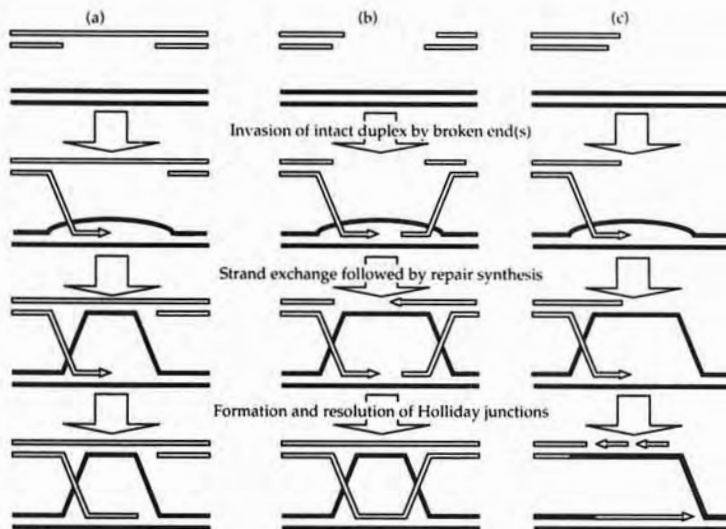
(2) During **strand transfer**, a single DNA strand is transferred from one duplex to the other. The first strand transfer marks the **initiation of recombination** as it invades the homologous duplex and (if the recipient duplex is intact) displaces a resident strand. This process may generate a short region of **heteroduplex DNA**: duplex DNA comprising strands from different parental molecules which may contain base mismatches reflecting sequence differences (different alleles) in the parental duplexes. If the recipient duplex is intact, the displaced resident strand is able to pair with the free strand of the initiating duplex. The two transferred strands cross each other, forming a structure termed a **cross bridge**, **cross branch** or **Holliday junction**. The site of the Holliday junction may move in relation to its original position by progressive strand exchange between duplexes. This is **branch migration**, and may increase or decrease the amount of heteroduplex DNA.

(3) **Repair** and **resolution** are shown in Figures 25.2 and 25.3; they do not occur in a fixed order as this depends upon the recombining partners and the availability of appropriate enzymes. Repair refers to three different processes. In the simplest case, the recombining duplexes are intact (i.e. there is no genetic information missing from either duplex) and repair involves *religation of the broken strands*. This is **conservative recombination**. The Holliday junction can then be resolved in either of two planes, as shown in Figure 25.2, to generate one of two products — a **patch** of heteroduplex DNA in a nonrecombinant background, or a **splice** of heteroduplex DNA with recombination of flanking markers.

However, if genetic information is missing from either duplex (i.e. if there is a single-strand gap, or a break) DNA repair synthesis *replaces the missing information* using information from the homologous duplex as a template (Figure 25.3). Recombination including the synthesis of new DNA is **nonconservative recombination**. In the extreme case where an entire chromosome segment is missing, resolution of the Holliday junction yields a replication fork which can duplicate the missing segment (also see Mutagenesis and DNA Repair). The third type of DNA repair, *mismatch repair* of



**Figure 25.2:** Resolution of the Holliday junction in either of two planes, generating different products. Only one resolution pathway generates a molecule which is recombinant for flanking markers A and B, although both pathways generate a region of heteroduplex DNA, known as a **splatch** (generic of patch and splice).



**Figure 25.3:** Homologous recombination as a mechanism to repair damaged DNA. Repair of single-strand gaps (a) and double-stranded breaks (b) involves new replication across the lesion using a strand from the undamaged duplex as a template. Completion of the replication is followed by strand ligation, forming Holliday junctions which can be resolved as shown in Figure 25.2. Recombination involving a partial chromosome (c) generates a Holliday junction intermediate, which is resolved as a replication fork. Arrows represent the direction of new DNA synthesis.

*heteroduplex DNA*, is generally random in its direction (cf. *post-replicative mismatch repair*) and may cause gene conversions (see below).

**Molecular basis of homologous recombination.** In *E. coli*, proteins responsible for all the major stages of recombination — synapsis, strand exchange, branch migration and Holliday junction resolution — have been identified (Box 25.1). The remaining requirement for recombination, single-stranded DNA, appears to be provided by a number of overlapping pathways.

The **RecBCD pathway** is the major source of recombinogenic substrates in *E. coli*. The products of the genes *recB*, *recC* and *recD* combine to form a large complex, the **RecBCD enzyme** or **exonuclease V**, which has three activities: (1) ATP-dependent exonuclease activity; (2) ATP-enhanced endonuclease activity; (3) ATP-dependent helicase activity.

Recombination is stimulated by *cis*-acting sites termed **chi** (GCTGGTGG), which occur approximately one every 5 kbp in the *E. coli* genome and represent preferential cleavage sites for the endonuclease. RecBCD cuts the strand containing the chi sequence approximately five bases to the 3' side in an orientation-dependent manner. Chi sites are thus *recombination hotspots* (q.v.). Blunt-ended linear duplex DNA is the substrate for the helicase activity. The enzyme binds to the DNA and progressively unwinds it, producing paired single-stranded loops in its wake. These loops are occasionally cleaved by the endonuclease activity of the enzyme, and following cleavage further unwinding generates the single-stranded tail necessary for homologous recombination.

Although the RecBCD pathway is important for recombination, *recBC* mutants retain up to 10% homologous recombination activity, indicating the presence of other pathways for producing recombinogenic DNA. Additionally, almost full recombination proficiency can be restored to *recBC* mutants by mutations at other loci, notably *sbcA*, *sbcB* and *sbcC*. The analysis of mutant strains with the genotype *recBC sbcBC* has facilitated the identification of further recombination genes. These have overlapping functions, and collectively define the **RecF pathway**. Another system, revealed by studying *recBC sbcA* mutants, involves upregulation of the *recE* gene. These recombination genes and their

**Table 25.2:** *E. coli* recombination genes involved in the production of recombinogenic DNA

Gene	Function
RecBCD pathway <i>recB</i> , <i>recC</i> , <i>recD</i>	RecBCD is an ATP-dependent exonuclease and helicase (exonuclease V) which binds to double-strand breaks and unwinds the DNA introducing nicks as it does so, preferentially at sites called chi (q.v. <i>recombination hotspots</i> ). The RecBCD enzyme thus generates substrates upon which RecA can act
RecF pathway <i>recF</i> , <i>recO</i> , <i>recR</i> <i>recN</i> <i>recJ</i> <i>recP</i> <i>recQ</i>	Facilitate strand pairing Role in double-strand break repair 5'–3' single-strand nuclease Unknown Helicase
RecE pathway <i>recE</i> <i>recT</i>	Exonuclease VIII Homologous strand pairing protein which may work in concert with RecE

proposed functions are shown in Table 25.2. Many are SOS-inducible (q.v. *SOS response*), suggesting that their normal role in the cell may be recombination-mediated DNA repair. Some bacterial strains used for molecular cloning (see Recombinant DNA and Molecular Cloning) are deficient for all recombination genes to avoid, for example, recombination within large plasmids, or between plasmids and the chromosome.

## 25.2 Homologous recombination and genetic mapping

**Linkage.** The term **linkage** has three distinct meanings in genetics. Firstly, linkage (or linking) describes the coiling of double-stranded DNA and is used to quantify topological parameters of duplex DNA (see Nucleic Acid Structure). Secondly, linkage indicates when a particular gene (and by extension, its associated phenotype) is found on a certain chromosome, e.g. genes and the characters they control associated with the X-chromosome are described as X-linked (q.v. *sex-linked inheritance*). Finally, **genetic linkage** refers to genes which tend to be inherited as a unit rather than assorting independently. This reflects close *physical linkage*, i.e. the genes are physically joined together on the same chromosome. Terms associated with genetic linkage are defined in Table 25.3.

**Linkage and meiosis in eukaryotes.** In sexual reproduction, a diploid cell undergoes a specialized form of division, **meiosis**, in order to halve the number of chromosomes and produce haploid gametes for fertilization. Meiosis involves two sequential divisions without an intervening round of DNA replication (c.f. *mitosis*). The first division is the **reduction division**, where the homologous chromosomes segregate to opposite poles of the spindle. The second division is the same as a mitotic division, where chromatids are segregated. The first meiotic division differs from mitosis in several ways.

(1) Meiosis I usually has a protracted prophase, divided into five separate stages — **leptonema**, **zygonema**, **pachynema**, **diplonema** and **diakinesis**<sup>1</sup> — characterized by the behavior and appearance of the chromosomes. The chromosomes first become visible at the leptotene stage.

(2) During the zygotene stage, homologous chromosomes become aligned. This process, termed **synapsis**, is poorly characterized and involves a proteinaceous **synaptonemal complex** which forms between the two pairs of sister chromatids. Synapsis may be initiated at the *telomeres* (q.v.). The four-strand structure, containing two homologous chromosomes, is a **bivalent**.

<sup>1</sup>Prophase meiosis I is divided into five stages named leptonema, zygonema, pachynema, diplonema and diakinesis. The terms leptotene, zygotene, pachytene and diplotene are widely used, but these are strictly adjectives.

**Table 25.3:** Some terms associated with genetic linkage

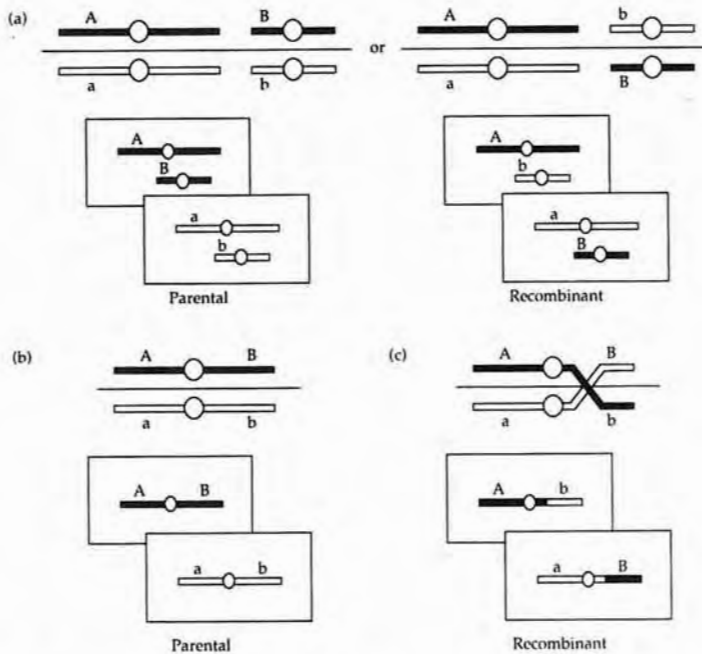
Term	Definition
<b>Linkage</b>	Close physical association between two loci on a chromosome, preventing independent assortment
<b>Phase</b>	The nature of linkage between alleles in a double heterozygote. The phase of two loci is in <b>coupling</b> (or in the <b>cis-configuration</b> ) if both dominant alleles and both recessive alleles segregate together, or in <b>repulsion</b> (or in the <b>trans-configuration</b> ) if the dominant allele of each gene segregates with the recessive allele of the other
<b>Haplotype</b>	Specific order and configuration of alleles on one chromosome
<b>Synten</b>	(1) <b>Syntenic genes</b> lie on the same chromosome. They may or may not demonstrate linkage — widely separated genes are effectively unlinked by the effect of <i>multiple cross-overs</i> (q.v.) (2) <b>Synten</b> is a term used to describe genes arranged in the same order on chromosomes of different species. Corresponding segments of chromosomes from different species are <b>syntenic regions</b> (q.v. <i>comparative genomics</i> )
<b>Linkage group</b>	A collection of genes which demonstrate linkage and therefore map together. In physical terms, a linkage group is equivalent to a chromosome
<b>Linkage equilibrium/disequilibrium</b>	For two or more polymorphic loci in a population, the equal representation of all combinations of alleles in the gametes is termed <b>linkage equilibrium</b> . A pair of alleles at a single locus reaches equilibrium in one generation in a randomly mating population. Where two or more loci are considered simultaneously, there is initially <b>linkage disequilibrium</b> where particular combinations of alleles are more common than others. Equilibrium is established over several generations if the genes are unlinked, but takes longer if the genes are linked because it is dependent on cross-over events occurring between them. Hence, the closer the linkage between two loci, the longer disequilibrium persists in the population. Linkage disequilibrium in natural populations thus indicates a mutation which has arisen recently or selection for a particular haplotype
<b>Pseudolinkage</b>	The situation where loci on separate chromosomes appear to be linked. This is seen in translocation heterozygotes, and can also occur when a tetraploid species undergoes diploidization, forming alternative homologous pairings between four related chromosomes

(3) In most organisms, homologous recombination occurs between the synapsed chromosomes and is required for proper segregation. This occurs at the pachytene stage and involves large, protein complexes termed **recombination nodules** which contain recombination enzymes.

(4) The mechanics of segregation are unique to meiosis I, involving the resolution of recombination rather than centromere division. The synaptonemal complex breaks down during the diplotene stage and homologous chromosomes remain associated due to **chiasmata** (points where crossing over has occurred). **Desynapsis** continues during diakinesis, where the chromosomes become fully condensed. The chromosomes then align on the metaphase plate and the resolution of recombination intermediates marks the onset of anaphase, where homologous chromosomes are segregated.

If two markers are located on separate chromosomes (i.e. if they are not physically joined), they undergo independent assortment according to *Mendel's Second Law* (q.v.). This can be seen in a double heterozygote, where the **parental** and **recombinant** combinations of alleles are recovered with equal frequency (Figure 25.4). Conversely, if the two markers are located on the same chromosome, the alleles found on each homolog would be expected to segregate as a unit because they are physically joined together. One might therefore expect to recover the parental combination of alleles in all meiotic products, but such **total linkage** happens only rarely because **crossing-over** usually occurs between aligned homologous chromosomes during meiotic prophase. A **cross-over** is a site





**Figure 25.4:** Linkage and chromosome segregation. Homologous chromosomes are shown pairing at meiotic prophase (paternal chromosomes in black, maternal in white). Only one strand is visible in each chromosome at this stage, although there are two chromatids, each carrying a duplex DNA molecule. The individual undergoing meiosis is heterozygous at two loci, A and B, and the parental (input) haplotypes are AB (paternal) and ab (maternal). Boxes represent daughter cells. (a) If two genes are located on separate chromosomes, they will assort independently. The parental (AB, ab) and recombinant (Ab, aB) combinations of alleles arise with equal frequency, reflecting random orientation of the chromosome pairs at the metaphase plate (thin line). (b) If two genes are located on the same chromosome, the parental combination of alleles might be expected 100% of the time because alleles of each parental haplotype are physically joined together (linkage). However, due to homologous recombination between synapsed chromosomes (c), the recombinant combination of alleles may be generated.

of chromosome breakage and strand exchange — a Holliday junction in molecular terms — leading to homologous recombination between the parental chromosomes and the production of recombinant combinations of alleles (Figure 25.4).

**Linkage and genetic mapping.** Homologous recombination during meiosis occurs essentially at random — this is an oversimplification, but is generally applicable (c.f. *recombination hotspots and coldspots*) — thus the probability of recombination occurring between any two heterozygous markers increases the further apart they lie on the chromosome. The *recombination frequency* (q.v.), measured as the proportion of recombinant offspring of a given cross, therefore reflects the physical distance between the markers. This principle underlies the meiotic mapping of eukaryotic genomes, discussed in detail elsewhere (see *Gene Structure and Mapping*), although as cloning technology becomes more advanced, the use of genetic maps is being progressively replaced by brute-force physical mapping methods. Linkage can also be exploited for gene mapping in other ways, for example q.v. *radiation hybrid mapping*, *HAPPY mapping*, *interrupted mating mapping*, *cotransformation* and *cotransduction mapping*.

**Mitotic recombination.** Although homologous recombination in eukaryotes usually occurs during meiosis (when homologous chromosomes are aligned), synapsis and exchange also occur at other times, especially in species such as *Drosophila*, where homologous chromosome pairs remain

associated in somatic cells. Any recombination occurring in the absence of meiosis (i.e. in the absence of sexual reproduction) is termed **parasexual exchange**, and this is the form of genetic exchange predominantly seen in bacteria (see Gene Transfer in Bacteria). Parasexual exchange in eukaryotes occurs between *organelle genomes* (q.v.) and between nuclear chromosomes in somatic cells (**mitotic recombination**).

Mitotic recombination between homologous chromosomes can lead to **mitotic segregation**, where heterozygous loci segregate into somatic daughter somatic cells generating reciprocal homozygous clones identifiable by their distinct phenotypes. The first evidence for mitotic recombination came from *Drosophila*, where adjacent patches of reciprocally homozygous tissue (**twin spots**) were observed in a heterozygous background. Like other forms of aberrant chromosome behavior, mitotic recombination can be stimulated by exposure to X-rays, which generate double-strand breaks in DNA. This suggests that the process may be a by-product of DNA repair. Mitotic recombination between homologous chromosomes can be used to derive genetic maps. These are colinear with meiotic maps, but the calculated genetic distances are not the same, indicating that at least some of the factors promoting meiotic and mitotic recombination are distinct. This is supported by the identification of genes in yeast which are dedicated to either meiotic or mitotic recombination (Box 25.1)

A second form of mitotic recombination occurs not between homologous chromosomes, but between identical chromatids — **sister chromatid exchange** (SCE). This is also stimulated by damage to DNA, and its relative predominance over the nonsister chromatid exchange discussed above reflects the close physical association of chromatids throughout the latter half of the cell cycle (q.v. *harlequin staining*, *DNA repair disorders*).

## 25.3 Random and programmed nonreciprocal recombination

**Nonreciprocal recombination.** Reciprocal recombination is both symmetrical and conservative, i.e. information is *exchanged* between duplexes and is neither lost nor gained. Thus, if recombination occurs between parental chromosomes carrying alleles *AB* and *ab*, the reciprocal recombinants are *Ab* and *aB*. **Nonreciprocal recombination** involves a unidirectional transfer of information. An example of nonreciprocal recombinants arising from the above cross would be *Ab* and *AB*, where information has been transferred unidirectionally from *A* to the *a* allele and the *a* allele has been lost. This type of recombination is also called **gene conversion**, because it appears that one allele has been converted to the other. Nonreciprocal recombination, or gene conversion, occurs under two circumstances.

Firstly, consider the fate of the heteroduplex DNA which arises during strand invasion and branch migration. When the Holliday structure is resolved, heteroduplex DNA, which contains base mismatches, can be subject to *mismatch repair* (q.v.). The repair can be in either direction because meiotic strand exchange occurs well after replication and the cell can no longer discriminate between the parent and daughter strands (c.f. *postreplicative mismatch repair*). To be involved in a gene conversion event, a locus must lie close to the site of strand exchange within the scope of branch migration, and the phenomenon is therefore most commonly observed when considering the recombination of closely linked markers. Because repair is nondirectional, gene conversion may be obscured in a population of meiotic products where equal representatives of each repair process may be found. It is therefore most easily observed in fungal tetrads, which represent individual meioses (q.v. *tetrad analysis*).

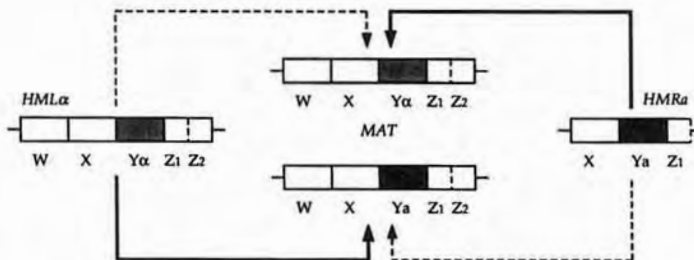
Secondly, gene conversion occurs during the recombinational repair of gaps and breaks. Single-stranded gaps may be filled by a strand from a homologous duplex prior to repair (Figure 25.3) which, as above, generates heteroduplex DNA that may be repaired in either direction. Double-stranded breaks are often targets for exonucleases, and there may be substantial loss of information prior to repair (Figure 25.3). In this case, conversion is always in the same direction, i.e. towards the homologous (undamaged) chromosome, because this is the only information available. The repair

of chromosome breaks by gene conversion occurs where DNA becomes damaged by agents in the environment, but it is also exploited as a mechanism of genetic rearrangement. Transposons which mobilize conservatively exploit gene conversion to increase their copy number (q.v. *passive transposition*, *mobile introns*). The switching of gene cassettes e.g. to specify yeast mating type or immunoglobulin structure, also occurs by this process (see below).

**Information cassette switching.** Nonreciprocal recombination can be programmed to vary information present at a specific gene locus. This mechanism of gene regulation, termed **switching**, requires an active locus where information can be expressed and silent loci where alternative forms of the information, termed **cassettes**, are stored in a constitutively unexpressed state. Nonreciprocal recombination between a silent locus and the active locus is induced by a double-strand break at the active locus which initiates recombinational repair and gene conversion.

Cassette switching is a form of passive transposition which is guided by the cell. It differs from active transposition and other forms of information exchange, such as the use of alternative antibiotic resistance cassettes in bacterial integrons, because the mechanism of recombination is different. In both active transposition and integron modulation (which involves site-specific recombination), the locus where the cassette is integrated is initially empty. In the case of switching, there is always information present at the active locus, i.e. switching is always a replacement process rather than an integration process.

**Mating-type switching in yeast.** Mating-type switching in yeast is a simple form of *differentiation* (q.v.). In *Saccharomyces cerevisiae*, the two mating types are termed *a* and  $\alpha$ . The mating-type phenotype, which reflects the particular pheromone and receptor expressed by the cell, is determined at the *MAT* locus, which may express either the *MATa* or *MAT $\alpha$*  alleles in a mutually exclusive manner. In homothallic yeast (which switch mating type at high frequency), a functional allele at the *HO* locus encodes an endonuclease with a long target site, which specifically cleaves DNA at the *MAT* locus, generating a double-stranded break. The *HO* gene is regulated by *SWI5*, a transcription factor synthesized in a cell-cycle-dependent manner, linking mating-type switching with cell-cycle regulation (see The Cell Cycle). Recombination can occur between the *MAT* locus and either of two silent information cassettes, *HML $\alpha$*  and *HMRa*, located several kilobases away from *MAT* on the left (5') and right (3') sides, respectively. The structures of *MAT*, *HML $\alpha$*  and *HMRa* are similar, as shown in Figure 25.5. There are four regions of homology, and the *a* and  $\alpha$  alleles differ only in the Y region,



**Figure 25.5:** Structure and activity of the yeast-mating-type locus *MAT* and the silent information cassettes *HML $\alpha$*  and *HMRa*. The three loci are structurally similar, each comprising a series of homologous sequence blocks, although *HMRa* is shorter than *HML $\alpha$* . The mating type is determined by the Y region present at the *MAT* locus. Recombination between the silent information cassettes and *MAT* usually occurs through the pathways indicated by thick arrows, which results in a mating-type switch. About 10% of recombination events involve the pathways shown by broken arrows, and the *MAT* allele is replaced by the same allele. Recombination is initiated by a double-stranded break at the border of Y and Z, generated by *HO* endonuclease. Both the expression and endonuclease cleavage of the silent information cassettes is prevented by their repressed chromatin structure.

which contains regulatory and coding information. The 24 bp target site for HO endonuclease overlaps the border between Y and Z and is identical in the silent-mating-type cassettes and the *MAT* locus itself. The silencing mechanism is not intrinsic to the silent-mating-type cassettes, but lies outside in flanking silencer regions which are targets for **SIR proteins** which modulate chromatin structure (see Chromatin).

Switching is initiated by the double-stranded break induced by HO, and nonreciprocal recombination involves *MAT* and either of the silent cassettes. Heterologous recombination occurs 90% of the time, suggesting that the structure adopted by the Y region demonstrates substrate preference. Mutations which disrupt silencing allow the silent cassettes to act as recipients in gene conversion. Thus the mechanism of gene silencing, the heterochromatinization of the silent-mating-type cassettes, is also responsible for protecting them against the endonuclease.

**Antigenic variation in trypanosomes.** Antigen switching describes a defense mechanism used by trypanosomes and *Borrelia* to avoid the immune responses of their vertebrate hosts. This mechanism involves changing the structure of a protein which covers the entire surface of the cell and is the only antigen exposed to the host. In trypanosomes, this is the **variable surface glycoprotein (VSG)**, whilst in *Borrelia*, it is the **variable major protein**. Switching occurs frequently during an infection, so that pathogens recognized and destroyed by antibodies are replaced by a second wave with different antigenic properties. This antigen switching process can occur many times during an infection, producing a series of **variant antigenic types (VATs)**.

*Trypanosoma brucei brucei* has many genes and gene segments representing the VSGs, many carried on nonessential minichromosomes which seem to have no other purpose than to carry VSG genes. There are probably thousands of VSG sequences in the genome but only one hundred or so are ever expressed, and the collection of potential antigenic types thus produced is termed a **serodeme**. Of the expressible genes, only one, the **expression-linked copy** or **expression-associated gene**, is active at any one time. The expression-linked copy is always found at one of several telomeric **expression sites**, although the position of these sites is not in itself important as transcriptionally silent **basic copies** are found both at telomeric and internal sites. Antigenic switching occurs in two ways: firstly gene conversion can occur, involving a basic copy being transferred to an expression site. This is mechanistically similar to yeast mating-type switching (see above), and recombination involving the large number of VSG gene segments further increases the potential variability of surface proteins. Secondly, switching can be achieved by shutting down one expression site and activating another. The precise mechanisms involved in these processes are not understood, but regions controlling the activity of expression sites have been characterized and may, like yeast mating-type switching, involve the modulation of chromatin structure.

## 25.4 Site specific recombination

**The site-specific recombination reaction.** Site-specific recombination requires short, specific DNA sequences in the donor and target molecules and is catalyzed by dedicated proteins, **recombinases**, recognizing these sequences. Site-specific recombination therefore differs from homologous recombination in its requirement for a specific recombinogenic sequence, and because homology between the donor and target molecules is not a criterion for recombination. Depending upon the system involved and the orientation of recombining sites, site-specific recombination can mediate three types of reaction:

- (1) intermolecular site-specific recombination results in **integration** or **fusion**;
- (2) intramolecular site-specific recombination between direct repeats of the recombining sites results in **excision** or **resolution**, the reverse of integration and fusion;
- (3) intramolecular site-specific recombination between inverted repeats of the recombining sites results in **inversion**.



Site-specific recombination systems fall into two classes. The  $\lambda$  integration/excision reaction ( **$\lambda$  integrase family**) has relaxed topological requirements and can undergo any of the three reaction types described above.  $\lambda$  integrase-like systems include the FLP/FRP recombinase system in the *Saccharomyces cerevisiae* 2 $\mu$  plasmid and the integration of antibiotic resistance cassettes at integron sites in the *E. coli* genome (see Mobile Genetic Elements). The Tn3 resolution reaction (**Tn3 resolvase family**) requires precise topological structure and the reaction products obtained are precisely defined. Tn3 resolvase-like systems include the invertible genome segments which control phase variation in *Salmonella typhimurium* and host range in bacteriophage Mu. Well-characterized site-specific recombination systems are discussed in Box 25.2.

## 25.5 Generation of immunoglobulin and T-cell receptor diversity

**The role of recombination in antigen receptor diversity.** The vertebrate immune system produces two types of antigen-recognizing protein: immunoglobulins (Igs), which may be membrane-bound as B-cell receptors (BCRs) or secreted as antibodies, and T-cell receptors (TCRs) (Box 25.3). Antigen-recognizing proteins may be generated for any conceivable antigen, i.e. the immune system must recognize between  $10^6$  and  $10^8$  different molecular configurations. This remarkable diversity reflects both underlying diversity in the corresponding genes (**germline diversity**), and the ability of individual lymphocytes to rearrange gene segments by site-specific recombination to create new combinations (**somatic diversity**). The **idiotypic diversity** of antigen specificity reflects this somatic recombination activity and is unique to individual lymphocytes.

Igs and TCRs also demonstrate **allotypic diversity** (diversity between individuals of a species reflecting the presence of allelic variants, i.e. polymorphism) and **isotypic diversity** (diversity present in all individuals of a species, usually reflecting the expression of closely related genes or alternative splice variants). In the immunoglobulin heavy chain locus, a novel type of isotypic diversity arises, once again by site-specific recombination. This determines the structure of the constant part of each Ig (i.e. the region concerned with effector functions rather than antigen recognition).

**Idiotypic diversity by V(D)J recombination.** The antigen-binding specificity (idiotypic diversity) of Igs and TCRs is determined by the variable regions, which are assembled from V, (D) and J segments by **somatic recombination**, i.e. recombination occurring in somatic pre-B and pre-T cells rather than in the germline. Of the three hypervariable complementarity-determining regions of the molecules, CDR1 and CDR2 are determined solely by the V-segment, and CDR3, which shows most variability, by the V(D)J assembly. Diversity is generated not only by the particular combinations of V-, (D-) and J-segments chosen, but also by their imprecise joining — variation at the coding joints can be

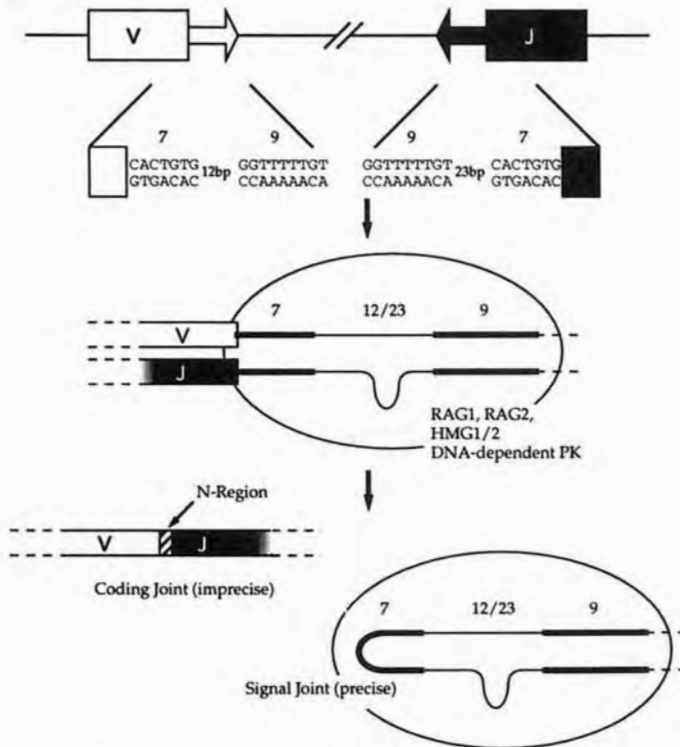
**Table 25.4:** Mechanisms of idiotypic diversity generation in antigen-receptor genes

Diversity mechanism	Molecular basis
Combinatorial diversity	The use of different V, (D) and J segments The use of multiple D segments
Junctional diversity	Exonucleolytic degradation at the coding joint leading to variation in the position of the junction between the gene segments
N-region diversity	Addition of random nucleotides at coding joint by terminal deoxynucleotidyl transferase
V-gene replacement	Further recombination between an assembled but unproductive V(D)J exon, and remaining V segments
Somatic hypermutation	Point mutation of residues in a variable region to facilitate affinity maturation; occurs in immunoglobulin loci only (see Mutation and Selection)

Further diversity is generated at the protein level by the combination of different heavy and light chains (Igs) or  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$  subunits (TCRs).

generated by both insertion and deletion of nucleotides (Table 25.4). Further diversity is produced by the generation of random mutations throughout the variable region (q.v. *somatic hypermutation*). Where diversity segments are found, D–J joining occurs first, followed by the joining of V-segments to the DJ assembly; the joining reaction is preceded by, and is dependent upon, transcription of the pre-rearranged gene. The segments used in a given rearrangement are chosen randomly, although there may be preference for certain combinations.

**Mechanism of V(D)J recombination.** V(D)J recombination is a site-specific recombination reaction occurring at **recombination signal sequences (RSSs)** flanking the V-, D- and J-segments. RSSs are found downstream of V-segments, upstream of J-segments, and on both sides of D-segments, where they occur. The RSS comprises a palindromic heptamer and an AT-rich nonamer with the consensus sequences CACTGTG and GGTTTTTGT. The heptamer is always adjacent to the gene segment and is separated from the nonamer by a nonconserved spacer region of either 12 bp or 23–24 bp (corresponding to one and two helical turns, respectively). Recombination only takes place between RSSs with different sized spacers, the **12–23 rule**. Alignment of two RSSs within a synaptic complex containing the proteins RAG1 and RAG2 results in cleavage of the DNA at the junction of each heptamer and gene segment (Figure 25.6). Transesterification forms a **coding joint** between the two gene segments and a **signal joint** between two heptamer motifs. The signal joint is precise, and the circularized DNA segment is excised from the genome and degraded. The coding joint, conversely, is imprecise. Up to ten nucleotides may be removed from either coding end by an unspecified exonuclease activity, or added at the joint by terminal deoxynucleotidyl transferase (these untemplated residues comprise the **N-region**).



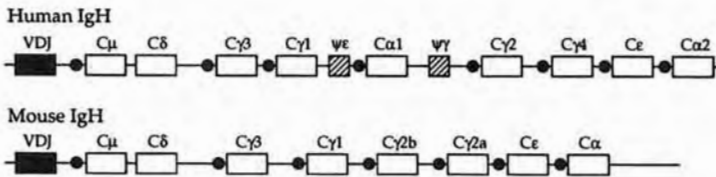
**Figure 25.6:** The mechanism of VJ recombination in the IgL loci. Synapsis of recombination signal sequences (RSSs) with different sized spacers in a complex containing RAG1, RAG2 and HMG proteins allows DNA cleavage between the RSS and the V- and J-segments. This forms a precise signal joint (which is discarded along with the residual recombination complex) and an imprecise coding joint (which may be subject to exonucleolytic degradation and the insertion of untemplated **N** nucleotides).

RAG1 and RAG2 proteins are sufficient to confer V(D)J recombinase activity on nonlymphoid cells in culture, and can perform efficient V(D)J recombination *in vitro* using cloned RSSs. RAG1 possesses a *helix–turn–helix* motif (q.v.) and binds to the nonamer sequence of the RSS, regardless of spacer size. RAG2 is recruited to RAG1–RSS complexes with greater efficiency if there is a 12 bp spacer, suggesting a basis for the 12–23 rule. RAG protein binding is stimulated by HMG1 and HMG2 (q.v. *high mobility group proteins*) which sharply bend the DNA and may facilitate interaction between different *cis*-acting sites. After cleavage, the synaptic complex of RAG and HMG proteins remain bound to the signal joint, whereas the coding joint is released, and this may protect the signal joint from exonuclease and terminal deoxynucleotidyl transferase activities.

The resolution stage of V(D)J recombination is carried out not by specialized recombinases, but by the general double-strand break repair (DSBR) machinery common to all cells (see Mutagenesis and DNA Repair). The link between V(D)J recombination and DSBR has been established by the sensitivity of mouse SCID cells<sup>1</sup> to X-rays, and the failure of V(D)J recombination in DSBR-defective cells transfected with RAG1 and RAG2 cDNA. The mouse *Scid* gene encodes a component of a multimeric **DNA-dependent protein kinase** which is recruited to damaged DNA; other components of this enzyme are encoded by the X-ray-sensitive XRCC genes.

**Allelic and isotypic exclusion.** Individual B-cells and T-cells are **monospecific**, i.e. they produce immunoglobulins or receptors of a single antigenic specificity. However, because lymphocytes are diploid and there are two immunoglobulin light-chain loci, it is in theory possible to produce up to eight types of functional immunoglobulin and four types of T-cell receptor by combining different polypeptides. The observed monospecificity reflects **allelic exclusion** (expression of one allelic copy of each gene only) and, in the case of the IgL loci, **isotypic exclusion** (expression of only one light-chain gene). The exclusion mechanism is not fully understood, but it is random and is activated by antigen binding to a B- or T-cell receptor. The most likely mechanism is that the first productive rearrangement of each locus allows the synthesis of a receptor, and that its activation by antigen binding represses further V(D)J recombination by feeding back through BCR or TCR signal transduction.

**Isotypic diversity by class switching.** The effector functions of immunoglobulins are determined by the constant region of the immunoglobulin heavy chain (C<sub>H</sub>). There are five isotypic classes of immunoglobulin: IgM, IgD, IgG, IgE and IgA, and subclasses of both IgG and IgA. The immunoglobulin class synthesized by a B-cell is determined by the particular C-gene segment which lies downstream of the recombined VDJ exon. The C-gene segments representing each class and subclass (eight in mice and nine in humans) are arranged in tandem (Figure 25.7). Class switching



**Figure 25.7:** Organization of the human and mouse IgH C-gene segments, with filled circles representing switch regions. The human IgH locus contains two pseudogenes and spans nearly 1.2 Mbp of DNA. Recombination initially occurs between the C<sub>μ</sub> switch and any other switch 3' to it. Subsequently, further switch recombination can occur in the 5'→3' direction.

<sup>1</sup>SCID is severe combined immunodeficiency, i.e. deficiency for both the cell-mediated and humoral immune systems. SCID can be caused by many defects, which result in the failure of B-cell and T-cell development. In humans, 30% of SCID cases are caused by mutations in the RAG1 or RAG2 genes which control the early stages of V(D)J recombination. In mice, the major form of SCID involves a defect in DSBR which controls the later stages of V(D)J recombination, and the effects are not restricted to lymphocytes.

involves recombination between **switch regions (S regions)**, which lie upstream of all the C-gene segments except C $\delta$ , a process which deletes the intervening DNA and brings the appropriate C-gene segment adjacent to the VDJ exon; the excised DNA is released as a circle. The recombined VDJ gene segment is initially expressed with C $\mu$ , generating immunoglobulins of isotype IgM. Switch recombination can occur between the switch region upstream of C $\mu$  and the switch region located upstream of any of the 3' C-gene segments except C $\delta$  (class switching from IgM to IgD occurs not by genomic rearrangement, but by alternative RNA processing, causing the C $\mu$  gene segment to be spliced out at the RNA level. RNA processing at the IgH locus is also used to generate the secreted and membrane-bound isoforms of immunoglobulins). Several sequential switch recombination events can occur, always in the 5'–3' direction, generating antibodies of different isotypes without affecting their antigenic specificity. Switch regions are 1–10 kbp sequences consisting mainly of simple sequence tandem repeats. Unlike VDJ recombination, switch recombination can occur at many sites within the switch region. A number of *cis*-acting elements upstream of and within switch regions have been identified as protein binding sites, and although several binding proteins have been identified (including NF- $\kappa$ B, NF-Sm (SNUP), LR1 and Pax-5), their role in recombination and its regulation is unclear. Switching appears not to be programmed, but the particular rearrangement carried out can be regulated by cytokines, which therefore specify the production of certain immunoglobulin classes and subclasses. Transcription precedes recombination, but it is the subsequent splicing that is crucial for switching to occur. This may reflect splicing factors recruiting recombination proteins to the DNA during transcription, the formation of a triple-stranded RNA–DNA triplex structure, or the placing of the DNA in the correct nuclear compartment for switch recombination to occur. Switching is accompanied by mutation in the DNA flanking the switch junction, suggesting that error-prone DNA synthesis forms part of the recombination mechanism. Switch recombination may thus involve a *copy-choice mechanism* (q.v.).

**Immunoglobulin diversity in other vertebrates.** The mechanisms of immune system diversity discussed above relate specifically to mammals, and in particular to mice and humans. In other vertebrates, and even in other mammals, there are subtle differences in the mechanisms involved. In the chicken, for instance, although the immunoglobulin loci are divided into noncontiguous segments similar to those of mammals, only one particular copy of each type of segment is ever expressed, and the diversity of immunoglobulin structure arises by gene conversion involving multiple silent copies of the V-, (D-) and J-segments. These are formally identified as pseudo-genes, but are no different in principle from the silent mating-type cassettes of yeast and VSG cassettes of trypanosomes (see above). In lower vertebrates, there appears to be considerable pre-existing germline diversity. Again, there are V-, D- and J-segments, but in many cases, these appear to be already rearranged in the germline, either already joined together or associated in such a close manner that only a single type of rearrangement is possible. Although the potential for combinatorial diversity is therefore limited, the coding joint is imprecise — as it is in mammals — and considerable junctional diversity arises by deletion and insertion of nucleotides.

## 25.6 Illegitimate recombination

**End-joining, aberrant replication and recombination.** Illegitimate recombination describes the promiscuous recombination mechanisms, requiring little or no homology between recombining partners, usually representing normal cellular processes using incorrect substrates. Illegitimate recombination often causes mutation by gene disruption, and is prominent in human genetic diseases and cancer. Some mechanisms are summarised in Table 25.5.



**Table 25.5:** Mechanisms of illegitimate recombination

Mechanism	Molecular basis
Illegitimate end joining	This reaction occurs frequently in eukaryotes, which repair double-strand chromosome breaks by direct religation of the ends. Repair by ligation is thought to have evolved because the large and highly repetitive genomes of higher eukaryotes make homology searching difficult and error-prone. The abundance of noncoding DNA reduces the danger of gene disruption when this process does go wrong and, as DNA is packaged into chromatin, it is possible that broken chromosome ends are held together by higher-order structure. When this repair system acts on illegitimate ends, however, it can cause translocations and other rearrangements. It also allows transfected linear DNA to be integrated into the genome (q.v. <i>transfection</i> , <i>transgenic animals</i> ). End joining is less common in prokaryotes which cannot ligate DNA without sticky ends
Illegitimate replication	This occurs in repetitive DNA and in DNA which forms stabilizing secondary structures (hairpins, cruciforms). The primer strand jumps out of register with the template and generates insertions and deletions. Repeats and self-complementary motifs stimulating illegitimate replication are often mutation hotspots (q.v. <i>fidelity of replication</i> , <i>trinucleotide repeat syndromes</i> )
Illegitimate strand exchange	All three classes of normal recombination — homologous, site-specific and transpositional — generate single strands as reaction intermediates. Illegitimate strand exchange occurs when these free strands become joined to the wrong partner, e.g. a single strand which just happens to be in the vicinity of the reaction. This can also occur with systems not usually involving recombination, such as <i>topoisomerases</i> (q.v.) and the <i>nicksases</i> (q.v.) which initiate replication in plasmids
Unequal crossing over	This occurs if homologous duplexes synapse out of register due to the presence of repetitive DNA. Following resolution, it generates an insertion in one duplex and a deletion in the other. Although unequal crossing over is unorthodox, the mechanism is normal homologous recombination rather than illegitimate recombination — homology is required for it to occur
Aberrant site recognition	Where recombination relies on the recognition of a specific sequence by a protein, the chance presence of a similar sequence (a <b>cryptic recognition site</b> ) in nearby DNA sometimes results in an aberrant reaction involving extra DNA. This type of recombination is responsible for most aberrant excision events involving transposons and site-specific episomes
Illegitimate V-(D)-J joining	Illegitimate recombination between RSS elements in the immunoglobulin and TCR gene loci and cryptic elements located elsewhere in the genome is responsible for a spectrum of chromosome aberrations often associated with lymphoid tumours. Additionally, tumours can arise if the recombinase genes, usually expressed only during early B- and T-cell development, become activated ectopically. Cryptic RSS elements are also found within the antigen receptor loci themselves. The <b>kappa deleting element</b> is one such cryptic site whose involvement in recombination causes deletion of the IgL $\kappa$ constant region, a process termed <b>RS recombination</b> in mice. Similarly, the <b>delta deleting element</b> causes loss of the TCR $\delta$ chain. In SCID mice, aberrant exonuclease activity occurs where DJ joining is blocked, and deletion occurs on both sides of the DJ coding joint in IgH, TCR $\beta$ and TCR $\gamma$ loci, resulting in an inability to produce mature B- and T-cells

**Box 25.1:** Genes for homologous recombination

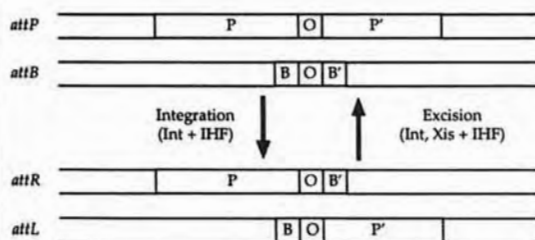
**Recombination proteins in *E. coli*.** **RecA** is a multifunctional protein facilitating synapsis and strand exchange, as well playing the primary role in SOS repair. *recA* mutants neither carry out recombination nor induce the SOS response (q.v.) and are thus exquisitely sensitive to DNA damaging agents. RecA can form a complex with single- or double-stranded DNA, although the single strand is a more efficient substrate. RecA polymerizes around the DNA to form spiral, **presynaptic filaments** within which the DNA is extensively unwound (50% longer than normal). The RecA-DNA complex can interact with duplex DNA to form a ternary complex and if a region of homology is found, strand transfer is initiated. Ternary complexes also form in the absence of sequence similarity, and this is thought to reflect an ill-fated **search for homology** by the protein. The protein filament has a definite polarity, coincident with the 5'→3' polarity of the invading and displaced strands. The protein promotes base pairing between the single strand and its complementary strand in the homologous duplex, first by forming a ternary complex containing RecA and both recombining partners, and then by displacing the resident strand of the recipient duplex as a tail or loop. It is thought that this reaction involves triple-stranded DNA and possibly unorthodox base pairing (see Nucleic Acid Structure). Strand invasion and displacement occurs in the 5'→3' direction with respect to the invading and displaced strands. Initially, matching must be perfect, but once strand transfer has started, mismatches can be tolerated, leading to the generation of heteroduplex DNA. Once strand transfer has been initiated, RecA

protein can catalyze the transfer of the displaced strand to the unpaired strand in the initiating duplex, thus generating a Holliday junction. RuvAB protein promotes branch migration in the direction of RecA strand exchange, and RecG promotes branch migration in the opposite direction. RuvC cleaves Holliday junctions and generates patch- and splice-type recombinants with equal frequency (also q.v. *D-loop mutagenesis*).

**Recombination proteins in eukaryotes.** The isolation of yeast mutants deficient in DNA damage repair and blocked during meiosis has been useful for the identification of recombination genes. The *RAD52 epistasis group* (q.v.) of genes is involved in recombinational repair. Predominant amongst these is *RAD51*, which is homologous to *E. coli recA*. *RAD51* homologs have been isolated from many eukaryotes and the *S. cerevisiae RAD51* gene is inducible by DNA damage. There appear to be several *RAD51* homologs in each species. In for example *S. cerevisiae*, the *DST1* gene encodes a *RAD51*-like product whose activity is restricted to meiotic cells. Other genes have mutant phenotypes, indicating a role in both mitotic and meiotic recombination. *RAD50* encodes a DNA-binding protein which functions in the repair of double-strand breaks in mitotic cells, and induces and processes such breaks in meiotic cells. The product of the *RAD52* gene is essential for both mitotic and meiotic recombination, the repair of double-strand DNA breaks and mating-type switching; it interacts with *RAD51*. *RAD54* encodes a helicase required for recombination and other forms of DNA repair.

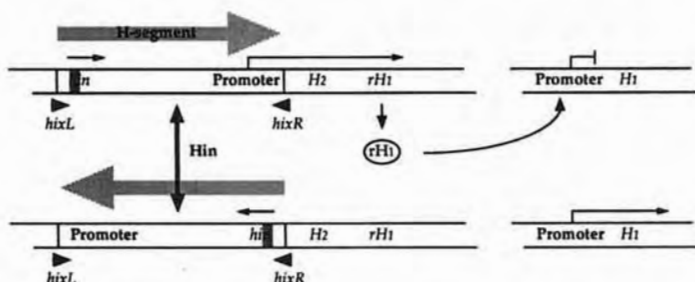
**Box 25.2: Site-specific recombination systems**

**Integration and excision of bacteriophage  $\lambda$ .** Two  $\lambda$ -encoded proteins, Int and Xis, are required for integration and excision (for a discussion of their expression and regulation during the  $\lambda$  infection cycle, see Viruses). These act in conjunction with host proteins IHF and FIS to mediate site-specific recombination between sequences termed **attachment sites**, the bacteriophage attachment site *attP* and the bacterial attachment site *attB*. Int and IHF are required for integration, whilst Xis is required in addition to these two proteins for excision (FIS stimulates, but is not essential for excision). Strand exchange reactions occur in sequence, first at the left site of the attachment region, followed by branch migration to the right-hand side where a second strand transfer occurs. These reactions proceed through a covalent intermediate where the DNA is bound to Int.



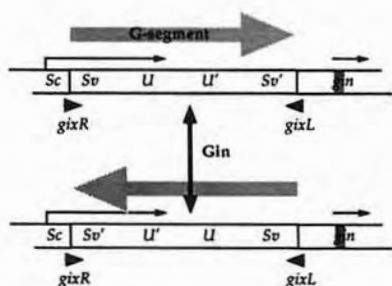
**Phenotypic variation by site-specific inversion.** In several systems, e.g. phase variation in *Salmonella typhimurium* and host-range variation in bacteriophage Mu, alternative patterns of gene expression rely on inverting a small genetic element. This inversion is facilitated by site-specific recombination systems of the Tn3 resolvase class using inverted repeats flanking the invertible element. The mechanisms of phase variation and Mu host-range variation are shown in the figures below, although similar systems control host range in bacteriophage P1 and the inversion of a small DNA segment of unknown function in the *E. coli* chromosome. The recombinases involved in these systems, Gin, Cin, Hin and Pin, are homologous and functionally interchangeable, although the *cis*-acting elements involved in the inversions are not. In the Gin/Cin systems of bacteriophages Mu and P1, the gene for the recombinase also contains a *cis*-acting enhancer of inversion and is located outside the invertible segment, whereas genes involved in the variation are within it. In the Hin system of *S. typhimurium*, the recombinase gene and enhancer are within the invertible segment whilst the genes involved in variation lie outside.

Site-specific recombination in bacteriophage  $\lambda$  integration and excision. The phage attachment site, *attP*, has the structure POP', where P and P' are complex elements containing binding sites for Int, Xis, IHF and FIS, and O is the region of overlap lying between strand-transfer sites. The bacterial attachment site *attB* is simpler and has the structure BOB', where B and B' are binding sites for Int. Site-specific recombination generates hybrid sites *attL* (BOP') and *attR* (POB') which flank the  $\lambda$  prophage.



**Phase variation in *Salmonella typhimurium*.** The flagella of *S. typhimurium* are composed mainly of the protein flagellin, which occurs as two antigenically distinct forms, H1 (specific phase antigen) and H2 (group phase antigen). H1 and H2 are encoded by distinct genes which are never coexpressed as the *H2* locus is tightly linked to a third gene *rH1*, which encodes a repressor of *H1* expression. The promoter for *H2* and *rH1* is located on invertible **H-segment** lying between *hixL* and *hixR* sites which allow site-specific inversion under control of Hin. In one orientation, promoter drives expression of *H2* and *rH1*, repressing *H1*. In the other, *H2* and *rH1* lack the promoter, thus *H2* is not expressed and *H1* expression is derepressed. The host protein FIS stimulates inversion by binding to an enhancer within the *hin* gene.

Continued



Host-range variation in bacteriophage Mu. There are two tail fiber genes in bacteriophage Mu (S and U). The S gene has a constant part (Sc) lying outside the invertible G region and alternative variable parts (Sv, S'v) which lie within it. Alternative versions of U (U, U') also lie within the G region. Inversion of the G-region is catalyzed by the protein Gln which facilitates recombination between the flanking *gixR* and *gixL* sites. The host protein FIS stimulates inversion by binding to an enhancer outside the G region, within the *gin* gene.

**Resolution by site-specific recombination.** Following replicative transposition of transposon Tn3 from one genome to another, a cointegrate is generated which contains a direct repeat of the transposon (q.v. *replicative transposition, cointegrate*). Site-specific recombination occurs between sites termed *res* within the transposon, to resolve the cointegrate into two DNA elements each carrying one copy of Tn3. This recombination reaction is catalyzed by the transposon encoded **resolvase**

protein. Unlike  $\lambda$  integration and excision, resolvase activity is directional. Fusion (integration) reactions are not catalyzed by this enzyme, reflecting the specific topological requirements of the system. It has been shown that Tn3 resolution *in vitro* always produces reaction products of the same topological state. The reaction proceeds via two-strand exchange reaction involving a covalently bound DNA-resolvase intermediate.

Another site-specific recombination controlling resolution is the **Cre-loxP** system of bacteriophage P1. This is entirely independent of the *Cin-cixL/R* system described above, and its primary function is to resolve plasmid dimers, facilitating accurate partition during cell division (phage P1 exists as a low-copy-number plasmid, and in its lysogenic state is propagated like the F-plasmid; see Plasmids, Viruses). The Cre recombinase acts on direct repeats of the *loxP* recombination site to facilitate resolution, but can also catalyze circularization of the linear phage genome and (rarely) may integrate the plasmid into the host genome. The FLP-FRP system of the yeast  $2\mu$  plasmid is also thought to function to resolve plasmid multimers, and may play a role in the increase of copy number. Unlike Tn3 resolvase and Cre, however, FLP recombinase belongs to the  $\lambda$  integrase family and catalyzes inversions and fusions as well as resolution. Of these systems, FLP-FRP and Cre-loxP have been widely exploited as tools for genome manipulation and gene regulation (q.v. *transgenic animals, knock-out mouse*).

#### Box 25.3: Antigen receptors and their genes

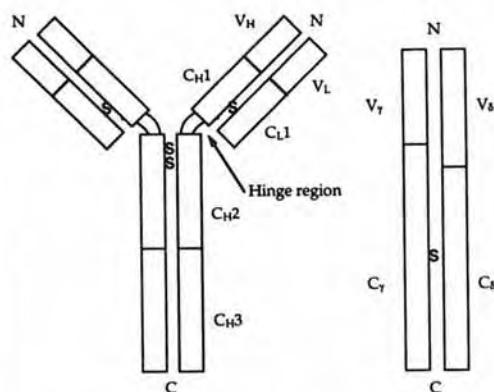
**The proteins. Immunoglobulins and T-cell receptors** are antigen-binding proteins synthesized by B-lymphocytes and T-lymphocytes, respectively. Immunoglobulins may be displayed on the cell surface as **B-cell receptors** or secreted as **antibodies**: both types of molecule have a fundamental heterodimeric structure, two identical **heavy chains** and two identical **light chains**, stabilized by disulfide bonds. There is one type of heavy chain (IgH) but two types of light chain (IgK and IgL), either of which may contribute to the immunoglobulin molecule. There are four types of TCR monomer ( $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$ ) and these form two types of heterodimer (the  $\alpha\beta$ -TCR and the  $\gamma\delta$ -TCR).

Both immunoglobulin and TCR chains have a similar organization comprising, an N-terminal **vari-**

**able region** which is concerned with antigen binding and a C-terminal **constant region**, which is concerned with effector functions (those common to all molecules of a given subtype, regardless of antigenic specificity). The overall structure of an immunoglobulin is Y-shaped due to the heavy-chain hinge regions: the fork of the Y contains the variable regions of both heavy and both light chains and binds to the antigen in a pincer movement. The variable region consists of moderately variable **framework regions** and **hypervariable regions** or **complementarity determining regions (CDRs)**, the former holding the latter in the correct three-dimensional configuration to interact directly with the antigen. There are several different classes and subclasses of immunoglobulin which have different



roles in the immune system, and the class is specified by the structure of the constant region. Alternative constant regions are also found in TCR $\beta$  and TCR $\gamma$  molecules. The structures of a generic immunoglobulin and the  $\gamma\delta$ -TCR, indicating the position of constant and variable regions, are shown below.



**The genes.** The mammalian immunoglobulin and TCR genes are unique in that each must be assembled in somatic cells from a series of alternative non-contiguous building blocks provided by the germline genome. This so-called **germline configuration** is nonfunctional, and gene rearrangement is essential for the production of functional receptors and antibodies.

The constant regions of each molecule are specified by **C segments** which lie at the 3' end of each locus (in the immunoglobulin loci, the C segments are arranged in a tandem array representing the constant region of each isotypic subclass). The variable regions of the IgH and TCR  $\beta$  and  $\gamma$  genes are assembled from three types of gene segment — **V (variable) segments**, **D (diversity) segments** and **J (junctional) segments**. Arrays of V-, D- and

J-segments are provided, allowing variable regions of different structure to be generated by **combinatorial joining**. The variable regions of the IgL and T cell  $\alpha$  and  $\delta$  genes are assembled from V and J segments only. The number and nature of the different segments differ between the different loci, and those from the mouse and human genomes are shown in the table below. Generally, the immunoglobulin genes show a relatively simple organization, with the various gene segments lying in the same order in which they are assembled. The T-cell receptor genes have a more complex structure. The **TCRD** gene lies embedded within the **TCRA** gene and the V segments are mixed. Additionally, V segments lie both upstream and downstream of the corresponding C-gene segments, and rearrangements involving downstream V-gene segments cause large inversions.

Species	Locus and number of segments				Comments
	V	D	J	C	
<i>Man</i>					
IgH	86	~30	9	9 (+2 $\psi$ )	
IgL $\kappa$	76	—	5	1	
IgL $\lambda$	52	—	7	7	JC combinations predetermined
<i>Mouse</i>					
IgH	>1500	12	4	8	
IgL $\kappa$	>200	—	4 (+1 $\psi$ )	1	
IgL $\lambda$	2	—	4	4	JC combinations predetermined

$\psi$  indicates a pseudogene.

## Further reading

- Cameriniotero, R.D. and Hsieh, P. (1995) Homologous recombination proteins in prokaryotes and eukaryotes. *Annu. Rev. Genet.* 29: 509–552.
- Eggleston, A.K. and West, S.C. (1996) Exchanging partners — recombination in *Escherichia coli*. *Trends Genet.* 12: 20–26.
- Hagmann, M. (1997) RAGged repair: What's new in V(D)J recombination. *Biol. Chem.* 378: 815–819.
- Lichten, M. and Goldman, A.S.H. (1995) Meiotic recombination hotspots. *Annu. Rev. Genet.* 29: 423–444.
- Rao, B.J., Chiu, S.K., Bazemore, L.R., Reddy, G. and Radding, C.M. (1995) How specific is the first recognition step of homologous recombination? *Trends Biochem. Sci.* 20: 109–113.
- Shinagawa, H. and Iwasaki, H. (1996) Processing the Holliday junction in homologous recombination. *Trends Biochem. Sci.* 21: 107–111.
- Stavnezer, J. (1996) Antibody class switching. *Adv. Immunol.* 61: 79–146.

**This Page Intentionally Left Blank**

## Chapter 26

# Replication

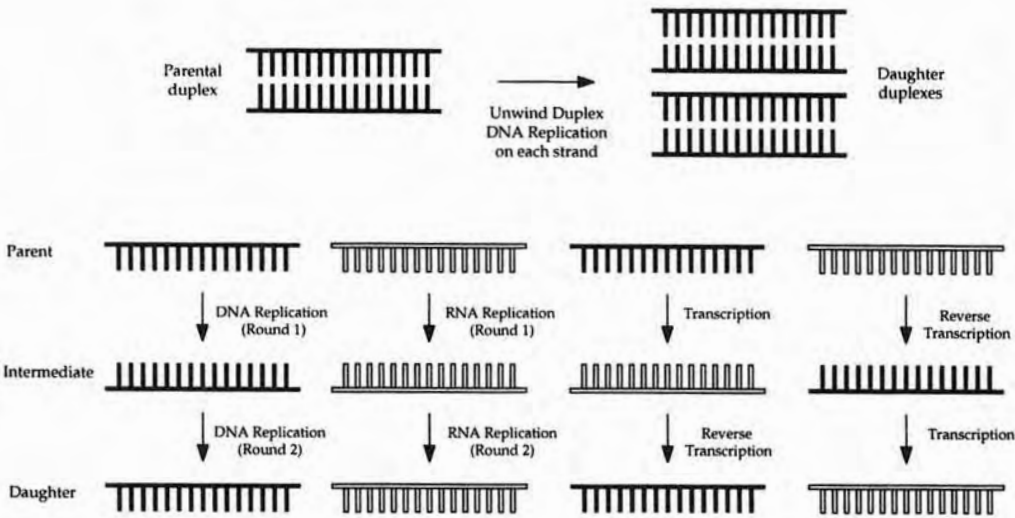
### Fundamental concepts and definitions

- **Replication** can be broadly defined as genome duplication, an essential process for the propagation of cellular genomes and those of 'molecular parasites' — viruses, plasmids and transposable elements. The genome to be duplicated is the **parental** genome, and the copies are **daughter** genomes.
- All cellular genomes are double-stranded DNA. However, viruses and other genetic elements with single-stranded DNA or RNA genomes must also replicate, and the propagation of the agents responsible for the *transmissible spongiform encephalopathies* (q.v.) can also be thought of as a type of replication occurring at the protein level.
- At the biochemical level, replication is defined as a template-directed nucleic acid synthesis reaction where the template and nascent (growing) strand are the same type of nucleic acid. This differs from transcription and reverse transcription where one strand is DNA and the other RNA (Box 26.1). The replication of double-stranded DNA begins with the **parental duplex** which separates into two strands; each can act as a template to generate two identical **daughter duplexes** in a one-step process, **direct replication** (Figure 26.1). The two daughter duplexes are **sister duplexes** with respect to each other (hence terms such as *sister chromatid exchange*). However, because each newly synthesized strand is *complementary* to the parental strand rather than identical to it, the replication of a single-stranded genome must proceed through a two-step process, the first synthesis reaction producing an intermediate of opposite sense to the parent strand, which can itself act as a template to generate *replicas* of the parental molecule. Such **indirect replication** and may involve any of the four types of nucleic acid synthesis reaction (Figure 26.1).
- Replication is a polymerization reaction and can be divided into stages of initiation, elongation and termination. The elongating replication center requires the coordination of many different enzyme activities, collectively described as the **replisome**.

### 26.1 Replication strategy

**Models for the replication of DNA.** The replication of cellular DNA was originally conceived as two models: conservative and dispersive. Following **conservative replication**, the parental DNA remains unchanged and gets passed to one daughter cell, whereas newly synthesized DNA gets passed to the other. Following **dispersive replication**, new DNA synthesis is interstitial, and each daughter cell receives a mixture of parental and newly synthesized DNA.

In a third **semiconservative replication** model, proposed by James Watson and Francis Crick following their elucidation of the double-helical structure of DNA, the parental *strands* remain unchanged, but the duplex is separated into two halves. Each parental strand acts as a template for replication and the daughter duplexes have one parental strand and one daughter strand each. The semiconservative model holds for cellular DNA, but the single-stranded genomes of viruses and some plasmids replicate conservatively — the structure of the (single) parental strand is conserved following replication. [The Meselson–Stahl experiment. In 1958, Matthew Meselson and Franklin Stahl showed that the replication of bacterial chromosomal DNA was semiconservative. *E. coli* were grown for many generations in a medium containing  $^{15}\text{N}$ , so that their DNA became universally labeled with the isotope (heavy DNA). The cells were then shifted to a medium containing normal  $^{14}\text{N}$  and DNA was isolated from cells after one and two rounds of replication. The DNA was ana-



**Figure 26.1:** Replication of single- and double-stranded genomes. The replication of double-stranded DNA genomes involves a single round of DNA replication and no intermediate is required. The replication of single-stranded genomes can involve any type of nucleic acid synthesis reaction because an intermediate is involved. Black strands are DNA, white strands are RNA.

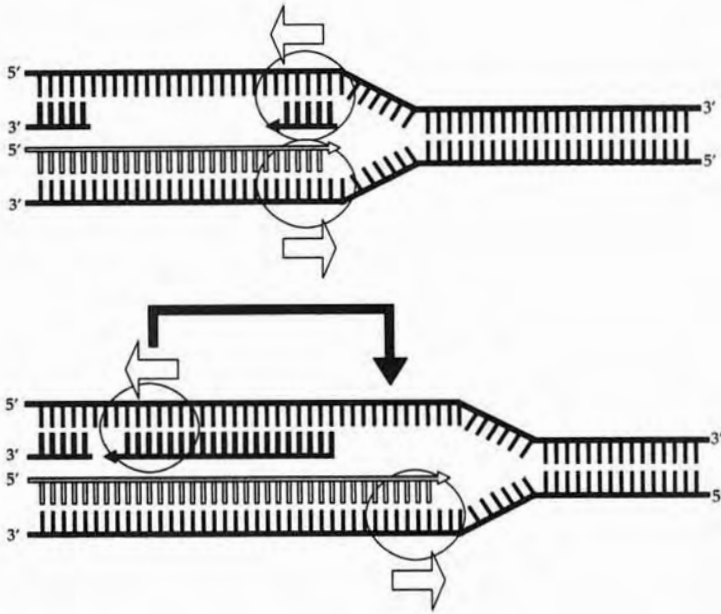
lyzed by buoyant density centrifugation, which discriminates between heavy DNA, normal light DNA and intermediate DNA containing one heavy strand and one light strand. After one round of replication, the DNA was all of intermediate density, and after two, there were equal amounts of intermediate density and light DNA. These results were consistent only with the semiconservative model. All cellular genomes replicate semiconservatively. In eukaryotes, semiconservative replication can be demonstrated by the incorporation of bromodeoxyuridine into chromosomal DNA. Incorporation over two rounds of replication allows sister chromatids to be discriminated on the basis that one has bromodeoxyuridine incorporated in both strands, and the other in only one strand. The chromatids then stain differently (q.v. *harlequin staining*).

**Semidiscontinuous replication.** Watson and Crick's semiconservative model of DNA replication predicted the existence of a **replication fork**, a dynamic Y-shaped structure with a barrel composed of parental duplex DNA and arms composed of daughter duplex DNA, each daughter duplex consisting of one parental and one daughter strand (Figure 26.2). At the center of the fork, the parental duplex would be unwound and nucleotides would be added to the growing daughter strands. The existence of replication forks has been confirmed directly by incorporating radioactive nucleotides into replicating bacterial DNA and observing the intermediate structures by electron microscopy. However, this model reveals a paradox which can be summarized as follows:

- (1) cellular DNA replication is semiconservative;
- (2) both daughter strands are extended simultaneously;
- (3) the strands of the parental duplex are *antiparallel* (q.v.);
- (4) DNA polymerases extend DNA only in the 5'→3' direction.

How can simultaneous 5'→3' elongation of both daughter strands occur at a replication fork when the parental templates have opposite polarity? This can be achieved by **semidiscontinuous replication**, where one strand is extended continuously and the other is synthesized discontinuously as a collection of short fragments. The mechanism of semidiscontinuous DNA replication can be formally expressed as the **leading strand – lagging strand model**. The **leading strand** is the nascent strand which is synthesized continuously in the direction of fork movement because its 3'





**Figure 26.2:** A model for semidiscontinuous DNA synthesis at the replication fork. The replisome contains two DNA polymerases which start in the same place. As the duplex unwinds, leading-strand synthesis is continuous and in the direction of fork movement. This reveals a portion of the lagging-strand template, and a second DNA polymerase synthesizes a short Okazaki fragment in the opposite direction to fork movement. When it reaches the previous Okazaki fragment the enzyme dissociates and returns to the original position where a further length of template has been uncovered. Note that although the relative directions of the two enzymes are different, they need not necessarily be physically separated. The same result can be obtained by looping the retrograde template around the enzyme, or by pulling the retrograde template backwards past the enzyme to create a loop. Black stands are parental DNA, white is the leading strand, and gray is the lagging strand. Circles represent DNA polymerase (the direction of synthesis and polymerase movement is shown by small and large arrows, respectively). For clarity, RNA primers are not shown.

end is exposed to the DNA polymerase. The leading strand template is thus the **forward template**. The **lagging strand** is the nascent strand which is synthesized discontinuously in the opposite direction to fork movement because its 5' end, the end which cannot be extended, is exposed to the DNA polymerase. The lagging-strand template is thus the **retrograde template**. The mechanism can be summarized as follows: as the replication fork moves forward and the leading strand is extended, a portion of retrograde template is exposed. DNA polymerase can then synthesize a small fragment of DNA, an **Okazaki fragment**, by moving backwards over the template in relation to the fork progression. The lagging strand is so called because the leading strand must be synthesized first to uncover the corresponding portion of lagging strand template. The enzyme dissociates from the template when it reaches the previously synthesized Okazaki fragment, by which time a further portion of retrograde template has been exposed. The enzyme can then reinitiate and synthesize a new Okazaki fragment. By repeating this back-stitching process over and over, the lagging strand would appear to grow in the 3'→5' direction (Figure 26.2). Because DNA polymerase cannot initiate *de novo* strand synthesis, each Okazaki fragment needs to be individually primed. The dissociation–reassociation cycle therefore requires a *priming* step (q.v.).

Evidence supporting the leading strand–lagging strand model includes the presence of DNA primase at the replication fork in both bacterial and eukaryotic replisomes. Also, both replisomes are asymmetrical, reflecting the presence of one highly processive DNA polymerase for leading-strand synthesis and one distributive DNA polymerase for lagging-strand synthesis. Pulse chase experiments

confirm that 50% of nascent DNA in cellular replication occurs as low molecular weight fragments.

**Displacement replication.** Cellular genomes and those of many DNA viruses and plasmids replicate using the semidiscontinuous mechanism described above, in which both daughter strands are extended simultaneously and replication forks are required. Other double-stranded DNA replicons use a distinct strategy where only one strand is initially used as a template, and continuous extension of the nascent strand displaces the other parental strand. This is **displacement replication** and the dynamic structure representing the site of continuous nascent strand synthesis is a **displacement fork**. In many such replicons (e.g. plasmid ColE1), the displaced strand eventually acts as a template for discontinuous synthesis, and the overall mechanism is identical to semidiscontinuous replication. In others (such as mitochondrial DNA and the adenovirus genome), continuous synthesis of the leading strand uncovers a second origin on the displaced parental strand, which facilitates continuous synthesis in the opposite direction. In this case there is no need for Okazaki fragments and there is no lagging strand; the process may be defined as **continuous replication**.

A special type of displacement replication is facilitated by nicking one strand of a circular double-stranded genome and extending the free 3' end. In this case, strand extension will displace the resident parental strand, but continued displacement can occur by circling the template more than once, leading to the extrusion of a concatemeric strand, which may then act as a template for second strand synthesis. This **rolling-circle replication** strategy is favored by many bacteriophages and by most plasmids of Gram-positive bacteria.

**Replication of single-stranded genomes.** The replication of single-stranded DNA and RNA genomes always involves a **replicative intermediate** of opposite sense to the genome (sometimes termed an **antigenome**). The replicative intermediate may act as a distinct template for daughter genome synthesis, or the genome and antigenome may stay associated in a double-stranded **replicative form** from which daughter genomes are produced by displacement replication. In the special case of the retroviruses, the parental RNA genome is destroyed once the DNA antigenome has been synthesized, and is replaced by a second DNA strand — daughter genomes are produced by transcription from this double-stranded provirus. For further discussion see Viruses.

## 26.2 The cellular replisome and the enzymology of elongation

**The replisome.** The replisome (replication complex, center, machine etc.) is the dynamic complex of enzymes and other proteins found at the replication fork during elongation. Despite the elegant simplicity of the underlying mechanism of replication, the logistics of the operation require many different enzyme activities to be coordinated for continued, accurate DNA synthesis. There is a formidable energy requirement to unwind the supercoiled DNA, and in eukaryotes, DNA is organized into chromatin which must be negotiated by the replisome (see Chromatin). The components of the cellular replisome and their functions are listed in Table 26.1, and the bacterial and eukaryotic replisome components are compared in Table 26.10. The replisome assembles from its components during initiation and is only found at the replication fork — it does not exist as a separate entity in the cell. Simpler replicons may require fewer replisome components, the minimal requirement being a single polymerase enzyme.

**Replication-deficient mutants.** The elucidation of the components of prokaryotic and eukaryotic replication systems has depended on the isolation of DNA replication-deficient mutants. Replication is an essential process, so the proteins have been identified through the use of *conditional mutants* (q.v.), mostly temperature-sensitive. In bacteria, replication mutants fall into two categories: quick stop and slow stop. **Quick-stop mutants** do not finish the current round of replication when shifted to the nonpermissive temperature, and identify genes critical for elongation. **Slow-stop mutants** finish the current round of replication, but fail to reinitiate. They identify genes required

**Table 26.1:** Generic components of the cellular replisome

Replisome component	Function during replication
DNA helicase	Unwinds DNA ahead of replication fork
DNA ligase	Joins fragments of repaired lagging strand
DNA polymerase	DNA synthesis, repair of gaps in lagging strand. The replisome may contain several distinct forms of DNA polymerase, one for leading-strand synthesis, one for lagging-strand synthesis and one for lagging-strand repair
DNA primase	Primes Okazaki fragment synthesis
DNA topoisomerase	Releases torsional strain caused by helicase activity. Decatenates linked circles following replication (q.v. <i>DNA topology</i> )
RNaseH	Removes RNA primers from lagging strand
Single-stranded binding proteins	Stabilizes single-stranded regions of replication fork. May interact with other replisome components to stimulate their activity

Topoisomerases are not found at the replication fork, but are nevertheless indispensable for replication in torsionally constrained DNA. See Table 26.10 for the specific components of the bacterial and eukaryotic replisome.

**Table 26.2:** Some *E. coli* replication-deficient mutants and the functions of each locus

Locus	Phenotype	Function
<i>dnaA</i>	Slow-stop	Initiator protein
<i>dnaB</i>	Quick-stop	Helicase
<i>dnaC</i>	Quick- and slow-stop alleles	Forms complex with DnaB helicase and assists loading at origin
<i>dnaE</i> ( <i>polC</i> )	Quick-stop	$\alpha$ subunit of pol III holoenzyme
<i>dnaG</i>	Quick-stop	Primase
<i>dnaN</i>	Quick stop	$\beta$ unit of pol III holoenzyme
<i>dnaQ</i>	Quick stop and mutator alleles	$\epsilon$ subunit of pol III holoenzyme
<i>dnaT</i>	Slow-stop	Primosome component
<i>dnaX</i>	Quick stop	$\gamma$ and $\tau$ subunits of pol III holoenzyme
<i>gyrA</i>	Quick- and slow-stop alleles	DNA gyrase subunit $\alpha$
<i>gyrB</i>	Quick- and slow-stop alleles	DNA gyrase subunit $\beta$
<i>lig</i>	Repair deficient	DNA ligase
<i>ori</i>	Lethal ( <i>not conditional</i> )	Origin of replication
<i>polA</i>	Repair deficient	DNA polymerase I
<i>rnhA</i>	Stable replication	RNase H
<i>ssb</i>	Quick-stop	Single-strand binding protein
<i>ter</i>	None	Terminus of replication
<i>tus</i>	None	Termination protein

This table lists those *cis*-acting elements and genes whose products function at the replication fork or at initiation or termination. Genes affecting general aspects of DNA metabolism or competence for the replication of phage or virus genomes are not listed.

for either initiation or termination. Some genes have both quick-stop and slow-stop alleles and therefore act at more than one stage. Several also have *mutator alleles* (q.v.); they influence the fidelity of DNA replication or *DNA repair* (q.v.). Notably, many replication-deficient mutants are also deficient in other aspects of DNA metabolism (recombination, repair, transcription), indicating that some genes have a general role in the processing of DNA. Table 26.2 provides a list of *E. coli* replication-deficient mutants which identify replisome components.

**Table 26.3:** Intrinsic enzyme activities associated with DNA polymerases

Activities of DNA polymerases	Function
<i>Activities of all DNA polymerases</i>	
Template binding activity	Substrate recognition
Nucleotide binding activity	Substrate recognition
5'→3' polymerase activity	DNA synthesis
Pyrophosphorylase activity	DNA synthesis
<i>Activities associated with some DNA polymerases</i>	
3'→5' exonuclease activity	Proofreading
5'→3' exonuclease activity	Primer excision and repair
Primase activity	Primer synthesis

**Table 26.4:** Comparison of the properties of the DNA polymerases of *E. coli*

	pol I	pol II	pol III
5'→3' polymerase	✓	✓	✓
3'→5' exonuclease	✓	✓	✓
5'→3' exonuclease	✓	x	x
Structure	Polypeptide	Polypeptide	Multimeric complex
<i>E. coli</i> gene	<i>polA</i>	<i>polB</i>	<i>dnaE</i> , <i>dnaN</i> , <i>dnaQ</i> , <i>dnaX</i> , other subunits unassigned
Function	Principle repair polymerase, primer excision	Error-prone repair polymerase (SOS inducible)	Principle replication polymerase

**DNA polymerases.** Enzymes catalyzing DNA synthesis on a DNA template are **DNA polymerases**. They perform two primary functions in the cell: the synthesis of DNA during genome replication, and the resynthesis of missing DNA following damage or recombination, and following primer excision from the lagging strand. In both prokaryotes and eukaryotes, specialized DNA polymerases are dedicated to replication and repair functions, the former sometimes being termed **DNA replicases**. All DNA polymerases possess a **5'→3' polymerase activity** and a **pyrophosphorylase activity**, which together facilitate DNA synthesis. Unlike RNA polymerases, DNA polymerases are unable to initiate *de novo* strand synthesis and therefore require a primer. Most DNA polymerases also possess further intrinsic activities (Table 26.3). The most important is a **3'→5' exonuclease activity**, which is the basis of proofreading (see Mutagenesis and DNA Repair for discussion of the enzymology of replication fidelity).

**The DNA polymerases of *E. coli*.** The *E. coli* genome encodes three DNA polymerases (**DNA polymerases I, II and III**, or **pol I, pol II and pol III**). The properties of these enzymes are summarized in Table 26.4. Pol I and Pol II are single polypeptides whose primary role appears to be DNA repair. Pol I (also known as **Kornberg polymerase**) is the predominant polymerase activity in the cell and possesses a unique **5'→3' exonuclease activity** which facilitates primer excision from the lagging strand during repair synthesis *in vivo* (q.v. *nick translation*). Although all the enzymatic activities of pol I lie on a single polypeptide of 109 kDa, each arises from a specific domain and proteolytic cleavage can generate a large C-terminal fragment (**Klenow fragment**, **Klenow polymerase**) which lacks the 5'→3' exonuclease activity and is useful for *in vitro* applications where excision would be undesirable (e.g. q.v. *random priming*, *in vitro* mutagenesis). Klenow polymerase is produced commercially by expressing a truncated *polA* gene.



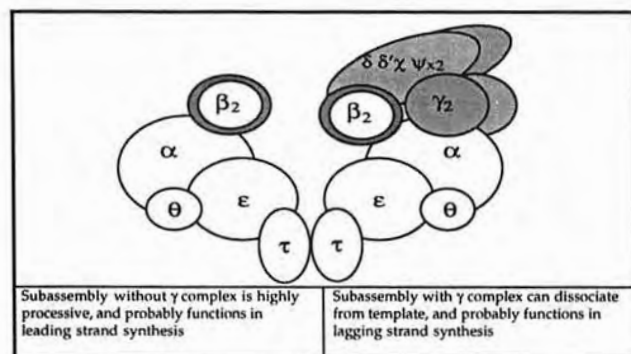
**Table 26.5:** Subunits of the *E. coli* DNA polymerase III holoenzyme and their proposed functions

Subunit	Gene	Properties and proposed function
<b>Core subunits</b>		
$\alpha$	<i>dnaE</i> ( <i>polC</i> )	5'→3' polymerase activity, required for DNA synthesis
$\epsilon$	<i>dnaQ</i> ( <i>mutD</i> )	3'→5' exonuclease activity, required for proofreading
$\theta$	Unassigned	Function uncertain, may help to assemble other subunits
<b>Accessory subunits</b>		
$\tau$	<i>dnaX</i> <sup>a</sup>	DNA-dependent ATPase, required for initiation. Promotes dimerization
$\gamma$	<i>dnaX</i> <sup>a</sup>	Associates with four peptides (see below) to form a DNA dependent-ATPase known as the <b><math>\gamma</math>-complex</b> , required for initiation, facilitates $\beta$ subunit binding
$\delta, \delta', \chi, \psi$	Unassigned	Associate with $\gamma$ to form the $\gamma$ complex (see above)
$\beta$	<i>dnaN</i>	'Sliding clamp' which increases processivity of the holoenzyme. $\beta$ binds to DNA to form a preinitiation complex, a process which requires the ATP-dependent activity of the $\gamma$ complex. <i>dnaN</i> is induced by the SOS response (q.v.)

<sup>a</sup> $\tau$  and  $\gamma$  are encoded by a single gene, *dnaX*, which was previously divided into two overlapping ORFs termed *dnaX* and *dnaZ*, with  $\gamma$  representing the N-terminal two-thirds of  $\tau$ . A *cotranslational frameshift* (q.v.) allows synthesis of  $\gamma$ .

Pol II is a minor component of the cell during normal growth but is inducible by the *SOS response* (q.v.). It appears that this enzyme allows nucleotide incorporation opposite *AP sites* (q.v.), i.e. lesions which stall pol I and pol III — it thus facilitates *translesion synthesis* (q.v.).

The principle replicative DNA polymerase of *E. coli* is Pol III which, unlike the other enzymes, is a multisubunit complex, the **Pol III holoenzyme** (Table 26.5). The holoenzyme functions as a heterodimer of complexes at the replication fork, with each monomer seeing to the synthesis of one daughter strand. *In vitro*, the  $\alpha$ ,  $\epsilon$  and  $\theta$  subunits associate to form the core enzyme, which contains the essential enzyme activities. Addition of the other subunits promotes dimerization and increases the processivity of the enzyme. The assembly of the holoenzyme *in vivo* occurs as follows: the  $\beta$  subunit functions as a dimer and forms a ring or clamp which can slide along single-stranded DNA. This is a processivity factor which keeps the core enzyme attached to the template. The  $\beta$  subunit is loaded onto the template-primer by the  $\gamma$  complex, an ATP-dependent process, to form the **preinitiation complex**. The loading of the  $\beta$  subunit allows the core enzyme to bind, and addition of the  $\tau$  subunit facilitates dimerization. The structure of the completed holoenzyme is shown in Figure 26.3. The holoenzyme is symmetrical except for the  $\gamma$  complex, which is associated with only one of the monomers. The  $\gamma$  complex is required for both loading and unloading the  $\beta$  subunit from DNA, and

**Figure 26.3:** Heterodimeric assembly of *E. coli* pol III to facilitate simultaneous leading- and lagging-strand synthesis.

hence controls the processivity of the enzyme. The presence of the  $\gamma$  complex allows the  $\beta$  subunit to disassociate from the template primer when the polymerase encounters the 5' end of a previously synthesized Okazaki fragment on the retrograde template. The lagging-strand core enzyme will thus be released from the holoenzyme when it has completed the Okazaki fragment and can reassociate with a preinitiation complex which has loaded at the next available template primer, provided by DNA primase.

The DNA polymerases of other bacteria are similar to those of *E. coli*, and in fact all DNA polymerases can be grouped and classified according to the structure of six conserved domains. The most remarkable polymerases are those of thermophilic and hypothermophilic prokaryotes, which can catalyze DNA synthesis at temperatures above 100°C, greatly in excess of the  $T_m$  of DNA. In such organisms, the melting of DNA is likely to occur passively, but the polymerases must demonstrate an extraordinary capability to clamp the primer on the template in order to extend it. *Taq* DNA polymerase, originally isolated from *Thermus aquaticus*, is widely used in the *polymerase chain reaction* (q.v.) and is produced commercially by overexpression in *E. coli*. It is homologous to *E. coli* pol I, although it lacks 3'→5' exonuclease activity and therefore does not carry out proofreading. Other thermostable DNA polymerases do possess proofreading activity (e.g. *Pfu* polymerase from *Pyrococcus furiosus*).

**Eukaryotic DNA polymerases.** Eukaryotic cells contain four nuclear DNA polymerases and a fifth which is responsible for organelle genome replication (Table 26.6). The nuclear enzymes are DNA polymerases  $\alpha$ ,  $\beta$ ,  $\delta$  and  $\epsilon$ . DNA polymerases  $\alpha$  and  $\delta$  are responsible for chromosomal replication. DNA polymerase  $\alpha$  has an associated primase activity but no 3'→5' exonuclease activity, whereas DNA polymerase  $\delta$  has a proofreading capability. DNA polymerase  $\delta$  binds an accessory factor called **proliferating cell nuclear antigen (PCNA)**, a *cyclin* (q.v.) analogous to the *E. coli* pol III  $\beta$  subunit in that it acts as a sliding ring to increase enzyme processivity. DNA polymerase  $\delta$  synthesizes both the leading and lagging strands. The function of DNA polymerase  $\alpha$  is to extend the RNA primer on the lagging strand and provide the template for an accessory factor, **replication factor C (RF-C)** whose role may be analogous to that of the *E. coli*  $\gamma$ -complex, i.e. to load DNA polymerase and control the processivity of replication on the lagging strand. The role of DNA polymerase  $\epsilon$  is unclear. It is structurally very similar to DNA polymerase  $\delta$  but does not associate with PCNA. It may have a role in the replication fork, or it may be involved in DNA repair, like DNA polymerase  $\beta$ .  $\beta$  is the smallest of the five enzymes and the one with the lowest fidelity and processivity. DNA polymerase  $\gamma$  (similar to *E. coli* pol I) is responsible for the replication of mitochondrial DNA, and a similar enzyme has been isolated from plant chloroplasts.

**Table 26.6:** Properties and proposed functions of eukaryotic DNA polymerases

Mammalian name	$\alpha$	$\beta$	$\gamma$	$\delta$	$\epsilon$
Yeast name	pol1	pol4	polM	pol3	pol2
Yeast gene	<i>POL1</i>	<i>POL4</i>	<i>MIP1</i>	<i>POL3</i>	<i>POL2</i>
Location	Nuclear	Nuclear	Mitochondrial	Nuclear	Nuclear
Number of subunits	4	1	2	2	>1
5'→3' polymerase	✓	✓	✓	✓	✓
3'→5' exonuclease	x	x	✓	✓	✓
Primase	✓	x	x	x	x
Associated factors	None	None	None	PCNA	None
Processivity	Moderate	Low	High	High with PCNA	High
Function	Lagging-strand priming	Repair polymerase	Organelle polymerase	Principle replicative polymerase	Unknown

**DNA primases.** DNA primases are enzymes which synthesize RNA primers (q.v.) for DNA synthesis (they can use both NTPs and dNTPs as substrates and mixed primers are common). Primases perform two functions. Firstly, they initiate leading-strand synthesis at the origin (this needs to be done only once); secondly, they facilitate the repetitive initiation of Okazaki fragments during elongation. For cellular genomes, the same priming strategy is used for both functions (see section on primers and priming below).

In *E. coli*, DNA primase is encoded at the *dnaG* locus. For the replication of *oriC*-type replicons (e.g. the chromosome itself), the primase depends on the DnaB helicase for efficient priming activity. Primase and helicase form a functional complex, the **primosome**. Phage replicons have different primase requirements. The bacteriophage G4 origin is recognized by primase directly and there is no primosome. Replicons of the  $\phi$ X class (e.g. bacteriophage  $\phi$ X174) require a complex primosome, which contains six **prepriming proteins**: DnaB (helicase), DnaC, DnaT, PriA, PriB and PriC. PriA is required for recognition of the origin (or primosome assembly site, *pas*) but the precise functions of PriB, PriC and DnaT are unclear. These factors assemble to form a **preprimosome** recognized by the DnaG primase. In both types of origin, the primosome is a mobile complex which translocates along the DNA in an ATP-dependent manner in the same direction as fork movement (i.e. in the opposite direction to lagging-strand synthesis) laying down RNA primers for each Okazaki fragment.

In eukaryotes, two polypeptides associated with DNA polymerase  $\alpha$  possess primase activity. This is a fundamental difference between the eukaryotic and bacterial replisomes, i.e. a single principle replicative DNA polymerase in *E. coli* (pol III) extends both strands, whereas in eukaryotes a distinct enzyme extends the primers on the lagging strand, and only this needs to be associated with primase. Cellular primases and helicases may interact, but they are distinct enzymes. However, several eukaryotic viral encoded proteins have both helicase and primase activities.

**DNA helicases.** DNA helicases are enzymes which translocate along single strands of DNA and use energy derived from ATP hydrolysis to break hydrogen bonds and separate duplex molecules. Helicases are required for many cellular processes (q.v. e.g. *nucleotide excision repair*, *homologous recombination*, *transcriptional termination*, *conjugation*, *Ti plasmid*), and are essential during replication to provide single-stranded templates — they are the first components to join the replisome. Helicases usually translocate in one direction only along DNA and are classified according to their 5'→3' or 3'→5' polarity. In topologically constrained DNA, helicases work together with a topoisomerase to relieve torsional strain. The properties of some bacterial and eukaryotic replicative helicases are listed in Table 26.7.

**DNA topoisomerases.** DNA topoisomerases are enzymes catalyzing the interconversion of DNA topoisomers (different topological forms of DNA, see Nucleic Acid Structure: Box 16.2) and are required during replication to relax the torsional strain generated when helicases unwind the duplex; they perform similar functions during transcription and recombination. They also remove knots and resolve catenanes (interlocked circles) which arise during replication and recombination. In *E. coli*, many reactions involving nucleic acids (including replication, transcription, homologous recombination and the transposition of some mobile elements) are favored by negative *supercoiling* (q.v.). In eukaryotes, topoisomerases are part of the nuclear scaffold and facilitate the organization of DNA into functional domains (see Chromatin); they also play a major role in the separation of sister chromatids during mitosis.

Topoisomerases are divided into two functional classes (Table 26.8). The major difference is the reaction mechanism: class I topoisomerases cleave only one strand and can thus catenate/decatenate only substrates containing a nick; class II topoisomerases cleave both strands and can therefore catenate/decatenate covalently closed circles. Most topoisomerases catalyze the relaxation of supercoils, but cannot actively introduce supercoils into relaxed DNA. Exceptions are *E. coli* **DNA gyrase**, which can generate negative supercoils, and **reverse gyrase** isolated from *Sulfolobus acidocaldarius*,

**Table 26.7:** Replicative helicases and their functions

Helicase	Polarity and function
<i>Bacterial replicative helicases</i>	
DnaB protein	5'→3' polarity. Major helicase in <i>E. coli</i> chromosomal replication; <i>dnaB</i> has a quick-stop phenotype (Table 26.2) and is the only helicase required in <i>in vitro</i> systems
PriA protein (n', Y)	3'→5' polarity. Component of the primosome
Rep	3'→5' polarity. Absolutely required for rolling circle replication. Role in cellular replication unclear
<i>Eukaryote replicative helicases<sup>a</sup></i>	
DNA helicase A	Enzymes isolated from yeast and calf thymus. Copurifies with DNA polymerase $\alpha$ :primase
DNA helicase $\delta$	5'→3' polarity. Copurifies with DNA polymerase $\delta$
DNA helicase E	3'→5' polarity. Copurifies with DNA polymerase $\epsilon$
RF-A associated helicases	Several enzymes from different species, either copurify with RF-A or are stimulated by it
RF-C associated helicase	Copurifies with replication factor C

In bacteria, the roles of replicative and other helicases are well characterized. In eukaryotes, more than 30 helicase activities have been described and additional putative helicases have been identified in the yeast genome sequence, but few have had a precise function assigned.

<sup>a</sup>Helicases which have been copurified with replication proteins.

**Table 26.8:** Topoisomerases of *E. coli* and eukaryotes

	Class I topoisomerase	Class II topoisomerase
Cleavage	One strand	Both strands
$\Delta L$	Steps of 1	Steps of 2
Mechanism	Enzyme binds noncovalently to DNA. One strand cleaved and 3' phosphate group covalently attached to active site tyrosine residue. Intact strand passed through break. Broken strand religated.	Enzyme binds noncovalently to DNA. Both strands cleaved and the 5' phosphate groups covalently bound to active site tyrosine residues (this involves a 4 bp stagger for both gyrase and eukaryotic topoisomerase II). An intact duplex is passed through the gap (facilitated by a conformation change involving ATP binding). Broken duplex religated and DNA released.
ATP required	No	Yes
<i>E. coli</i>	topo I ( $\omega$ protein) topo III Relaxes negative supercoils	Gyrase (topo II). Introduces negative supercoils topo IV. Decatenates linked circles
Eukaryotes	Topoisomerase I. Relaxes positive and negative supercoils Topoisomerase III. Relaxes negative supercoils (weak activity), no decatenation	Topoisomerase II. Relaxes positive and negative supercoils Topoisomerase IV. Relaxes negative and probably also positive supercoils

$\Delta L$ , change in linking number (q.v.).

which can generate positive supercoils. The role of gyrase is likely to be to maintain DNA in an energetically favorable state. The function of reverse gyrase is less clear; it may concern keeping the chromosome in a duplex conformation at the high temperatures in which *S. acidocaldarius* lives (also q.v. *thermostable polymerase*).



**Table 26.9:** Properties of single-stranded DNA binding proteins in *E. coli* and eukaryotes

	<i>E. coli</i>	Eukaryotes
Protein	SSB (single-strand binding protein)	RP-A (replication protein A) or HSSB (human SSB)
Structure	Homotetramer of 19 kDa subunits	Heterotrimer of 70 kDa, ~30 kDa and ~15 kDa subunits
Function	Stabilizes single-stranded regions during replication, recombination and repair Directs priming to origin of M13-related genomes Associates with PriB in primosome complex (possible priming function)	Stabilizes single-stranded regions during replication, recombination and repair Interacts with pol $\alpha$ : primase to prevent nonspecific priming events Interacts with transcription factors, repair protein XP-A and several helicases (possible specific roles in transcription and repair?)
Properties	ssDNA-specific but no sequence specificity. Cooperative binding	ssDNA-specific, partial sequence specificity. Activity modulated by phosphorylation
Genes	<i>ssb</i>	<i>RP-A1, RP-A2, RP-A3</i>

**Single-stranded binding proteins.** Single-stranded binding proteins (SSBs) are replication accessory proteins lacking enzymatic activity, but required for efficient activity of other enzymes in the replisome (Table 26.9). SSBs perform many functions in the cell concerning the stability of single-stranded regions of DNA (e.g. q.v. *homologous recombination*; see also *hnRNP proteins*). In replication, this involves stabilizing the melted origin, sustaining the activity of helicases, removing secondary structures from the DNA template (q.v. *frameshift fidelity*) and the inhibition of nuclease activity. The proteins have a high affinity for single-stranded DNA but not for double-stranded DNA nor for RNA. They bind cooperatively to DNA and coat it with a protein polymer. There is some base composition preference but little sequence-specificity to the binding.

SSBs also interact directly with various components of the replisome to stimulate their activity. In both *E. coli* and eukaryotes, the SSB interacts with primase or components of the primosome to facilitate specific priming activity. *E. coli* SSB also directs the priming of single-stranded genomes by covering all available DNA except the origin, which is characterized by a hairpin secondary structure. The SSBs encoded by bacteriophages T4 and T7 directly stimulate the activities of the phage-encoded DNA polymerases.

**Nucleases.** Nucleases are enzymes which digest nucleic acids by hydrolyzing phosphodiester bonds. Nucleases are **deoxyribonucleases (DNases)** if their substrate is DNA and **ribonucleases (RNases)** if their substrate is RNA. **Exonucleases** require a free end and digest the molecule stepwise, whereas **endonucleases** are able to hydrolyse internal phosphodiester bonds and can therefore use a covalently closed circular template as a substrate. **Excinucleases** facilitate the release (excision) of an oligonucleotide fragment. Nucleases vary widely in their substrate specificity. Exonucleases may be single- or double-strand-specific, may have a specific polarity, and may require a particular end structure to initiate digestion. Endonucleases may also possess single- or double-stranded substrate preference, and may demonstrate various degrees of sequence specificity; a **nickase** is an endonuclease which cleaves one strand of a duplex. Nucleases play roles in many systems (e.g. q.v. *restriction endonucleases*, *UvrABC nuclease*, *AP endonucleases*, *RecBCD nuclease*, *retroviruses*, *conjugation*, *cDNA synthesis*, *recombinant DNA*, *RNA processing*, *transposase*, *integrase*, *ribozyme*, *spliceosome*).

Several sources of nuclease activity are required during replication. Generally, the recycling of nucleotides by DNases and RNases provides substrates for both replication and transcription (**salvage pathways**), and mutations in genes controlling these pathways generate *quick-stop* phenotypes

(q.v.). Three specific nuclease activities are needed during replication. Proofreading requires a 3'→5' exonuclease activity, which is intrinsic to the DNA polymerase enzymes (discussed above). Primer removal requires the combined activity of RNaseH (which specifically digests the RNA strand from a DNA:RNA hybrid) and a 5'→3' exonuclease activity. In *E. coli*, the latter activity is intrinsic to the principle repair enzyme, DNA polymerase I, whereas in eukaryotes it is supplied by the exonuclease MF-1. RNaseH is not sufficient to remove primers in *E. coli*, and presumably in eukaryotes, because it is unable to cleave phosphodiester bonds linking ribonucleotides to deoxyribonucleotides. RNaseH digestion of primers thus leaves at least a single ribonucleotide residue attached to the Okazaki fragment, and because of the promiscuity of DNA primase, there may be several internal ribonucleotide residues needing to be excised. RNaseH also processes the leading-strand primer for ColE1 plasmid replication (see Plasmids).

**DNA ligases.** DNA ligases are enzymes catalyzing phosphodiester bond formation between adjacent nucleotides in double-stranded DNA. The 5' nucleotide must have an intact phosphate group and the 3' nucleotide an intact hydroxyl group. In bacteria, DNA ligases require a NAD cofactor, whereas in eukaryotes and archaea, ligases require ATP. In both cases, the cofactor supplies an adenylate group which becomes covalently linked to the active site of the enzyme. This group is then transferred to the 5' nucleotide which is subsequently attacked by the 3' hydroxyl group of the adjacent nucleotide to form a phosphodiester bond. DNA ligases control the final stage in all DNA repair pathways, the sealing of nicks remaining on one strand of a double-stranded DNA. Mammals appear to possess at least four ligase activities, DNA ligase I being the principle enzyme involved in lagging-strand repair (also q.v. *bacteriophage T4 DNA ligase*).

**Comparison of the bacterial and eukaryotic cellular replisomes.** Important parallels exist between the structures and functions of the bacterial and eukaryotic replisomes: their asymmetrical organization, the presence of a highly processive polymerase and a distributive polymerase to synthesize leading and lagging strands, respectively, and the similarity of their enzymology. There are also important distinctions, e.g. the strategies for priming and repairing the lagging strand. These properties are compared in Table 26.10. Bacteria and eukaryotes also differ significantly with respect to how DNA replication is organized in the context of the cell division cycle — see The Cell Cycle for discussion.

### 26.3 Initiation of replication

**General principles of initiation.** Genome replication begins at one or more *cis*-acting sites termed **origins of replication** (or simply, **origins**). These increase the efficiency of initiation by providing a site for the assembly of the protein factors and enzymes required for replication, which would otherwise bind randomly to DNA. The initial, static assembly of proteins, an **orisome**, becomes a replisome when it begins to move away from the origin. Origins also provide a target for the regulation of replication, and thus initiation is the predominant stage of replication control. The initiation of replication is controlled by dedicated initiation factors, but may be influenced by other proteins and by properties of the DNA itself. Origins can be identified physically by examining labelled replication intermediates using electron microscopy (Box 26.2), by labelling and isolating nascent DNA chains, or by separating DNA structures by 2D electrophoresis. Like other *cis*-acting elements, origins can be functionally mapped by *in vitro* mutagenesis (q.v.).

In bacterial chromosomes and the genomes of viruses and plasmids, a single origin is used to initiate replication<sup>1</sup>. Such elements thus constitute a single **replicon**, which is defined as a unit of

<sup>1</sup>Note that many plasmids and viruses may possess several alternative origins, but only one is used per replication cycle. Double-stranded DNA genomes which undergo continuous replication (e.g. mitochondrial DNA) possess two origins, one on each strand, and may be thought of as two replicons

**Table 26.10:** Components and properties of the bacterial and eukaryotic replisomes

Replisome property/component	Bacteria	Eukaryotes
<i>Comparison of general properties</i>		
Origins	Single	Many
Rate of elongation	100 kbp min <sup>-1</sup>	2 kbp min <sup>-1</sup>
Okazaki fragments	1–2 kbp	100–200 bp
Priming strategy	Primase generates primer Extended by pol III	Primase generates primer Extended by pol $\alpha$ Completed by pol $\delta$
Replicative polymerase	Heterodimer with monomers of differing processivity	Separate enzymes for each strand, with differing processivities
Topology	Balance between relaxing and winding, net negative supercoil	Negative supercoil constrained in nucleosome structure. Higher order structure influences topology
<i>Comparison of replisome components</i>		
Replicative polymerase	pol III holoenzyme	pol $\delta$ /pol $\alpha$
Processivity factor (clamp)	$\beta$ subunit	PCNA
Clamp loading factor	$\gamma$ complex	RF-C
Primase	DnaG	pol $\alpha$ : primase
Helicase	DnaB (DnaC required for loading)	? several candidates
Primer removal	RNaseH and pol I	RNaseH1 and MF-1 nuclease
Lagging-strand repair	pol I and DNA ligase	pol $\delta$ /pol $\epsilon$ and DNA ligase I
Topoisomerase	DNA gyrase	topo II
Single-strand binding	SSB	RF-A

**Table 26.11:** Modes of replication, reflecting origin location and the nature of helicase loading

Replication mode	Mechanism	Example
<i>Semidiscontinuous mechanisms</i>		
Coupled bidirectional	Helicases loaded onto both strands and two replication complexes are assembled	<i>oriC</i>
Sequential bidirectional	One helicase loaded at initiation and promotes unidirectional replication until a second priming site is uncovered which can initiate in the opposite direction	R6K plasmid
Unidirectional	Only one helicase is loaded at initiation and discontinuous strand synthesis thus arrests at the origin	ColE1 plasmid
<i>Displacement mechanisms</i>		
Unidirectional displacement	Origin is melted then cleaved by an endonuclease before helicase is loaded	pT181

replication under common *cis*-control. The larger chromosomes of eukaryotes have many origins and may be thought of as tandemly arranged replicons.

**The nature of origins and initiation strategy.** The mode of replication (Table 26.11) depends on the nature and distribution of origins and initiation strategy. In double-stranded DNA genomes, the origin is where the duplex is initially unwound. In most cases, unwinding is followed by primer synthesis, although for rolling-circle replication, one strand is nicked to provide a primer terminus.

For single-stranded DNA genomes, neither unwinding nor nicking are necessary and the origin is simply the site of primer synthesis. In each case, the precise position of the origin corresponds to the beginning of nascent DNA synthesis and is therefore where the primer terminus is generated. In practice, however, the origin often cannot be defined exactly: primer length varies over a short range (8–10 nucleotides) and, because of the promiscuity of DNA primase, the primer may be composed of mixed ribonucleotide and deoxyribonucleotide subunits. In RNA replication, the origin is analogous to a *promoter* (q.v.), i.e. a site where RNA replicase can bind to its substrate and initiate synthesis.

For the semidiscontinuous replication of double-stranded DNA genomes (all cellular genomes and those of many plasmids and viruses), initial opening of the duplex at the origin establishes either one or two replication forks. Unwinding is usually mediated by a specific initiation protein which recruits helicase and primase, although in some systems (e.g. ColE1-related plasmids) RNA polymerase performs this role as well as that of primer synthesis. Helicase is required for continued unwinding of the parental DNA, and an important function of initiator proteins is to recruit helicase to the replication fork (occasionally, the initiator protein itself has helicase activity, e.g. the SV40 T antigen). At most origins, helicases are loaded onto both strands to give two replication forks, and replication is bidirectional. In some plasmid systems, a single helicase is loaded and replication is unidirectional. Usually, the leading strand is primed and extended for some distance before the first Okazaki fragment is primed on the retrograde template.

For continuous replication, a single priming event is required (one on each strand for double-stranded genomes). The priming site of a single-stranded DNA genome is often recognized by its secondary structure, or for linear plasmid and virus genomes, the origin is the 3' end of each strand. In double-stranded genomes, a displacement fork is established which allows continuous strand synthesis in one direction, progressively displacing a resident parental strand. At some stage, a further origin on the displaced strand is revealed, allowing second-strand synthesis to begin. This strategy is used during the replication of mitochondrial DNA and the replication of adenovirus and several related linear plasmids with terminal proteins.

**Bacterial origins of replication.** The origin of the *E. coli* chromosome<sup>1</sup> is designated *oriC* and is essential for normal replication. The minimal sequence is 245 bp long, and contains a series of tandemly repeated elements, or *iterons*, comprising four nonomers and three 13-mers. The nonomers are binding sites for the DnaA initiator protein and are spread over a 'recognition site'; the three 13-mers are located together in an adjacent region which is generally AT-rich. Mutant alleles of *dnaA* have a slow-stop phenotype, indicating that the function of the DnaA protein is restricted to initiation. DnaA binds cooperatively to its recognition site in an ATP-dependent manner, so that 30 DnaA molecules are eventually present. DnaA binding promotes duplex unwinding at the three 13-mers which have low torsional stability, and the DnaBC helicase complex is then loaded, a process stimulated by RNA polymerase acting at two nearby promoters (also q.v. *bacterial artificial chromosome*). This initiation sequence is summarized in Table 26.12.

A striking feature of *oriC* is the presence of 14 Dam methylation sites. Hemimethylated origins, which would arise directly following the initiation of replication, are unable to promote reinitiation and bind to components of the cell membrane. It is possible, therefore, that *DNA methylation* (q.v.) contributes to the control replication timing during cell division. The structural arrangement described above is strongly conserved in the origins of other Gram-negative bacteria. Both the sequence of the iterons and their spatial relationship appears to be important. In Gram-positive bacteria, the sequences are different but the overall architecture remains the same.

The organization of rolling-circle origins is distinct from that of  $\theta$  origins although they possess binding sites for initiation proteins and AT-rich elements which are melted. The melting uncovers a

<sup>1</sup>Bacterial plasmid and bacteriophage replication origins are discussed elsewhere — see Plasmids, Viruses.



**Table 26.12:** Stages in the initiation of replication in *E. coli*

Initiation stage	Characteristics
Initial complex	DnaA binds to nonmer DnaA boxes No ATP required
Open complex	DNA melted in 13-mer region Requires ATP, and HU protein which controls DNA structure (q.v. <i>nucleoid</i> )
Prepriming complex	DnaBC helicase loaded. DnaC released from the complex
Priming complex	DNA unwound by DnaB. Primase and pol III added
Elongation	Topoisomerase and SSB proteins recruited, elongation begins

**Table 26.13:** Components of simple eukaryotic origins

Origin component	Function
<i>ori-core</i>	
Origin recognition element (ORE)	Site for initial binding of initiation factors — facilitate DNA unwinding and loading of replisome components
DNA unwinding element (DUE)	Site where DNA is unwound and where replication machinery enters the duplex
AT-rich element	A motif associated with most simple origins which has T-rich and A-rich strands. Easily undergoes bending which may facilitate DNA melting at DUE and/or interaction between initiation proteins and the ORE
<i>ori-aux</i>	
Auxiliary elements	Transcription factor binding sites which enhance the efficiency of replication initiation if bound by transcription factors. It is possible that the control of replication using transcription factors allows replication to be regulated in a cell-type-specific and developmentally programmed fashion (and differential activation could control early and late replication of transcriptionally active and repressed chromatin). Transcription factors could enhance replication by facilitating the binding of initiator proteins, by altering their activity, or by modulating chromatin structure or the structure of DNA. In mitochondrial DNA, they may support transcription through the origin to generate a primer

site for strand cleavage, so generating the primer terminus. Despite this difference, the overall initiation process is similar in all bacterial replicons, involving origin recognition and unwinding followed by loading of helicase.

**Origins in simple eukaryotic systems.** Origins in simple eukaryotic systems (plasmids, viruses and unicellular eukaryotes) are similar to prokaryotic origins in their modular nature (Table 26.13). Typically, they comprise a core sequence (*ori-core*) containing an origin recognition element, a DNA unwinding element and an AT-rich element. Outside the core there are one or more auxiliary elements which act as binding sites for transcription factors. These are thought to increase the efficiency of replication initiation in the same way that they affect transcriptional initiation, i.e. by interacting with core components of the system, in this case the **orisome** (q.v. *basal transcriptional apparatus*). Simple origins can function as **autonomously replicating sequences (ARSs)**, i.e. sequences which when linked to any other fragment of DNA confer upon it the ability to replicate autonomously in the cell from which the ARS was derived (q.v. *yeast cloning vectors*).

**Origins in chromosomes of multicellular organisms.** In contrast to simple origins, origins in metazoan chromosomes have been difficult to characterize. Functional studies have suggested that the length of the DNA, rather than any specific sequence, is the critical factor for replication. However, the physical analysis of origins by identifying sequences at replication initiation by PCR has shown specific sequences to be involved. The site at which DNA synthesis begins, a **primary origin** (or **origin of bidirectional replication** — most primary origins are bidirectional), may be flanked by a number of **secondary origins**. The primary origin is usually 0.5–2 kbp in length, whereas secondary origins may span up to 50 kbp, delineating the so called **initiation zone**.

Of the known metazoan origins of replication, only some demonstrate ARS activity when subcloned in a plasmid, and it is thought that chromatin structure may therefore play an important role in origin function. A number of transcription factor binding sites are usually found flanking origins, although transcription itself is not an essential prerequisite for initiation. Rather, the factors provide an open chromatin domain for initiation factor assembly, and may also directly assist the assembly process — this is directly analogous to their function in transcription. Potential DNA unwinding elements have been found in several origins, although the AT-rich elements of simple origins have not been described. There are sites with hyphenated dyad symmetry which have the potential to form cruciform structures whose function is unclear. There are also sites for attachment to the *nuclear matrix* (q.v.).

A striking feature recently described for two hamster origins is a **densely methylated island** (DMI) of 100–500 bp where all the deoxycytidine residues of both strands are methylated. This is not the typical CpG methylation pattern observed in mammals, as the context of the methylated cytosines appears unimportant (c.f. *CpG island*). The role of methylation in these origins is unclear, but it may influence chromatin structure or transcription factor binding, or it may regulate replication timing, as in bacteria.

## 26.4 Primers and priming

**The rationale for priming.** The requirement for priming in DNA synthesis may reflect the evolution of proofreading. Proofreading is facilitated by the inability of DNA polymerases to extend a primer that is not paired with the template, but at the very beginning of strand synthesis, there is no primer to extend so the enzyme would have to tolerate some inaccuracy in order to begin synthesis at all. For leading-strand synthesis, mutations would accumulate at the origin. However, for lagging-strand synthesis, a 'primerless' mechanism would introduce periodic mutations throughout the genome. Priming initiation with a short stretch of RNA allows the replisome to identify these inaccurate regions where initiation has occurred, and replace them with DNA when a DNA primer terminus becomes available. Priming thus preserves the integrity of the genome.

**Priming strategies.** RNA primers synthesized by DNA primase are used to prime both leading- and lagging-strand synthesis in all cells. The primer, **initiator RNA** (iRNA), is usually about 10 nucleotides long. In prokaryotes, the primer is directly extended by DNA polymerase III, whereas in eukaryotes, 4–5 deoxyribonucleotide residues are added to each primer by DNA polymerase  $\alpha$  (this is termed **initiator DNA**, iDNA), and the DNA primer terminus is extended by DNA polymerase  $\delta$ . Short RNA primers synthesized by DNA primase are the universal priming mechanism for lagging-strand synthesis. Noncellular replicons, however, have developed a series of alternative strategies for initial priming of the leading strand (Table 26.14).

## 26.5 Termination of replication

**Completing the replication of circular templates.** The replication of circular genomes may terminate at a specific sequence, a **terminus region** (defined *ter*), as for the *E. coli* chromosome and some plasmids. Alternatively, replication may simply cease when two replication forks approaching each

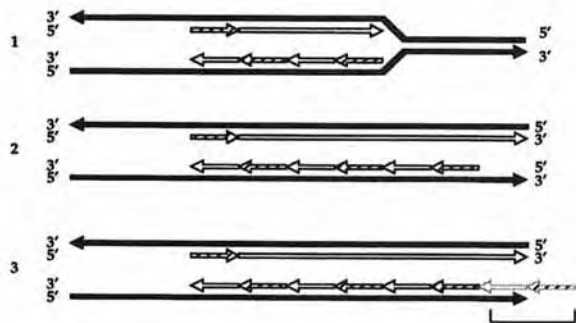
**Table 26.14:** Mechanisms for priming leading-strand synthesis at an origin of replication

Priming mechanism	Details	Example
RNA primer	Short nascent oligoribonucleotides synthesized by DNA primase	Cellular DNA
	Specific transcripts synthesized by RNA polymerase	Mitochondrial DNA, ColE1 plasmid
	Annealing of a preformed tRNA primer	Retroviral reverse transcription
Hairpin-priming	Single-stranded template with inverted terminal repeats may fold back to form a hairpin	Parvovirus
Endonucleolytic priming	Endonuclease cleavage of a dsDNA generates a nick. The exposed 3'-OH group can be used as a primer terminus	Rolling-circle replication of many Gram-positive bacterial plasmids
Invasive priming	A 3' end may be introduced into an intact duplex by homologous recombination	T4 late replication
Terminal protein priming	Some linear replicons have terminal nucleotide-binding proteins which facilitate priming	Adenovirus
Oligonucleotide priming	Anneal oligonucleotides to a template for <i>in vitro</i> extension	<i>Polymerase chain reaction, random priming, cDNA synthesis, etc. (q.v.)</i>

other from opposite directions meet, as occurs in the SV40 genome. The terminus region in the *E. coli* chromosome comprises two pairs of inverted repeats. These bind a protein factor, Tus, which blocks the advance of the replication fork, a process facilitated by inhibiting helicase activity. The terminus region divides the replicating *E. coli* genome into two halves, termed **replichores**, each beginning at the bidirectional origin and ending at the terminus. The ability of *ter*-Tus to block replication is orientation-dependent. The advantage of a specific termination strategy is unclear: deletion of the *ter* site or *tus* gene in *E. coli* has no effect. Termination may provide an opportunity for regulating the decatenation of interlocked rings.

**Completing the replication of linear templates.** Linear DNA genomes require special strategies to complete their 5' ends. DNA polymerases, which can only synthesize DNA in the 5'→3' direction and require a preformed primer, cannot complete the extreme 5' ends of each lagging strand because there is nothing upstream for the primer to bind (*Figure 26.4*). Without special strategies for completion, the chromosome would shorten with each round of replication. Circular genomes have an advantage in this respect, and many linear replicons solve the termination problem by adopting a circular conformation for part of their replication cycle (e.g. bacteriophage  $\lambda$ ). A number of further, elegant strategies have evolved for completing lagging-strand synthesis, and these are listed in *Table 26.15*.

Eukaryotes solve the replication problem by adding preformed blocks of nucleotides to the chromosome ends to form *telomeres* (q.v.). The telomere repeat sequences are added by a specialized enzyme, *telomerase* (q.v.), to maintain chromosome length. The recently generated telomerase *knock-out mouse* (q.v.) is phenotypically normal, but mutations begin to show their effects after six generations, when the existing telomeres are eventually fully deleted. As well as the effect of deleting terminal genes, the mutant mice also show chromosome mutations, e.g. fusions and translocations, reflecting another function of telomeres — to distinguish natural chromosome ends from random breaks. For further discussion of telomeres, see *Chromosome Structure and Function*.



**Figure 26.4:** The replication problem at the end of linear templates. (1) A replication fork proceeding towards the right of the diagram has a leading strand (upper) and a lagging strand (lower). Parental strands are black, nascent strands are white, primers are hatched; note that there is a single primer on the leading strand and multiple primers for the lagging strand. (2) The 3' end of the leading strand reaches the 5' end of the template. Meanwhile, the final lagging-strand primer is synthesized, priming the final Okazaki fragment on the lagging strand. (3) There is nothing for the next primer to bind, so the final segment of lagging strand is never made. The chromosome remains as shown in (2).

**Table 26.15:** Strategies for completing the 5' ends of linear genomes

Strategy	Mechanism and example
Concatenation	Linear genomes possess redundant termini which allow circularization or concatenation. These structures can be cleaved to generate single genomes with 5' overhanging termini which can be filled by conventional DNA synthesis, e.g. bacteriophages T7 and $\lambda$
Terminal protein priming	Viruses which initiate strand synthesis with terminal proteins do not need a specific termination mechanism: they initiate at the 5' end of each strand and can complete up to the 3' end, e.g. adenovirus, bacteriophage $\phi 29$
Hairpin priming	Another priming strategy which initiates strand synthesis from the extreme 5' end of the strand, e.g. parvoviruses
Covalently sealed ends	Some viruses, which are superficially double stranded and linear, have covalently sealed ends so that melting generates a single-stranded circle which can be replicated like a circular replicon, e.g. viroids
Telomeres	Enzymes termed telomerases add oligonucleotides to the ends of linear chromosomes. Although extreme 5' sequences are lost in this method, post-replicative telomerase activity can replenish the telomeres so that no actual genes are lost in successive round of replication (see Chromosome Structure and Function)

26.6 The regulation of replication

**Temporal control of replication.** All cells coordinate genome replication with cell division to prevent gain or loss of DNA (*see* The Cell Cycle), whilst the replication of plasmids is often copy-number-dependent (*see* Plasmids). Given the much slower rate of elongation and the much larger genome of eukaryotes, rapid replication is increased by increasing the number of functional origins per chromosome (e.g. during the cleavage stage of development in *Xenopus*). Replication in eukaryotes is temporally controlled so that not all origins are activated simultaneously and different regions of the genome are replicated in a regulated temporal sequence. The temporal control of replication broadly correlates with genetic activity, i.e. housekeeping genes replicate early whereas cell-type-specific genes generally replicate early in the cells where they are expressed and late in cells where



they are transcriptionally silent. The temporal control is developmentally regulated: many cell-type-specific genes are late-replicating in the embryo but switch to early replication when the appropriate cell type differentiates; X-linked genes switch from early to late replication when the X-chromosome is inactivated (q.v. *X-inactivation*).

The regional nature of replication timing can be investigated by pulse-labeling chromosomes during the S phase. Labeling early replicating DNA with bromodeoxyuridine produces a pattern of bands which correlates with light G-bands and DNase I-sensitive regions (q.v. *chromosome banding*, *isochore*). Early replicating DNA thus appears to identify gene-rich, active areas of the chromosome. High resolution banding of prometaphase chromosomes reveals a pattern of sub-bands dividing each major band into regions 1–2 Mb in length. Replicons in mammalian chromosomes are 50–300 kb in length, suggesting that each **replication time zone** contains a number of synchronized origins, and this has been confirmed directly by observing tandem origins by autoradiography.

Translocations have shown that the timing of replication is not an intrinsic property of any particular origin, but is position-dependent. Therefore, temporal control may be mediated by *cis*-acting sites. Several candidate elements have been identified, including telomeric sequences and the *locus control region* of the  $\beta$ -globin locus (q.v.). In mammals, replication forks with similar timing are clustered together in the nucleus as **replication foci**, probably due to common interaction with the *nuclear matrix* (q.v.). A relationship between replication timing, chromatin structure, DNA methylation and transcriptional activity has thus been established. However, the causal relationships between the components are not fully understood.

#### Box 26.1: Nucleic acid synthesis

**Template-directed nucleic acid synthesis.** In **template-directed synthesis**, a preexisting nucleic acid acts as the template for the assembly of nucleotides to form a new strand, and dictates the order in which the bases are incorporated by complementary base pairing. This is the basis of *information transfer* between nucleic acids. There are two types of nucleic acid, DNA and RNA, and there are therefore four types of nucleic acid synthesis reaction, each catalyzed by a specific **template-dependent nucleic acid polymerase** (or **synthetase**), as shown below. These enzymes are unique in that the template as well as the enzyme itself determines substrate specificity.

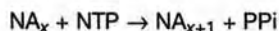
As used in the table, the terms DNA polymerase and RNA polymerase are misleading. A **DNA polymerase** is strictly any enzyme with DNA synthesis activity (i.e. the term encompasses both reverse transcriptases and terminal deoxynucleotidyl transferases as well as the DNA polymerases involved in DNA replication). Similarly an **RNA polymerase** is any enzyme with RNA synthesis activity (including transcription enzymes, viral RNA replicases, DNA primase and polyadenylate polymerase). DNA polymerase may also be called **DNA replicase**, and RNA polymerase may also be called **transcriptase**. Individual enzymes may possess multiple polymerase activities, e.g. most

Copying process	Enzyme: systematic nomenclature (common name)
DNA→DNA ( <b>DNA replication</b> )	DNA-dependent DNA polymerase ( <b>DNA polymerase</b> )
RNA→RNA ( <b>RNA replication</b> )	RNA-dependent RNA polymerase ( <b>RNA replicase</b> )
DNA→RNA ( <b>transcription</b> )	DNA-dependent RNA polymerase ( <b>RNA polymerase</b> )
RNA→DNA ( <b>reverse transcription</b> )	RNA-dependent DNA polymerase ( <b>reverse transcriptase</b> )

*Continued*

reverse transcriptases possess additional DNA-dependent DNA polymerase activity.

**Polymerase action.** All template-directed nucleic acid synthesis reactions involve a structure termed a **template primer**. The primer is the nascent strand, and is so called because it primes its continued elongation by providing a terminus for extension; the template strand supplies the information for primer elongation. The substrates for the reaction are nucleoside triphosphates: elongation involves nucleophilic attack by the terminal 3' hydroxyl group of the primer strand on the  $\alpha$ -phosphate group of the incoming nucleotide. This forms a 5'→3' phosphodiester bond and eliminates pyrophosphate, as shown.



Addition of excess pyrophosphate to an *in vitro* system causes reversal of the reaction (**pyrophosphorylisis**). *In vivo*, pyrophosphate is removed from the cell by conversion to inorganic phosphate, making the reaction essentially irreversible. Extension by formation of sequential 5'→3' phosphodiester bonds thus causes chain growth exclusively in the 5'→3' direction, which is known as **tail growth**. The polymerase reaction may be **processive** (if many polymerization steps occur without release of the enzyme) or **distributive** (if the enzyme dissociates after every addition). Polymerases are often found in association with other proteins which enhance processivity by modulating enzyme or template structure.

**Differences between DNA and RNA synthesis.** Although all nucleic acid synthesis reactions are similar in mechanism, DNA synthesis differs from RNA synthesis in three important aspects. Firstly, different substrates are used — deoxyribonucleoside

triphosphates for DNA synthesis and ribonucleoside triphosphates for RNA synthesis. Secondly, at initiation, RNA polymerase reactions can begin *de novo*, i.e. the enzyme can insert the first nucleotide opposite the template strand without a preexisting primer. The first nucleotide in an RNA strand thus retains its 5' triphosphate moiety, and acts as the primer for further elongation. Conversely, DNA polymerase reactions cannot begin *de novo* and require a preexisting primer. All DNA polymerases (including reverse transcriptases) need a primer, with the single exception of mitochondrial reverse transcriptase from *Neurospora crassa*. In cells, the primer is a short RNA molecule synthesized by a dedicated enzyme within the replication center. For other genomes, a variety of priming strategies are employed (Table 26.14). Primers increase the fidelity of DNA replication but complicate the completion of the 5' end of linear products (see main text for further discussion). Finally, because DNA synthesis is primarily a means of genome duplication, whereas RNA synthesis is primarily a means of gene expression, many (but not all) DNA polymerases possess associated exonuclease activities which allow proof-reading — the identification and removal of mispaired nucleotides at the primer terminus. RNA polymerases do not possess this activity.

**Untemplated nucleic acid synthesis.** In **template-independent synthesis**, nucleotides are added to the ends of preexisting nucleic acids without template instructions. Such reactions fall into two classes: specific reactions, where the substrate is chosen by the enzyme itself, and nonspecific reactions, where any nucleotides can be used generating random sequences. Some template-independent enzymes and their cellular functions are listed below.

**Terminal deoxynucleotidyl transferase (TdT)**

Responsible for adding extra nucleotides into the junctions during V-D-J recombination (see Recombination); adds deoxyribonucleotides nonspecifically to DNA with terminal 3' hydroxyl groups (also q.v. *labeling*, *homopolymer tailing*)

**Polyadenylate polymerase (PAP)  
mRNA guanylyltransferase**

Responsible for adding adenosine residues to the 3' end of eukaryotic mRNAs (q.v. *polyadenylation*)

Responsible for adding 7-methylguanosine cap to the 5' end of eukaryotic mRNAs (q.v. *capping*)

**Polynucleotide phosphorylase**

Bacterial nonspecific ribonucleotide polymerase. Unique in using nucleoside diphosphates rather than triphosphates as substrates

**Box 26.2: Replication strategy and replication intermediates**

**Replication intermediates.** Replication intermediates are the structures observed when a genome is partially replicated (c.f. *replicative intermediate*). These can be seen by incorporating radioactive nucleotides into replicating DNA, then isolating the replicon and observing the structure by autoradiography. The analysis of replication intermediates allows the mechanism of replication to be determined, and can also be used to map origins of replication.

**$\theta$  and  $\sigma$  structures.** Circular replicons of dsDNA can form four major types of intermediate, divided into  $\theta$ -type and  $\sigma$ -type structures.  **$\theta$ -type structures** are generated by initiation at internal origins. The classic  **$\theta$ -structure** resembles the Greek letter  $\theta$  and is indicative of semidiscontinuous replication, which may be termed  **$\theta$ -replication** or **Cairns replication**. A similar structure is formed by internal displacement replication, although the displaced strand is unlabeled so the intermediate appears as a simple circle during the first round of replication and a  $\theta$ -structure in the second round. This is a **displacement loop (D loop)**, similar to that seen

during recombination with a free DNA strand (see Recombination).  **$\sigma$ -type structures** also come in two varieties. The classic  **$\sigma$ -structure** resembles the Greek letter  $\sigma$  and is indicative of rolling-circle replication, with the tail of the letter representing the extruded concatemeric strand. Rolling-circle replication is thus sometimes termed  **$\sigma$ -replication**. In some replicons undergoing rolling-circle replication, the displaced strand remains attached to the nascent strand by a protein, and forms a double-loop intermediate, a **ariat structure**.

**Intermediates in linear replicons.** Early replication by internal initiation generates a **replication eye** or **replication bubble** whether the replicon is circular or linear (c.f. *transcription bubble*). In linear replicons, however, internal bubbles can meet to generate larger bubbles, or they can reach the end to generate a **Y-structure**. A Y-structure is also generated by internal displacement as one resident strand is peeled off, although like D-loops, these are not seen during first round replication because the displaced strand is unlabeled.

**References**

- Baker T.A. and Wickner S.H. (1992) Genetics and enzymology of DNA replication in *E. coli*. *Annu. Rev. Genet.* 26: 447–477.  
 De Pamphilis M.L. (Ed.) (1996) *DNA Replication in Eukaryotic Cells*. Cold Spring Harbor Press, New York.  
 Kornberg A. and Baker T.A. (1992) *DNA Replication*. W.H. Freeman, New York.

**Further reading**

- Chong, J.P.J., Thommes, P. and Blow, J.J. (1996) The role of MCM/P1 proteins in the licensing of DNA replication. *Trends Biochem. Sci.* 21: 102–106.  
 Sousa, R. (1996) Structural and mechanistic relationships between nucleic acid polymerases. *Trends Biochem. Sci.* 21: 186–190.  
 Wold, M.S. (1997) Replication protein A: A heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Annu. Rev. Biochem.* 66: 61–92.  
 Wynford-Thomas, D. and Kipling, D. (1997) Telomerase: Cancer and the knockout mouse. *Nature* 389: 551–552.

**This Page Intentionally Left Blank**



## Chapter 27

# RNA Processing

### Fundamental concepts and definitions

- **RNA processing** describes the structural and chemical maturation of newly synthesized RNA molecules. The modifications occur during transcription (**cotranscriptional modification**) and afterwards (**posttranscriptional modification**); they may be essential for RNA function or may represent a mechanism of gene regulation. Such reactions fall into ten categories, as shown in Table 27.1.
- An RNA molecule copied from a DNA template is a **transcript**. During transcription, it is a **nascent transcript**, and when transcription is complete it is a **primary transcript** — an exact copy of the DNA from which it was transcribed. After any modification, the product is a **mature transcript** and may no longer be an exact copy of the DNA.
- Not all RNA is modified: bacterial mRNAs are rarely processed (indeed protein synthesis usually initiates before transcription is complete), and eukaryotic 5S rRNA is transcribed as a molecule with mature ends. The precursor of fully processed and functional RNA is termed **pre-RNA**. Eukaryotic **pre-mRNA** is also described as **heterogeneous nuclear RNA (hnRNA)** because, unlike the other forms of nuclear RNA, it has a great size diversity, reflecting varying gene sizes and the presence of partially processed molecules.
- In eukaryotes, transcription and RNA processing do not occur 'free' in the nucleus, but are localized at discrete foci in the nuclear matrix. Certain mRNA processing factors (specifically the spliceosome and polyadenylation enzymes) are attached to the C-terminal tail of RNA polymerase II in the elongation complex, so that transcription and processing are directly linked. The localization of mRNA processing complexes in the nuclear matrix may facilitate export. RNA is associated with proteins in the nucleus and (if appropriate) the cytoplasm, to form ribonucleoprotein particles.

**Table 27.1:** Categories of RNA processing reactions

Processing reaction	Examples
Cleavage	Release of rRNA and tRNA from polycistronic transcripts Termination of eukaryotic mRNA transcription
Exonucleolytic degradation	Processing rRNA and tRNA to generate mature ends
Nucleotidyl transfer	Transfer of CCA trinucleotide to 3' end of some tRNAs
Chemical modification of bases	Occasional methylation in mRNA and rRNA Extensive base modification in tRNA and snRNA
Nucleotide excision and replacement	Hypermethylation of guanosine residues in tRNA to produce queuosine and wyosine
Capping	Addition of 7-methylguanosine to 5' end of eukaryotic mRNA
Polyadenylation	Addition of polyadenylate tail to 3' end of most eukaryotic mRNAs and a few bacterial mRNAs
Splicing (transesterification)	Removal of most introns (usually in <i>cis</i> , but occasionally in <i>trans</i> )
Splicing (ligation)	Removal of tRNA introns
Editing	Changing the information carried in mRNA by base modification, insertion and deletion of residues from the coding region

### 27.1 Maturation of untranslated RNAs

**tRNA cleavage and maturation.** Some tRNA genes may be transcribed singly, others as part of a polycistronic unit. In *E. coli*, some tRNA genes are cotranscribed with rRNA genes in a common

operon. Depending upon its source, **pre-tRNA** may be subject to up to seven distinct processing reactions. In polycistronic tRNA, or mixed tRNA and rRNA transcripts, individual tRNAs are released by cleavage at the mature 5' end (usually a guanidylate residue). In *E. coli* this is the function of ribonuclease P. The immature 3' end is processed by exonucleolytic degradation (possibly involving ribonuclease D) until the trinucleotide motif CCA is reached. If there is no CCA motif, the trinucleotide is transferred to the 3' end of the molecule by a **tRNA nucleotidyltransferase** (the product of the *E. coli* *cca* gene). All mature tRNAs end with CCA. In eukaryotes and the archaea, some tRNA genes contain introns which must be spliced out, which involves cleavage, end-modification and RNA ligation (q.v. *tRNA nuclear introns*). Pre-tRNA is subject to further processing in the form of base modification: most of the bases in tRNA are *major bases* (q.v.), but approximately 10% become modified during tRNA synthesis, usually by posttranscriptional chemical modification *in situ*. The hypermodified guanosine derivatives queuosine and wyosine, however, are inserted by specific nucleotide excision and replacement reactions analogous to *nucleotide excision repair* in DNA (q.v.). Uridine residues in the small nuclear RNAs are also extensively modified; hence the alternative term **U-RNA**.

**rRNA cleavage and maturation.** Both bacteria and eukaryotes synthesize polycistronic rRNA transcripts. In *E. coli* there are seven rRNA operons (*rrn*) which contain the genes for all three rRNA species as well as certain tRNA genes. The 16S and 23S **pre-rRNA** segments form a stem loop structure by complementary base pairing. This appears to be a substrate for ribonuclease III, which cleaves in the stem to release the individual precursors. The mature ends are generated by exonucleolytic processing.

In mammals, the 5S rRNA, which is transcribed as a monocistronic unit by RNA polymerase III, is transcribed as a mature length molecule and requires no processing. RNA polymerase I synthesizes a 45S pre-rRNA containing the remaining 5.8S, 18S and 28S rRNAs. Occasional ribose moieties are methylated throughout the polycistronic transcript and most are retained in the mature rRNAs, suggesting their role may be to define which regions of the transcript are to be retained and which are to be discarded. The 45S rRNA associates with a ribonucleoprotein complex in the nucleolus as it is synthesized, and this is where processing occurs. The complex is termed a **processosome** or **snorp** (the latter a colloquialism for **small nucleolar ribonucleoprotein**, **snoRNP**, and contains the U3 snRNA). The sizes of the processing intermediates suggest that endonucleolytic cleavage generates both mature endpoints and rough divisions which must be processed further. As well as undergoing cleavage reactions, some lower eukaryotic rRNAs contain self-splicing introns which must be removed to generate functional rRNA.

## 27.2 End-modification and methylation of mRNA

**mRNA processing in bacteria and eukaryotes.** In bacteria, mRNA is generally unstable and is seldom modified — it is often synthesized, translated and degraded in the space of a few minutes.

Conversely, eukaryotic mRNA is generally stable and undergoes extensive processing in the nucleus before export. Eukaryotic pre-mRNA may undergo several types of processing reaction: end-modification by capping and polyadenylation; internal modification by splicing and occasionally RNA editing; chemical modification by methylation of internal adenine residues (the function of this last process is unknown). Eukaryotic pre-mRNA does not exist as naked RNA in the nucleus, but is associated with a number of abundant proteins to form **heterogenous ribonucleoproteins (hnRNPs)**. Cross-linking has helped to characterize many of these proteins; they possess RNA-binding motifs (see Nucleic Acid-Binding Proteins) but bind to RNA with differing specificity, reflecting preferences for certain types of base composition. The proteins may act in the same way as *single-stranded DNA-binding proteins* (q.v.) to remove secondary structure and facilitate interactions with components of, for example, the splicing machinery. They may also act as specific docking sites for RNA processing factors, such as splicing proteins.

**Capping.** As soon as transcription commences, nascent eukaryotic mRNAs are **capped** by the addition of an inverted guanosine triphosphate residue at the 5' end (there are no known examples of capped transcripts in bacteria). This rapid reaction is catalyzed by the enzyme **guanylyl transferase** (mRNA **guanylyltransferase**) and generates an unusual 5'→5' phosphodiester bond. The enzyme may be associated with a component of the RNA polymerase II initiation complex, because not only mRNAs but also RNAP II-transcribed snRNAs are capped (RNAP III-transcribed snRNPs, such as U6 snRNA, are not). The structure (a 5' end cap or simply a cap) is then methylated at position G<sup>7</sup> by the enzyme **guanine methyltransferase**, generating the **type zero cap** which is predominant in yeast. In higher eukaryotes, a further methyl group is transferred to position O<sup>2'</sup> of the ribose moiety in the next residue (originally the first residue in the transcript, corresponding to position +1) to generate a **type 1 cap**. If this residue is adenine, the base may also be methylated at position N<sup>6</sup>. In some species, the subsequent residue (position +2) is also methylated (**type 2 cap**), again at position O<sup>2'</sup> of the ribose.

The 5' cap is essential for several RNA functions in eukaryotes. It is required for export through nuclear pores; it is essential for ribosome binding — which explains why almost all eukaryotic mRNAs are monocistronic (see Protein Synthesis, but cf. *trans-splicing*, *internal ribosome entry site*) — and it also prevents 5' RNA degradation. The capping reaction can be used to regulate protein synthesis, a strategy utilized by some animals during egg maturation. Most RNA viruses cap their genomes and mRNAs, whilst the *picornaviridae*, whose infection strategy exploits their lack of cap-dependence, block the 5' end of their genome with a viral protein. The *orthomyxoviridae* (e.g. influenza virus) do not cap their genome segments but steal preformed caps from host mRNAs, a transesferification process which has been termed **capsnatching**.

**Polyadenylation.** The precise mechanism of RNAP II transcriptional termination in eukaryotes is not understood, and occurs up to several kilobases downstream of the mature 3' end of the transcript. The 3' end is generated by endonucleolytic cleavage, which is usually followed by **polyadenylation**; i.e. the addition of a variable number (usually approximately 200) of adenylate residues to generate a **polyadenylate** or **poly(A) tail**. In higher eukaryotes this occurs 10–30 nt downstream of a highly conserved **polyadenylation site** (AAUAAA). Polyadenylation sites in yeast genes show more sequence variation.

Cleavage and polyadenylation are carried out by a multisubunit complex comprising a trimeric **cleavage polyadenylation specificity factor** (CPSF) recognizing the polyadenylation site, an endonuclease comprising two **cleavage factors** which carries out the cleavage reaction, the enzyme **polyadenylate polymerase** (PAP) which catalyzes the addition of adenylate residues, and several other uncharacterized components. These are thought to comprise part of the RNAP II elongation complex and associate with the phosphorylated C-terminal tail of the enzyme. Initial polyadenylation is slow because PAP dissociates after adding each adenylate residue. However, after a short oligoadenylate sequence has been generated, a further component, **polyadenylate binding protein** (PABP), attaches to the tail and increases the processivity of PAP. PABP, through an unknown mechanism, controls the maximum length of the polyadenylate tail.

The precise role of polyadenylation is not clear. It may influence transcript stability, and in certain cases has been shown to play a critical role in translation. The *Drosophila bicoid* mRNA, for instance, is not translated until after fertilization when three proteins facilitate the extension of the polyadenylate tail. A few eukaryotic transcripts are not polyadenylated, the most notable examples being the histone mRNAs and the genomes of certain plant viruses. A secondary structure adopted by the histone transcripts is responsible for 3' end maturation, which involves U7 snRNA and associated proteins. While most bacterial mRNAs lack polyadenylate tails, short and relatively unstable oligoadenylate tails have been identified in some species, e.g. *Methanococcus vannielii*.

Polyadenylation can be exploited for *in vitro* RNA manipulation or purification by using synthetic oligo-dT sequences which selectively hybridize to them. The cellular RNA purified in this

manner, consisting mostly of mRNA, is termed the **poly(A)<sup>+</sup> RNA fraction**, the remainder (rRNA, tRNA, etc.) comprising the **poly(A)<sup>-</sup> fraction**. It is useful to be able to purify mRNA because it is a minority component (<5%) of the total RNA in the cell (q.v. *transcript analysis*, *cDNA cloning*, *RT-PCR*).

27.3 RNA splicing

**Introns and splicing reactions.** RNA splicing describes the precise removal of introns, the noncoding elements predominating in the genes of higher eukaryotes, but also found in those of lower eukaryotes and occasionally bacteria (q.v. *intron*, *interrupted genes*). Introns are classified according to splicing mechanism, of which there are four types (Table 27.2). With the exception of the pre-tRNA introns, all splicing mechanisms involve a pair of sequential transesterification reactions. In the first reaction, a nucleotide carrying a free hydroxyl group attacks the phosphodiester bond joining the intron and the upstream exon. This liberates the 3' end of the exon which carries its own hydroxyl group. In the second reaction, the free hydroxyl group attacks the phosphodiester bond joining the intron to the downstream exon. The second reaction joins the exons together and ejects the free intron, which is usually degraded. The three types of transesterification introns differ in their structure, the source of the initial hydroxyl group donor, the nature of the intermediate formed and whether the reaction is autocatalytic or requires a splicing apparatus to be supplied in *trans*. Autocatalytic introns are termed **self-splicing introns** and may be regarded as *ribozymes* (q.v.).

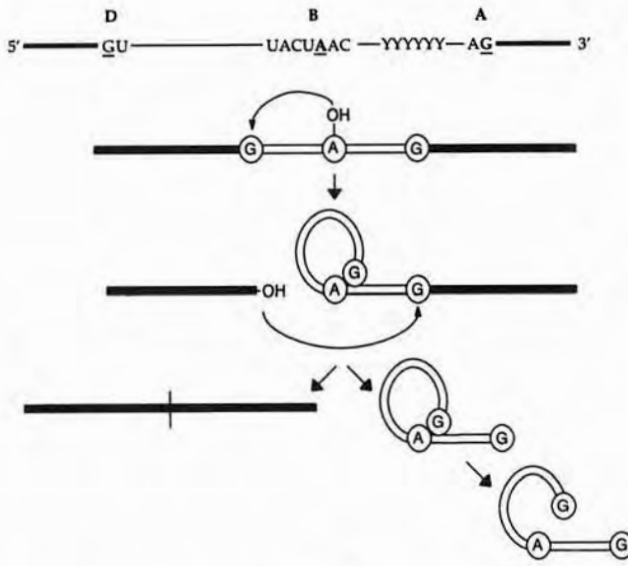
**Nuclear pre-mRNA introns.** Most introns found in eukaryotic nuclear genomes are unable to splice

Table 27.2: Properties of different classes of intron and summaries of their splicing mechanisms

Intron class	Characteristics of intron and splicing mechanism
<i>Transesterification introns</i> — Splice recognition sites within intron	
Nuclear pre-mRNA introns	Structurally diverse Initial donation of free hydroxyl group made by adenosine residue at internal branch site Lariat intermediate formed Large <i>trans</i> -acting splicing assembly required
Self-splicing introns class I	Conserved secondary structure Initial donation of free hydroxyl group made by guanosine nucleotide cofactor No lariat formed Autocatalytic splicing
Self-splicing introns class II	Conserved secondary structure Initial donation of free hydroxyl group made by adenosine residue at internal site Lariat intermediate formed Autocatalytic splicing
Self-splicing introns class III	Similar to group II introns but smaller (100–120 bp), containing a restricted number of domains
Twintrons	Multiple embedded self-splicing introns, often group II or mixed group II and group III. Usually spliced in a particular order
<i>Nontransesterification introns</i> — Splice recognition involves exon structure	
Nuclear tRNA introns	Splicing mechanism similar to tRNA maturation — involves cleavage followed by ligation No intermediate formed — intron excised as linear fragment Several processing enzymes required in <i>trans</i>

Note that nuclear mRNA introns and class II self-splicing introns use very similar splicing mechanisms although only the latter are autocatalytic.





**Figure 27.1:** Splicing through a lariat intermediate. Top panel shows the typical structure of a nuclear pre-mRNA intron with the donor (D), branch (B) and acceptor (A) sites indicated. Upstream from the acceptor site is a polypyrimidine tract which is the recognition site for one of the splicing factors. Lower panel shows the splicing reaction, with the terminal guanidylate residues of the intron and the active adenylate residue of the branch site enclosed in circles. This pathway is also followed by class II self-splicing introns. For nuclear introns, the reaction is catalyzed by a *trans*-acting spliceosome, whilst for class II introns, the RNA itself is catalytic.

autocatalytically and are identified as **nuclear pre-mRNA introns**. These are highly divergent in size and structure, but several short conserved elements have been identified which are *cis*-acting sites for the control of splicing (Figure 27.1). Of these, the **donor site** (5' site, **left splice site**), the **acceptor site** (3' site, **right splice site**) and the **branch site** are directly involved in the splicing reaction. The donor and acceptor sites possess the highly conserved consensus sequences GU and AG, respectively (the **GU-AG rule**) embedded within a weaker consensus. These sites determine the intron boundaries, and mutations which alter or delete them disrupting splicing (q.v. *aberrant splicing*, below). The branch site in yeast has the sequence UACUAAC; in higher eukaryotes the sequence is poorly conserved except for the penultimate adenylate residue (shown in *italic* in the yeast sequence) which is the initial hydroxyl group donor.

The splicing reaction occurs in the following way (summarized in Figure 27.1). The free 2' hydroxyl group carried by the active adenylate residue at the branch site attacks the phosphodiester bond linking the upstream exon to the intron. This first transesterification reaction joins the internal adenylate to the guanidylate residue of the 5' splice site through a 5'→2' phosphodiester bond, and generates a lasso-shaped structure termed a **lariat**. The 3' hydroxyl group of the upstream exon then attacks the phosphodiester bond linking the intron to the downstream exon. This second transesterification reaction at the 3' splice site ligates the exons together and releases the intron as a **lariat intermediate**, which is linearized (**debranched**) and degraded.

The splicing reaction is catalyzed by a 40–60S ribonucleoprotein complex, the **spliceosome**, which forms from its components upon pre-mRNA during transcription. The spliceosome consists of **small nuclear ribonucleoproteins (snRNPs, snurps)**, each containing several proteins and one or two **small nuclear RNA (snRNA)** molecules. As discussed above, the snRNAs are often termed U-RNAs because of their modified uridine content, and are identified as U1, U2 RNA, etc. The snRNPs are named according to the particular species of U-RNA they contain. Those whose role in

**Table 27.3:** Splicing of nuclear pre-mRNA introns — the splicing pathway and roles of the principle components

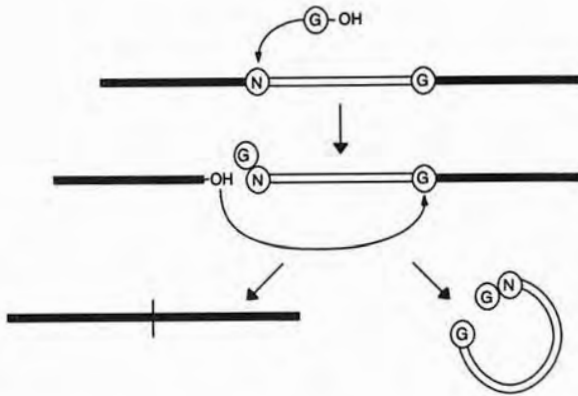
Presplicing complex	Principle components and events
E complex (early or commitment complex)	U1 and U2AF U1 binds to donor site Splicing factor U2AF binds to pyrimidine-rich sequence near acceptor site (this is required for subsequent binding of U2)
A complex	U1, U2 and U2AF U2 binds to branch site ATP hydrolysis required
B1 complex	U1, U2, U4/U6, U5 and U2AF U4/U6/U5 binds to presplicing complex U5 interacts with 5' exon, U4/U6 binds to U2
B2 complex	U2, U4/U6, U5 and U2AF U1 released from presplicing complex U5 repositioned to intron U6 binds to donor site ATP hydrolysis required
C1 complex	U2, U5, U6 and U2AF U4 released U5 repositioned to acceptor site First transesterification reaction occurs ATP hydrolysis required
C2 complex	Second transesterification reaction occurs Splicing factors ejected with lariat intermediate

The spliceosome does not exist in the cell as a preformed complex, but is assembled in stages, termed **presplicing complexes**, upon pre-mRNA.

splicing has been characterized are U1, U2, U4/U6 (which contains two snRNAs) and U5. Each snRNP contains between 10 and 20 individual proteins, some of which are common to all the snRNPs. Other proteins transiently associated with the spliceosome, but not themselves snRNP components, are termed **splicing factors**. The splicing pathway and roles of the principle spliceosome components and splicing factors are summarized in *Table 27.3*.

One remarkable aspect of nuclear pre-mRNA splicing is the precise manner in which most genes are spliced, given the generic nature of the splice signals. All splice sites are essentially the same, and therefore any 5' site can in principle join to any 3' site, but in most genes with multiple introns there is usually no **exon skipping** and only the principle splice sites, not *cryptic sites* (q.v.), are used (c.f. *alternative splicing*, below). Investigation of splicing intermediates has revealed a preferred splicing pathway for many genes. This suggests that splice-site choice may be an intrinsic property of the transcript, perhaps reflecting changes in secondary structure as splicing occurs (i.e. the order and specificity of splicing may be controlled by making illegitimate splice sites unavailable by sequestration into secondary structure, but once a splice has been successfully completed, a change in conformation reveals further splice sites). The precise basis of recognition remains unknown, but commitment may be mediated by splicing factors known as **SR proteins** (because they are serine/arginine-rich), which interact with both the U1 snRNP and U2AF. Essentially, the interactions facilitating commitment may be mediated in two ways: the SR proteins could span the intron and define the segment of RNA to be removed (**intron bridging**), or they could span the exon and define the segment of RNA to be retained (**exon definition**). It is likely that since exons are relatively small compared with introns, exon definition may be the mechanism used. In support of this theory, U1 snRNP stimulates the binding of U2AF to an upstream 3' splice site.

**Group II self-splicing introns.** Group II introns are found in plant and lower eukaryote organelle genomes. Like group I introns (see below) they splice autocatalytically, but unlike group I introns the



**Figure 27.2:** Self splicing of group I introns. See text for detailed mechanism.

splicing mechanism is closely related to that used by nuclear pre-mRNA introns. Splicing is initiated by the 2' hydroxyl group of an internal adenosine residue which attacks the phosphodiester bond linking the intron and upstream exon, generating a lariat and a free exon with a terminal 3' hydroxyl group. This group then attacks the phosphodiester bond linking the intron to the downstream exon in a second transesterification reaction which joins the exons together and releases a lariat intermediate (Figure 27.1). The autocatalytic activity of group II introns arises from a characteristic secondary structure comprising six stem-loop domains which bring the exons close together. The secondary structure of group II introns is similar to the structures adopted by the branch site of nuclear introns and the U2 and U6 snRNAs, suggesting that nuclear introns may have evolved from class II introns by transferring the information responsible for splicing from the intron itself to a *trans*-acting regulatory complex. This has allowed nuclear pre-mRNA introns to diversify in size and structure, whereas group II introns have remained homogeneous (for discussion of intron evolution, see Protein, Structure Function and Evolution).

**Group I self-splicing introns.** Group I introns are found in mitochondrial genomes and, more rarely, in the nuclear genomes of unicellular eukaryotes (e.g. the rRNA genes of *Tetrahymena thermophila*). The rare introns of prokaryotic systems are also group I introns (e.g. in the bacteriophage T4 thymidylate synthase gene). Group I introns are autocatalytic and use a guanosine-containing nucleotide cofactor to provide the free hydroxyl group. The splicing reaction proceeds as follows. The free 3' hydroxyl group of the cofactor attacks the phosphodiester bond linking the intron and upstream exon, extending the intron by adding guanosine to its 5' end and producing a free exon with a terminal 3' hydroxyl group. In the second transesterification reaction, the upstream exon attacks the phosphodiester bond joining the intron to the downstream exon, ligating the exons and releasing the intron as a linear fragment with a 5' terminal guanidylate residue (Figure 27.2). This fragment may become circularized by a third transesterification reaction in which the 3' end of the intron becomes joined to the guanidylate residue.

The specificity of the reaction is controlled by the highly conserved secondary structure of group I introns, which comprises nine hairpins, of which three are directly involved in the recognition of exon sequences. In particular, hairpin P1 comprises the distal end of the upstream exon and the first few bases of the intron. This is termed the **internal guide sequence (IGS)** because it was originally thought to juxtaposition the exons by pairing with the proximal regions of both, and thus to be the sole determinant of splicing specificity.

**Intron-encoded proteins.** Some introns of classes I and II contain open reading frames encoding proteins which facilitate intron splicing. Such proteins are termed **maturases**, and although they are

required for splicing, they are not catalytic (they presumably have a structural role). Other introns encode functions concerned with intron mobility (see Mobile Genetic Elements). Class I introns often encode endonucleases which cleave DNA specifically at the site of intron insertion, allowing passive transfer of the intron to intronless alleles of the gene in the same cell by repair-mediated recombination. This process is termed **homing** and the introns are known as *homing introns* (q.v.). Class II introns may encode proteins with reverse transcriptase and endonuclease activity, allowing them to mobilise like a *retroposon* (q.v.). This process is termed **retrohoming**. Some intron-encoded proteins have both homing and maturase activities.

In higher eukaryotes, entire genes (some containing their own introns) may be embedded within an intron of a larger gene, either in the same orientation or reversed (q.v. *overlapping genes*, *nested genes*, *T-cell receptor genes*). Introns may also contain transcriptional regulatory elements (q.v. *enhancer*, *silencer*).

**Nuclear tRNA introns.** A unique class of introns is found in nuclear tRNA genes. Splicing proceeds not by sequential transesterification reactions, but by cleavage followed by ligation. The intron is removed from the pre-tRNA by an unusual endonuclease reaction which generates a 5'-hydroxyl group and a 3' cyclic phosphate group. The ends are processed separately to generate conventional 5'-phosphate and 3'-hydroxyl termini for ligation. The modified ends, which are juxtaposed by intramolecular base-pairing in the tRNA, are then joined by RNA ligase. Other features of the tRNA introns are also unusual. There appears to be no conservation in the sequence or structures of the tRNA introns, i.e. recognition of the intron is facilitated by exon structure. In yeast, although introns in different tRNA genes are unrelated, they are all found in homologous positions — one nucleotide downstream of the anticodon. The introns do contain sequences complementary to the anticodon which allows them to pair with the anticodon loop. This secondary structure may be required for correct splicing: mutations disrupting it reduce splicing efficiency.

**Generating diversity through splicing.** For many genes, splicing is an invariant processing step. The same mature transcript is generated in all cells where the gene is transcribed, and a single product is synthesized. This is **constitutive splicing**.

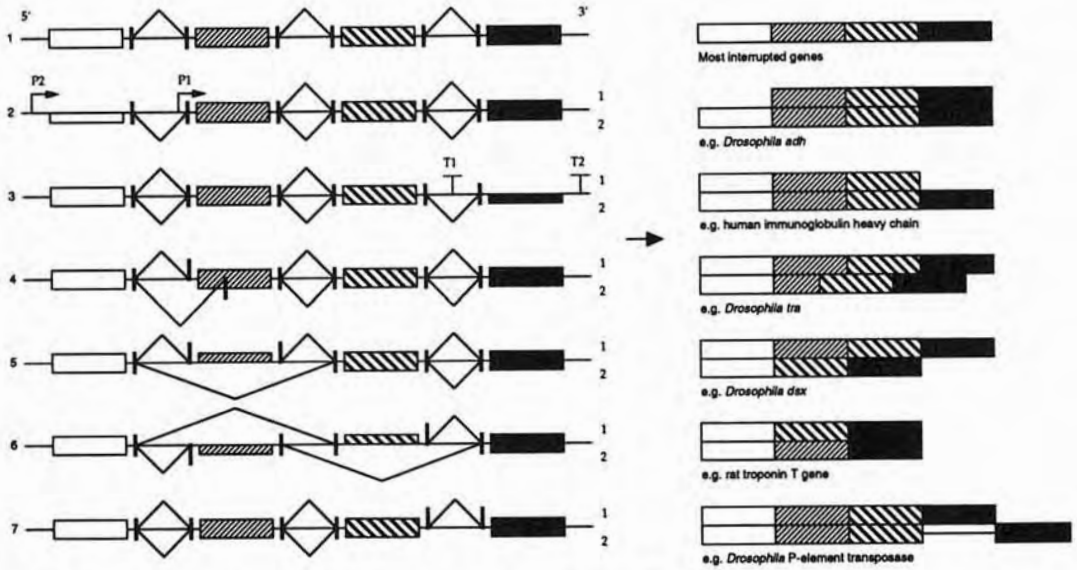
For other genes, RNA splicing may be used as a mechanism of gene regulation, i.e. a gene may be expressed differently in two tissues even though it is transcribed in exactly the same manner and at the same rate. Such regulation occurs in two ways.

In the first case, **processing or discard regulation**, the primary transcript is processed in some tissues and left unspliced in others. In some lower eukaryotes (e.g. sea urchins) this is a major regulatory mechanism, with most genes transcribed in most tissues, and the decision as to which genes are translated involving regulation of processing. In *Drosophila*, a similar mechanism is used to control synthesis of *P-element* transposase (q.v.) and restrict *P-element* transposition to the germline, although in this case only a single intron is left unspliced. The transposase gene contains three introns which, in germ cells, are spliced out to yield a mature transcript encoding functional transposase. In other tissues, however, the third intron is left intact, and as it contains an in-frame stop codon, it causes truncation of the transposase protein and prevents somatic *P-element* transposition. Similar strategies are used for endogenous genes, e.g. the mammalian GABA<sub>A</sub> receptor  $\epsilon$  subunit pre-mRNA is mis-spliced in all tissues except brain.

In the second case, **differential or alternative splicing**, the primary transcript can be processed in different ways by alternative usage or definition of exons. The family of related mature transcripts generated produces different gene products, which may be termed **splice variants** or **splice isoforms**. Alternative splicing occurs in several ways (Figure 27.3):

- (1) omission of 5' exon(s) by alternative promoter usage;
- (2) omission of 3' exon(s) by alternative polyadenylation site usage (in both these cases, regulation is at the level of transcription: differential RNA processing is a secondary consequence of the use of alternative transcriptional control elements);





**Figure 27.3: Forms of alternative splicing.** Primary transcripts and splicing pathways are shown on the left, and mature transcripts are shown on the right. Exons are represented by blocks and introns by horizontal lines. Diagonal lines represent splicing reactions, with alternative splicing reactions shown above and below the primary transcript and alternative products shown in corresponding positions. (1) Constitutive splicing; (2) alternative promoter usage generates different products from the *Drosophila* alcohol dehydrogenase gene in larvae and adults; (3) differential polyadenylation site usage generates secreted and membrane-bound forms of the immunoglobulin heavy chain; (4) alternative 3' splice site usage produces functional and nonfunctional forms of the *Drosophila transformer* mRNA; (5) exon skipping generates different forms of the *Drosophila doublesex* mRNA; (6) mutual exclusion of exons 16 and 17 generates alternative forms of the rat troponin T mRNA; (7) failure to splice out a single intron generates functional and nonfunctional forms of the *Drosophila* P-element transposase — this can be regarded as a single intron processing or discard regulation.

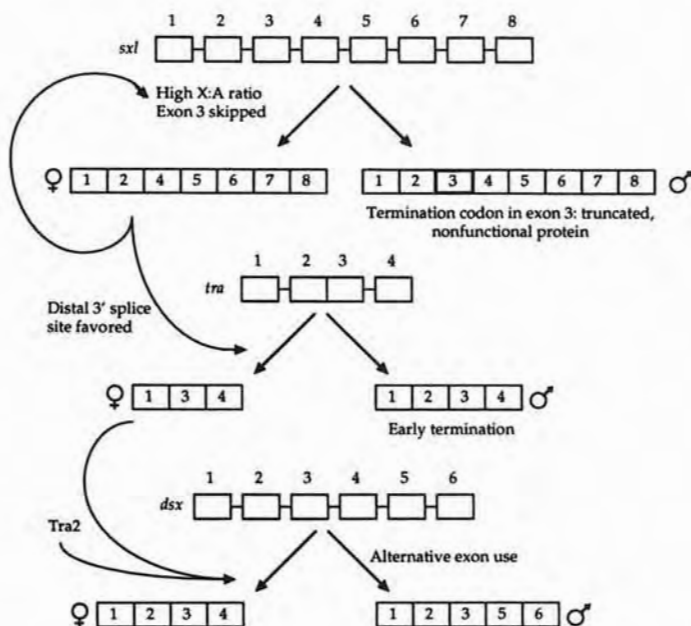
- (3) choice between alternative 5' or 3' splice sites in exon definition;
- (4) choice between alternative 5' or 3' splice sites to skip an exon;
- (5) mutual exclusion of exons.

These processes are regulated at the level of RNA processing and the factors involved are discussed below.

**Control of alternative splicing.** As discussed above, the commitment to splicing is initiated by the U1 snRNP and the splicing factor U2AF, together with SR-proteins which interact with them. The alternative splicing mechanism directs these splicing components to specific sites on the pre-mRNA, either by stimulating or inhibiting the use of particular splice sites. Two types of splice control can be distinguished: one acts constitutively and is based on the quantity of particular splicing factors; the second acts in a tissue-specific manner and involves dedicated splicing control proteins.

Quantitative control often occurs where alternative 5' or 3' sites are used for differential splicing. For example, analysis of differential splicing in the SV40 T/t antigen gene identified an alternative splice factor (ASF), an SR-protein which was shown to be identical to the constitutive splice factor SF2. Subsequently, it was shown that where alternative 5' splice sites are available, higher concentrations of ASF/SF2 favor the use of the most proximal site, whereas lower concentrations favor the use of the most distal site. A second factor, hnRNPA1, has the opposite effect to SF2. Differences in the relative levels of the two factors in different cells can explain cell-type-specific differences in the predominance of the two splice isoforms.

The quantitative model cannot adequately explain situations where splice isoforms are produced in a strict cell-type-specific manner. Many situations where this occurs have been described in



**Figure 27.4:** Sex determination in *Drosophila*. Alternative splicing of exon 3 of the *sex lethal* (*sxl*) gene generates a functional protein in the female and a truncated protein in the male. *Sxl* is a splicing factor which promotes productive splicing of its own gene and of the downstream gene *transformer* (*tra*). Under the influence of *Sxl*, *tra* pre-mRNA splicing favors a distal 3' splice site and omits exon 2 which causes truncation of the protein in the male-specific default pathway. *Tra* is also a splicing factor, which together with a further splicing factor *Tra2* alters the male-specific default splicing of the *doublesex* (*dsx*) gene and initiates female-specific differentiation. The initial production of *Sxl* is dependent on the *balance mechanism* (q.v.) of sex determination in *Drosophila* (see Development: Molecular Aspects).

eukaryotes. In *Drosophila*, perhaps the most remarkable example of alternative splicing is the hierarchy of splice control signals controlling sex determination (Figure 27.4). Tissue-specific splicing factors promote the use of, or specifically inhibit, particular splice signals by binding *cis*-acting elements (**splice enhancers** and **splice repressors**) in the primary transcript, in a manner which is analogous to transcriptional regulation (see Transcription). Tissue-specific splice regulators have been identified using genetic screens in *Drosophila*, but a number of tissue-specific variants of the general splicing machinery have also been identified in mammals, including the protein SmN, which is restricted to heart and brain.

**Unusual splicing pathways.** Aberrant splicing may occur when a wild-type splice site is destroyed by mutation or if a new splice arises by the same mechanism. **Cryptic splice sites** are weak consensus sites which may be revealed by loss of a genuine site, and several cryptic sites are often revealed at once, resulting in the production of several types of aberrant mature transcripts where exon sequences have been deleted or intron sequences included. Many forms of thalassemia have been traced to mutations in splice sites, which generate aberrant globin transcripts (see Box 15.1).

Another unusual form of splicing is **trans-splicing**, where exons are joined from two separate RNAs, i.e. the primary transcripts of two or more distinct genes (the normal process is **cis-splicing**). *Trans*-splicing occurs in some chloroplast genomes, where *divided genes* (q.v.) are represented by dispersed exons in different orientations (see Organelle Genomes, The Gene). *Trans*-splicing also occurs in the nuclear genomes of lower eukaryotes: trypanosomes provide the extreme example where the same 35 nt 5' leader sequence, the **spliced leader RNA** (SL RNA) is added to all mRNAs by this process. *C. elegans* also adds a 22 nt leader to several mRNAs, including three actin tran-

scripts. In this case, alternative *trans*-splicing allows polycistronic pre-mRNAs to be converted into several forms of mature transcript where each gene can be translated (q.v. *operon*). Generally, the *trans*-spliced leader has a 5' splice site but no 3' acceptor site, and the 3' gene has an upstream unpaired acceptor site. SL RNA adopts a secondary structure with three stem loops and a single-stranded region which is thought to function in the same way as the U1 snRNP. Influenza virus *cap-snatching* (q.v.) is also a form of *trans*-splicing.

## 27.4 RNA editing

**RNA editing.** Under most circumstances, the information in a protein-encoding gene is unaltered when it comes to be translated. The gene may be interrupted by introns which are spliced out, and the information in the gene may be used selectively, the actual sense of the information is not changed. **RNA editing** is a co- or posttranscriptional mechanism which alters the information contained within the exon sequences of mRNA. This process is largely restricted to organellar genomes, although there are several examples of minor editing in mammalian nuclear mRNAs. Genes subjected to extensive RNA editing are termed **cryptogenes**: the structure of the gene product cannot be deduced from the genomic DNA sequence. The significance of RNA editing in evolutionary terms is unknown, although the predominance of pyrimidine insertions in major editing processes indicates that it may have evolved as a mechanism to introduce pyrimidines into purine-rich sequences. Editing can have important functional consequences, e.g. in mammals; it determines the properties of some ion channels and G-protein-coupled receptors. There are four categories of RNA editing (Table 27.4).

## 27.5 Post-processing regulation

**RNA export and subcellular localization.** In eukaryotes, mRNA is synthesized in the nucleus and must be transported to the cytoplasm for translation. It is thought that the restriction of RNA processing complexes to discrete foci in the nucleus plays a direct role in the subsequent export of RNA through nuclear pores. The exact mechanism of export is still not fully understood. It is known that there is selective transport of mRNPs, and that export is dependent upon ATP hydrolysis. It is also known that the 5' cap plays a major role in nuclear export (uncapped transcripts such as rRNA and U6 snRNA are not exported), and that the presence of spliceosome components blocks export (thus preventing the translation of partially spliced transcripts). Some hnRNP proteins are removed from the transcript before export, and some dissociate following transport and return to the nucleus to be reused. Processed mRNA associates immediately with ribosomes as it leaves the nucleus.

RNA export from the nucleus represents a potential regulatory target in eukaryotes, but to date only viral RNA has been shown to be controlled in this manner. The best-characterized system is the HIV genome, where splicing and export are controlled by the **Rev protein** and *cis*-acting elements, **Rev response elements**, in the introns. Rev appears to facilitate the export of RNA with bound spliceosomes, and thus allows partially spliced HIV genomes to be exported. It is not clearly understood how Rev circumvents the normal nuclear inhibition of this process (see Box 30.3).

Once in the cytoplasm, mRNAs associated with ribosomes may diffuse freely, or may be targeted to a particular region of the cell. Partially translated transcripts encoding secreted proteins are often transported to the membrane of the rough endoplasmic reticulum so that the polypeptide can be translocated into the lumen of this organelle (q.v. *signal peptide*). Whereas this relies upon a localization signal in the polypeptide, in other cases the signal is carried by the RNA itself (**RNA targeting**). Certain transcripts become associated with the cytoskeleton, and are localized to specific regions of the cell. The latter is a common phenomenon during animal development as it provides a mechanism to localize positional signals as *cytoplasmic determinants* (q.v.) in the egg (e.g. *bicoid* and *nanos* mRNA in *Drosophila* development, *vg-1* mRNA in *Xenopus* development; see Development: Molecular Aspects). In other cases, localization can be directly related to function

**Table 27.4:** Four RNA editing mechanisms

Editing process	Properties	Examples	Mechanism
Simple editing	Single residue conversions Posttranscriptional	C→U transition in the mammalian apolipoprotein mRNA	Modification by specific cytidine deaminase
		A→G transition in mRNA for mammalian glutamate receptor subunits and serotonin receptors. Editing also occurs in introns	Modification by dsRNA deaminase converts adenosine to <i>inosine</i> (q.v.). Three genes have been identified
		C→U transitions in plant organelles	Unknown
Insertional editing	Insertion of single nucleotides or small runs of nucleotides Cotranscriptional	G insertions during transcription of the paramyxovirus P gene	Transcriptional strand slipping
Pan-editing	Insertion/deletion of multiple uridine residues Posttranscriptional	U insertions/deletions in trypanosome kinetoplast mRNA (q.v. <i>kinetoplast DNA</i> )	Editing sequence provided by external antisense <b>guide RNA (gRNA)</b> which pairs with <b>pre-edited mRNA</b> in a ribonucleoprotein particle, the <b>editosome</b> , and identifies positions to be edited as mismatches. gRNA has polyuridylyate tail which supplies uridine residues for insertion. Editing has 3'→5' polarity
	Insertion of multiple cytidine residues	C insertions in at least four <i>Phisarum polycephalum</i> mitochondrial mRNAs	Unknown
Polyadenylation editing	Addition of adenosine residues at end of transcript to complete stop codons	Polyadenylation of several vertebrate mitochondrial mRNAs (see <i>Organelle Genomes</i> )	Pre-mRNA lacking a stop codon is polyadenylated, with the first one or two adenylate residues providing the missing information

For more details of pan-editing, q.v. *kinetoplast DNA*.

(e.g. the localization of  $\alpha$ -actin and  $\beta$ -actin mRNAs during myoblast differentiation) or may be essential for cell survival (e.g. translation of myelin basic protein in the wrong location is lethal). In each case, the localization signal is found in the 3' UTR of the transcript and associates with one or more proteins required for its localization.

**mRNA turnover.** The abundance of a particular transcript is controlled both by its rate of synthesis and its stability, which reflects its rate of degradation. The stability of mRNA determines how quickly the steady-state levels of the mRNA change when the rate of transcription is altered, and



thus how much is available for protein synthesis. Like other forms of gene regulation, RNA stability can be constitutive or regulated. Changes in the rate of mRNA degradation, which can be expressed as the **mRNA half-life**, can effect rapid and transient alterations to the abundance of a particular mRNA without any change in transcriptional activity.

**mRNA turnover and retroregulation in bacteria.** Bacterial mRNAs have short half-lives, in the order of several minutes for the most stable transcripts, which allows the rate of protein synthesis to be altered rapidly in response to the environment by regulating the rate of transcription. mRNA degradation in prokaryotes is mediated by RNA endonucleases (also called **RNA restriction enzymes**) and 3'→5' exonucleases. The secondary structure of mRNA is important in the determination of stability, with the most stable transcripts possessing multiple hairpins and stem-loop structures in the 3' untranslated region which may protect the transcript from exonuclease activity. Transcripts which contain endonuclease target sites are particularly unstable. The specific enzymes which degrade RNA in bacteria are not well characterized, although mutations which disrupt the *E. coli* ribonuclease E protein induce a 2–3-fold increase in RNA stability.

The close association of transcription, protein synthesis and degradation in bacteria permits an unusual form of gene regulation, termed **retroregulation**, where RNA degradation is regulated at the level of transcriptional termination. Gene expression depends upon whether or not a *cis*-acting element located downstream of the gene is transcribed. If it is, the nascent RNA adopts a structure favoring rapid degradation and translation is prevented. If transcription terminates prior to this site, the RNA is relatively stable and protein synthesis proceeds. Retroregulation is used, for example, by bacteriophage  $\lambda$  to control the expression of its integrase gene (see Box 30.1).

**mRNA turnover in eukaryotes.** Eukaryotic mRNAs are generally much more stable than bacterial transcripts. The half-life of yeast mRNA ranges from ~5 to ~45 minutes, and metazoan mRNA is even more stable, with an average half-life of about 10 hours, reflecting the relatively constant environment of cells in multicellular organisms. The polyadenylate tail present on most mammalian mRNAs appears to confer stability by binding the PABP, which maintains tail length. Deadenylation or depletion of PABP results in rapid mRNA degradation in mammalian cells, but in yeast cells, the presence of PABP appears to be a signal for degradation, so its precise role is unclear. In eukaryotes, exoribonuclease degradation of the polyadenylate tail is the first stage in mRNA degradation. Histone mRNAs, which lack polyadenylate tails, are degraded by a specific exonuclease. There is also evidence for endoribonucleases active in eukaryotic cells.

Some eukaryotic regulatory proteins are required in transient bursts, and their mRNAs are consequently unstable like bacterial transcripts. These include the transcripts of many *immediate early genes*, the genes induced by signal transduction cascades and required to produce a short-lived regulatory responses (e.g. *c-fos*, *c-jun*; see The Cell Cycle, Oncogenes and Cancer, Signal Transduction). Many unstable eukaryotic mRNAs contain specific instability elements, often **AU-rich elements (AUREs, AREs)** such as multiple copies of the sequence AUUUA, generally located in the 3' UTR (e.g. in interleukin 1, interferon  $\beta$  and *c-fos* mRNAs), although in several cases, within the coding region (e.g. in  $\beta$ -tubulin and *c-myc* mRNAs). Although the mechanism by which instability is conferred is not understood, AUREs have the ability to form stem-loop structures, suggesting that factors which influence RNA folding may regulate RNA stability. Instability elements appear to be relatively independent, as they can confer instability on heterologous mRNA when inserted into the 3' untranslated region. However, there may be some dependence on secondary structure and/or context as similar elements have been identified in a number of stable transcripts such as neuron-specific enolase mRNA.

The stability of a few eukaryotic mRNAs can be regulated by *trans*-acting factors which bind to the instability elements. This occurs in the transferrin receptor mRNA which, in the presence of excess intracellular iron levels, is degraded, presumably by the same process which controls degra-

dation of the constitutively unstable transcripts described above. When iron levels fall, however, a protein factor binds to an **iron response element (IRE)** in the transcript and prevents degradation. IREs are 30 bp motifs which form stem-loop structures with instability elements in the stem. The presence of the IRE-binding protein is thought to block access to the instability elements. In other transcripts, the same IRE is used to regulate translation (*see Protein Synthesis*).

### Further reading

- Bachelierie, J.P. and Cavaillie, J. (1997) Guiding ribose methylation of rRNA. *Trends Biochem. Sci.* 22: 257–261.
- Bass, B.L. (1997) RNA editing and hypermutation by adenosine deamination. *Trends Biochem. Sci.* 22: 157–162.
- Chabot, B. (1996) Directing alternative splicing — cast and scenarios. *Trends Genet.* 12: 472–478.
- Chen, C.Y.A. and Shyu, A.B. (1995) AU-rich elements — characterization and importance in messenger-RNA degradation. *Trends Biochem. Sci.* 20: 465–470.
- Copertino, D.W. and Hallick, R.B. (1995) Group II and group III introns of twintrons — potential relationships with nuclear premessenger RNA introns. *Trends Biochem. Sci.* 18: 467–471.
- Guo, Z.J. and Sherman, F. (1996) 3'-end-forming signals of yeast mRNA. *Trends Biochem. Sci.* 21: 477–481.
- Kable, M.L., Heidmann, S. and Stuart, K.D. (1997). RNA editing: Getting U into RNA. *Trends Biochem. Sci.* 22: 162–166.
- Ross, J. (1996) Control of messenger-RNA stability in higher eukaryotes. *Trends Genet.* 12: 171–175.
- Sarker, N. (1997) Polyadenylation of mRNA in prokaryotes. *Annu. Rev. Biochem.* 66: 173–197.
- Tarn, W.Y. and Steitz, J.A. (1997) Pre-mRNA splicing: The discovery of a new spliceosome doubles the challenge. *Trends Biochem. Sci.* 22: 132–137.
- Valcarcel, J. and Green, M.R. (1996) The SR protein family — pleiotrophic functions in pre-messenger-RNA splicing. *Trends Biochem. Sci.* 21: 296–301.
- Wahle, E. and Keller, W. (1996) The biochemistry of polyadenylation. *Trends Biochem. Sci.* 21: 247–250.

## Chapter 28

# Signal Transduction

### Fundamental concepts and definitions

- Cells respond to their environment by reorganizing their structure, regulating the activity of proteins and altering patterns of gene expression. The stimulus for such responses is termed a **signal**, and may be a small molecule, a macromolecule or a physical agent, such as light. Signals interact with the responding cell through molecules termed **receptors**.
- Small molecules often act as diffusible signals. In unicellular organisms, diffusible signals may be environmental in origin or may be released from other cells (e.g. yeast mating-type pheromones, cAMP in *Dictyostelium*). In metazoans, signals may be released from nearby cells and diffuse over short distances (**paracrine signaling**), or they may be released from distant cells and reach their target through the vascular system (**endocrine signaling**). Macromolecular signals are often associated with the extracellular matrix or displayed on the surface of neighboring cells (**juxtacrine signaling**). A molecular signal that binds to a receptor is termed a **ligand**.
- Signals may be processed in three ways. Certain chemical signals may penetrate the plasma membrane of the cell and interact with internal receptors (e.g. steroids, nitric oxide). Most signals, however, are hydrophilic molecules remaining outside the cell. These interact with transmembrane (membrane-spanning) or membrane-associated receptors and cause a change of receptor structure. The interaction may result in **signal transport**, i.e. the signaling molecule is internalized (either by carriage caused by the conformation change of the receptor, by the creation of a pore, e.g. in the case of ion channels, or by receptor-mediated endocytosis). Alternatively, the conformational change in the receptor may induce enzyme activity inside the cell which mediates downstream effects while the ligand remains on the outside (**signal transduction**). Physical stimuli may also interact with receptors or may mediate their effects directly. Light stimulates the G-proteins linked to rhodopsin and cone opsin receptors when photons cause a conjugated light-sensitive molecule 11-*cis* retinal to change to the all-*trans* conformation. Conversely, the response to heat shock and similar stresses is mediated directly by increases in denatured protein in the cell.
- Signal transduction involves pathways of sequential enzyme activation and modulation of the levels of small molecules termed second messengers. This allows the amplification of the original signal (direct diffusion and signal transport provide only a linear response). Signal transduction pathways can *converge* and *diverge*, allowing multiple stimuli to generate similar responses, and individual signals to effect different responses. Further diversity is generated by different responses to the length and intensity of the stimulus, and the cell-specific synthesis of different receptors and signaling components. Signal transduction pathways are subject to complex regulatory networks, and the response depends upon a balance of opposing forces in the cell.
- Delivery of the signal involves the activation or repression of transcription factors, enzymes and structural components of the cell, usually by altering their state of phosphorylation.

### 28.1 Receptors and signaling pathways

**Cell-surface receptors and their ligands.** Cells respond to a diverse range of signals through an equally diverse range of receptors. In metazoans, many signals are small polypeptides. Locally acting (paracrine) polypeptide signals are termed **growth factors** (or **cytokines** in the hematopoietic system) and can be assigned to families according to structural or functional similarities. These

molecules are concerned not only with cell growth, but also with differentiation, motility and other cellular functions. **Hormones** are endocrine signaling molecules and many are peptides like growth factors, but there are also large glycoprotein hormones and steroid hormones, the latter able to diffuse directly through the plasma membrane of the cell and interact with cytoplasmic and nuclear receptors. Another important class of peptide signaling molecules is the **neuropeptides**, which mediate neurotransmission and neuromodulation in the central and peripheral nervous systems. Cells can also respond to numerous small molecules such as amino acids, nucleotides and bioactive amines, as well as macromolecules embedded in the extracellular matrix or displayed on the surface of adjacent cells, which control cell motility and adhesion. Some receptors bind **contra-receptors** on adjacent cells, so that the interaction between cells induces reciprocal signaling.

Receptors show less overall diversity than their ligands because the cell has a smaller repertoire of responses than it does stimuli. Thus many ligands effect the same type of response (e.g. growth arrest, proliferation, transcription of specific genes), and this is achieved by channeling many signals into common pathways. Thus, whereas there is great diversity in the ligand-binding domain of any class of receptor, receptors as a whole can be grouped into a small number of families with common effector structure and signaling mechanisms. A current problem for researchers is to find out why there are so many subtypes of receptors, all apparently doing the same thing. The structure and activity of several of the more common receptor classes is discussed below.

**Ion channels.** Ion channels are water-soluble pores in the cell membranes whose activity is controlled by opening and closing, thus allowing or preventing the movement of ions and other small molecules in or out of the cell. Ion channels are closed as a default state but open in response to a specific signal. Some are **ligand-gated**, i.e. they open in response to the binding of a particular ligand (e.g. glutamate and  $\gamma$ -aminobutyric acid (GABA) receptors). Others, particularly those found in neurons, are **voltage-gated**, i.e. they open in response to the electrical changes associated with an action potential. Finally, there are **second-messenger-gated** ion channels, which respond to second messengers in the cell, e.g. calcium ions, cyclic nucleotides and lipids (see *second messengers* below). Ion channels may allow the passive movement of ions along an electrochemical gradient, or may actively pump ions against such a gradient, a process which requires ATP hydrolysis.

The transmembrane region of ion channels comprises several *amphipathic helices* (q.v.) which pack together so that hydrophobic residues surround a hydrophilic core. Some channels are single, large proteins with a number of similar domains. Others are multisubunit proteins. Neurotransmitter-gated ion channels have been particularly well characterized and comprise five similar subunits, each of which possesses a four-helix transmembrane domain. One of the four helices is amphipathic, and the pore of the channel is lined by the amphipathic helices from each subunit. The opening and closing of ion channels is mediated by conformational changes stimulated by the gating mechanism (i.e. binding of a ligand or voltage changes). The particular arrangement of charged residues at the entrance to the pore, and lining it, control the ion selectivity of the channels.

**G-protein-coupled receptors.** The **G-protein-coupled receptors** (GPCRs) are single polypeptides whose central, hydrophobic region forms a seven-span transmembrane domain. The external, N-terminal region is often the ligand-binding domain, although some ligands (e.g. adrenaline) bind within the membrane domain. The internal, C-terminal region is associated with a trimeric guanine nucleotide binding protein, or **G-protein**. The genes for several hundred GPCRs have been cloned and characterized. The receptors demonstrate remarkable conservation considering the diversity of their ligands, which include small molecules such as adrenaline and serotonin, peptides such as glucagon, the neurokinins and the opioids, and large glycoproteins such as follicle-stimulating hormone. Further GPCRs are stimulated by odorants and light. GPCRs are also widely distributed — as well as the hundreds of vertebrate receptors identified, GPCRs are responsible for the transduction of yeast mating-type signals, and the response to cAMP in *Dictyostelium discoideum*.



**Table 28.1:** G-protein  $\alpha$ -subunit families, major subtypes, their effectors and the second-messenger responses activated

Family	Major subtypes	Function of effector	Response of second messenger	Example
$G_s$	$G_s$	Stimulates adenylate cyclase and $Ca^{2+}$ channels	cAMP increases, $Ca^{2+}$ channels open	Histamine ( $H_2$ receptor)
	$G_{olf}$	Stimulates adenylate cyclase	cAMP increases	Olfaction receptors
$G_i$	$G_{i1-3}, G_0$	Inhibits adenylate cyclase	cAMP decreases	$\delta$ opioid receptor
		Stimulates $K^+$ channels	Change in membrane potential	$\alpha_2$ adrenergic receptor
$G_q$	$G_t$	Inhibits $Ca^{2+}$ channels	$Ca^{2+}$ channels close, cGMP decreases	Rhodopsin
	$G_{gust}, G_z$	Stimulates cGMP-PDE	Unknown	Taste receptors
	$G_q, G_{11}, G_{14-16}$	Activates phospholipase C- $\beta$	Ins(1,4,5) $P_3$ increases, DAG increases	Tachykinin receptors
				Some glutamate receptors
$G_{12}$	$G_{12}$	Stimulates/inhibits	Change in pH	
	$G_{13}$	$Na^+/H^+$ exchange		

Abbreviations: cA(G)MP, cyclic adenosine (guanosine) monophosphate; PDE, phosphodiesterase; Ins(4,5,6) $P_3$ , inositol-4,5,6-trisphosphate; DAG, diacylglycerol.

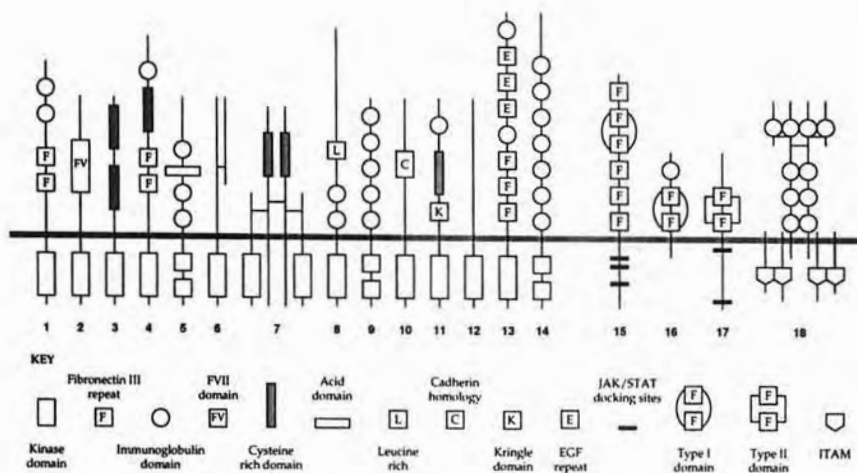
Two putative GPCRs have also been identified in viral genomes, although their functions are unknown. A membrane protein with similar structure, but no sequence homology, is also found in the halophilic bacterium *Halobacterium halobium*.

G-proteins are heterotrimeric complexes (comprising  $\alpha$ ,  $\beta$  and  $\gamma$  subunits) which associate with GDP in their inactive form. Activation of the receptor by ligand binding stimulates an interaction which increases the rate at which GDP dissociates from the G-protein. Dissociation allows GTP to replace GDP because it is more abundant than GDP in the cell, and this causes the  $\alpha$ -subunit to separate from the  $\beta\gamma$  dimer. G-proteins have intrinsic GTPase activity, and slowly hydrolyze their cognate GTP thus inactivating themselves, whereupon they reassemble into trimers.

More than 20 different types of  $\alpha$ -subunit and also multiple types of the  $\beta$ - and  $\gamma$ -subunits have been identified in mammals: potentially hundreds of combinatorial trimers can form, some of which can be coexpressed in a particular cell type. The  $\alpha$ -subunit is often the primary activator of downstream effector molecules, and four major families of G-proteins have been identified whose  $\alpha$ -subunits interact with different types of effector and have different effects on the availability of intracellular second messengers (Table 28.1). The  $\beta\gamma$  dimers also mediate downstream responses, e.g. the STE1 and STE2 proteins of *S. cerevisiae* activate a kinase which links G-protein signaling into the MAP kinase pathway (see below). The precise response to G-protein activation is governed both by the particular  $\alpha$ ,  $\beta$ , and  $\gamma$  subtypes and the particular isoforms of downstream targets. For example, the various isoforms of adenylate cyclase show differential responses to  $G_s$  and  $G_i$  regulation.

**Receptor tyrosine kinases.** Receptor tyrosine kinases (RTKs) are single membrane spanning receptors with intrinsic, ligand-activated protein tyrosine kinase activity on their cytoplasmic domains. Over 50 RTKs have been characterized. They are divided into 14 major families based on structural motifs found in the extracellular domain (Figure 28.1). The tyrosine kinase domain is strongly conserved across all families, although in the PDGFR, FGFR and VEGFR subfamilies the domain is divided by an internal sequence which binds to other signaling molecules.

In the absence of the ligand, RTKs are monomeric and have no kinase activity. Although not all receptor activation mechanisms have been investigated, activation is probably brought about in



**Figure 28.1:** Structure of receptors with intrinsic or associated tyrosine kinase activity. 1–14 are examples of the different subfamilies of receptor tyrosine kinases. 1, ARK; 2, DDR; 3, EGFR (epidermal growth factor receptor); 4, Eph; 5, FGFR (fibroblast growth factor receptor); 6, HGFR; 7, insulin receptor; 8, NGFR (nerve growth factor receptor TrkA); 9, PDGFR (platelet-derived growth factor receptor); 10, c-RET; 11, ROR1; 12, c-Ryk; 13, TIE; 14, VEGFR. 15–17 are examples of different subfamilies of cytokine receptors. 15, gp130; 16, IL-6R (interleukin-6 receptor). These are class I receptors. 17, IFNAR1 (interferon- $\alpha$  receptor), a class II receptor; 18, an immunoglobulin B-cell receptor.

each case by oligomerization, which stimulates kinase activity and **autotransphosphorylation** (the phosphorylation of one receptor monomer by another). Many RTKs are active as dimers, dimerization being brought about by ligand binding, often because the ligands themselves are dimeric and can bind two receptors simultaneously (e.g. PDGF). The insulin receptor family is anomalous in this respect as the receptors are constitutively dimeric, the two subunits being held together by disulfide bonds. In this case, ligand binding is likely to induce a conformational change facilitating activation of the kinase domain.

Once the receptor is activated, RTK signaling can be initiated in two ways: (a) by phosphorylating downstream targets, and (b) by recruiting signaling complexes including proteins which specifically recognize phosphotyrosine residues. Several domains of other proteins interact specifically with phosphotyrosine, including the SH2 domain and the PTB domain (see below). The insulin receptor family is again slightly different from other RTKs in that many, if not all, of its downstream reactions are mediated by a small protein, insulin receptor substrate-1 (IRS-1), which is phosphorylated by the insulin receptor tyrosine kinase. Major signaling pathways initiated by RTK activation include the Ras–Raf–MAP kinase pathway, and the phospholipase C- $\gamma$ -activated second-messenger system, discussed below.

**Receptors with associated tyrosine kinase activity.** A number of receptors phosphorylate target proteins upon activation but possess no intrinsic enzyme activity. Such receptors, which include many cytokine receptors and the receptors which process antigens in the immune system, recruit and activate cytoplasmic protein tyrosine kinases upon ligand binding.

There are two major families of cytokine receptors. Class I receptors include most hematopoietic and immune system cytokine receptors and three ubiquitous receptors: gp130,  $\beta c$  and  $\gamma c$ . The distantly related class II receptors include the interferon receptors and the interleukin 10 receptor. Receptors of each class are identified by conserved motifs in both the extracellular and intracellular domains of the molecule (Figure 28.1).

**Table 28.2:** Signal transduction components of the class I and class II cytokine receptors

Receptor complex and recruitment motif		JAKs	STATs
<i>Class I receptors</i>			
EPOR	XXYLVLV	Jak2	STAT5
G-CSFR		Jak1, Jak2	STAT3
GHR		Jak2	STAT1
GM-CSFR ( $\beta$ c)		Jak1, Jak2	STAT5
IL-2Rb ( $\gamma$ c)	DAYLSL, DAYCTF	Jak1, Jak3 (also Lck, Syk)	STAT5
IL-3R ( $\beta$ c)		Jak1, Jak2	STAT5
IL-4R ( $\gamma$ c)	XGYKPFG, GYKAFS	Jak1, Jak3	STAT6
IL-5R ( $\beta$ c)		Jak1, Jak2	STAT5
IL-6R (gp130)	XXYXPQX XXYXXQ	Jak1, Jak2, Tyk2	STAT1, STAT3
IL-7R ( $\gamma$ c)		Jak1, Jak3	STAT5
LIFR (gp130)		Jak1, Jak2, Tyk1 (also Yes, Hck)	STAT3 >> STAT1
<i>Class II receptors</i>			
IFNAR1	XXYXXQ	Tyk2	STAT1, STAT2
IL-10R		Jak1, Tyk2	STAT1
INFAR2		Jak1	STAT1, STAT3
INFR		Jak1, Jak2	STAT1, STAT3

For class I receptors which heterodimerize with one of the ubiquitous receptors gp130,  $\beta$ c or  $\gamma$ c, the particular common component is indicated in parentheses; the others act as homodimers. Some cytokine receptors interact with tyrosine kinases which do not belong to the Janus family — the alternative kinases are also indicated in parentheses. For some receptors, the phosphotyrosine motifs which recruit STATs have been identified by mutation.

Abbreviations: EPOR, erythropoietin receptor; G-CSFR, granulocyte colony stimulating factor receptor; GHR, growth hormone receptor; GM-CSFR, granulocyte macrophage colony stimulating factor receptor; IL, interleukin; LIFR, leukemia inhibitory factor receptor; IFN $\alpha$ ( $\gamma$ )R, interferon  $\alpha$  ( $\gamma$ ) receptor; JAK, Janus kinase; STAT, signal transducers and activators of transcription.

Ligand binding induces receptor dimerization which juxtaposes the intracellular domains of two monomers. Many class I receptors function as homodimers, whereas others form heterodimers with the ubiquitous receptors p130,  $\beta$ c and  $\gamma$ c. Heterodimerization is nonpromiscuous, so the hematopoietic cytokine receptors can be divided into three functional subfamilies depending upon which of the ubiquitous molecules each interacts with. Within each subfamily there is a degree of functional redundancy, reflecting the role of the common receptor monomer. Class II receptors are formed from multiple subunits, many of which are uncharacterized. However, activation of the class II receptors also involves oligomerization.

The intracellular domain of each receptor is constitutively associated with a protein tyrosine kinase of the Janus family (JAKs; Table 28.2). Dimerization stimulates autotransphosphorylation of reciprocal JAKs, which become activated and phosphorylate the receptor itself. Phosphotyrosine residues on the receptors then recruit STATs (signal transducers and activators of transcription) through their SH2 domains (Table 28.2). Phosphorylation of the STATs induces dimerization and translocation to the nucleus, where they act as transcriptional regulators.

Antigen receptors are unique because of their idiotypic diversity generated through somatic recombination (q.v. *V(D)J recombination*), and in the case of the immunoglobulins, *somatic hypermutation* (q.v.). They are multicomponent receptors, with ligand-binding subunits of variable structure associated with invariable subunits whose function is to recruit signaling complexes upon activation. The invariable subunits of all antigen receptors contain a motif known as an **ITAM** (**immune tyrosine activation motif**; see Figure 28.1) which interacts with, and activates, a cytoplasmic protein

**Table 28.3:** The TGF- $\beta$  superfamily of growth factors, their class I and class II receptors and downstream targets where known

Family	Ligands	Type I receptor	Type II receptor	Targets where known
TGF- $\beta$	TGF- $\beta$ 1/5 TGF $\beta$ -2 TGF $\beta$ -3	T $\beta$ R-I (Alk-5)	T $\beta$ R-II	Smad 2 (Smad 3)
Activin/inhibin	Activin Several inhibins	ActR-I, Act-IB, Atr1	Act-RII, ActRIIB Punt (Atr-II)	
BMP (DPP)	Dpp/ BMP-2	Punt (Atr-II)	Saxophone, Thick veins	MAD Smad 1 (Smad 5)
(OP-1)	BMP-4 BMP-5 Vgr-1/BMP-6 OP-1/BMP-7	BMPR1A, BMPR1B	BMPR-II	
(GDF-5)	OP-2/BMP-8 GDF-5/CDMP-1 GDF-6/CDMP-2 GDF-7	ActR-I, BMPR-1A, BMPR-1B	Act-RII, Act-RIIB, BMPR-II, C14	
(BMP-3)	BMP-3 GDF-10			
Other BMPs	Vg-1/GDF-1 Dorsalin/Screw/ BMP-9 Nodal GDF-3/Vgr-2			
Unclassified	GDNF GDF-9 MIS		C14	

There is a degree of cross-talk between different pathways, as can be seen, e.g., by binding of activin receptors by both activins and BMPs. The receptors for many TGF- $\beta$  molecules have yet to be identified, and there are several orphan receptors.

Abbreviation: TGF- $\beta$ , transforming growth factor- $\beta$ ; BMP, bone morphogenetic protein; Dpp, decapentaplegic (*Drosophila* BMP-2 homolog); OP, osteogenic protein; CDMP, cartilage-derived morphogenetic protein; GDF, growth/differentiation factor; GDNF, glial-derived neurotrophic factor; MIS, Mullerian inhibiting substance.

tyrosine kinase. Of the many nonreceptor tyrosine kinases known (see *Figure 28.3*), Lck, Fyn and Zap are critical for T-cell receptor signaling, Blk, Fyn, Lyn, Syk and members of the Tec family for B-cell receptor signaling, and Csk kinase plays an apparent regulatory role.

**Receptor serine/threonine kinases.** Receptors for members of the TGF- $\beta$  superfamily of signaling molecules (*Table 28.3*) are classified according to size and ligand-binding specificity. A number of type I and type II receptors are known to possess an intrinsic serine/threonine kinase activity. Type I receptors typically possess a smaller cysteine-rich extracellular domain and a smaller C-terminal tail than the type II receptors, and also have a unique GS domain (glycine and serine rich) adjacent to the kinase domain.

Signal transduction involving receptor serine/threonine kinases may require receptor oligomerization. TGF- $\beta$  initially binds to the class II TGF- $\beta$  receptor, which is a constitutively active kinase. Binding recruits a class I TGF- $\beta$  receptor, which is activated by phosphorylation on serine residues and can itself phosphorylate downstream targets. Other signaling molecules in the superfamily can bind directly to type I receptors, notably the bone morphogenetic proteins.



Several of the downstream targets for TGF- $\beta$  signaling are cyclin-dependent kinase inhibitors (see *The Cell Cycle* and *q.v. tumor suppressor genes*), which, by inactivating cyclin-dependent kinases, prevent phosphorylation of the retinoblastoma protein and block progression into the S-phase. Activated type I receptors also phosphorylate proteins of the **SMAD family** on the C-terminal consensus motif SSV/MS. SMADs then form heterodimers and translocate to the nucleus, where they interact with DNA-binding proteins and influence transcription, or bind to DNA directly. More than 10 SMAD proteins have been identified in vertebrates and invertebrates, some of which are pathway-specific. Smad4 and the homologous *Drosophila* protein appear to be common to both TGF- $\beta$  and BMP signaling pathways. Smad6 and Smad7 are thought to be inhibitors of SMAD phosphorylation, thus acting as feedback controls.

**Other membrane receptors.** The four receptor types discussed above represent nearly all known receptor molecules involved in signal transduction in animal cells. Some signals, however, are transmitted through receptors with distinct signaling mechanisms. The atrial natriuretic peptides, for instance, bind receptors with intrinsic guanyl cyclase activity. These increase the intracellular levels of cGMP and activate protein kinase G (PKG). A number of receptors with protein tyrosine phosphatase activity have also been described, including the leukocyte common antigen CD45 receptor. The tumor necrosis factor receptor family signal using **death domains** which bind to cytoplasmic proteins with similar domains (*q.v. apoptosis*). For other receptors well characterized at the genetic level, the signal transduction mechanisms remain far from clear (e.g. Notch and Delta which control *lateral inhibition* (*q.v.*) in the nervous system).

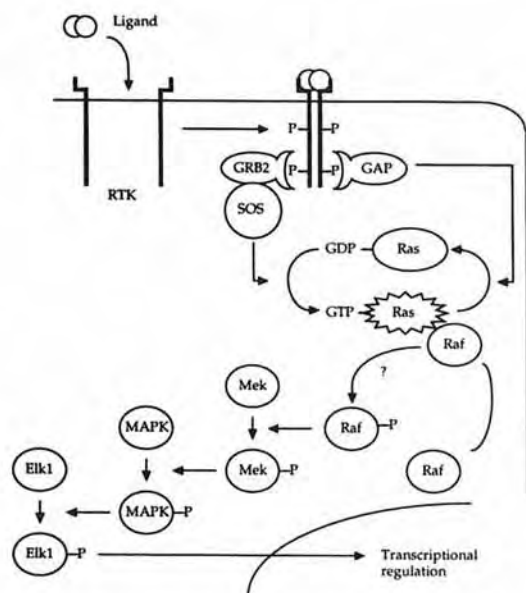
**Signal transduction through cytoplasmic and nuclear receptors.** Most receptors for extracellular signals are on the cell surface because the signaling molecules are hydrophilic and unable to penetrate the plasma membrane. Other molecules can cross the plasma membrane directly and interact with cytoplasmic or nuclear receptors, and these often initiate simple signal transduction pathways. Examples of such signals include the steroid and thyroid hormones and retinoic acid, which are lipid-soluble, and the gas nitric oxide (*q.v. receptor guanylate cyclases*).

The steroid hormone receptor superfamily includes receptors for steroid and thyroid hormones, and vitamins A and D and their derivatives, including the important developmental molecule retinoic acid. Interaction with the ligand causes a conformational change (**transformation**) which stimulates DNA binding activity. Some receptors are located in the cytoplasm (e.g. the glucocorticoid receptor), and transformation allows nuclear translocation. Others are already present in the nucleus. The transformed receptors are transcription factors and interact with DNA through a highly conserved zinc-binding domain (see *Nucleic Acid-binding Proteins*). They can be divided into several families based on the architecture of the recognition site. The activity of some receptors can be regulated by phosphorylation as well as ligand-binding (e.g. the progesterone and estrogen receptors), allowing integration with other signaling pathways.

## 28.2 Intracellular enzyme cascades

**Intracellular signaling.** The activation of receptor tyrosine kinases or receptor-associated tyrosine kinases has been discussed as a mechanism for transducing a signal across the cell membrane without transferring the ligand into the cell. Since the ultimate targets of the signal transduction pathway are usually not found at the plasma membrane, there must be further *intracellular* signal transduction. This is often mediated by a **cascade** of sequential enzyme activation: one signaling component phosphorylates and activates a second component, which performs the same process on a third, and so on until the ultimate target is reached. The signaling process may also be inhibitory.

Some signaling pathways are simple. The JAK-STAT and TGF- $\beta$ /SMAD pathways have already been described; both involve the phosphorylation of receptor-associated proteins which then translocate to the nucleus and act as transcription factors or their components. Other pathways



**Figure 28.2:** The Ras-Raf-MAP kinase cascade. Ligand binding stimulates dimerization and autotransphosphorylation of receptor tyrosine kinases. Proteins with SH2 domains recognize the phosphotyrosine residues and recruit signaling complexes. The adaptor protein GRB2 has SH2 and SH3 domains; the latter binds the guanine nucleotide exchange factor (GNEF) SOS. This stimulates the activation of membrane-bound Ras which recruits Raf. Raf becomes phosphorylated by an unknown membrane-associated kinase, and then induces a phosphorylation cascade involving Mek and MAP kinase. MAP kinase phosphorylates a number of latent transcription factors (Elk-1 is used as the example in the figure). The various kinase activities are inhibited by cytoplasmic phosphatases, and Ras activity is inhibited by GTPase-activating proteins (GAPs) which are also recruited by RTK signaling, thus switching off the signal once it has been transduced.

involve many steps and provide ample opportunity for branching and integration. The interaction of different pathways is reflected by the recurring motifs found in signaling molecules, which allow interactions with other proteins. Some of these motifs are discussed in *Box 28.1*.

**The Ras pathway.** Ras family proteins are small membrane-associated proteins whose activity depends upon a bound guanosine nucleotide cofactor. Ras itself is a central component in many signal transduction pathways. The related proteins Rac and Rho are involved in the control of cell mobility and cytoskeletal organization. Ras cycles between active (GTP-associated) and inactive (GDP-associated) states, a process regulated by enzymes termed **GAPs** (GTPase activating proteins) and **GNRPs** or **GNEFs** (guanine nucleotide releasing proteins, guanine nucleotide exchange factors). The GAPs act by stimulating the intrinsic GTPase activity of Ras thereby inactivating it, whereas GNRPs have the opposite effect. Both GAPs and GNRPs interact with RTK signaling complexes through SH2 or SH3 domains (*Figure 28.2; Box 28.1*); thus signal transduction can either stimulate or inhibit Ras activity, depending upon the receptor and the abundance of each enzyme in an active form. Activated Ras interacts with and stimulates the activity of downstream proteins, the best-characterized of which is the serine/threonine kinase Raf (linking to the MAP kinase pathway). Activation is brought about by recruiting Raf to the plasma membrane (Raf is normally a cytoplasmic protein) where it may interact with other proteins, possibly protein kinases. Recombinant Raf protein which is constitutively membrane-bound is also constitutively active. Thus, bringing Raf to the plasma membrane is sufficient to activate its kinase activity. Few other targets for Ras have been identified; one likely candidate is the enzyme phosphoinositide 3-kinase (PI(3)K) (see *second messengers*, below).

**Table 28.4:** The RTK-Ras pathway linked to the MAP kinase pathway through phosphorylation of Mek by Raf

Species	<i>S. cerevisiae</i>	<i>S. pombe</i>	<i>C. elegans</i>	<i>D. melanogaster</i>	Mammals
Receptor			Let-23	Sevenless	Many
Adaptor			Sem-3	Drk	Grb2
GNRF				Sos	SOS
Ras			Let-60	Ras1	Ha-Ras, Ki-Ras, N-Ras
Raf			Let45	D-Raf	c-Raf, A-Raf, B-Raf
Mek	STE7	Byr1	Mek-2	D-Sor	Mek1, Mek2
MAPK	FUS3, KSS1	Spk1	Mpk-1/ Sur-1	Rolled	Erk1, Erk2
Targets	STE12, FAR1		Lin-1, Lin-31	Pointed Rsk,	c-Fos

Yeasts lack RTKs, but signals are transduced through G-protein-coupled receptors which activate the MAP kinase pathway through Mek kinase. The yeast pathways are transducing mating pheromone signals, the *C. elegans* pathway is involved with vulval specification and the *Drosophila* pathway is involved with the specification of the R7 photoreceptor cell during eye development. The mammalian pathway has many redundant components.

Three Ras genes have been characterized in humans (Ha-ras, Ki-ras, N-ras), all three of which have been identified as cellular and/or viral oncogenes (see Oncogenes and Cancer).

**The MAP kinase signaling cascade.** MAP kinase (mitogen-activated protein kinase, also known as Erk, extracellular-signal regulated kinase) is a serine/threonine protein kinase activated by many growth factors. The pathway to MAP kinase activation involves Ras and Raf, with Raf phosphorylating a kinase upstream of MAP kinase, termed MAP kinase kinase (MAPKK) or Mek (MAPK/Erk kinase) (Figure 28.2). MAP kinase requires phosphorylation of tyrosine and threonine residues to become activated, and Mek is an example of a **dual specificity kinase**, with the combined activities of both a tyrosine kinase and a serine/threonine kinase.

The MAP kinase pathway channels mitogenic signals (signals driving cell proliferation) from the cell surface to the nucleus, and many components of the pathway are therefore oncogenic when inappropriately activated (see Oncogenes and Cancer, The Cell Cycle). The activation of MAP kinase causes it to be translocated to the nucleus, where it phosphorylates and activates a number of transcriptional regulators including Elk-1, C/EPB $\beta$  and c-Myc (see below). MAP kinase also phosphorylates a further kinase called Rsk which translocates to the nucleus and may activate other transcription regulators, such as the serum response factor (SRF).

The MAP kinase pathway is highly conserved in eukaryotes. As well as mediating an important growth response in mammals, pathways with homologous components are found in *Drosophila* and *C. elegans* (where they control cellular differentiation; see Development: Molecular Aspects) and in yeast, where the function is environmental monitoring (Table 28.4). In vertebrates, there is considerable redundancy of signaling components in the MAP kinase pathway, with multiple genes encoding the adaptors, Ras activating and inhibiting proteins, Ras and Raf, Mek and MAP kinase itself. These components may have varying substrate specificities and some are cell-type-specific or developmentally regulated, allowing them to play specific roles in signal transduction.

**Stress-activated kinases.** Stress-activated protein kinases (SAP kinases) are serine/threonine kinases, related to MAP kinase, which activate the transcription factor c-Jun in response to stresses such as UV-irradiation and inflammatory cytokines, but only poorly in response to growth stimulation. The SAP kinase pathway is also active during T-cell development and can induce proliferation in some cells. There are at least eight SAP kinase proteins, derived from three genes by alternative

splicing. The individual SAP kinases are named Jnk-1, Jnk-2, etc. reflecting their substrate specificity: they phosphorylate the N-terminal Ser63 and Ser73 residues of Jun (hence Jun N-terminal kinase) as well as other transcription factors such as Elk-1 and ATF-2.

Like the MAP kinases, SAP kinases require tyrosine and threonine phosphorylation for activation. Several families of putative MAP kinase-related molecules have now been identified, which are discriminated on the basis of their dual phosphorylation sites. MAP kinases, for instance, are identified by the sequence Thr-Pro-Tyr, whereas SAP kinases are identified by the sequence Thr-Glu-Tyr. The dual specificity kinases lying upstream of the SAP kinases are designated Sek (SAPK/Erk kinase) and are closely related to the yeast proteins STE7 (*S. cerevisiae*) and Byr1 (*S. pombe*). A protein isolated in mammals on the basis of its homology to STE11 and Byr2, which are upstream activators of STE7 and Byr1, was initially termed MAP kinase kinase (or Mek kinase) because it was able to phosphorylate Mek *in vitro* and *in vivo* when overexpressed. It is now thought that Mek kinase does not phosphorylate Mek under physiological conditions and is specific to the Sek-SAPK pathway. Mammalian homologs of the proteins found upstream of Mek kinase are beginning to be identified.

**Networks of cytoplasmic kinases and phosphatases which regulate signaling information.** As discussed above, responses to signals depend not only upon the nature of the signal itself (which reflects the availability of different signaling components) and the state of the responding cell, but also on the duration of the signal. This reflects the time taken to switch a signal off once activated. The 'damping-down' of signal transduction pathways is essential so that signaling mechanisms do not become saturated, and for every active component of a signaling pathway, there is also an inhibitor. Inhibitors can be thought of as acting in several ways.

- (1) When signal transduction pathways are activated, one of the components activated is usually an inhibitor of the same pathway. The inhibitor acts slightly later than the activator so that the signal is switched off, but not before a pulse of information transfer.
- (2) The launch of a signal in the first instance may depend upon the balance of activators and inhibitors of the pathway.
- (3) The purpose of some signals is to inhibit the transduction of others, through the activation of specific inhibitors.

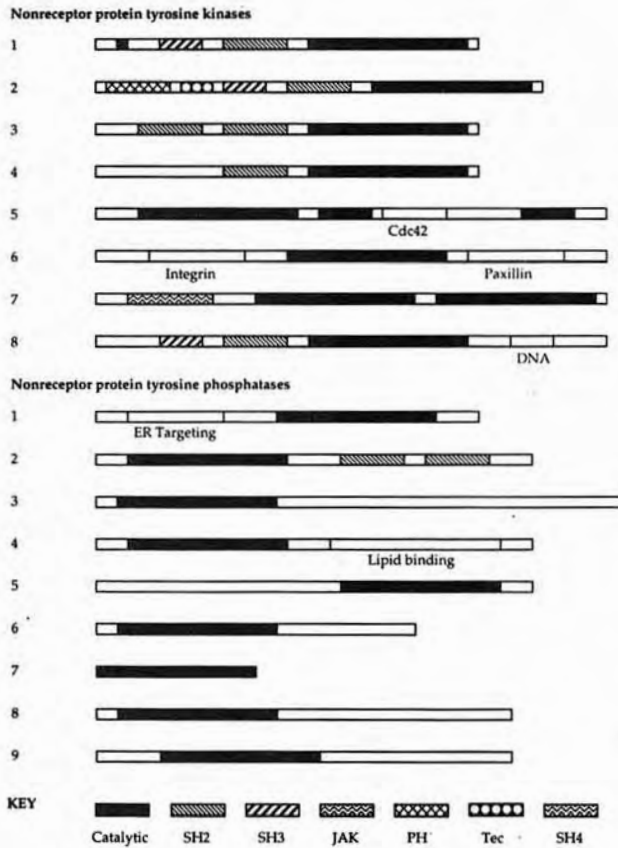
The cell can thus be thought of as a processing centre where information arriving at its surface is converted into the synthesis and activation of different molecules with opposing activities, such as kinases and phosphatases (Figure 28.3). Signaling pathways are controlled by a complex system of cross-links and feedback loops which rely on the balance between opposing forces in the cell. The arrival of new information disrupts the equilibrium maintained in the cell and causes, for example, a particular kinase to become transiently more active than the opposing phosphatase. This allows a burst of kinase activity culminating in the activation of a given transcription factor or enzyme before the signal is shut off by feedback and a new equilibrium is established.

### 28.3 Second messengers

**The second-messenger concept.** Cells respond to a diverse range of signals, and require a huge repertoire of receptors. However, the range of responses is much smaller. Many of the signals arriving at the cell surface, for instance, cause the cell to either divide or withdraw from the cell cycle. Others induce the expression of characteristic groups of genes which protect the cell from stress. For this reason, early signaling pathways converge into a small number of intracellular signaling networks, and this allows the cell to convert the complex information arriving at the cell surface into simple biochemical signals in the cytoplasm. The molecules involved in this process are termed **second messengers**.

**Cyclic nucleotide second messengers.** The earliest second messengers to be discovered were the cyclic nucleotides cAMP and cGMP. The levels of these molecules in the cell are controlled by the

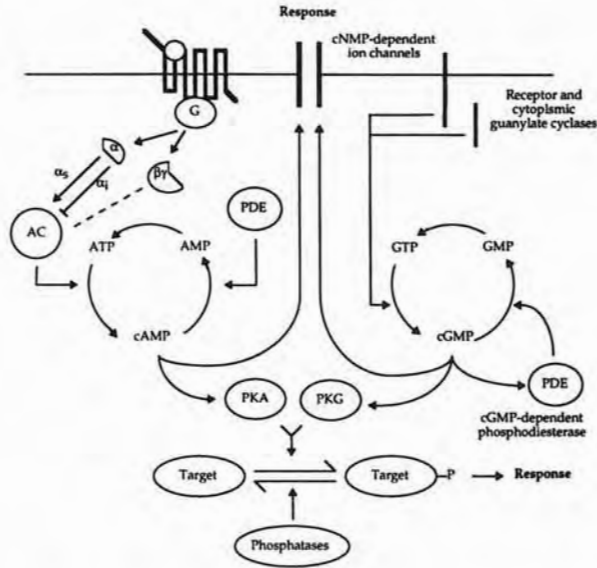




**Figure 28.3:** Cytoplasmic protein tyrosine kinases and phosphatases, which are responsible for the complex regulatory pathways controlling intracellular signaling. Kinases are grouped into families based on domain structure and size. '1' represents the domain structure common to four families (Csk, Sm, Brk/Rak and Lyn/Lyk/Blk) which differ in size; only the Lyn/Lyk/Blk family possesses the SH4 domain. 2, Tec family (Tec, Src, Yes, Fyn); 3, Syk/Zap family; 4, Fes family; 5, Ack; 6, Fak; 7, Janus family; 8, Abl/Arg. Phosphatases are classified according to the structure of their noncatalytic domains: 1, PTP1B; 2, SHPTP; 3, PTP1H; 4, MEG 2; 5, PTP-PEST; 6, YOP 2b; 6, VH1; 7, Cdc 25A; 8, MKP-1. 6–8 are dual specificity phosphatases; MKP-1 is thought to be the MAP kinase phosphatase which inactivates MAP kinase. Specific binding domains are shown where appropriate.

opposing activities of nucleotidylate cyclases, which catalyze the reaction  $\text{NMP} \rightarrow \text{cNMP}$ , and cyclic nucleotide phosphodiesterases (PDEs) which catalyze the reverse reaction. Several early signaling pathways influence the activity of nucleotidylate cyclases. G-protein-linked receptors of the  $G_s$  and  $G_i$  families stimulate and inhibit adenylate cyclase, respectively, whereas transducin ( $G_t$ ) stimulates cGMP PDE activity. A small number of cell surface receptors have intrinsic guanylate cyclase activity. There are also **receptor guanylate cyclases** that bind nitric oxide via a heme group.

Downstream of the cyclic nucleotides are effectors dependent on cyclic nucleotides for their activity: protein kinase A (PKA), which is cAMP-dependent, and protein kinase G (PKG), which is cGMP dependent, are serine/threonine kinases with a variety of substrates (see below). There are also cyclic nucleotide-gated ion channels and feedback pathways involving cyclic nucleotide-dependent PDEs. Other PDEs are regulated by phosphorylation or by calmodulin (q.v. *calcium signaling* below). The control of cyclic nucleotide levels in the cell is summarized in Figure 28.4.

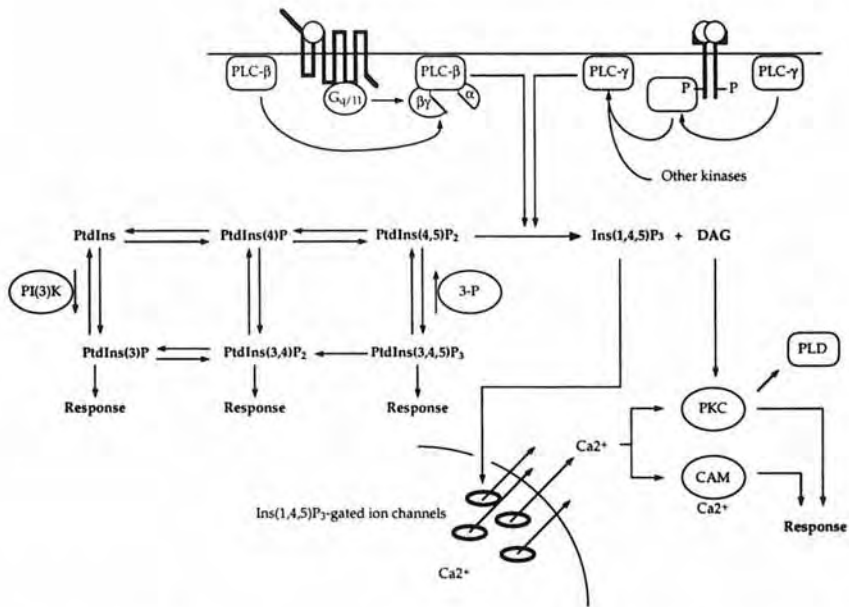


**Figure 28.4:** Cyclic nucleotide second messengers. cAMP is synthesized by adenylyl cyclase which is stimulated by G-protein  $\alpha_s$  subunits and inhibited by G-protein  $\alpha_i$  subunits ( $\beta\gamma$  subunits may also influence AC activity); it is removed from the cell by cAMP phosphodiesterases (PDEs). cGMP is synthesized by guanylate cyclase domains of membrane-spanning or cytoplasmic receptors. Both cAMP and cGMP activate cyclic nucleotide-dependent kinases and ion channels. cGMP also activates a cGMP-dependent PDE in a negative feedback loop.

Protein kinase A phosphorylates a wide range of proteins in the cell including many enzymes, receptors, ion channels and transcription factors. One example of the latter is the transcription factor CREB which binds to a *cis*-acting element termed the cAMP response element (see below). A smaller number of targets for PKG have been determined, including several G-proteins, and the  $\text{Ins}(1,4,5)\text{P}_3$  receptor. PKA and PKG function as homodimers, although PKA exists as an inactive tetramer with catalytic and regulatory subunits, the latter being released upon cAMP-binding. The catalytic and regulatory subunits of PKG are part of a single polypeptide chain. There are several isoforms of PKA and PKG, which may display differential substrate specificities. In addition, there is a degree of cross-talk between the enzymes, as PKA can be activated by cGMP and PKG by cAMP in certain cells. The activities of PKA and PKG are reversed by a collection of protein phosphatases with varying substrate specificities and complex transcriptional and posttranscriptional regulatory pathways. The response of the cell to a particular signal depends upon the balance of specific kinase and phosphatase activities.

**Lipids as second messengers.** A well-characterized second-messenger system involves products of the hydrolysis of a minor phospholipid component of the inner cell membrane, **phosphatidylinositol-4,5-bisphosphate ( $\text{PtdIns}(4,5)\text{P}_2$ )**. Activation of G-protein-linked receptors associated with  $\text{G}_q$  family proteins ( $\text{G}_q$ ,  $\text{G}_{11}$ ,  $\text{G}_{12}$ ) stimulates the activity of phospholipase C- $\beta$  (PLC- $\beta$ ). PLC enzymes cleave  $\text{PtdIns}(4,5)\text{P}_2$  into two components, **inositol-1,4,5-trisphosphate ( $\text{Ins}(1,4,5)\text{P}_3$ )** and **1,2-diacylglycerol (DAG)**. Another PLC isoform, PLC- $\gamma$ , is activated by RTK signaling by recruitment to the membrane through its SH2 domain. Both  $\text{Ins}(1,4,5)\text{P}_3$  and DAG are important second messengers (Figure 28.5).

$\text{Ins}(1,4,5)\text{P}_3$  mediates calcium release from the endoplasmic reticulum (ER) or sarcoplasmic reticulum (SR) by activating  $\text{Ins}(1,4,5)\text{P}_3$ -gated  $\text{Ca}^{2+}$  channels (q.v. *calcium signaling*, below). Its effects can be mimicked by **ionophores**: ion transporters such as A23187, which act like constitutively open



**Figure 28.5:** Phospholipid and calcium ion second messengers in the cell. Phospholipase C (PLC) converts the lipid membrane component PtdIns(4,5)P<sub>2</sub> into inositol-4,5,6-trisphosphate (Ins(1,4,5)P<sub>3</sub>) and diacylglycerol (DAG). Different forms of PLC are stimulated by G<sub>q</sub> proteins and phosphotyrosyl residues on activated receptor and nonreceptor kinases. Ins(1,4,5)P<sub>3</sub> activates Ins(1,4,5)P<sub>3</sub>-dependent calcium channels in the endoplasmic reticulum (ER) membrane releasing calcium ions into the cytoplasm which cooperate with DAG to activate phosphokinase C (PKC). This has many substrates, including phospholipase D (PLD). Ca<sup>2+</sup> also binds to calmodulin (CaM) which activates Ca<sup>2+</sup>/CaM-dependent kinases. Phosphatidylinositol (PtdIns) can be phosphorylated at several positions and five phosphorylated derivatives can be interconverted. The enzyme phosphoinositide 3-kinase (PI(3)K) adds phosphate groups to the 3D position, initiating a number of downstream responses. This process is reversed by the enzyme 3-phosphatase (3-P).

calcium channels. Some Ins(1,4,5)P<sub>3</sub> is further phosphorylated to inositol-1,3,4,5-tetrakisphosphate (Ins(1,3,4,5)P<sub>4</sub>), whose effects are largely uncharacterized. Ins(1,4,5)P<sub>3</sub> is also able to interact specifically with various proteins, including phospholipase C-δ and SOS, thus initiating alternative signaling cascades. DAG, in combination with calcium and another membrane phospholipid phosphatidylserine, activates several isoforms of calcium-dependent protein kinase (protein kinase C, PKC). This effect can be mimicked by plant-derived **phorbol esters** which activate PKC directly. There are at least 11 PKC isoforms, grouped into four major families based on domain structure. They are cell-type-specific and differentially localized within the cell. Gene knockout, overexpression and selective inhibition studies have shown that they have overlapping substrate specificities and mediate different effects in different cells, but as such techniques disrupt the *in vivo* subcellular localization, it has been difficult to establish the specific targets of each enzyme under physiological conditions. One substrate of PKC is phospholipase D (PLD), which hydrolyzes membrane phospholipids although at the terminal phosphodiester bond, producing phosphatidic acid (PLD can also be activated by G-proteins).

Another lipid intracellular signaling system involves phosphatidylinositides phosphorylated at the D3 position. The enzyme **phosphoinositide 3-kinase (PI(3)K)** can add a D3-phosphate group to basic PtdIns as well as PtdIns(4)P and PtdIns(4,5)P<sub>2</sub> to generate PtdIns(3)P, PtdIns(3,4)P<sub>2</sub> and PtdIns(3,4,5)P<sub>3</sub>, respectively (Figure 28.5); there are three classes of PI(3)K which can use different subsets of phosphatidyl inositides as substrates. These lipids are absent in unstimulated cells but

accumulate following extracellular signals and oncogenic transformation. They bind specifically or with differing affinities to a number of downstream effectors, recruiting them to the membrane and initiating secondary signaling cascades. PtdIns(3,4,5)P<sub>3</sub> appears to be the most important of these second messengers: it interacts with multiple targets through SH2 and PH domains (e.g. Src, PI(3)K, phosphokinase C- $\epsilon$  and - $\lambda$  isoforms, and specific PtdIns(3,4,5)P<sub>3</sub>-dependent protein kinases PDK-1 and PDK-2). The targets of PtdIns(3,4)P<sub>2</sub> overlap with PtdIns(3,4,5)P<sub>3</sub>, but PtdIns(3)P has been shown to bind with great specificity to Adaptin, a protein involved in protein trafficking.

Lipid second messengers may also signal between cells. Phospholipase A<sub>2</sub> synthesizes arachidonic acid and lysophospholipids, which are intracellular second messengers. However, they are also intermediates in the production of prostaglandins and leukotrienes. There are nine different classes of prostaglandins, synthesized and secreted by many different cell types. They act as local (paracrine) signals and, unlike most lipid signaling proteins, bind to receptors on the cell surface.

**Calcium ions as second messengers.** The level of calcium ions in the cell controls diverse processes including metabolic processing, proliferation and specialized functions such as membrane excitability and muscle contraction. The calcium signaling pathway involves calcium-binding proteins which, when activated by calcium, associate with inactive proteins and stimulate them thus effecting a number of downstream pathways. The level of Ca<sup>2+</sup> in the cytoplasm is maintained at a low concentration by active export, but is up to 10000-fold higher in the extracellular fluid and in certain intracellular compartments. Upstream signals originating at cell-surface receptors induce the release of Ca<sup>2+</sup> into the cytoplasm by opening calcium channels, thus increasing the abundance of activated calcium-binding proteins and inducing downstream signaling pathways (Table 28.5). Two major pathways stimulating calcium release are the G-protein-linked receptors associated with G<sub>q</sub> family proteins, and the RTKs. Both these receptors stimulate phospholipases which increase levels of Ins(1,4,5)P<sub>3</sub> (Figure 28.5), leading to the opening of Ins(1,4,5)P<sub>3</sub>-gated Ca<sup>2+</sup> channels in the ER membrane.

There are many different calcium-binding proteins in the cytoplasm. Although most act as buffers (an additional mechanism for reducing cytoplasmic Ca<sup>2+</sup> concentration), a few act as calcium monitors and signaling proteins. The **annexins** are calcium-dependent membrane-binding proteins which may reorganize cytoskeletal elements of the cell; they also inhibit phospholipase A<sub>2</sub>. The major regulatory calcium-binding proteins belong to the EF-hand superfamily (see Proteins: Structure, Function and Evolution) and include **calmodulin (CaM)** and troponin C. Whereas CaM is a ubiquitous and multifunctional protein, troponin C is muscle-specific and controls the interaction

**Table 28.5:** Modulation of calcium ion levels in the cytoplasm

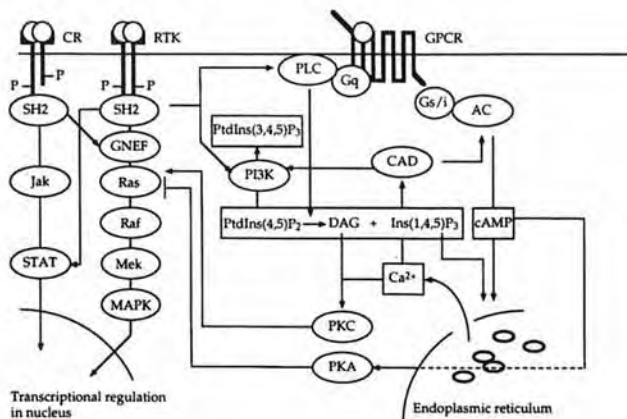
Ca <sup>2+</sup> transport	Mediators
Import from extracellular space	Voltage-gated Ca <sup>2+</sup> channels release Ca <sup>2+</sup> into cell following membrane depolarization Ligand-gated Ca <sup>2+</sup> channels release Ca <sup>2+</sup> into cell following ligand binding (e.g. NMDA receptors) CRAC channels (calcium release activated Ca <sup>2+</sup> channels which are stimulated by Ca <sup>2+</sup> release from ER)
Import from ER/SR	Ins(1,4,5)P <sub>3</sub> gated channels Ryanodine receptors which release Ca <sup>2+</sup> upon activation by cyclic adenosine diphosphate ribose (cADPR) derived from NAD <sup>+</sup> Ca <sup>2+</sup> can activate release of Ca <sup>2+</sup> by stimulating both the above channels
Export	Na <sup>+</sup> /Ca <sup>2+</sup> exchange — uses the energy derived from transporting Na <sup>+</sup> along its electrochemical gradient to export Ca <sup>2+</sup> against its electrochemical gradient Ca <sup>2+</sup> ATPases in plasma membrane and ER/SR which actively export Ca <sup>2+</sup> from cytoplasm



between actin and myosin during muscle contraction. CaM activates a number of protein kinases (CaM kinase II, elongation factor kinase), phosphatases (calcineurin) and cytoskeletal elements, and also activates  $\text{Ca}^{2+}$  ATPases, thus promoting  $\text{Ca}^{2+}$  removal from the cytoplasm. CaM also interacts with the components of other second-messenger systems including isoforms of adenylate cyclase, IP(3)K and nitric oxide synthetase.

**Cross-talk.** Several well-defined signaling pathways and second-messenger systems have been described in the preceding paragraphs, but as discussed, there is extensive interaction between the pathways so that individual stimuli can activate several pathways in the cell and different signals can produce the same effects. This can occur by several mechanisms.

- (1) **Receptor–ligand promiscuity**, where ligands activate multiple receptors or many ligands activate the same receptor, or where receptors comprise multichain oligomers with distinct signaling specificities.
- (2) **Divergence**, where the stimulus of a receptor activates two parallel pathways. An example is the activation of the Ras–Raf–MAP kinase pathway and phospholipase C- $\gamma$  (which increases levels of the second messengers Ins(1,4,5) $\text{P}_3$  and DAG) by RTK activity. This is mediated by different signaling molecules possessing domains, in this case the SH2 domain, which can interact with the activated receptor.
- (3) **Cross-talk**, where one pathway branches off and interacts with another (*Figure 28.6*). All the major signaling pathways in the cell use protein kinases and phosphatases. There is a high level of interaction between them, e.g. protein kinase A, which is activated by G-protein-mediated increases in cAMP levels, inactivates Ras. Conversely, protein kinase C, which is calcium-dependent, stimulates Ras. The major second-messenger systems of the cell are also interdependent: Ins(1,4,5) $\text{P}_3$  induces calcium transport from the ER and DAG cooperates with calcium to activate protein kinase C. The calcium-dependent molecule CaM regulates the activities of PI(3) kinase and adenylate cyclase, which control the levels of Ins(1,4,5) $\text{P}_3$  and



**Figure 28.6:** Examples of cross-talk in signaling pathways. The MAP kinase pathway can be activated by RTK signaling through Ras and Raf, but Ras may also be activated by cross-talk from cytokine receptors and protein kinase C (PKC), which is regulated by calcium. Some RTKs may also directly regulate STATs (e.g. EGF receptor). Ras is inactivated by protein kinase A (PKA) which is activated by cAMP. The major secondary-messenger systems of the cell are also interdependent, with Ins(1,4,5) $\text{P}_3$  and cAMP both influencing calcium release, and calcium controlling the activity of phosphoinositol 3-kinase (PI(3)K) and adenylate cyclase (AC) through kinases dependent on the calcium-binding protein calmodulin (CaM). The cytoplasm contains numerous other kinases and phosphatases which link different signaling pathways into a complex network. Proteins are represented by circles, second messengers by rectangles. CR, cytokine receptor; RTK, receptor tyrosine kinase; GPCR, G-protein-coupled receptor.

cAMP, respectively. cAMP can activate ion channels and hence influence the levels of  $\text{Ca}^{2+}$  in the cytoplasm.

With these myriad interconnections superimposed upon a regulatory network of broad specificity kinases and phosphatases, it is a wonder that any signaling specificity is maintained at all; the mechanisms of signaling specificity are a major topic of current research. However, cells synthesize only a subset of the many signaling molecules that have been described, so that individual cells can respond only to certain signals and can respond through pathways restricted by the particular components and active in the cell. Another mechanism which could be used to regulate signal response is the global regulation of gene expression. Thus activated transcription factors in the nucleus would find different arrays of genes available to them through selective epigenetic silencing (see Chromatin, DNA Methylation and Epigenetic Regulation).

## 28.4 Signal delivery

**Response to signals.** Diverse signals arriving at the cell surface are transduced into cascades of sequential kinase activity, or the production of second messengers which themselves exert their effects by modulating the activity of cellular kinases. The kinase cascade is controlled by protein phosphatases whose abundance and activity are subject to complex regulation. Signal delivery can be defined as the end of the signal transduction pathway, where the components of the cell which are targets for the initial signal are phosphorylated. In this respect, phosphate groups are a universal signaling currency causing changes to protein shape and therefore activity, although the same effect can be brought about by other forms of modification (q.v. *histone acetylation, protein modification*). Signals mediate their ultimate effects in two ways.

- (1) By modulating the activity of a protein already present in the cell. Protein kinase A, for instance, phosphorylates several enzymes with key roles in metabolism (e.g. in muscle cells it phosphorylates (and inhibits) glycogen synthase but leads to the activation of glycogen phosphorylase, thus switching the cell to net glucose use). Cytoskeletal proteins are also regulated by calmodulin-dependent kinases (e.g. neuromodulin, Tau; also q.v. *M-phase kinase*).
- (2) By modulating the activity of a transcriptional regulator or translational regulator and thus influencing gene expression.

These effects can be distinguished by blocking *de novo* gene expression with inhibitors of transcription or protein synthesis.

**Activation of transcription factors.** Many signal transduction pathways terminate at the nucleus. The downstream targets may be components of the cell-cycle machinery, allowing control of cell growth or transcription factors, which elicit specific patterns of gene expression through *response elements* (q.v.) on target genes. Many transcription factors have been shown to be activated by specific signaling pathways, and in many cases, the diversity of response is increased by the availability of different isoforms of transcription factor components. A selection of well-characterized pathways is shown in Table 28.6. Within about six hours of receptor stimulation, approximately 100 immediate early genes are activated. These include some proto-oncogenes (e.g. *c-jun*, *c-fos*), genes for many other transcriptional regulators (e.g. *ets*, *srf*, hormone receptors), and structural proteins (fibronectin, actin). Many of the genes are activated in response to both growth stimulatory signals (e.g. RTK–Ras–Raf–MAP kinase) and growth inhibitory signals (e.g. TGF- $\beta$ –CDI).

**Table 28.6:** Signal delivery to transcription factors

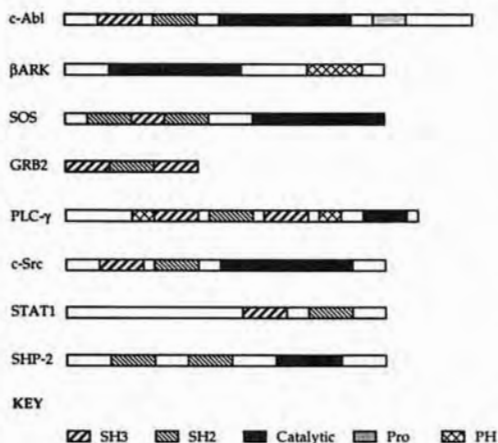
Transcription factor	Signal delivery mechanism (activation)
AP1 (c-Jun)	Dephosphorylation of DNA-binding domain by inhibition of certain kinases by Rsk-1, which is activated by MAP kinase
(c-Fos)	Phosphorylation of Ser-63 and Ser-73 in transactivation domain by SAP kinase Induction of c-fos gene expression by activation of Elk-1 (see separate entry) Phosphorylation of Thr-232 in transactivation domain by FRK, a Ras-activated kinase distinct from MAP kinase (Phosphorylation of C-terminal serine/threonine residues causes c-Fos to silence transcription of the c-fos gene)
CREB	Phosphorylation at Ser-133 by PKA allows interaction with accessory protein CREB binding protein (CBP) which interacts with initiation complex
CREM	Phosphorylation on Ser-117 required for transactivation
Dorsal (NF- $\kappa$ B)	Phosphorylation by Pelle kinase stimulates release of inhibitory factor Cactus (I $\kappa$ -B) revealing nuclear localization sequence and allowing translocation to the nucleus
Elk-1/SRF	Phosphorylation of Elk-1 on C-terminal Ser residues by MAP kinase (a) allows construction of a stable complex with SRF and other regulatory components, and (b) stimulates the transactivation domain
p53	DNA damage causes increased translation of p53 mRNA and increased protein stability
STATs	Phosphorylation stimulates dimerization and translocation to the nucleus.
Steroid receptor superfamily	Ligand-binding causes conformational change which allows DNA binding Some receptors also regulated by phosphorylation on serine residues.

**Box 28.1:** Src homology and other conserved signaling domains

**Domain structure of signaling proteins.** Many signal transduction proteins contain conserved domains facilitating interaction between the various components of signal transduction pathways. These were first identified as common modules found in nonreceptor tyrosine kinases of the Src family, and were termed **Src-homology (SH) domains**. SH1 is the kinase domain itself and is the most highly conserved module. SH2 is found in many tyrosine kinases and other molecules (see figure) — it is a phosphotyrosine-binding motif, i.e. it allows proteins to bind to phosphorylated tyrosine residues such as those found on activated RTKs. SH3 is a

further binding domain, preferentially interacting with proline-rich motifs. SH4 is a short domain regulating the myristoylation of glycine residues and thus allowing membrane docking (only the Src family kinases and certain isoforms of Abl contain an SH4 domain). Several other conserved signaling modules have been identified, including the phosphotyrosine-binding (PTB) domain, which is distinct from SH2, and the pleckstrin homology (PH) domain, whose precise function is unclear, but which involves lipid binding and interactions with  $\beta$ -subunits of G-proteins through WD40 motifs.

*Continued*



The figure shows diversity amongst signaling proteins with conserved domains SH2, SH3, Pro (proline-rich, putative SH3 recognition domains) and PH (Plekstrin homology). c-Abl and c-Src are kinases, PLC-γ is a phospholipase, SHP-2 is a phosphatase, βARK is a receptor tyrosine kinase and STAT1 is a transcription factor, SOS is a guanine nucleotide exchange factor and GRB2 is an adaptor protein. Note that the PH domain of PLC-γ is interrupted by the SH3-SH2-SH3 module.

### Further reading

- Barford, D. (1996) Molecular mechanisms of the protein serine/threonine phosphatases. *Trends Biochem. Sci.* 21: 407–412.
- Bourne, H.R. (1997) How receptors talk to trimeric G proteins. *Curr. Opin. Cell Biol.* 9: 134–142.
- Cohen, C.B., Ren, R. and Baltimore, D. (1995) Molecular binding domains in signal transduction proteins. *Cell* 80: 237–248.
- Cohen, P.T.W. (1997) Novel protein serine/threonine phosphatases: Variety is the spice of life. *Trends Biochem. Sci.* 22: 245–251.
- Fauman, E.B. and Saper, M.A. (1996) Structure and function of the protein tyrosine phosphatases. *Trends Biochem. Sci.* 21: 413–417.
- Houslay, M.D. and Milligan, G. (1997) Tailoring cAMP-signalling responses through isoform multiplicity. *Trends Biochem. Sci.* 22: 217–224.
- Jan, L.Y. and Jan, Y.N. (1997) Receptor-regulated ion channels. *Curr. Opin. Cell Biol.* 9: 155–160.
- Lohmann, S.M., Vaandrager, A.B., Smolenski, A., Walter, U. and de Jonge, H.R. (1997) Distinct and specific functions of cAMP-dependent protein kinases. *Trends Biochem. Sci.* 22: 307–312.
- Marshall, C.J. (1995) Specificity of receptor tyrosine kinase signalling: Transient versus sustained extracellular signal-regulated kinase activation. *Cell* 80: 179–186.
- Morrison, D.K. and Cutler, R.E. Jr. (1997) The complexity of Raf-1 regulation. *Curr. Opin. Cell Biol.* 9: 174–179.
- Newton, A.C. (1997) Regulation of protein kinase C. *Curr. Opin. Cell Biol.* 9: 161–167.
- Pellegrini, S. and Dusanter-Fourt, I. (1997) The structure, regulation and function of the Janus kinases (JAKs) and the signal transducers and activators of transcription (STATs). *Eur. J. Biochem.* 248: 615–633.
- Robinson, M.J. and Cobb, M.H. (1997) Mitogen-activated kinase pathways. *Curr. Opin. Cell Biol.* 9: 180–186.
- Singer, W.D., Brown, H.A. and Sternweis, P.C. (1997) Regulation of eukaryotic phosphatidylinositol-specific phospholipase C and phospholipase D. *Annu. Rev. Biochem.* 66: 475–509.
- Spiegel, S., Foster, D. and Kolesnick, R. (1996) Signal transduction through lipid second messengers. *Curr. Opin. Cell Biol.* 8: 159–167.
- Sprang, S.R. (1997) G protein mechanisms: Insights from structural analysis. *Annu. Rev. Biochem.* 66: 639–678.
- Ten Dijke, P., Miyazono, K. and Heldin, C.-H. (1996) Signalling via hetero-oligomeric complexes of type I and type II serine/threonine kinase receptors. *Curr. Opin. Cell Biol.* 8: 139–145.
- Toker, A. and Cantley, L.C. (1997) Signalling through the lipid products of phosphoinositide-3-OH kinase. *Nature* 387: 673–676.
- Van Haesebroeck, B., Leeyers, S.J., Panayotou, G. and Waterfield, M.D. (1997) Phosphoinositide 3-kinases: A conserved family of signal transducers. *Trends Biochem. Sci.* 22: 267–272.
- Whitman, M. (1997) Feedback from inhibitory SMADs. *Nature* 398: 549–551.
- Wittinghofer, A. and Nassar, N. (1997) How Ras-related proteins talk to their effectors. *Trends Biochem. Sci.* 21: 488–491.



## Chapter 29

# Transcription

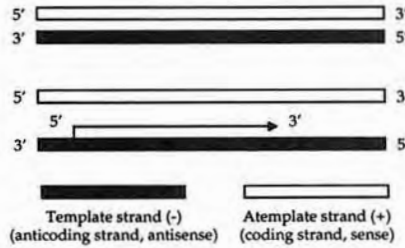
### Fundamental concepts and definitions

- **Transcription** is the synthesis of RNA using DNA as a template. *In vivo*, transcription is the first level of gene expression and the predominant level of gene regulation. Additionally, RNA primers for cellular DNA replication are generated by transcription, and transcription plays a major role in the replication cycle of the *retroid viruses* (q.v.). The components of several bacterial and eukaryotic transcription systems have been defined, allowing the *in vitro* transcription, and coupled *in vitro* transcription and translation of cloned genes.
- Unlike *DNA replication* (q.v.), transcription is asymmetric — only one strand of the DNA is used as a template (Figure 29.1). The nascent RNA is analogous to the *leading strand* (q.v.) in DNA replication, i.e. it is transcribed continuously. Each RNA molecule is a **transcript**, and the region of DNA from which it was transcribed is a **transcription unit** (q.v. *gene, cistron, operon*). The first nucleotide in the transcript is defined as position +1 of the transcription unit. The nucleotide immediately preceding this on the corresponding DNA strand is defined as position -1; there is no position 0.
- Enzymes which catalyze transcription are termed **(DNA-dependent) RNA polymerases**. Unlike DNA polymerases, RNA polymerases initiate strand synthesis *de novo* (i.e. without primers) and do not proofread their transcripts (see Box 26.1).
- Successful transcription requires RNA polymerase to be recruited to a *cis*-acting **promoter** site upstream of the active gene. Prokaryotic RNA polymerases bind to DNA directly, whereas eukaryotic enzymes require other proteins to form an initiation complex at the promoter. The efficiency of RNA polymerase recruitment may be influenced by transcription factors acting either positively or negatively, and binding at the promoter, or at more distant enhancer (positive) or silencer (negative) sites, which interact with the initiation complex by the looping out of intervening DNA.

### 29.1 Principles of transcription

**Stages of transcription.** Like other polymerization reactions, transcription is divided into three stages: **initiation**, where the RNA polymerase binds to the DNA and begins RNA synthesis, **elongation**, where the RNA strand is extended and the majority of RNA synthesis takes place, and **termination**, where elongation ceases and the transcript dissociates from the template. The detailed mechanisms differ between prokaryotes and eukaryotes, but similar principles apply.

At initiation, RNA polymerase recognizes and binds to a *cis*-acting element, the **promoter**, located close to the transcriptional start site of the gene. Initial binding generates a **closed promoter complex**. Successful initiation converts this assembly into an **open promoter complex** where the DNA is locally unwound. Transcription begins with the insertion of the first ribonucleotide (usually a purine nucleotide). The end of initiation is signified by **promoter clearance**, where the RNA polymerase moves away from the promoter site without dissociating, freeing the promoter for further initiation events. Promoter clearance occurs only if the open promoter complex is stable, and usually follows a number of **abortive initiations** where short transcripts are generated. This is a general property of RNA polymerases and appears to be required for *de novo* strand synthesis. Initiation is usually the rate-limiting step in transcription, and is the primary level of gene regulation in both prokaryotes and eukaryotes. Regulatory factors which bind to *cis*-acting sites surrounding the promoter, and also at more distant sites, interact with the initiation proteins to



**Figure 29.1:** Nomenclature of DNA strands in a transcription unit. It is necessary to discriminate between the two DNA strands because only one acts as the template for transcription. The **template strand** is complementary to the transcript, whereas the **nontemplate strand** carries the same sequence as the transcript, except that thymidine replaces uridine. In protein-coding genes, the nontemplate strand carries the same information (codons) as the mRNA and is described as the **coding strand, sense strand\*** or **(+) strand**, whereas the template strand is the **anticoding strand, antisense strand\*** or **(-) strand**; the anticoding strand has the same sequence as antisense RNA. The sense of the genomic DNA strand often switches because adjacent genes differ in their orientation. Only in certain virus genomes are all genes in the same orientation, so that the entire genome can be designated as (+) or (-) sense (also q.v. *ambisense RNA*, *nonsense codon*, *missense mutation*, *countertranscript RNA*). \*There is some disagreement as to the use of the terms sense and antisense when referring to DNA strands. The definition given here is the logical and popular one (i.e. sense DNA strand = mRNA sequence; antisense DNA strand = antisense RNA sequence) although this differs from its original usage.

modulate either initial binding of the enzyme, the formation of the stable open promoter complex or the efficiency of promoter escape.

Once successful initiation has been achieved, the RNA polymerase moves processively along the template synthesizing nascent RNA in the 5'→3' direction. The double-stranded DNA is unwound ahead of the elongation complex and rewound behind it; in eukaryotes, this involves disruption of the nucleosome structure of DNA (see Chromatin). The transcript is thus paired with the template only transiently. The dynamic, transiently melted structure representing the site of transcription is a **transcription bubble**. The elongation rate is usually constant, although it may be perturbed by secondary structure in the template (q.v. *cotranscriptional regulation*, *transcription-coupled repair*). In prokaryotes, continued elongation may depend upon concurrent protein synthesis (q.v. *attenuation*).

Transcription may terminate by several different mechanisms: secondary structures in the nascent transcript, termination sites in DNA or cleavage of the transcript. Termination of transcription involves release of the enzyme and any associated factors, and liberation of the nascent transcript.

**Three components of transcriptional activity.** Transcriptional activity is the rate of RNA synthesis, e.g. the number of full-length transcripts synthesized per minute. Initiation is usually rate-limiting, so transcriptional activity directly reflects the efficiency of transcriptional initiation. The control of initiation is divided into three components: basal, constitutive and regulated.

**Basal components** serve primarily to recruit RNA polymerase to the start of transcription, *permitting* the initiation of RNA synthesis.

**Constitutive components** control the efficiency of initiation in the absence of regulation and therefore dictate the default level of transcriptional activity of each gene, allowing different genes to be expressed at different rates to suit the fundamental needs of the cell.

**Regulatory components** alter initiation efficiency, allowing the transcriptional activity of individual genes to be modulated in response to changes in the environment. In multicellular organisms, the regulation of transcription also plays a major role in specifying and maintaining differentiated cells.

These three components of transcriptional control differ between prokaryotes and eukaryotes, as discussed below (see Table 29.1).

**Table 29.1:** Comparison of prokaryotic and eukaryotic transcriptional initiation control

	Bacteria	Eukaryotes
Basal promoter element	-10 and -35 sequences (context at start site and downstream also contribute to initiation)	Class I: Core promoter Class II/some Class III: Usually initiator and/or TATA box Class III: Internal bipartite promoter
Recognition	$\sigma$ -factor of RNA polymerase	TBP and/or other GTFs
Control of constitutive expression	Depends entirely on basal promoter sequence, context and architecture	Depends both on basal promoter structure and the presence of constitutive transcriptional activators bound at <i>cis</i> -acting sites in promoter/enhancers
Regulated expression	Controlled predominantly by proteins binding to regulatory elements lying adjacent to or overlapping promoter. Distant enhancers rare	Class I genes are constitutive. Class II and III gene expression controlled by proteins binding to regulatory elements in upstream promoter and distant enhancer/silencer elements which interact with the promoter by looping out the intermediate DNA
Strategies	Functionally related genes often clustered as operons and cotranscribed. Close spacing of genes allows regulation by countertranscription. Rapid response strategies are important, hence predominance of allosteric modulation in transcription factor regulation, antisense RNA control, unstable mRNA and linked regulation of transcription and protein synthesis	Functionally related genes often dispersed (c.f. <i><math>\beta</math>-globin cluster</i> , <i>Hox genes</i> ) with overlapping sets of regulatory elements — absence of operon organization (see text for exceptions), but some examples of coordinated <i>cis</i> -regulation (e.g. <i>q.v. locus control region</i> ). Genes widely spaced — countertranscription uncommon (c.f. <i>parental imprinting</i> .) Both rapid and slow responses common. Rapid responses usually linked to signal transduction, hence predominance of covalent modification in transcription factor regulation. Slow responses include differentiation, often involving <i>de novo</i> synthesis of transcription factors

TBP, TATA-binding protein; GTF, general transcription factor.

## 29.2 Transcriptional initiation in prokaryotes — basal and constitutive components

**RNA polymerases.** In prokaryotes, a single RNA polymerase transcribes all genes. The *E. coli* polymerase has a tetrameric **core enzyme** containing  $\alpha$ - and  $\beta$ -type subunits with the stoichiometry  $\alpha_2\beta\beta'$ . This is sufficient for transcriptional elongation, but initiation requires a further subunit termed  $\sigma$ , which completes the **holoenzyme**. The  $\sigma$ -factor has two functions: it recognizes the promoter and it converts the closed promoter complex into an open promoter complex. Once transcription is initiated, the  $\sigma$ -factor dissociates from the holoenzyme. The core enzyme can bind to DNA in the absence of  $\sigma$ , but with low efficiency and no specificity. The primary function of the  $\sigma$ -factor is thus to increase the binding efficiency of RNA polymerase at the promoter and decrease nonspecific binding.

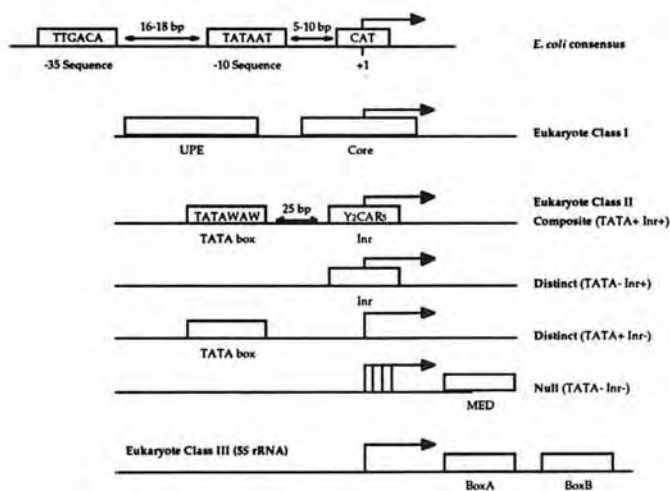
A single  $\sigma$ -factor ( $\sigma^{70}$  in *E. coli*) initiates the transcription of most genes, but other  $\sigma$ -factors, recognizing variant promoters, may be synthesized to induce the coordinated expression of specialized genes (e.g. those involved in the heat shock response). Some bacteriophages (e.g. T4)

encode their own  $\sigma$ -factors which subvert the host core enzyme into transcribing the phage genes. Others (e.g. T3, T7) encode their own RNA polymerases, which are single polypeptides with great affinity for the phage promoters.

Other bacteria contain similar RNA polymerases to *E. coli* although the number of subunits varies. Conversely, the archaean RNA polymerases are similar to the eukaryote enzymes.

**Basal transcriptional initiation.** Transcriptional initiation in bacteria begins with the binding of the RNA polymerase holoenzyme to the promoter, situated at the 5' side of the gene (this binding is facilitated by the  $\sigma$ -factor, as discussed above). Initially, the enzyme binds loosely and reversibly to duplex DNA (**loose binding**) as it searches for the promoter sequence — this is the closed promoter complex. When the correct sequence is recognized, the DNA at the promoter site is locally unwound. The interaction between the enzyme and DNA then becomes irreversible (**tight binding**) and characterizes the open promoter complex. Transcription begins *de novo*, usually with GTP or ATP, and unlike the subsequent nucleotides, the initial residue retains its triphosphate moiety. Processive transcription often fails after a short oligonucleotide has been synthesized, and the enzyme returns to the start (abortive initiation). Successful initiation usually follows the assembly of 10 or more nucleotide subunits on the template, by which time the enzyme has cleared the promoter. The  $\sigma$ -factor then dissociates from the core enzyme.

**Constitutive control — promoter architecture.** Bacterial promoters are relatively simple. *E. coli* promoters consist of two consensus motifs, the **-10 sequence (Pribnow's box)** and the **-35 sequence** (Figure 29.2), which interact directly with the  $\sigma$ -factor. The sequence of each motif and their relative positioning are critical for successful initiation. Mutations in the -35 sequence affect the efficiency of initial binding, whereas mutations in the -10 sequence affect the rate of open complex formation. This indicates that the -35 site is the initial recognition motif for the  $\sigma$ -factor causing the -10 sequence to be unwound, a process facilitated by its multiple weak A:T bonds (q.v. *thermal melting*). The distance between the two sites represents a single turn of helix, allowing the two motifs to interact with the  $\sigma$ -factor simultaneously. Mutations which alter this spacing change the efficiency of initiation by affecting the ability of  $\sigma$  to interact with both motifs; the alteration of DNA topology or conformation has similar effects. The sequence at the start site itself can also influence the efficiency of



**Figure 29.2:** Basal promoter structure. Architecture of the *E. coli* consensus promoter, and eukaryotic class I, class II and class III promoters. There are four types of eukaryotic class II promoter which may or may not possess a TATA box and an initiator consensus. The class III promoter is that of the 5S rRNA gene. Other class III promoters are reminiscent of class II promoters and may possess a TATA box and initiator-like element.



initiation and often includes the trinucleotide motif CAT, with the central purine corresponding to position +1 of the transcript. Additionally, the efficiency of promoter clearance is modulated by the nature of the first fifty or so bases in the transcribed region.

The efficiency of 'default' transcriptional initiation in *E. coli* varies through three orders of magnitude, reflecting deviations from the consensus promoter sequence. The weakest promoters lack a -35 sequence altogether, the default expression rate is close to zero, and additional activator proteins are required to facilitate RNA polymerase binding (q.v. *catabolite repression*). The different  $\sigma$ -factors of *E. coli* recognize different consensus sequences, although the bipartite architecture appears to be broadly conserved. Similar promoters are found in many other bacteria, although there may be variation in the relative positions of the conserved motifs.

### 29.3 Transcriptional initiation in eukaryotes — basal and constitutive components

**RNA polymerases.** Eukaryotes possess three DNA-dependent RNA polymerases, each responsible for the transcription of different classes of gene; their properties are summarized in Table 29.2. The enzymes can be distinguished by their differing sensitivities to  $\alpha$ -amanitin, a fungal toxin. The subunit structure of the eukaryotic RNA polymerases is relatively poorly characterized, although each has 10 or more subunits, some of which are common to all three enzymes. The largest subunits of each polymerase are homologous to each other and to the  $\alpha$ ,  $\beta$  and  $\beta'$  subunits of *E. coli* RNA polymerase. The  $\beta'$ -like subunit of RNA polymerase II has a flexible C-terminal domain which plays an important role in transcriptional initiation and elongation (see below) and binds splicing and polyadenylation factors (see RNA Processing). There is no counterpart to the bacterial  $\sigma$ -factor, and the eukaryotic RNA polymerases are consequentially unable to recognize or bind to their promoters without the assistance of additional proteins.

**Overview of basal transcriptional initiation.** Eukaryotic RNA polymerases require accessory protein factors, the **transcription initiation factors (TIFs)** or **basal or general transcription factors (GTFs)**, to assist them in the recognition of the promoter. The GTFs must bind to the DNA first to form a complex which recruits RNA polymerase to the DNA at the transcriptional initiation site. Recruitment of the enzyme and further basal factors to the DNA forms the **preinitiation complex**. This may be converted from a closed to an open DNA configuration by the activities of one or more of the GTFs. The GTFs, in concert with the RNA polymerase itself, comprise the **basal apparatus** of transcriptional initiation.

The GTFs associated with the three eukaryotic RNA polymerase enzymes are distinct, but contain common components. The most significant of these is the **TATA-binding protein (TBP)**, which makes sequence-specific contacts with DNA in certain types of promoter (see Nucleic Acid-binding Proteins).

**Table 29.2:** General properties of the eukaryotic RNA polymerases

Enzyme	Function	Sensitivity
RNA polymerase I	Transcription of the 45S rRNA precursor, a single polycistronic unit containing 5.8S, 18S and 28S rRNA genes ( <b>class I genes</b> )	Insensitive to $\alpha$ -amanitin, sensitive to actinomycin D
RNA polymerase II	Transcription of all protein encoding genes and most genes for small nuclear RNAs ( <b>class II genes</b> )	Inhibited by $\alpha$ -amanitin
RNA polymerase III	Transcription of tRNA genes, 5S rRNA genes and genes encoding U6 snRNA and the various scRNAs ( <b>class III genes</b> )	Moderately sensitive to $\alpha$ -amanitin depending on species

**Overview of constitutive control.** The basal initiation complexes formed by the eukaryotic RNA polymerases support only minimal transcriptional activity: in the absence of other interacting proteins, they assemble slowly and are unstable. The efficiency of initiation is increased by interaction with protein factors bound at *cis*-acting sites surrounding the promoter and, in the case of class II genes, at distant sites termed enhancers. These **transcription factors** provide a suitable environment for the assembly of the initiation complex and/or interact with its components to stimulate their binding or their activity, i.e. they are **transcriptional activators**. The presence of some transcription factors in an active form is regulated, and such factors contribute to the regulation of their target genes in response to external stimuli etc. Others are active in all cells and thus contribute to the constitutive transcriptional activity of the gene. A major distinction between transcriptional initiation in bacteria and eukaryotes is therefore that the bacterial RNA polymerase holoenzyme can efficiently initiate transcription on a consensus promoter *without the assistance of further components*, whereas eukaryotic RNA polymerases require GTFs for promoter recognition and assembly of the initiation complex, and need constitutive regulatory factors for the assembly process to be efficient and stable enough to sponsor transcription, *even on strong consensus promoters*. This reflects the general negative regulatory strategy in eukaryotic genomes — there are lots of genes and most are inactive most of the time, so gene expression is usually mediated by selective activation. There are also examples of active repression, however (q.v. *silencer*, *repression domain*).

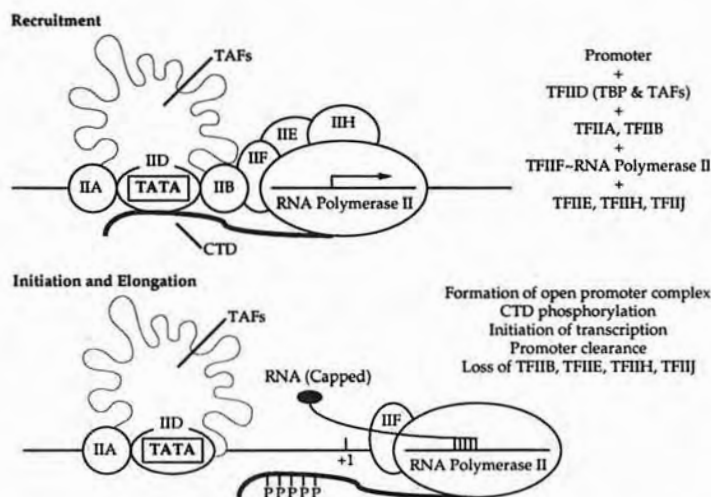
**Initiation of RNA polymerase I transcription.** There is a single class I transcription unit in eukaryotic genomes, encoding several ribosomal RNAs as a polycistronic transcript. The control of RNA polymerase I transcription is simple. The vertebrate RNA polymerase I promoter is bipartite, consisting of a **core promoter** which surrounds the transcriptional start site and an **upstream control element** (UCE) about 100 bp 5' to the start site. The core promoter alone is sufficient for basal transcription, but the presence of the UCE strongly increases transcriptional activity. A protein called UBF1 binds to GC-rich motifs in both elements and serves as a recognition point for a tetrameric complex SL1, which contains the TATA box-binding protein TBP and other components required to load RNA polymerase onto the DNA. The activity of UBF1 is influenced by interaction with the *retinoblastoma protein* (q.v.), coupling rRNA synthesis to the cell cycle. In some lower eukaryotes, the SL-1 homolog itself binds to DNA and there is no UBF1.

**Initiation of RNA polymerase II transcription.** RNA polymerase II promoters are the most diverse, reflecting the great variety of regulatory elements controlling the complex expression patterns of the many protein-encoding genes. Generally, such promoters can be divided into the **basal promoter** located at the transcriptional start site and the collection of **upstream promoter elements**, comprising both constitutive and regulatory motifs. Basal promoters show some variability, but can be placed into four major classes. However, the upstream promoter elements of different genes appear to have distinct configurations of constitutive and regulatory elements, allowing great diversity of transcriptional regulation (see later). The basal promoter consists of an **initiator** (**Inr**) sequence, which is found at the transcriptional start site and has the consensus YYCARR, and/or a motif termed the **TATA box** (also known as the **Goldberg-Hogness (GH) Box**) positioned at about -25, which has the consensus TATAWAW. These motifs are similar to the bacterial start site (CAT) and -10 sequence (TATAAT), respectively, although the relative positions differ. The initiator and TATA box serve as recognition sites for the general transcription factors and function to position the initiation complex accurately at the start site. Some class II genes lack both a TATA box and initiator and may rely on downstream elements for binding. These promoters usually have multiple transcriptional initiation sites (**multiple start sites**), reflecting a lack of accurate polymerase localization. The four types of class II basal promoter are shown in Figure 29.2. There is no evidence that any particular promoter type is consistently more efficient than any other, and mutations in the conserved motifs have different effects in different genes. This suggests that the context of the

promoter and its relationship with other proximal elements is probably important for the efficiency of transcriptional initiation.

The first stage of transcriptional initiation is the recognition of the TATA box by the TFIID factor, which comprises the TATA-binding protein TBP and a variety of **TBP-associated factors (TAFs)**. TBP makes sequence-specific contacts with DNA at the TATA box though the minor groove (see Nucleic Acid-binding Proteins). TAFs are important for two reasons: they may recognize non-TATA elements in the core promoter and may thus play an important role in the recognition of initiator-only promoters and null promoters (Figure 29.2), and they also allow TFIID to respond to transcriptional activators and silencers. The particular assembly of TAFs at the basal promoter therefore dictates how the initiation complex interacts with upstream activators, enabling the basal promoter itself to demonstrate some cell-type specificity. For instance, the lymphoid-specific expression of the terminal deoxynucleotidyl transferase gene and the myeloid-specific expression of the interferon- $\gamma$  gene are both deregulated if their basal promoter architecture is altered. Similarly, the promoter usage changeover between embryonic and adult flies in the *Drosophila* alcohol dehydrogenase gene depends on the structure of the initiator element.

Further multimeric factors then bind sequentially to the promoter (Figure 29.3). TFIIA is the first to be recruited and is responsible for blocking the binding of factors such as DR1, which inhibit TFIID activity. TFIIB then binds and acts as a bridging protein for the recruitment of TFIIF. TFIIF carries the RNA polymerase II enzyme into the complex. TFIIB and TFIIF may together promote specific interactions between RNA polymerase II and the start site. The enzyme facilitates the recruitment of TFIIE, which in turn recruits TFIIH (and stimulates its activity) and TFIIJ. TFIIH is essential in the preinitiation complex: it possesses helicase activity which controls promoter melting, and kinase activity which phosphorylates the C-terminal domain (CTD) of RNA polymerase II. Phosphorylation releases the enzyme from the initiation complex and facilitates promoter clearance, leaving a residual initiation complex at the promoter. TFIIH also phosphorylates other basal components and different forms of the protein are involved in *transcription-coupled repair* (q.v.). Like bacterial RNA polymerases, RNA polymerase II may make abortive initiations before promoter clearance. The first nucleotide inserted is usually a purine, which is modified immediately by the enzyme mRNA guanyltransferase to generate a distinctive *cap* (q.v.). The transcription start site in RNA polymerase II genes is thus sometimes termed the **cap site**.



**Figure 29.3:** Stages in the activation of transcription at RNA polymerase II promoters.

**Initiation of RNA polymerase III transcription.** RNA polymerase III promoters fall into three categories. The 5S rRNA and tRNA genes, and genes encoding several scRNAs possess **internal control sites (ICS)**, promoters which lie downstream of the start of transcription and become transcribed into RNA (Figure 29.1). 5S rRNA genes have **type I promoters** and require the binding of three GTFs (TFIIIA, TFIIB and TFIIIC) to load the RNA polymerase. The tRNA genes have **type II promoters** and require the binding of only TFIIB and TFIIIC. **Type III promoters** control snRNA gene expression and resemble typical RNA polymerase II promoters, possessing TATA boxes and RNA polymerase II-like regulatory motifs; these promoters are thus more complex than type I and II promoters and can drive cell-type-specific gene expression. TFIIB and other accessory factors are required for polymerase recruitment on type III promoters. TFIIB is the common component in each type of RNA polymerase III promoter. It contains the TATA-binding protein and in each case recruits the RNA polymerase to the initiation complex. In type III promoters, TFIIB interacts directly with the DNA, whereas on type I and II promoters it binds to the preinitiation complex by protein-protein interactions with the other GTFs. This is analogous to the different assembly pathways of TATA-box promoters and TATA-less promoters of RNA polymerase II genes.

## 29.4 Transcriptional initiation — regulatory components

**Principles of transcription factor activity.** The concept of transcription factors binding to sites surrounding the basal promoter or beyond, and controlling the assembly, stability or activity of the initiation complex, was introduced above in the context of eukaryotic constitutive transcriptional activators. In both prokaryotes and eukaryotes, the basal promoter may also be surrounded by short sequence motifs which are binding sites for *regulatory* transcription factors. These proteins employ the same operational mechanisms as constitutive transcription factors, but may act either positively, to increase the rate of transcriptional initiation (**transcriptional activators**), or negatively to decrease it (**transcriptional repressors**). Furthermore, unlike the constitutive factors, their presence or activity is limited, e.g. by external signals, enabling the genes they control to be similarly regulated.

In principle, transcription factors act in four ways:

- (1) by binding to DNA and interacting with the basal initiation apparatus;
- (2) by binding to DNA and altering its structure or conformation;
- (3) indirectly, by binding to DNA and interacting with another regulator to influence its activity;
- (4) through protein-protein interactions (i.e. without binding to DNA at all), either directly with components of the basal apparatus or indirectly with other regulators.

Transcription factors in bacteria and eukaryotes can be divided into a small number of families based on their domain structure. They may possess several domains, typically a DNA-binding domain and an activation (or repression) domain, but often also a dimerization domain, and domains for interaction with effector molecules or signal transduction components. DNA-binding domains are not discussed in this chapter (see Nucleic Acid-binding Proteins).

The transcriptional activity of a given gene thus depends upon its invariant, intrinsic properties, which set its default expression rate (i.e. the structure of its regulatory elements and, in eukaryotes, the presence of constitutive activators) and variable extrinsic factors (i.e. the availability of regulatory transcriptional activators and repressors). Housekeeping genes are controlled predominantly by promoter architecture and constitutive transcription factors, whereas inducible or repressible genes, and genes expressed in a cell-type-specific or developmentally restricted manner, are also controlled by regulatory transcription factors.

**Regulatory elements — promoters and enhancers.** Regulatory *cis*-acting elements are often found near the basal promoter, facilitating local interactions with RNA polymerase (in bacteria) or the preinitiation complex (in eukaryotes). Most bacterial regulatory elements flank the promoter, whereas in eukaryotes, constitutive and regulatory transcription factors usually bind at sites located 5' to the basal



promoter, in the upstream promoter region. As well as these proximal sites, additional *cis*-acting elements may be found hundreds or even thousands of base pairs away from the gene they control. Such distant regulatory elements are termed **enhancers** if they act positively, or **silencers** if they act negatively; they are common in eukaryotes but rare in bacteria — e.g. in *E. coli*, enhancers are associated only with genes transcribed under the control of the  $\sigma^{54}$ -factor. Enhancers act similarly to promoters in that they are *cis*-acting elements which bind regulatory and (in eukaryotes) constitutive transcription factors, and interact with the basal initiation complex to control transcriptional initiation. However, because of their distance, such interactions are mediated by the looping out of the intervening DNA, allowing enhancers to act in a position- and orientation-independent manner — hence they may be located upstream, downstream or even within the gene they regulate. Promoter regulatory elements are more restricted in this respect, and promoters thus function in one orientation only. Some interactions occur by adjacent binding so that the position of the regulatory element in the promoter is essential for its function — most bacterial operators and activator/initiator sites fall into this category. Other interactions require local bending and folding of the DNA round the initiation complex and the position of the regulatory element is more flexible — this applies to many eukaryotic transcription factors bound to the upstream promoter region. Table 29.3 gives a summary of *cis*-acting sites and *trans*-acting factors in transcription.

**Locus control regions.** A locus control region (LCR) is a eukaryotic *cis*-acting element, usually located a considerable distance from the genes it regulates, which is essential for transcriptional activity because it establishes an independent chromatin domain. The importance of the LCR is demonstrated by the analysis of globin gene expression in transgenic mice, and from this type of experiment, the operational definition of an LCR is derived. The  $\beta$ -globin gene may be introduced into cultured erythroid cells by transfection and is maintained episomally for a short time. If the construct contains the  $\beta$ -globin promoter and enhancers, it may be expressed at high levels. Conversely, if the same construct is introduced into the mouse genome, it may demonstrate only minimal activity, even in erythroid cells, because it is inhibited by the adjacent chromosome regions. This *cis*-acting inhibition is alleviated if the construct is joined to the locus control region, which allows the  $\beta$ -globin transgene to be expressed in the erythroid lineage of the transgenic mice. The LCR not only allows integration position-independent expression, but the transgene expression levels are also directly proportional to copy number. However, the LCR has no effect on transcription in transfected cells, i.e. it is not simply an enhancer.

Like the promoters and enhancers, the globin LCR contains many binding sites for transcription factors, some constitutive and some erythroid-specific. Because it does not affect the transcription rate in cultured cells but does so *in vivo*, the LCR is thought to act by modulating chromatin structure (episomally maintained DNA is not packaged into chromatin and is therefore free of the constraints of *cis*-repression). How the LCR mediates its effect is unknown, but it is in some way connected with the temporal regulation of the genes of the  $\beta$ -globin cluster (see Box 29.1 for further discussion).

**Bacterial transcription factors that interact with RNA polymerase.** Many transcription factors function by directly influencing initiation through interaction with one of the components of the initiation complex. In bacteria, such transcription factors interact with RNA polymerase itself. In *E. coli*, a simple model for transcriptional repression in the *lac* operon involves the Lac repressor binding to the major operator site which overlaps the promoter, directly blocking the access of RNA polymerase to the promoter. Alternatively, both the Lac repressor and RNA polymerase may bind to the DNA, and interaction between the repressor and the enzyme could prevent open complex formation. There are examples of both strategies in other operons. There is an obvious overlap between the promoter and operator in the *aroH* and *trp* operons, suggesting that blocking/displacement is the likely mechanism of inhibition, whereas adjacent binding of CAP and RNA polymerase is responsible for activation of transcription at the *lac* and *gal* operons under conditions of catabolite

**Table 29.3:** Summary of the *cis*-acting elements and *trans*-acting factors which control transcriptional initiation in bacteria and eukaryotes (+), positively acting; (-), negatively acting

Element	Definition and function
<i>Bacteria</i>	
<b>Promoter (+)</b>	In bacteria, the site where RNA polymerase binds. In <i>E. coli</i> and many other bacteria, the promoter consists of two motifs at positions -35 and -10, recognized by the $\sigma$ -factor of RNA polymerase. Equivalent to eukaryotic <i>basal promoter</i>
<b>Initiator/activator (+); operator (-)</b>	Positively/negatively acting regulatory elements in bacteria which usually overlap or abut the promoter and influence RNA polymerase binding or activity
<b>Enhancer (+)</b>	A distant, positively acting regulatory element. Bound factors interact with the promoter by looping out of the interstitial DNA
<i>Eukaryotes</i>	
<b>Promoter (+)</b>	In eukaryotes, a control site proximal to the gene which contains two components: a <b>basal promoter</b> where RNA polymerase binds (equivalent to the bacterial <i>promoter</i> ), and an <b>upstream promoter</b> . The latter contains both <b>constitutive elements</b> and <b>regulatory elements</b> , which together control transcriptional initiation. Promoters can extend some distance from the gene, but act in an orientation-dependent manner. Most promoters are found immediately 5' to the gene, although RNA polymerase III promoters may lie within the transcribed region ( <b>internal control sites</b> )
<b>Enhancer (+); silencer (-)</b>	Enhancers are positively acting control sites in eukaryotes, often containing both constitutive and regulatory elements, which can increase the transcriptional activity of their target gene by up to 1000-fold from very great distances (<50 kbp). Enhancers are orientation- and position-independent because they interact with the basal apparatus by looping out interstitial DNA, but do not show promoter activity because they lack basal components. Enhancers may contain functionally clustered groups of binding sites termed <b>enhanccons</b> . They may regulate more than one promoter, resulting in enhancer competition in some systems. Silencers are negatively acting control sites with the same operational properties as enhancers
<b>Upstream activator site (+); upstream repressor site (-) (UAS, URS)</b>	Enhancer- or silencer-like sites found in yeast, which function in a similar manner but are incapable of operating downstream of the gene they control
<b>Locus control region (+)</b>	Distant regulatory elements controlling gene expression by establishing an open chromatin domain. Alleviate <i>cis</i> -repression caused by chromatin silencing, allowing position-independent and copy number-dependent transgene expression, but do not enhance transcriptional activity in transient transfection assays in the manner of a classic enhancer
Transcription factors (TFs)	Definition and function
<i>Bacteria</i>	
<b>Sigma factor</b>	Subunit of RNA polymerase which recognizes the promoter
<b>Transcriptional activators/repressors</b>	Proteins which bind to initiator/activator and operator sites and influence RNA polymerase binding/activity
<i>Eukaryotes</i>	
<b>Basal TF</b>	A transcription factor such as TFIID which is part of the preinitiation complex at the basal promoter and is required for RNA polymerase
<b>General TF</b>	

Continued

<b>Transcription initiation factor (TIF)</b>	binding or activity. Nomenclature according to the system TFab, where <i>a</i> is the cognate RNA polymerase and <i>b</i> reflects the order of discovery. Hence TFIID was the fourth transcription factor discovered in the RNA polymerase II initiation complex.
<b>Constitutive TF</b>	A transcription factor such as Sp1 which is present in all cells and has a general positive role in gene expression. Binds to constitutive elements in the upstream promoter and enhancer
<b>Transcriptional activators/repressors</b>	A transcription factor such as MyoD1 whose presence in an active form is itself regulated. Binds to regulatory elements in the upstream promoter and enhancer and facilitates regulated gene expression

repression (see Box 29.2). Positive interactions also occur between RNA polymerase at the promoter and transcription factors bound at bacterial enhancers by the looping out of intervening DNA. RNA polymerase binds to the promoter of the *glnA* gene to form a closed promoter complex, but open promoter complex formation requires interaction with NtrC transcriptional activators bound at twin enhancers located 150 bp away. Direct physical contact between the activators and RNA polymerase is required for DNA unwinding. In certain cases, the activity of a transcription factor can influence the specificity of RNA polymerase for its promoter. The progress of sporulation in *B. subtilis* is controlled by the regulated synthesis of several distinct  $\sigma$ -factors and additional transcription factors which control promoter specificity (see Box 6.2).

**Bacterial transcription factors that modulate DNA structure.** In prokaryotes, the topological properties of DNA can influence transcription. Cryptic operons can be derepressed by mutations in the gene encoding *DNA gyrase* (q.v.), or transposon insertion, presumably altering the level of supercoiling and allowing the DNA to be recognized by RNA polymerase. Furthermore, several transcriptional repressors function by isolating the promoter in a topologically constrained loop of DNA so that open promoter complex formation is inhibited. The dual promoters of the *gal* operon in *E. coli* are flanked by operators which cooperatively bind the Gal repressor. Repressor monomers bound at each site associate to form a dimer, isolating the promoter in a 114 bp DNA loop within which the duplex cannot be unwound. A similar loop is formed in the *lac* promoter by interaction between Lac repressor bound at the major operator and one of the two minor operators.

**Eukaryotic transcription factors that interact with the preinitiation complex.** In eukaryotes, most transcription factors directly interacting with the preinitiation complex are transcriptional activators. The analysis of transcription factor domain structure has shown that the **activation domains** (which contact the basal apparatus to facilitate transcriptional activation) can be assigned to a small number of families (see Box 29.3 and also q.v. *domain swap*), as for DNA-binding domains (see Nucleic Acid-binding Proteins). The acidic activation domain is highly interchangeable between transcription factors from different species, suggesting that all transcription factors of a given activation family interact with the same, highly conserved component of the basal apparatus. The phenomenon of **squelching** supports this conclusion — a gene is squelched when a transcription factor is overexpressed and not only activates all its target genes, but sequesters, by protein–protein interactions, components of the basal apparatus of genes it does not normally regulate, causing global downregulation. Activation domains of all three classes stimulate the *recruitment* of TFIIB to the initiation complex, and acidic activation domains induce a conformational change in TFIIB which increases its affinity for RNA polymerase. Activation domains also influence the *activities* of several components of the basal apparatus, including RNA polymerase itself. A multicomponent complex termed **mediator** has been identified in yeast and humans which assembles on the RNA polymerase II C-terminal domain to transmit signals from transcriptional activators. The TATA-binding protein is also a target, and activation is mediated through interaction with the TAFs. Different transcription factors have been shown to interact with different TAFs, and different TAFs assemble on promoters with different architecture. This allows the basal promoter to control

cell-type-specific gene expression, by dictating the availability of TAFs to interact with particular upstream transcriptional activators (Box 29.3). Several transcription factors are known to act as integrators of multiple signals: CBP, for example, is a giant integrator protein with tens of binding sites for different transcription factors, interacting by the looping of DNA. Transcriptional activators may also interact with several other GTFs, including TFIIF. Thus transcription is regulated by multiple factors acting on multiple targets in the preinitiation complex to increase its rate of assembly, its stability and its activity.

Unlike most bacterial repressors, which directly interact with RNA polymerase, most eukaryotic transcriptional repressors act passively by inhibiting the function of an activator (see below). Some direct repressors have been isolated, however, and these appear to mediate their function analogously to activators, i.e. by interacting directly with the basal initiation complex. These factors interact by protein-protein contacts, as is the case for DR1 which inhibits the assembly of the preinitiation complex, whereas others have discrete *cis*-acting sites in the promoter or in distant silencer elements and may act either to inhibit the assembly or activity of the preinitiation complex (e.g. the *Drosophila* pair-rule protein Even-skipped and the mammalian thyroid hormone receptor in the absence of its ligand).

**Eukaryotic transcription factors that alter DNA structure.** In eukaryotic promoters, the function of some transcription factors is to displace a nucleosome, thereby creating a *DNase I hypersensitive site* (q.v.). This may be a direct route to transcriptional activation, by permitting the binding of other transcription factors in the near vicinity previously blocked by nucleosome positioning, or it may generally facilitate the rearrangement of chromatin in the promoter and alleviate repression caused by higher-order chromatin structure. Examples of such factors include the ubiquitous GAGA factor in *Drosophila*, and the SWI/SNF complex which is associated with the C-terminal domain of RNA polymerase II in yeast and vertebrates (q.v. *histone acetylation*). Some transcription factors bound at enhancers and silencers can act similarly, by propagating a change of chromatin organization over a wide area. For example, the yeast mating-type cassettes *HML $\alpha$*  and *HMR $\alpha$*  (q.v. *mating-type switching*) are endowed with their own promoters, but are kept in an inactive state (and protected from endonuclease cleavage) by heterochromatinization directed by silencers located several kilobases away. A protein termed RAP1 recognizes these silencers (and telomeric silencing sites also) and recruits SIR3 and SIR4, two of the four **SIR proteins** (**silent information regulators**) which interact with histones H3 and H4 and may polymerize as part of the heterochromatinization process; they may also attach the silenced region to the nuclear matrix.

Some eukaryotic transcription factors function primarily by introducing bends into DNA, which facilitates the interaction of other components. Many are HMG-proteins, related to the ubiquitous HMG1/Y and HMG 14/17 proteins that play an important role in chromatin structure (see Chromatin). The tactical positioning of HMG-box transcription factors such as Sry (which controls sex-determination in mammals), Sox-2, Sox-3 and Sox-11 (which are involved in neural differentiation) and Lef-1 (which controls lymphocyte-specific gene expression) can organize DNA into a specific three-dimensional conformation wherein all the various transcriptional activators interact productively with each other (such a structure is an **enhanceosome**). This provides another example of how the positioning of a *cis*-acting site in a promoter may be critical for its activity — the repositioning of a DNA bend can disrupt the ordered packing of other bound proteins causing non-productive interactions and loss of transcriptional activity (also q.v. *V(D)J recombination*).

**Transcriptional regulation by transcription factor interactions.** Most eukaryotic transcriptional repressors act passively by interfering with the function of activators, rather than by directly regulating the preinitiation complex. One common mechanism of repressor activity is simply to occupy the binding site of an activator, or a site which overlaps or lies adjacent to it, in such a way that steric hindrance prevents the binding of the activator. In this case, the fate of the regulated gene depends upon competition between activators and repressors for binding, and upon the relative affinity



which each transcription factor has for its binding site. A weakly bound activator, for instance, could be displaced by a high-affinity repressor. An example of transcriptional regulation by competition for binding sites is provided by the *Drosophila even-skipped* gene, which is regulated by competition between the various gap proteins for occupation of upstream binding sites (see Box 6.5). In other cases, competition can occur between positively and negatively acting proteins for interaction with a transcriptional activator in solution, the result of the interaction dictating whether the activator can bind to the DNA, or whether it can activate transcription once bound. The simplest case of this type of interaction is the dimerization of transcription factors of the helix-loop-helix and leucine zipper families. Dimers can form containing either two active monomers (which bind to DNA and activate transcription), or one active and one inactive monomer, producing an inactive heterodimer which cannot bind to DNA. Examples include the inhibition of myogenic bHLH proteins by Id, which lacks a DNA binding domain, and the control of *Drosophila sex-determination* (q.v.) by the balance between active bHLH proteins such as Sisterless-a and inhibitory HLH proteins such as Deadpan.

Other mechanisms of indirect silencing are not based on competition, i.e. increasing the level of the activator has no effect. This is seen in the phenomenon of **quenching**, where an inhibitor binds to a site in DNA which obstructs the activity of a bound activator, often by blocking its access to the basal apparatus or masking its activation domain. By extension, it is probable that activators could also occasionally act in this manner, i.e. 'unquenching' bound repressors by obstructing their inhibition domains.

**Regulation of transcription factors.** Only under very exceptional circumstances are the *cis*-acting elements surrounding a gene themselves regulatable (q.v. *phase variation*, *mating-type switching*). More usually, transcriptional control depends on the regulated availability of active transcription factors. In principle, transcription factors can be regulated in two ways — by controlling their *availability* and controlling their *activity*. Both systems are used in bacteria and eukaryotes, but in bacteria the latter is predominant: most transcription factors are synthesized constitutively and regulated posttranslationally.

The regulated *de novo* synthesis of transcription factors is economical (resources are not wasted on the synthesis of proteins that will not be used), but there is a considerable delay between the decision to synthesize the factor and its appearance, at the correct levels, in the cytoplasm. Consequently, *de novo* synthesis as a mechanism of transcription factor regulation is useful where a long-term transition of transcriptional activity is required, but not where the cell needs to make a rapid and transient response to an external stimulus. It is therefore predominant in the regulation of developmental processes in multicellular organisms and in the maintenance of differentiated states. In bacteria, *de novo* synthesis is used to mediate long-term fundamental responses to the environment, e.g. sporulation in *B. subtilis*.

The regulation of transcription factor activity solves the regulation chain problem — if a transcription factor is regulated at the level of the transcription of its gene, then there must be another transcription factor to mediate this control, the problem of regulation has simply been shifted one link along the chain. Regulation of activity also allows a prompt response to external stimuli, especially where the activity of the transcription factor is directly coupled to the stimulus. In the most extreme case, the stimulus itself is the transcription factor: lactoferrin released by neutrophils in response to bacterial infection is taken up by surrounding cells and binds to DNA, acting as a transcriptional activator to induce transcription of a number of genes controlling defence against bacterial infection. One step away from this, the stimulus becomes an **allosteric modulator** of the transcription factor. This is very common in bacteria where metabolites in the environment control the transcription factors which regulate the genes involved in their synthesis or degradation. In the *lac* operon, allolactose, a metabolic product of lactose, binds to the Lac repressor and inactivates it, allowing transcription of the genes which encode lactose utilization enzymes (Box 29.2). Similarly,

in mammals, steroid hormones act as allosteric modulators of their receptors, which are also transcription factors. A further step away generates a signal transduction pathway — the stimulus, a signal, binds to a receptor which directly or indirectly activates a transcription factor. The JAK–STAT signal transduction pathway is direct — the signal, a cytokine, binds to a cell surface receptor and activates an intracellular JAK kinase which phosphorylates and activates a transcription factor of the STAT family. Other signals reach their transcription factor targets by a longer journey involving many components (see Signal Transduction).

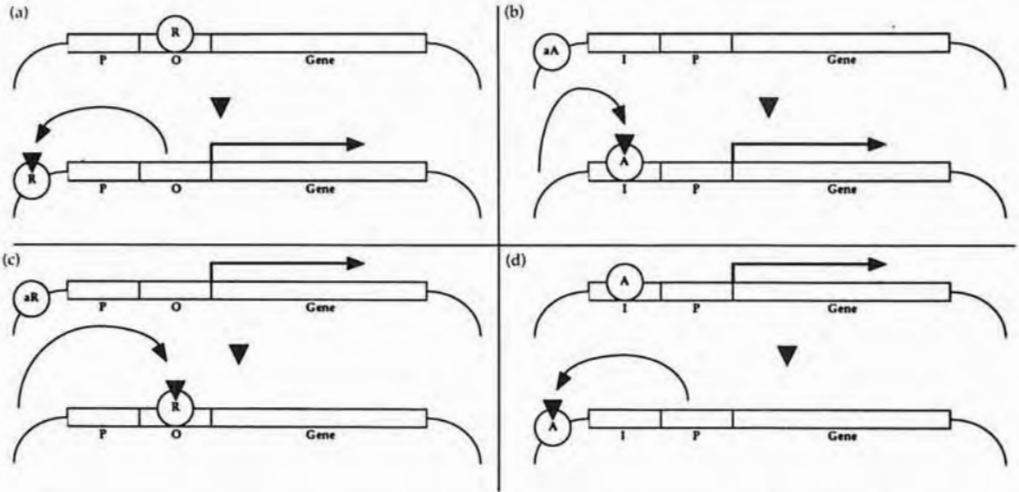
Signal transduction pathways exemplify another mechanism for controlling transcription factor activity: covalent modification. Many transcription factors in eukaryotes are activated (or inactivated) by phosphorylation, which can influence their ability to bind DNA (e.g. c-Jun), their ability to activate transcription (e.g. c-Fos, CREM), their ability to form dimers (e.g. STATs), and their ability to interact with other proteins which mediate their effects (e.g. CREB, which interacts with CREB-binding protein CBP when phosphorylated by protein kinase A). The same mechanism is seen in bacterial signal transduction pathways, e.g. NtrC, which binds the enhancer upstream of the *glnA* gene, must be phosphorylated in order to bind; it is phosphorylated by the NtrA kinase which is responsive to levels of nitrogen in the environment. Alternatively, the target of the pathway may be an inhibitor of transcription factor activity. The predominant example is the activation of NF- $\kappa$ B by the phosphorylation, and subsequent targeted degradation, of the inhibitory protein I- $\kappa$ B. In this case, the factor is regulated by both activation and availability. I- $\kappa$ B masks the nuclear localization signal of NF- $\kappa$ B and prevents its interaction with DNA by sequestering it in a different cellular compartment.

**Combinatorial and context-dependent regulation.** The effects of two different transcription factors acting simultaneously upon the same gene are often unexpected, given the effects of each acting independently. One example of this is **synergism**, where the transcription activity of a gene driven by two transcription factors is greater than the additive effects of each individual factor. The transcription factor Pit-1 acts synergistically with the estrogen receptor in the prolactin gene probably by increasing the activity, stability or recruitment of different components of the basal apparatus. In other cases, two transcription factors which independently activate a gene can in combination cause silencing. Such context-dependent regulation may depend upon the abundance of a transcription factor, or indeed the presence of specific regulatory sites in the DNA. The p53 factor provides an example of multiple-level context-dependent regulation: it is a strong activator of gene expression if there is a p53 recognition site available in the promoter, but if there is no site, it binds to TFIID directly and inhibits transcription. Additionally, while the WT1 protein is a transcriptional activator alone, it acts as a transcriptional repressor when bound by p53. Also binding specificity may change due to differential dimerization (this occurs for many of the *Drosophila* homeodomain transcription factors which bind to alternative sites when associated with cofactors).

## 29.5 Strategies for transcriptional regulation in bacteria and eukaryotes

**Characteristic regulatory strategies of bacteria.** Bacteria live in a dynamic environment where rapid and appropriate responses to varying levels of metabolites and other molecules is essential for survival. To streamline these responses, the bacterial genome is organized so that genes with a common function (e.g. those encoding enzymes in a common metabolic pathway) are often grouped together in units termed **operons** under common transcriptional *cis*-regulation (see Gene Structure and Mapping). The genes of the operon are transcribed into a polycistronic transcript and can thus be activated and repressed as a unit at the level of transcription. This depends on the ability of bacterial ribosomes to initiate at internal sites (see Protein Synthesis). The regulation of transcription occurs predominantly at two stages: **promoter control**, which regulates transcriptional initiation, and **attenuator control**, which regulates termination (attenuator control is discussed later in this chapter).

The rapid induction and repression of transcription is mediated by **allosteric control circuits**, the predominant form of promoter control in bacteria, where transcription factors are synthesized



**Figure 29.4:** Allosteric control circuits in bacteria. (a) **Negative inducible regulation.** Default state of gene is OFF because the regulatory element (**operator**) is bound by a negatively acting factor (**repressor protein**). Addition of a small effector molecule (**inducer**) causes a conformational change in the repressor which causes it to release the operator and bind nonspecifically to DNA. The gene thus becomes **derepressed** and is transcriptionally active. Example: the *lac* operon, a catabolic operon repressed in the absence of lactose, induction by allolactose, an analog of lactose. (b) **Positive inducible regulation.** Default state of gene is OFF because the positive regulatory factor (**activator protein**) is unable to bind the regulatory element (the **initiator**, or **activator**). In this state the activator protein is known as the **apoactivator** and binds nonspecifically to DNA. Addition of a small effector molecule (**coactivator**) causes a conformational change in the apoactivator which allows it to bind with great specificity to the initiator element thus allowing transcriptional initiation. Example: catabolite repression of the *lac* operon — CAP protein is apoactivator, binds DNA only in presence of cAMP (coactivator) to facilitate initiation. (c) **Negative repressible regulation.** Default state of gene is ON because the negative regulatory factor (the **repressor protein**) is unable to bind to the operator. In this state it is known as an **aporepressor** and binds nonspecifically to DNA. Addition of a small effector molecule (**corepressor**) causes a conformational change in the aporepressor which allows it to bind specifically to the operator and inhibit transcriptional initiation. Example: the *trp* operon, encoding enzymes involved in tryptophan synthesis. Tryptophan itself is the corepressor and facilitates the shutdown of its own synthesis (negative feedback). The Trp repressor also represses its own gene (**autogenous regulation**). (d) **Positive repressible regulation.** Default state of gene is ON because the regulatory element (**initiator**) is bound by a positively acting factor (**activator protein**). Addition of a small effector molecule (**repressor**) causes a conformational change in the activator which causes it to release the initiator and bind nonspecifically to DNA. The gene thus becomes repressed and is transcriptionally silent.

constitutively, but their activity is modulated by their interaction with small **effector molecules** in the environment, whose presence is indicative of the need for the regulated gene products. An example is the *lac* operon (Box 29.2). The genes encoding the transcription factors may be under **autogenous control** (i.e. regulated by their own product) or may be entirely unregulated (i.e. the rate of transcriptional initiation is dependent only upon promoter architecture). There are four types of allosteric circuit, as shown in Figure 29.4.

Long-term responses in bacteria are often mediated by synthesizing a new component of RNA polymerase which recognizes a distinct promoter structure. In *E. coli*, new  $\sigma$ -factors are synthesized in response to heat shock and nitrogen starvation; in *B. subtilis*, a cascade of  $\sigma$ -factors control sporulation (see Box 6.2).

Because bacterial genes are close together, a form of transcriptional repression can be exploited where transcription from one promoter can block transcription from another. This is termed **countertranscription**, and the antisense RNA produced by transcribing through a gene in the unorthodox direction is termed **countertranscript RNA**. Countertranscription is extensively exploited by



bacteriophage  $\lambda$  in the control of lysis and lysogeny (see Box 30.1); it may work by preventing transcriptional initiation at the normal promoter, by disrupting elongation, by forming triplex structures, or by antisense effects. An example from the *E. coli* genome is the CAP protein, which inhibits the transcription of its own gene by activating transcription from a counterpromoter within its transcription unit. Countertranscription also occurs occasionally in eukaryotes (q.v. *parental imprinting*).

**Characteristic regulatory strategies of eukaryotes.** In eukaryotes, all genes are inactive as a default state because they require the presence of constitutive transcriptional activators for expression. Rapid responses to external stimuli in eukaryotes are therefore predominantly mediated by positive inducible regulation, i.e. by the activation of a latent transcriptional activator. Allosteric regulation is pivotal in bacterial systems, but the integration of many signaling pathways in eukaryotes makes covalent modification of either the transcription factor itself, or a protein that interacts with it, the principle mechanism of transcription factor regulation in eukaryote inducible systems. Such transcription factors bind to so-called **response elements** in the target genes (Table 29.4). Long-term processes, such as differentiation in multicellular organisms, are controlled predominantly by the synthesis of new transcription factors. However, cell-type-specific and developmental transcription factors also function by binding to regulator elements in the promoters and enhancers of their target genes, and examples of these are also shown in Table 29.4.

A major difference in regulatory strategy between bacteria and eukaryotes is that eukaryotes lack operons. With few exceptions, all eukaryotic transcripts are monocistronic, reflecting the inability of eukaryotic ribosomes to initiate protein synthesis from an internal site. Some eukaryotic genes are clustered and under common *cis*-regulation (q.v. *locus control region*, *Hox genes*, *parental imprinting*) and a distant enhancer can interact with more than one promoter, allowing coregulation (this can be shown in the immunoglobulin locus by placing two promoters under the influence of the kappa enhancer). However, in these examples the genes are not cotranscribed. Consequently, eukaryotic genes which need to be coregulated are often dispersed but under common *trans*-regulation. Genes with common regulatory elements can be coactivated, and genes with overlapping sets of elements can be individually activated, which allows flexibility in transcriptional control. In eukaryotes a group of genes under common *trans*-regulation is a **gene battery** (the equivalent in bacteria is termed a **regulon**). A gene battery or regulon may consist of several to many dispersed genes and operons controlled by the same transcription factor. There are three occasions where polycistronic RNA occurs in eukaryotes: (1) ribosomal RNA genes, which are not translated and therefore do not use ribosomes; (2) bicistronic transcripts in *C. elegans*, which are subject to alternative *trans*-splicing to generate mature transcripts with caps upstream of each open reading frame; (3) polycistronic picornaviral genomes. The picornaviruses possess *internal ribosome entry sites* (q.v.), allowing ribosomes to bind upstream of each open reading frame (these elements have been extensively exploited for gene manipulation; q.v. *gene knockout*, *reporter gene*).

## 29.6 Transcriptional elongation and termination

**Elongation and polymerase processivity.** The biochemistry and kinetics of elongation are similar in bacteria and eukaryotes. Elongation is not smooth: the RNA polymerase pauses if it encounters an impediment, such as a secondary structure, and this may progress to **arrest** (where the enzyme loses contact with the end of the transcript) or **termination** (where it also loses contact with the template). In eukaryotes, a number of general **elongation factors** are associated with the enzyme during elongation and act to suppress pausing (e.g. elongin) and prevent arrest (e.g. TEFb, TFIIS). TFIIS (also known as SII) may prevent arrest by cleaving the nascent transcript, placing a new 3' terminus at the active site of the enzyme (the nascent strand may continue to elongate when the enzyme pauses, thus placing the reactive 3' hydroxyl group outside the active site). While elongation is constitutive for most eukaryotic genes, pausing and consequent termination can be regulat-



**Table 29.4:** A selection of regulatory motifs found upstream of eukaryotic genes involved in the control of constitutive, inducible, cell- or lineage-specific and developmentally regulated gene expression, and the transcription factors which recognize them

Consensus sequence	Motif name	Transcription factor (class)	Comments
<i>Constitutive sites</i>			
GCCAATCT	CAAT box	CTF/NF1 CP family C/EBP (bZIP)	<i>Distribution</i> Ubiquitous Ubiquitous Ubiquitous but high levels in liver
GGG CGG ATGCAAAT	GC box Octamer	Sp1 (zf) Oct1 (h) Oct2 (h)	Ubiquitous Ubiquitous B-lymphocytes
<i>Response elements</i>			
CNNGAANNCTCCNNG CCATATTAGG	HSE SRE	HSF (bZIP) SRF	<i>Stimulus</i> Heat shock Growth factors in serum
TTNCNNNA	IGRE	STAT 1	Interferon- $\gamma$
<i>Cell-type-specific elements</i>			
GATA ATATTCAT CANNG GGGACTTTCC TTYAGNACCRCGGASAGNRCC	E-box $\kappa$ B site NRSE	GATA-1 (zf) Pit-1 (POU) MyoD1 (bHLH) NF- $\kappa$ B NRSF/REST (zf)	<i>Cell type</i> Erythroid Pituitary Myoblasts Lymphoid cells Nonneural cells
<i>Developmental regulator sites</i>			
TCCTAATCCC		Bicoid (h)	<i>Developmental system</i> <i>Drosophila</i> AP axis specification
GCGGGGGGCC		Krox-20 (zf)	Vertebrate hindbrain development
TAATAATAATAATA		Antennapedia (h)	<i>Drosophila</i> homeotic gene
TCAATTAAATGA		Fushi tarazu (h)	<i>Drosophila</i> pair rule gene

Transcription factors are classified by DNA-binding/dimerization domain where known (bHLH, basic helix-loop-helix; bZIP, basic leucine zipper; H, homeodomain; POU, POU domain; zf, zinc finger). C/EBP (CAAT/enhancer binding protein) binds to two distinct sites, the CAAT box and the enhancer core sequence TGTGGWWWG. Abbreviations: CTF/NF1, CAAT transcription factor/nuclear factor 1; CP, CAAT-binding protein; HSE/F, heat shock element/factor; SRE/F, serum response element/factor; IGRE, interferon- $\gamma$  response element; STAT, signal transducer and activator of transcription; NRSE/F, neural restrictive silencer element/factor.

able: c-Myc protein levels in myeloid cells are regulated predominantly at the level of transcriptional elongation and termination. In mature granulocytes, full-length *c-myc* mRNA is synthesized and translated, whereas in undifferentiated cells, transcription terminates in the first exon. Pausing also regulates transcription of the HIV genome, with a low level of mostly truncated transcripts produced during early infection and a high level of mostly full-length transcripts produced subsequently. In this example, regulation is mediated by a viral-encoded *trans*-acting factor called Tat which binds to a *cis*-acting element in the RNA called Tar. Tat increases the frequency of initiation as well as alleviating polymerase pausing.

The size difference between bacterial and higher eukaryotic genes means that while the completion of bacterial transcription is rapid, the transcription of the larger eukaryotic genes takes several hours — 16 h in the case of the 2.5 Mbp human dystrophin gene, the largest gene known. Recent

evidence suggests that elongating RNA polymerase II is associated both with the polyadenylation apparatus and a spliceosome as it translocates along the DNA, both complexes interacting via the C-terminal domain. Thus the dystrophin gene and others are spliced cotranscriptionally.

**Termination of transcription in bacteria.** In bacteria, transcription terminates at discrete sequences, **terminators (t)**, and can involve two mechanisms, both of which respond to a signal in the transcript itself rather than in the gene.

The most common is **intrinsic termination (p-independent termination)**: transcripts adopt particular secondary structures, causing RNA polymerase to pause and the DNA–RNA hybrid nucleic acid to dissociate. A common terminator motif is a GC-rich inverted repeat followed by a poly-U sequence: the GC-rich palindrome forms a hairpin, stalling the elongation complex, leaving the DNA and RNA paired by a run of weak A:U base pairs which cause the RNA polymerase to dissociate from the template. It is not clear exactly how this dissociation occurs. It is unlikely, as originally thought, that the weakness of the multiple A:U base pairs itself causes dissociation. An alternative model suggests that interaction between RNA polymerase and its product is the pivotal determinant: the weak pairing may facilitate displacement of the nascent RNA terminus from the template and its loss from the active site of the enzyme. This causes arrest, and the conformation adopted by the polymerase at this point may allow the hairpin loop to displace it.

The alternative termination mechanism, **p-dependent termination** is rare in the bacterial chromosome but common for phage. It requires a protein called  $\rho$  (**rho**) which binds to free RNA and separates it from DNA by interacting directly with the RNA polymerase, perhaps by wedging itself into the DNA:RNA hybrid at the active site of the enzyme.  $\rho$  does not bind directly to the hybrid duplex but to a specific site in the transcript, which is probably identified by a cytidine-rich and guanosine-poor region between 50 and 100 nucleotides in length. It is thought that  $\rho$  may translocate along the transcript towards the elongation complex and release the transcript from the DNA, although there has been no direct proof that the molecule translocates along RNA, and there is evidence that  $\rho$  remains attached to its initial binding site. In the **hot pursuit model**, the elongating transcript outruns  $\rho$ , and  $\rho$  fails to catch up to the transcription bubble until the polymerase pauses. It is not known what causes the enzyme to pause at p-dependent terminators. In some cases, a hairpin structure has been identified (e.g. the  $t_{R1}$  terminator in bacteriophage  $\lambda$  and in the *E. coli* *his* operon), but much of the pausing activity is retained when these structures are removed. **Antitermination** (q.v.) usually occurs at p-dependent terminators.

**Termination of transcription in eukaryotes.** The termination of transcription in eukaryotes is poorly characterized. Termination of RNA polymerase I transcription occurs at a site approximately 1 kb 3' to the end of the mature rRNA and involves the recognition of a specific *cis*-acting element. Termination of RNA polymerase III transcription occurs at sites similar to bacterial p-independent terminators (i.e. GC-rich sequences followed by polyuridine tracts), but the GC-sequence does not appear to form a secondary structure and termination occurs at the second uridylate residue, suggesting disassociation at a weak run of deoxyadenylate: uridylate residues is not the key mechanism. Most RNA polymerase II transcripts are processed by 3' cleavage and polyadenylation (q.v. *RNA processing*) so the intrinsic termination reaction is not clear; it is possible that transcription continues beyond the 3' end of the gene. However, recent evidence showing that the cleavage and polyadenylation apparatus is associated with the elongating RNA polymerase suggests that cleavage and polyadenylation may also be linked to arrest, and dissociation of RNA polymerase II from the template.

**The regulation of termination in bacteria.** Bacteria can regulate transcription at the level of termination either positively (by stimulating readthrough of a terminator site — a process termed **antitermination**) or negatively (by inducing premature termination — a process termed **attenuation**), both processes may be used in the context of operon organization to add or subtract reading frames from a polycistronic mRNA.

Antitermination occurs in the temporal control of bacteriophage  $\lambda$  gene expression and in the *E. coli* rRNA operon. The RNA polymerase is modified before reaching the terminator as it passes an **antitermination site** at which is bound an **antitermination protein** such as the bacteriophage  $\lambda$  N or Q proteins (see Viruses and Subviral Agents). There appears to be no essential site of action for the antiterminator protein: the phage antitermination sites *nutL*, *nutR* and *qut* are located at the promoter, at the terminator and upstream of the promoter, respectively. The nature of the modification is not precisely understood. In the case of the Q termination system, Q binds to RNA polymerase as it pauses at the promoter and is carried with it along the template. Similarly, the N protein forms a tight complex with RNA polymerase and other cellular proteins NusA, NusG, NusB and ribosomal protein S10 (NusE). They may act by directly controlling the kinetics of elongation, i.e. by acting as pausing suppressors, or they could specifically interact with the termination apparatus, by preventing arrest or directly controlling template structure.

Attenuation is a bacterial regulatory mechanism controlling the expression of several operons concerned with amino acid biosynthesis and, in *E. coli*, the *pyrBI* operon which encodes enzymes for pyrimidine synthesis. In attenuator control, transcriptional termination is controlled by the efficiency of protein synthesis (in the amino acid operons) or transcription itself (in the *pyrBI* operon), allowing transcription to be regulated by the availability of substrates for these two essential reactions.

For the amino acid operons, ribosome stalling due to the lack of particular amino acids causes inefficient translation of a short open reading frame encoding a **leader peptide**, upstream of the major open reading frames. Due to the influence of stalled ribosomes, the nascent mRNA adopts a structure favoring continued elongation, resulting in the synthesis of biosynthetic enzymes. Conversely, efficient translation of the leader when particular amino acids are plentiful causes it to adopt the structure of an intrinsic terminator, preventing synthesis of unnecessary enzymes. The leader ORF features several tandem codons for the amino acid, whose biosynthesis is controlled by the downstream genes, e.g. there are nine tandem histidine codons in the *his* operon leader. In the absence of histidine, the ribosome will stall at this attenuator site and transcriptional elongation continues, but if there is abundant histidine, the leader is translated and transcription is terminated; histidine biosynthetic enzymes are thus synthesized only when the cell senses a lack of histidine.

In the *pyrBI* operon, transcription is slowed in a region containing multiple tandem adenosine residues if there is limiting UTP, and this allows ribosomes to inhibit the formation of a terminator. When UTP is plentiful, transcription through the attenuator is rapid and the terminator site in the mRNA is exposed before the ribosomes can bind, resulting in termination.

#### Box 29.1: Eukaryotic gene regulation — $\beta$ -globin and the $\beta$ -globin cluster

**Globin gene organization.** The oxygen carrying protein hemoglobin consists of two  $\alpha$ -globin and two  $\beta$ -globin class polypeptides. The genes for the globin proteins are found in two clusters: the  $\alpha$ -globin cluster consists of the  $\zeta$ -globin and  $\alpha$ -globin genes, whereas the  $\beta$ -globin cluster consists of the  $\epsilon$ -globin,  $\gamma$ -globin,  $\delta$ -globin and  $\beta$ -globin genes (see figure below). The globin genes are developmentally regulated, as discussed in Box 15.1.

**Regulation of the  $\beta$ -globin gene: Promoter and enhancers.** The regulation of  $\beta$ -globin gene expression has been studied in detail by *in vitro* deletion experiments, and by mutation of each site in the

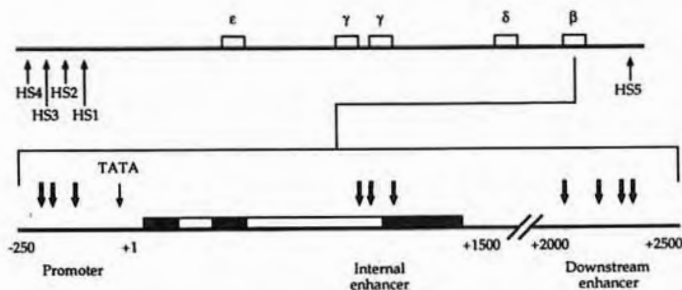
immediate 5' flanking region. The promoter spans about 250 bp and contains a TATA box and a CCAAT box, as well as binding sites for several ubiquitous transcriptional activators and the erythroid-specific factor GATA-1. The promoter is sufficient for minimal, cell-type-specific transcription, but high-level transcriptional activity depends on two enhancers, one partially spanning intron 2 and exon 3, and one located downstream of the gene. The intergenic enhancer contains three GATA-1 binding sites and is thus responsible for elevated tissue-specific expression of the gene. The downstream enhancer contains binding sites for constitutive and erythroid-

specific factors and is concerned with temporal regulation of  $\beta$ -globin gene expression.

**The  $\beta$ -globin cluster locus control region (LCR).** Located ~20 kbp upstream of the  $\epsilon$ -globin gene (the most 5' gene of the  $\beta$ -globin cluster) is a locus control region required for high-level expression of all the  $\beta$ -globin cluster genes in erythroid cells. Individuals with deletions spanning the LCR lose expression of all genes in the cluster ( $\epsilon\gamma\delta\beta$ -thalassaemia). Furthermore, transgenic mice with the human  $\beta$ -globin gene driven by its promoter and dual enhancers express the construct only minimally, whereas if the LCR is included, human  $\beta$ -globin is strongly expressed. The LCR contains four erythroid-specific DNase I hypersensitive sites (q.v.), representing nucleosome-free regions where transcription factors are located. There are many binding sites for positive regulators, including both constitutive factors and the erythroid-specific transcriptional activators GATA-1 and NF-E2. There are several models for the activity of the LCR. It is thought to establish an independent chromatin domain by forming a topologically isolated chromatin loop, and putative matrix associated regions have been identified in the LCR that can act as *boundary elements* (q.v.) in transgenic mice. The LCR is thus thought to protect the cluster from chromatin-position effects, explaining its ability to increase expression of integrated  $\beta$ -globin reporter genes in transgenic mice, but its inability to increase the expression of transfected  $\beta$ -globin genes, which are episomally main-

tained and thus not subject to position effects.

In addition to its role as a general positive regulator, the LCR also controls the temporal aspects of  $\beta$ -globin cluster gene expression. Further transgenic studies have shown the  $\beta$ -globin gene to be activated at different times during development depending on its distance from the LCR. Models explaining the distance-dependent activation, involve the propagation of a wave of chromatin remodeling across the cluster and/or the sequential interaction between the LCR and individual promoters. It is thought that the multiple transcriptional activators bound at the LCR may interact with transcription factors bound at individual promoter/enhancer complexes to form a so-called **holocomplex**, thus activating gene expression, perhaps by recruiting further essential activators. This is supported by the presence of DNase hypersensitive sites in globin gene promoters only in the presence of the LCR. There would be competition between individual promoters for LCR-interaction, as seen in enhancer competition (q.v. *parental imprinting*), and the outcome would depend on the relative stability of individual holocomplexes. In support of this model, a transcription factor related to the *Drosophila* Krüppel protein has recently been identified, which is essential for  $\gamma$ -globin to  $\beta$ -globin switching, and may therefore stabilize the  $\beta$ -globin-LCR holocomplex. Additionally, the levels of other transcription factors, including GATA-1, have been shown to influence the choice of promoter targeted by the LCR.



The human  $\beta$ -globin cluster (upper) consists of five genes whose order along the chromosome is reflected by the temporal order of activation. There is a locus control region located 20 kbp to the 5' side, characterized by four DNase I hypersensitive sites (HS1-4). The  $\beta$ -globin gene (lower) has three exons and two introns and is regulated by an upstream promoter and dual enhancers. Binding sites for the erythroid-specific transcription factor GATA-1 are shown by thick arrows. There are also numerous binding sites for GATA-1 and a second erythroid-specific protein, NF-E2, in the locus control region.



**Box 29.2: Bacterial gene regulation — the *lac* operon**

**Basic structure.** The *lac* operon contains three structural genes required for lactose catabolism (see figure and table below) which are coordinately induced in the presence of lactose. The three genes are arranged in the order *lacZYA* and are cotranscribed as a polycistronic message from a single promoter *P1*. There is also a second, latent promoter, *P2*, whose function *in vivo* is unknown. There are four regulatory elements in the *lac* operon. Approximately 60 bp upstream of *P1* is a positive regulatory element, the **activator site (AS)**, which binds the CAP-cAMP (catabolite activator protein-cAMP) complex. 11 bp downstream of *P1* is the **major operator**, *O1*, which binds Lac repressor. There are two further **minor operator sites** (*O2* and *O3*), at positions -82 and +401, which also bind Lac repressor. The Lac repressor is encoded by the *lacI* gene, which is immediately upstream of the operon and is constitutively expressed (i.e. there are no regulatory elements in its promoter). The adjacent position of the *lacI* gene is not important, however, and in other loci the regulator gene is located a great distance from the structural genes of the operon (e.g. the *gal* and *trp* operons).

Gene	Enzyme	Function
<i>lacZ</i>	$\beta$ -galactosidase	Lactose $\rightarrow$ $\beta$ -D-galactose + D-glucose
<i>lacY</i>	$\beta$ -galactoside permease	Transport of lactose into cell
<i>lacA</i>	$\beta$ -galactoside transacetylase	Unknown. Not required for lactose metabolism

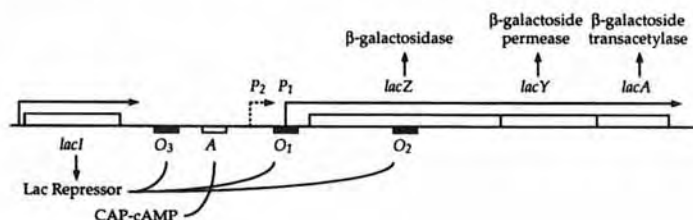
**The regulatory circuit.** The *lac* operon is repressed as a default state because the Lac repressor is synthesized and is able to bind the operators and inhibit transcription. Allolactose, a metabolic product of lactose, binds to the repressor and inactivates it by causing a conformational change in the DNA-binding domain. Thus, in the presence of lactose, the Lac repressor is inactivated and the lactose

catabolic genes derepressed — the *lac* operon is an *inducible operon* under *negative regulation*. However, promoter *P1* lacks a -35 box and is therefore weak. For significant transcriptional activity, the CAP-cAMP complex must bind at the activator site (AS) to directly assist the binding of RNA polymerase. CAP-cAMP activity depends upon the level of cAMP in the cell, which is inversely proportional to the level of glucose. Thus, the *lac* operon is only activated in the presence of lactose when the favored carbon source, glucose, is absent (**catabolite repression**). It is still not known how glucose regulates cAMP levels.

**Mutations which define functional elements.** Functional elements in the *lac* operon have been defined by mutation and *complementation analysis* (q.v.). The latter requires diploidy for the operon, and in bacteria this is achieved by introducing a second copy of the operon into the cell as part of a plasmid (q.v. *F' plasmid*). Mutations in the *lac* operon can either affect the expression of single genes or groups of genes, they can be dominant or recessive to wild-type alleles in the same cell, and can act in *cis* or in *trans*.

Mutations which affect the expression of single genes map to the structural genes themselves and result in individual loss of function; these are usually recessive because a wild-type copy of the gene in the same cell can supply functional product. Mutations in *lacZ* and *lacY* generate the Lac mutant phenotype (inability to utilize lactose), whilst mutations in *lacA* lack a phenotype because this gene is not essential for lactose utilization; it may be responsible for interconverting lactose analogs. Occasionally, dominant *LacZ* or *LacY* mutants arise because both encoded proteins function as multimers and mutant polypeptides can act in a dominant negative fashion.

Mutations which affect the expression of all three genes generally map to the regulatory components. Mutations in the promoter, activation site and operators are *cis-dominant* (q.v.). Those affecting the pro-



motor or activation site result in the Lac phenotype because transcription is disrupted, whereas operator mutants are often constitutive because the repressor is unable to bind. Mutations in *lacI*, which encodes the repressor protein, are *trans*-acting, as are mutations in the *crp* gene encoding the CAP protein and in the *cya* gene which encodes adenylate cyclase. Repressor mutants fall into three classes: *lacI<sup>-</sup>* is a recessive loss of function (constitutive) mutant — no repressor is produced; *lacI<sup>d</sup>* is a dominant negative (constitutive) mutation — the repressor is unable to bind the oper-

ator and sequesters wild-type repressor into inactive multimers; *lacI<sup>S</sup>* is a dominant uninducible mutation — the repressor binds to the operator but not the inducer — it is a **superrepressor**. Occasionally, mutations in the structural genes can have *cis*-dominant effects on the other structural genes of the operon — these are *polar mutations* (q.v.), which disrupt ribosome binding at downstream open reading frames, probably resulting from altered secondary structure in the mutated gene.

### Box 29.3: Activation and inhibition domains of eukaryotic transcription factors

**Families of activation domains.** Transcription factors often possess multiple domains with different domains carrying out different functions — DNA binding, transcriptional regulation, dimerization, interaction with cofactors, etc. The activation domains of eukaryotic transcription factors fall into a three major families based not on any sequence homology, but on their high proportions of particular amino acids. **Acidic activation domains** are the most common (e.g. herpes simplex virus VP16 transactivator, *S. cerevisiae* GAL4 and the mammalian glucocorticoid receptor) and contain up to 20% glutamic and aspartic acid residues. **Glutamine-rich activation domains** are found in several homeobox and POU domain transcription factors and in the constitutive transcriptional activator Sp1, whereas **proline-rich activation domains** are found in c-Jun and the CCAAT binding factor CTF/NF1. The engineering of artificial proteins has shown that the activity of glutamine-rich domains reflects solely the presence of multiple glutamine residues. Conversely, in acidic activation domains, there are many critical residues in addition to the acidic ones (although the strength of transcriptional activation does appear to reflect the number of acidic residues available).

**Position dependence.** The investigation of transcription factor activity in different *cis*-acting sites has shown that the three domains differ in their ability to stimulate transcriptional initiation. Transcription factors with acidic activation domains can activate

transcription when bound at the promoter or distant enhancer of a reporter gene. Conversely, transcription factors with glutamine- or proline-rich domains appear to be unable, or only weakly able, to activate transcription when bound at a distant enhancer.

**Target preference.** Transcription factors with different classes of activation domain interact with different components of the basal apparatus. The constitutive transcription factor Sp1, for instance, interacts with TAF<sub>II</sub>110 of TFIID (the nomenclature of the TAFs reflects the RNA polymerase they associate with and their molecular mass in kDa — TAF<sub>II</sub>110 is a 110 kDa TAF associated with RNA polymerase II). Conversely, several transcription factors with acidic activation domains interact predominantly with TAF<sub>II</sub>40.

**Inhibition domains.** There are few direct transcriptional repressors in eukaryotes compared to the vast number of activators. However, where such proteins have been identified, a specific inhibition domain has often been shown to confer transcriptional repression ability upon a heterologous protein. Several of the defined inhibition domains, including that of the *Drosophila* protein Even-skipped, are proline rich and lack charged residues. However, others have unique domains and it remains to be seen if, as more such factors are characterized, there will be easily recognized inhibition domain families as there are for the activation domains.

## Further reading

- Adhya, S. (1996) The *lac* and *gal* operons today. In: *Regulation of Gene Expression in Escherichia coli* (eds E.C.C. Lin and A.S. Lynch), pp. 181–200. R.G. Landes, Texas/Chapman & Hall, New York.
- Bjorklund, S. and Kim, Y.J. (1996) Mediator of transcriptional regulation. *Trends Biochem. Sci.* 21: 335–337.
- Busby, S. and Kolb, A. (1996) The Cap modulon. In: *Regulation of Gene Expression in Escherichia coli* (eds E.C.C. Lin and A.S. Lynch), pp. 255–280. R.G. Landes, Texas/Chapman & Hall, New York.
- Hannarose, W. and Hansen, U. (1996) Active repression mechanisms of eukaryotic transcription repressors. *Trends Genet.* 12: 229–234.
- Henkin, T.M. (1996) Control of transcriptional termination in prokaryotes. *Annu. Rev. Genet.* 30: 35–57.
- Kamakaka, R.T. (1997) Silencers and locus control regions: Opposite sides of the same coin. *Trends Biochem. Sci.* 22: 124–128.
- Magasanik, B. (1996) Regulation of nitrogen assimilation. In: *Regulation of Gene Expression in Escherichia coli* (eds E.C.C. Lin and A.S. Lynch), pp. 281–290. R.G. Landes, Texas/Chapman & Hall, New York.
- Novina, C.D. and Roy, A.L. (1996) Core promoters and transcriptional control. *Trends Genet.* 12: 351–355.
- Ptashne, M. and Gann, A. (1997) Transcriptional activation by recruitment. *Nature* 386: 569–575.
- Reeder, R.M. and Lang, W.H. (1997) Terminating transcription in eukaryotes: lessons from RNA polymerase I. *Trends Biochem. Sci.* 22: 473–477.
- Reines, D., Conaway, J.W. and Conaway, R.C. (1996) The RNA polymerase II general elongation factors. *Trends Biochem. Sci.* 21: 351–355.
- Rippe, K., Von Hippel, P.H. and Langowski, J. (1995) Action at a distance — DNA-looping and initiation of transcription. *Trends Biochem. Sci.* 20: 500–506.
- Roberts, J.W. (1996) Transcription termination and its control. In: *Regulation of Gene Expression in Escherichia coli* (eds E.C.C. Lin and A.S. Lynch), pp. 27–46. R.G. Landes, Texas/Chapman & Hall, New York.
- Roeder, R.G. (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* 21: 327–335.
- Svaren, J. and Horz, W. (1997) Transcription factors vs nucleosomes: Regulation of the *PHO5* promoter in yeast. *Trends Biochem. Sci.* 22: 93–97.
- Svejstrup, J.Q., Vichi, P. and Egly, J.M. (1996) The multiple roles of transcription/repair factor TFIIH. *Trends Biochem. Sci.* 21: 346–350.
- Uptain, S.M., Kane, C.M. and Chamberlin, M.J. (1997) Basic mechanisms of transcript elongation and its regulation. *Annu. Rev. Biochem.* 66: 117–172.
- Van Dromme, M., Gauthier-Rouviere, C., Lamb, N. and Fernandez, A. (1996) Regulation of transcription factor localisation — fine-tuning of gene expression. *Trends Biochem. Sci.* 21: 59–64.
- Verrijzer, C.P. and Tjian, R. (1996) TAFs mediate transcriptional activation and promoter selectivity. *Trends Biochem. Sci.* 21: 338–342.
- Wood, W.G. (1996) The complexities of  $\beta$ -globin gene regulation. *Trends Genet.* 12: 204–206.

## Websites

- TRANSFAC, a database which contains a list of known transcription factors and their binding sites — <http://transfac.gbf-braunschweig.de/TRANSFAC/browse/index.html>
- The Homeobox page, containing a listing and classification of all homeodomain proteins — <http://copan.bioz.unibas.ch/homeo.html>

**This Page Intentionally Left Blank**



## Chapter 30

# Viruses and Subviral Agents

### Fundamental concepts and definitions

- **Viruses** are small, noncellular parasitic organisms. They lack an intrinsic metabolism and depend on the host cell to provide the raw materials and enzymes for essential functions — replication, transcription and protein synthesis. Viruses have been identified in all types of cell. Viruses which infect bacteria are termed **bacteriophage** (or simply **phage**).
- The viral genome is either DNA or RNA, and is associated with proteins, usually in the form of a shell (or **capsid**), which may be surrounded by a proteolipid envelope. The genome encodes structural components of the capsid and often enzymes and other products required for successful completion of the infection cycle. The size of viral genomes ranges from about 1 kbp to 400 kbp, the smallest containing only a single gene and the largest more than 300. A complete virus particle is a **virion**.
- The infection of a cell by a virus is usually harmful, often because the virus interferes with host-cell functions by altering patterns of endogenous gene expression. In bacteria, this results in a slowing of growth and often cell lysis. In multicellular organisms, viral infection is often manifest as a localized or systemic disease.
- Viruses are classified according to morphological characteristics (shape and size of capsid and associated structures), physicochemical properties (nature of the genome, viral-encoded proteins, lipids and carbohydrate content) or biological characteristics (e.g. host range, mode of transmission, pathogenicity). There are several different types of capsid: the classic icosahedral capsid (which in phage may or may not be associated with a tail for attachment), spherical and bacilliform capsids, and helical (rod-shaped) capsids. The latter are unique in that their size depends on the nucleic acid of the virus — subunits condense around the genome. Other capsids are preformed and then filled with nucleic acid (**encapsidation**), i.e. they have a defined capacity. Some viruses have unique morphology (e.g. the lemon-shaped *Fuselloviridae*), whereas some are **pleomorphic** or **amorphic** (i.e. they lack a structurally defined capsid).
- Viral genome structure is diverse, reflecting an equally diverse spectrum of genome replication and gene expression strategies. Many viruses have evolved elegant and ingenious methods for the temporal regulation of gene expression during their infection cycle, and for the avoidance or neutralization of interference from the host cell. These are particularly evident in eukaryotic viruses which have to avoid the immune system, and overcome the effects of the cell cycle and the monocistronic environment to replicate and express their genes.
- Viruses have been extensively exploited as vectors for gene manipulation, cloning and gene transfer. Some of the most widely used viruses include bacteriophage  $\lambda$  (q.v. *molecular cloning, genomic library*), bacteriophage M13 (q.v. *DNA sequencing, phage display*), baculovirus, adenovirus, herpesvirus and vaccinia virus (q.v. *expression vectors, gene therapy*) and retroviruses (q.v. *transgenic mice*). In addition, strong viral promoters and enhancers have often been used to drive gene expression in plasmid vectors (e.g. promoters from bacteriophage  $\lambda$ , SV40, Rous sarcoma virus and cauliflower mosaic virus.)
- A number of organisms, defined here as **subviral agents**, have also been characterized. Like viruses, these are noncellular parasites, but they lack one or more of the definitive features of a virus, e.g. a self-encoded capsid, or the ability autonomously to infect a cell. The most intriguing subviral agent is the **prion**, the cause of transmissible spongiform encephalopathies. This organism is thought to be a pathological conformational isomer of a host-encoded neuronal protein with the ability to convert normal isomers into copies of itself.

**Table 30.1:** Stages of a productive viral infection (lytic or persistent)

Stage of infection cycle	Events
Attachment or adsorption	Virus particle attaches to receptor on the outside of the cell
Penetration	Internalization of the particle or part of it (at least the genome must enter the cell)
Uncoating	Release of the genome from the virion if the whole virion enters the cell. In eukaryotic viruses this may be followed by a transport phase where the partially uncoated virion is taken to its normal site of replication. For DNA viruses, this is generally the nucleus, and for RNA viruses, the cytoplasm
Early gene expression	Early genes are usually associated with replication and regulation of late gene expression
Replication	Production of many copies of the viral genome
Late gene expression	Late genes are associated with virion assembly and escape
Virion assembly	Building new capsids and packaging genome
Release	Release of progeny virus from cell

The two differ in the consequences of the release stage. Lytic infections cause host-cell destruction, whereas persistent infections result in continuous release of viruses from the cell.

### 30.1 Viral infection strategy

**Viral infection cycles.** The infection cycle (Table 30.1) begins with the introduction of the virus, or part of it, into the cell. Once inside, viral gene expression begins. Often, gene expression is organized as a cascade and viral genes can be divided into several temporal groups. So-called **early genes** are concerned with replicating the viral genome and regulating the **late genes**, which are concerned with synthesis of the progeny virus particles and genome packaging. One of the best characterized strategies is that of *E. coli* bacteriophage  $\lambda$ , which has three groups of temporally regulated genes (Box 30.1).

The consequences of viral infection vary according to the type of virus and host cell. In many cases, viral infection kills the cell — this is known as **lytic infection**, and phage engaged in this type of infection are described as **virulent**. Initial infection is followed by a **latent period** where the virus propagates, then release of the progeny virus from the cell occurs by **lysis** (breaking open). Alternatively, the virus may coexist with the cell and produce progeny by continued budding or extrusion — a **persistent infection**. Both lytic and persistent infections are *productive*, i.e. infectious progeny viruses are released (Table 30.1).

Many viruses are also capable of **latent infection**: the virus is maintained in the cell but does not produce infectious progeny. In bacteria, phage capable of latent infection are described as **temperate** and are said to *lysogenize* their host — the host is termed a **lysogen**, i.e. it can undergo lysis if the phage re-enters the lytic cycle (a process termed **induction**<sup>1</sup>). Many temperate phage lysogenize their host by integrating into the genome, the integrated phage being described as a **prophage**. Bacteriophage P1 is an exception in that the prophage is maintained episomally as a plasmid. Some eukaryotic viruses can also integrate, and are termed **proviruses**. However, an integrated virus is not necessarily latent. Bacteriophage Mu is not only virulent as a prophage, but it also *depends* upon integration for its virulence. The same is true of eukaryotic *retroviruses* (q.v.). Both bacteriophage Mu and the eukaryotic *retroviruses* are *transposable elements* (q.v.). Animal viruses may display tissue-specific latency, e.g. herpes simplex virus is latent in neurons but not in most other cells. Latency can

<sup>1</sup>The term *induction* is used in a number of different ways, each with the sense of 'turning something on': (1) the induction of the lytic cycle of temperate phage; (2) the induction of repressed genes in operons or generally switching on gene expression by an external stimulus (see Transcription); (3) the induction of one cell type by another to change its fate in development, i.e. by switching on genes involved in differentiation (see Development: Molecular Aspects).

be disrupted by altering the balance of host- and virus-encoded transcription factors. Lytic animal viruses may multiply with differing efficiencies in different cells. In some cases, no infectious particles are produced at all (**abortive infection**), reflecting a deficiency in that cell type for a product essential for virus replication, gene expression or genome packaging e.g. influenza virus infections are abortive in the absence of the host-encoded protease which cleaves the hemagglutinin precursor.

A final consequence of some viral infections in animal cells is neoplastic transformation. **Transforming infections** are caused by papovaviruses (e.g. SV40), adenoviruses and herpesviruses, each of which encodes products which deregulate the *cell cycle* (q.v.). Additionally, retroviruses, which can integrate into the host genome, can also transform their host cells by transducing or activating genes which stimulate cell growth. Such genes are termed *oncogenes* (q.v.).

**How the virus gets into the cell.** In many host-virus relationships, the virus gains entry into the cell by binding to a cell surface receptor. In animal viruses this is mediated by a domain on one of the viral coat proteins, termed an **attachment site**. Usually, this interaction is the sole determinant of the **host range**, and the virus receptors are often molecules whose normal cellular function concerns the immune response (e.g. MHC class I proteins act as receptors for adenovirus and SV40, CD21 is the receptor for Epstein-Barr virus and CD4 for HIV). Bacteriophages also adsorb to surface receptors (e.g. bacteriophage  $\lambda$  attaches to the maltose receptor), but some attach to other appendages such as the flagellum or conjugal pili, in the latter case being internalized when the pilus is retracted into the cell. The host range of these phages is thus determined by the presence and nature of these appendages (e.g. the f1 phages are described as male-specific because only 'male' bacteria possess pili (q.v. *conjugation*). Many bacteriophage are endowed with **fixation organelles**: specific structures such as tails and spikes which facilitate adsorption.

Once attached to a specific receptor, various strategies are used to internalize the virus. In some cases, the entire virion does not enter the cell. Many phage create a breach in the cell wall and eject their DNA into the cytoplasm, i.e. only the genome enters the cell. Animal viruses are taken up by endocytosis. Acidification of the resulting endosome results in the uncoating of nonenveloped viruses, but it is unknown how the uncoated virus enters the cytoplasm during the transport phase. Enveloped viruses fuse with the membrane of the endosome and are released into the cytoplasm.

Plant viruses need to cross the impermeable cuticle and the plant cell wall before entering the cytoplasm. For this reason, viral infection is often mediated mechanically by transfer from one damaged cell to another or via a vector such as an insect. Many plant viruses encode proteins which interact with the plasmodesmata and allow semicomplete virions to spread from cell to cell.

### 30.2 Diversity of replication strategy

**Diversity of viral genome structure.** Whereas all cells possess a genome of double-stranded DNA, usually circular in prokaryotes and linear in eukaryotes, viruses display much greater diversity in the nature of their genome. An important division of the viruses separates those with a DNA genome from those with an RNA genome; there are no viruses with mixed nucleic acid genomes.

Each type of nucleic acid brings with it specific advantages and disadvantages, especially for those viruses infecting eukaryotic cells (Table 30.2). DNA genomes are more stable because of the inherent properties of DNA (see Nucleic Acid Structure) and because they can exploit host encoded DNA-repair functions (see Mutagenesis and DNA Repair). DNA genomes have thus evolved to be larger and more complex than RNA genomes, with the most complex DNA viruses carrying many nonessential genes, i.e. genes whose functions are complemented by host cell proteins. It is not clear why viruses should have evolved to carry redundant functions, although they may confer an advantage upon the virus in different host-cell environments. For example, some eukaryotic viruses encode seven transmembrane domain receptors, which may alter host-cell signaling (see Signal Transduction).

**Table 30.2:** Advantages and disadvantages of DNA and RNA viruses

Advantages	Disadvantages
<b>DNA viruses</b>	
DNA is a stable genetic material	Viruses with linear genomes need strategies to complete replication of 5' ends (see Replication)
DNA is material of cellular genome therefore virus can exploit cellular replication machinery	(Eukaryotes only) Virus has to overcome the restriction of cellular replication to once per cell cycle
Virus can exploit DNA repair functions of cell	
<b>RNA viruses</b>	
No restriction to replication timing as cellular resources for RNA synthesis are constitutively available	RNA is not as stable as DNA, nor are there cellular functions for RNA proofreading or repair. Hence, RNA viruses are limited to a certain genome size by the natural frequency of disabling mutations. Many RNA viruses therefore exploit space conservation mechanisms: overlapping genes, differential splicing, translational frameshifting
RNA synthesis can initiate <i>de novo</i> , hence linear genomes do not require special replication strategies	Cells have no need for RNA replicases so RNA viruses must encode their own. (–) sense RNA viruses must carry the replicase in the virion because a (+) sense copy of the genome is required before further replicase can be synthesized (Eukaryotes only) Nearly all mRNAs are monocistronic because translation is initiated at the 5' cap. Polycistronic virus genomes have evolved strategies to circumvent this restriction, e.g. polyproteins, segmented genomes, expression of subgenomic RNAs and the use of internal ribosome entry sites

Eukaryote viruses are encumbered by the biology of the eukaryotic cell. DNA viruses must overcome the restriction of cellular replication to once per cell cycle, and do so either by encoding their own replication functions or by inducing the S-phase in the host cell. The simplest DNA viruses can only productively infect proliferating cells — either naturally proliferating cells or those superinfected with a virus that can induce proliferation. RNA viruses are not so restricted because they encode their own replicases. However, because of the monocistronic environment of the eukaryotic cell, they have evolved specialized strategies for the translation of their genes.

Both DNA and RNA viral genomes show great structural diversity (Table 30.3): they may be circular or linear, double- or single-stranded, or partially double-stranded. If single-stranded, they may be (+) **sense** (messenger sense) or (–) **sense** (antisense) or a combination (**ambisense**). Genomes may comprise a single chromosome, or they may be segmented. In the latter case, the segments may be packaged together or into separate particles (**multicomponent viruses**). Genomes may be unique, partially redundant or diploid. Many viruses, for example, possess redundant termini reflecting their replication and/or genome packaging strategy. The eukaryote RNA viruses may add 5' caps and 3' polyadenylate tails to their (+) strands, or one but not the other, or neither. Alternatively, the 5' terminus may be blocked by a different chemical group or by a protein.

**Baltimore classification of replication strategy.** The purpose of a virus infection is genome replication, and the production of many progeny virions. The universal **Baltimore classification** of viruses, devised in 1971 and based solely on replication strategy (and ignoring morphological properties and host range), divided viruses into five families based on the replication mechanisms known at the time. More categories are needed to include more recent discoveries (Table 30.4). The strategy for genome replication depends on the type of nucleic acid, its strandedness and its sense.



**Table 30.3:** Viral genome structure — a sample of viruses showing the type and configuration of their genomes and, for eukaryotic RNA viruses, the relation of the positive-sense strand-to-host mRNA

Virus family or genus	Genome structure	Other genome properties
<i>Adenoviridae</i>	dsDNA linear	
<i>Papovaviridae</i>	dsDNA circular	
<i>Caulimoviridae</i>	Partially dsDNA circular	
<i>Polydnaviridae</i>	dsDNA circular, supercoiled	Segmented (many)
<i>Parvoviridae</i>	ssDNA linear	Negative sense
<i>Inoviridae</i>	ssDNA circular	Positive sense
<i>Tobamovirus</i>	ssRNA linear	Positive sense Cap, no poly(A) tail
<i>Picornaviridae</i>	ssRNA linear	Positive sense No cap, poly(A) tail
<i>Geminiviridae</i>	ssDNA circular	Ambisense Some segmented and multicomponent
<i>Paramyxoviridae</i>	ssRNA linear	Negative sense
<i>Orthomyxoviridae</i>	ssRNA linear	Negative sense Segmented (8)
<i>Retroviridae</i>	ssRNA linear	Positive sense Cap and poly(A) tail Diploid
<i>Reoviridae</i>	dsRNA linear	Segmented (10–12)

**Table 30.4:** Extended Baltimore classification of viral replication strategy

Class	Strategy
I, dsDNA viruses	Semidiscontinuous replication (as for cellular genome) or by strand displacement (see Replication), e.g. SV40
IIa, ssDNA (+)	Synthesis of a double-stranded replicative form from which daughter ssDNA genomes can be produced by strand displacement, e.g. bacteriophage M13
IIb, ssDNA (–)	Synthesis of a double-stranded intermediate by hairpin priming, e.g. parvoviruses
III, dsRNA	These viruses carry RNA replicase in the virion allowing synthesis of daughter (+) RNA from the (–) RNA strand. The (+) RNA is packaged and complementary (–) RNA is synthesized to complete the genome, e.g. reoviruses
IV, ssRNA (+)	(+) RNA viruses have <b>infectious nucleic acid</b> because the replicase can be translated directly from the genome. Initial translation of replicase is followed by production of genome length (–) RNA which acts as the replicative intermediate for daughter (+) RNA synthesis, e.g. poliovirus
V, ssRNA (–)	(–) RNA viruses have noninfectious nucleic acid because replicase cannot be translated directly from the RNA genome. They must carry replicase into the cell as part of the capsid and generate (+) RNA which acts both as mRNA and as a replicative intermediate for daughter (–) RNA synthesis, e.g. influenza virus
VIa, RNA retroid viruses (retroviruses)	(+) sense RNA is reverse-transcribed into DNA. This is followed by second-strand cDNA synthesis and the integration of the dsDNA intermediate into the host genome. Transcription of the provirus yields full-length daughter (+) RNA for packaging, e.g. HIV
VIb, DNA retroid viruses	dsDNA is transcribed into full-length RNA replicative intermediate which is reverse-transcribed to yield a cDNA copy. Second-strand cDNA synthesis produces daughter dsDNA genomes for packaging. Does not necessitate integration into host genome, e.g. hepatitis B virus

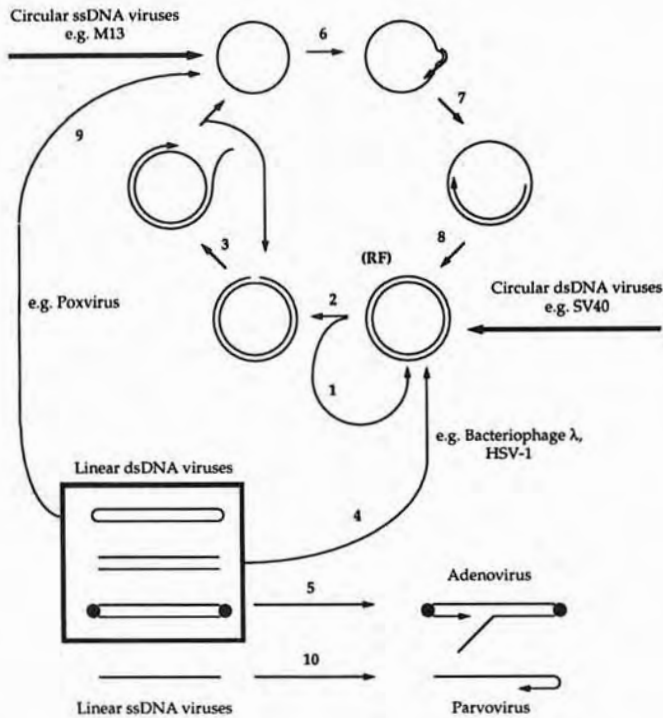
**Conventional replication of double-stranded DNA viruses.** Most dsDNA viruses replicate in a conventional manner using DNA polymerase and other replisome components which are encoded either by the virus itself or the host. Viruses with circular dsDNA genomes often replicate bidirectionally during early infection, but switch to *rolling circles* (q.v.) in late infection. For dsDNA bacteriophage such as  $\lambda$  and T7, late replication generates long concatemers which can be cleaved into linear, genome-sized fragments and packaged into phage heads (the linear genomes circularize after infection, using the cohesive ends produced by cleavage during packaging). Animal viruses with circular dsDNA genomes — the *Papovaviridae* (e.g. SV40), the *Baculoviridae* and the *Polydnaviridae* — replicate predominantly in a bidirectional manner so that circular genomes are packaged into capsids, although rolling circle-type structures have been observed in late SV40 infections. The *Polydnaviridae* are unique: their genome is represented by multiple, redundant circles of dsDNA which vary considerably in size; how these circles are produced has yet to be determined. Other viruses with circular dsDNA genomes use distinct strategies: e.g. bacteriophage T4 replicates as a concatemer and utilizes forked intermediates resulting from *invasive priming* (q.v.).

Viruses with linear dsDNA genomes may circularize upon infection (e.g. *Herpesviridae*) or may replicate using a displacement mechanism. As for chromosome replication, completing the 5' ends is difficult because terminal replication cannot be primed in the conventional manner (see Replication). Viruses do not employ the use of telomeres, however. They display a variety of alternative strategies, including the use of terminal proteins for priming (e.g. *Adenoviridae*), or covalently sealed hairpin ends (*Poxviridae*). The *Iridoviridae*, a family of viruses which infect lower vertebrates and invertebrates, use a unique mechanism which involves partial replication in the nucleus, followed by transport to the cytoplasm where replication continues to yield large concatemers which are processed to generate genome-length circularly permuted and terminally redundant molecules. The core strategies for DNA virus replication are shown in Figure 30.1.

The production of **concatemers**<sup>1</sup> for packaging is usual for DNA viruses which utilize rolling circle replication mechanisms. The concatemers contain many copies of the genome arranged head to tail and can be processed for packaging in two ways. In the first strategy, specific *cis*-acting sites are recognized by endonucleases which cleave the concatemer into genome-length fragments for insertion into the capsid. This is used, for example, by bacteriophage  $\lambda$  (the *cos* site), and results in each virus particle containing identical genomes which are nonredundant. The second strategy, the **headfull mechanism**, is used, for example, by bacteriophage P22 and by the eukaryotic frog virus 3. There may be a *cis*-acting site which allows the initiation of packaging, but once recognized, there are no specific sites which facilitate packaging of individual genomes. Rather, packaging continues until the capsid is full, then the DNA is cleaved. Generally, the capacity of the capsid is such that more than one genome of DNA can be packaged, and this results in **terminal redundancy** (the presence of repeated sequences at the genome termini) and **circular permutation** (where different genomes begin and end in different places, but contain the same loci). The two packaging strategies are compared in Figure 30.2 (also q.v. *generalized transduction*).

**DNA retroid viruses.** The *Hepadnaviridae* (e.g. hepatitis B virus) and the caulimoviruses (e.g. cauliflower mosaic virus) are dsDNA viruses which replicate by transcription and reverse transcription. In both cases, the double-stranded DNA genomes are discontinuous. The *Hepadnaviridae* possess circular but unclosed, partially double-stranded DNA genomes with one fixed-length, negative-sense **long strand**, L(-), and one variable length, positive-sense, **short strand**, S(+). Caulimovirus genomes are open circles with 3 or 4 single-stranded discontinuities, depending on the species. Replication begins with the transcription of a long positive-sense transcript, the **pregenome**, which

<sup>1</sup>Concatemers are head-to-tail arrays of covalently joined DNA sequences, and the process of generating them is **concatenation**. They are not the same as *catenanes*, which are interlinked DNA circles generated by *catenation* (q.v. *DNA topology*).

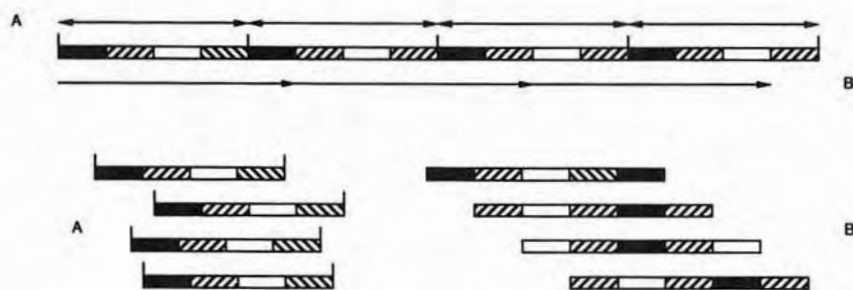


**Figure 30.1:** Replication strategies of DNA viruses (excluding retroviruses). Genomes may be double-stranded (ds) or single-stranded (ss), circular or linear. Viruses with a circular dsDNA genome often replicate in a conventional bidirectional manner (1), but a nick may be introduced at the origin of replication (2) and they may switch to rolling circle replication to generate concatemers (3). The circular dsDNA genome is also an intermediate for some viruses with linear dsDNA genomes (4), although others may use a displacement strategy, primed at terminal proteins (5). Viruses with circular ssDNA genomes also use a dsDNA intermediate, which is called a **replicative form**. A ssDNA circle is primed (6) and used to generate a replicative form by second-strand synthesis (7, 8). This may replicate in the manner of a conventional circular dsDNA genome to generate more replicative forms (1) or may produce single-stranded daughter genomes or concatemers by rolling circle replication (2, 3). Some linear dsDNA viruses have covalently sealed ends and may use a circular ssDNA intermediate (9). Finally, linear ssDNA viruses may replicate by hairpin priming (10). Examples of viruses using each of these strategies are shown.

is often longer than the genome and can be used both as mRNA and as a template to generate new (–) sense DNA, using the viral-encoded *reverse transcriptase* (q.v.). In hepadnaviruses, short-strand synthesis begins after degradation of the pregenome by reverse transcriptase-associated RNaseH activity, but its synthesis may be interrupted by host-cell lysis, the consequence of which is its variable length.

**Transposing bacteriophage.** Bacteriophage Mu, like a *transposon* (q.v.), replicates only within the genome of its host, and does so by repeated *replicative transposition* (q.v.); this is **transpositional replication**. Unlike other transposons, however, Mu encodes functions which allow it to excise destructively from the genome and package into a capsid, and infect other bacteria in the manner of a phage (see Mobile Genetic Elements for discussion of transposons).

**Replication of single-stranded DNA viruses.** Viruses with single-stranded genomes replicate via a double-stranded intermediate whose opposite sense, (antigenome) strand, acts as a template for the synthesis of new DNA genomes (Figure 30.1). Bacteriophage falling into this category (the *Inoviridae*, which include the filamentous phages M13 and f1, and the *Microviridae*) have positive-sense single-



**Figure 30.2:** Strategies for packaging concatemeric DNA. A concatemer is shown at the top of the figure divided into genome lengths (bidirectional arrows), with each genome arbitrarily divided into four sequence blocks (represented by different shading). The processing of the concatemer by cleaving at specific sites (A) generates nonredundant and nonpermuted genomes for packaging. The processing of concatemers by cleaving at nonspecific sites (B) but producing fragments of the same length which are greater than the size of the nonredundant genome generates cyclically permuted genomes with terminal redundancies.

stranded DNA genomes, and the antigenome intermediate must also act as a template for mRNA synthesis. The geminiviruses are a family of plant viruses with similar genome structure and replication mechanism, although their genomes are ambisense, and in some cases the viruses are segmented and multicomponent. The genome is circular and is termed the **viral** or **v-strand**, whereas the antigenome is the **complementary** or **c-strand**. Synthesis of the c-strand forms a double-stranded circle, the **replicative form (RF)**, which can replicate by a conventional bidirectional mechanism to generate copies of itself. Later, replication switches to an asymmetrical rolling-circle mechanism producing concatemeric single-stranded genomes for packaging (q.v. *DNA sequencing*).

Some animal viruses may replicate in a similar manner to the M13-like phage, but the best characterized ssDNA animal virus family, the *Parvoviridae*, uses a distinct strategy. The parvovirus genome is linear. Some species (e.g. mice minute virus) package specifically (–) sense DNA; others package both types of strand into individual particles with (+) strands representing 1–50% of viral progeny, depending on the species. The genome contains self-complementary repeats which prime opposite-sense strand synthesis by hairpin formation. The hairpins are cleaved by the viral-encoded protein, Rep, producing genome-length molecules for packaging. The hairpins adopt a specific T-shaped configuration so that a conventional hairpin is retained on the progeny genome when the parental genome is cleaved off. This allows continuous priming of new progeny strands and the formation of large concatemers.

**Replication of conventional single-stranded RNA viruses.** RNA viruses are generally single-stranded and are usually fully (+) sense or (–) sense (c.f. ssDNA genomes which may also be ambisense). (+) sense RNA genomes can, in principle, act as mRNA as soon as they have entered the cell, whereas (–) sense genomes cannot. Because cells generally do not encode RNA replicase functions, naked (+) RNA viral genomes transfected into cells are infectious because viral RNA replicase can be produced, whereas (–) RNA genomes are latent because a complementary copy of the genome must be synthesized before the replicase can be translated. (–) RNA viruses therefore carry replicase into the cell in their capsid, allowing synthesis of (+) RNA without prior viral gene expression. Most plant viruses are (+) RNA viruses.

(+) RNA viruses generally possess linear genomes. Those which infect eukaryotes (e.g. the *Picornaviridae*, coronavirus, tobacco mosaic virus) may resemble host mRNA in the possession of 5' caps and/or 3' polyadenylate tails, or may lack either or both structures. The 5' cap is essential for the translation and stability of cellular mRNA (see RNA Processing), so viruses which lack this



structure usually block the 5' end of their genome in an alternative manner, either with a different chemical group or with a viral encoded protein. The picornaviruses lack a cap and exploit the cells dependence on capped mRNA to block host protein synthesis by inhibiting the *cap binding protein* (q.v.) essential for ribosome loading. The translation of their own genes is mediated by a unique *internal ribosome entry site* (see below). Most (+) RNA viruses replicate using a (–) sense RNA replicative intermediate, which acts as a template both for genome duplication and mRNA synthesis. However, the retroviruses replicate using a unique strategy involving a dsDNA intermediate (see below). Eukaryotic (–) RNA virus tend to neither cap nor polyadenylate their genomes, and act as templates for both mRNA synthesis and production of the genome-length (+) RNA replicative intermediate, which itself is a template for further daughter (–) RNA genome strand synthesis.

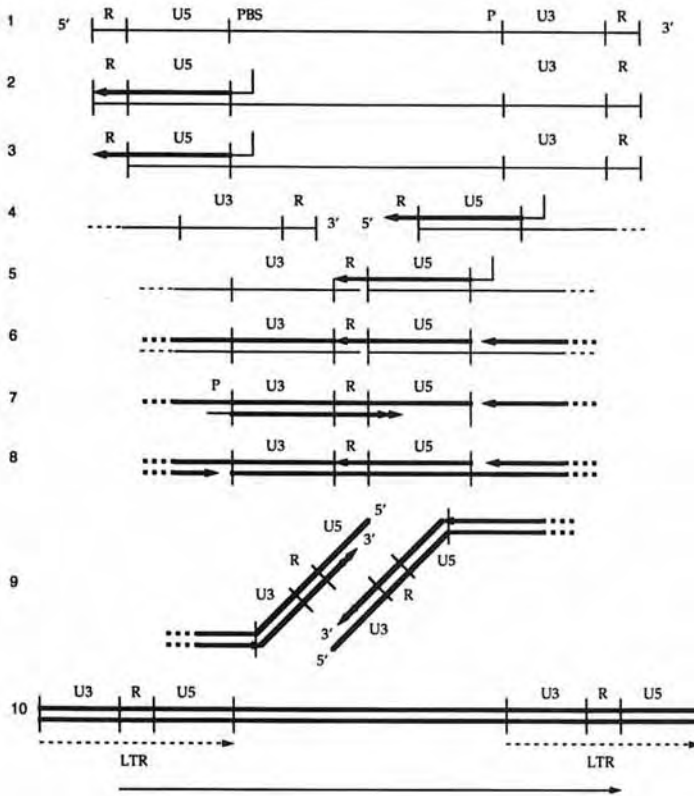
**Replication of double-stranded RNA viruses.** Several virus families possess dsRNA genomes, e.g. the *Partitiviridae*, which infect fungi and plants, the *Totiviridae*, which infect unicellular eukaryotes, bacteriophage of the *Cystoviridae* family, and the *Reoviridae*, which have a broad host range amongst multicellular organisms and are the best characterized. All families have linear genomes and all, except the *Hypoviridae*, are segmented — two or three segments for most of the families but 8–10 segments in the *Reoviridae*. The (+) strand of eukaryotic dsRNA viruses tends to be capped but not polyadenylated, whereas the (–) strand possesses neither modification. The *Reoviridae* use the (–) strand to generate (+) genome strands and mRNA. They therefore replicate in the manner of a typical (–) strand RNA virus; similarly, they carry replicase in the capsid. The (+) strands are packaged into partially formed capsids, wherein (–) strand synthesis occurs.

**Replication cycle of the retroviruses.** The replication strategy of the RNA retroviral viruses, the *Retroviridae*, is a unique and complex process involving sequential transcription, reverse transcription, RNA degradation, DNA synthesis and integration into the host genome (Figure 30.3). The central region of the virus genome is flanked by direct long terminal repeats (LTRs) which carry transcriptional control elements responsible for production of genome length RNA from the integrated provirus. The terminal redundancy of the retroviruses results from the remarkable replication strategy: a copy-choice replication mechanism involving two template switches. A family of eukaryotic transposable elements, the *retrotransposons* (q.v.), mobilize in a manner which is very similar to that of the retroviruses, and their genomes are similarly organized, usually differing in the absence of a functional *env* gene which encodes capsid proteins. The retrotransposons and retroviruses use the same mechanism for integration (see Mobile Genetic Elements). The human immunodeficiency virus (HIV) is a retrovirus and is responsible for the acquired immune deficiency syndrome (AIDS) (Box 30.2).

### 30.3 Strategies for viral gene expression

**Temporal control — regulatory cascades in viral infections.** Viral infection cycles are often divided into distinct phases, reflecting the expression of specific subsets of genes. **Temporal control** of viral gene expression is analogous to the control of gene expression during development: both involve cascades of transcriptional regulators, or regulators of other levels of gene expression (see Gene Expression and Regulation). In DNA viruses, temporal control is often achieved by expressing, as an early gene product, an essential regulator of later genes. The later-acting genes may encode further regulators, and by placing genetic regulators into such a dependent series, as many temporal steps can be added to an infection cycle as necessary. Many viruses divide their infection cycle into two or three such steps, and the lag between the individual phases reflects the time taken to synthesize and/or activate the appropriate regulators.

A diverse range of mechanisms are used by individual viruses to achieve these aims. As examples, the regulatory cascades of bacteriophage  $\lambda$ , human immunodeficiency virus (HIV) and herpes simplex virus (HSV) are discussed in Boxes 30.1, 30.2 and 30.3, respectively. The progression of the



**Figure 30.3:** Replication strategy of the retroviruses. RNA is shown as thin lines and DNA as thick lines. (1) The viral (+) RNA has a redundant terminal region (R) and unique 5' and 3' regions (U5 and U3), and is capped at the 5' end and polyadenylated at the 3' end (not shown). (2) (-) strand DNA synthesis is primed by a host-encoded tRNA which binds to a primer binding site (PBS) downstream from U5; (-) strand DNA synthesis runs to the 5' end of the RNA template. (3) The 5' redundant region of the RNA template is degraded by RNaseH. (4) The unpaired DNA at this region can then pair with a 3' redundant region from the same genome, or another genome. (5) The DNA hybridizes with the RNA allowing continued (-) strand DNA synthesis on a new RNA template. (6) The extension of the (-) DNA strand on the new template constitutes the 'first jump' and continues until it displaces the tRNA primer. (7) As the (-) DNA strand is extended, the RNA template is degraded by RNase H. Some fragments of RNA remain to prime (+) strand DNA synthesis, often at a polypurine tract (P) upstream of U3. (+) strand DNA synthesis is initiated and (8) switches to the initial template constituting the 'second jump'. (9) Completion of (+) and (-) strand DNA synthesis by displacement duplicates the terminal regions of the genome. (10) This generates the characteristic long-terminal repeats of retroviral genomes (broken arrows). The double-stranded cDNA copy of the retroviral RNA can integrate into the host genome at this point (q.v. *retrotransposition*). Transcription, initiated at a promoter within U3 and terminating at U5, produces the viral RNA or packaging as shown by the long arrow; it is equivalent to the RNA strand shown in (1). Transcription from the right-hand LTR can also occur and may activate adjacent host genes (q.v. *slow transforming retrovirus*).

bacteriophage  $\lambda$  lytic cycle involves three phases of gene expression regulated by antitermination, whereas the choice between lysis and lysogeny depends on the balance between two transcriptional regulators. HSV lytic infection is also controlled by transcriptional regulation, but the HIV infection cycle is controlled predominantly by regulating RNA splicing and export from the nucleus. Mechanisms used by other viruses include the synthesis of cascades of  $\sigma$ -factors (bacteriophage T7), and regulation of translation (adenovirus). Temporal regulation may also be exploited by RNA viruses, and is usually controlled by genome structure and the presence of replicase. Togaviruses, for example, have a bicistronic genome but only the 5' gene can be translated in early infection because

**Table 30.5:** Strategies for protein synthesis in polycistronic eukaryotic RNA viruses

Strategy	Mechanism and examples
Segmented genomes	Segmented RNA genomes allow individual genes to be placed on separate segments. Influenza virus has eight genes on eight negative-sense ssRNA segments and encodes 10 products (two by alternative splicing). This strategy also facilitates genetic mixing by reassortment
Polyproteins	The entire genome is expressed as a single polypeptide which is then cleaved to generate individual gene products. This strategy allows all gene products to be encoded by a single transcript and is exploited by the <i>Picornaviridae</i>
IRES motifs	Some picornaviral genomes contain the <b>IRES (internal ribosome entry site)</b> motif which allows ribosomes to bind in the absence of the <i>trans</i> -acting components facilitating cap recognition. The IRES adopts a secondary structure which recruits translation initiation factors in the absence of the cap binding protein. In recombinant constructs, the IRES can be placed upstream of individual reading frames for expression of polycistronic mRNA in eukaryotes
Discrete mRNAs	The <i>Rhabdoviridae</i> transcribe individual mRNAs corresponding to each viral gene from their polycistronic negative-sense genome. The transcriptase transcribes each gene sequentially, but with increasing likelihood of dissociation. This novel form of gene regulation during transcriptional elongation allows the five transcripts to be produced in decreasing levels
Nested subgenomic RNAs	The togavirus genome is bicistronic, and in the early phase of infection only the 5' ORF, encoding replicase, can be translated. Once replicase has been synthesized, it can initiate transcription of a subgenomic mRNA containing the downstream ORF only, which encodes capsid proteins as a polyprotein. Coronaviruses use a similar strategy involving seven subgenomic mRNAs corresponding to the seven genes. Each has a common leader sequence which may be used as a primer

the viral replicase (encoded by the 5' gene) is required to generate a smaller RNA from which the downstream gene (encoding capsid proteins) is translated. For ambisense viruses, the early genes can be translated from RNA produced from the genomic strand but late genes can be translated only from RNA produced from the antigenomic strand; hence replication must precede late gene expression.

**Protein synthesis by eukaryotic RNA viruses.** A major strategic problem for eukaryotic RNA viruses is how to translate their open reading frames in the monocistronic environment of the eukaryotic cell. Cellular mRNAs are almost exclusively monocistronic (*see* Transcription) because ribosome loading depends upon the 5' modified cap. This contrasts with the situation in bacteria, where ribosomes can enter mRNA at internal *Shine–Dalgarno* sequences (q.v.). The strategies employed by RNA viruses to circumvent this impediment are ingenious (*Table 30.5*). One of the most remarkable is demonstrated by the picornaviruses, which possess **internal ribosome entry sites (IRES motifs)** analogous to bacterial *Shine–Dalgarno* sequences. These have been extensively exploited in the expression of cloned genes as they allow coordinated expression under the same promoter, which is useful for targeted gene expression in transgenic organisms and has many applications in the use of reporter constructs (q.v. *transgenic mouse*, *gene knockout*, *reporter gene*).

### 30.4 Subviral agents

**Classification of subviral agents.** Subviral agents are noncellular organisms containing a genome but lacking one of the essential features of a virus (*Table 30.6*). The common property of viruses and

**Table 30.6:** Viruses and different classes of subviral agents and their discriminating characteristics

Agent	Distinguishing properties
<b>Virus</b>	DNA or RNA genome Genome packaged in protein coat, which mediates infection Genome codes for coat proteins and other functions required for infection cycle
<b>Viroid</b>	RNA genome Genome is naked (no protein coat), infection is mediated mechanically Genome appears not to encode proteins
<b>Satellite virus</b>	A virus whose replication is dependent on coinfection by a second 'helper' virus of a different type, which supplies missing functions in <i>trans</i> The satellite virus and helper virus are separate entities; they encode their own capsids and infect the cell independently
<b>Satellite nucleic acid</b>	(a) A DNA or RNA genome with virus-like coding properties but lacking the ability to encode its own capsid and therefore needing the late functions of a helper virus (b) A viroid-like RNA found in plants, packaged as a stowaway in the capsid of its helper virus and which depends on that helper virus for infection and replication functions. The helper virus is not dependent on the satellite
<b>Virusoid</b>	RNA found in plants, structurally and functionally similar to a viroid, and packaged into a viral capsid, but actually part of the virus genome and therefore not expendable like satellite RNA
<b>Virino</b>	Theoretical infectious particle which has a noncoding nucleic acid genome packaged in a protein coat derived from the host
<b>Prion</b>	Proteinaceous infectious particle resistant to treatments that destroy nucleic acids. Commonly taken to mean an infectious agent composed entirely of a host-encoded protein which can change to a pathogenic and 'self-replicating' conformation

subviral agents, which discriminates them from other self-replicating genetic elements such as plasmids and transposable elements, and from organelles such as mitochondria and chloroplasts, is their existence in a stable *extracellular* form, an existence which allows them to be infectious in their own right.

**Viroids.** Viroids are small single-stranded circular RNA agents which infect plants. They differ from RNA viruses in three major aspects: their size (they are one-quarter of the size of the smallest RNA virus, i.e. 250–400 bases); the genome does not encode any proteins and they are not encapsidated. Viroid infection is mediated mechanically.

About 30 distinct types of viroid have been identified: some generate disease symptoms and some are cryptic. Disease symptoms range from mild to lethal, and may result from interference with host RNA processing. All viroids undergo extensive intramolecular base-pairing to form rigid rod-like structures with conserved secondary structures, and two viroid 'families' have been recognized on this basis. In mixed infections, frequent recombination occurs between homologous domains of individual viroids to generate novel sequence variants.

Viroid replication requires host encoded RNA polymerase, which may have weak RNA replicase activity. Viroids are thought to replicate via rolling circles, producing concatemeric replicative intermediates which act as templates for genome synthesis. How single-stranded circular genomes are released from the double-stranded concatemers is poorly understood, although the avocado sun-blotch viroid can catalyze self-cleavage, and several of the structural features of viroids are also found in group I introns, suggesting autocatalytic monomerization in some species and perhaps host-dependent cleavage in others (q.v. *ribozymes*, *self-splicing introns*).



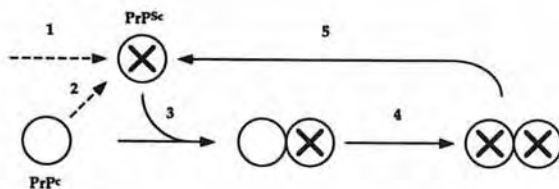
**Satellites.** Satellites<sup>1</sup> are subviral agents which depend on a **helper virus** for successful infection. There are two types of satellite: a **satellite virus** which encodes its own coat protein (and thus mediates infection and often other stages of its infection cycle without help), and **satellite nucleic acids**, which do not. Satellite viruses include the eukaryotic *Dependovirus* genus (e.g. adeno-associated viruses) and several ssRNA plant viruses with a diverse range of helper viruses. Satellite nucleic acids can be divided into two classes. Some behave like viruses, but require help with late functions and hence utilize the capsid proteins of the helper virus (e.g. bacteriophage P4, which uses the capsid proteins of bacteriophage P2, also q.v. *killer factors*). This type of satellite can be thought of as a plasmid which can be promoted to a virus by subverting the late functions of its helper virus to form infectious particles — some vectors used in recombinant DNA research work on similar principles (q.v. *phagemid*). Other satellites, found only in plants, resemble viroids in structure and likewise have no coding function, but hitch-hike within the capsid of a helper virus to mediate infection. These satellites depend on the helper virus for infection and replication, but the converse is not true; viral infections can occur both with and without the satellite (the presence of the satellite can influence the symptoms produced in the infected plant and reduce the yield of the helper virus, probably by competing for components and resources during replication; for example, cucumber mosaic virus infections of tomato plants are relatively mild, but in the presence of CMV-associated RNA, a satellite RNA, a lethal necrosis results. In other plants, however, the presence of the satellite RNA *reduces* the severity of the symptoms). In some viruses, viroid-like elements represent permanent components of the genome (i.e. the helper virus and satellite are mutually dependent for replication and infection). For example, the velvet tobacco mottle virus has a bipartite RNA genome consisting of RNA1 and RNA2, the latter resembling a viroid but actually being essential for viral infection and replication; in this situation the RNA2 is known as a **virusoid**. Conversely, in the closely related lucerne transient streak virus, RNA2 is a nonessential satellite.

**Agents of transmissible spongiform encephalopathies.** Transmissible spongiform encephalopathies (TSEs) are degenerative diseases of the nervous system which occur naturally in many mammals including humans (Table 30.7). The nature of the TSE agent is not fully understood, but many lines of evidence suggest that it is caused by a proteinaceous infectious particle, or **prion**. The presence of nucleic acid in prions has not been demonstrated, and prion isolates remain infectious following all forms of treatment which destroy nucleic acids, e.g. UV-irradiation, incubation with nucleases. However, prions exist in distinct strains, which suggests a genetic basis.

**Table 30.7:** Transmissible spongiform encephalopathies (TSEs) in different mammals

Species	TSE
Sheep and goats	Scrapie
Cattle	Bovine spongiform encephalopathy (BSE), 'Mad Cow Disease'
Cats	Feline spongiform encephalopathy (FSE)
Mink	Transmissible mink encephalopathy (TME)
Cervids (deer, elk)	Chronic wasting disease (CWD)
Humans	Creutzfeldt–Jakob disease (CJD) Gerstmann–Straussler–Scheinker syndrome (GSS) Kuru Fatal familial insomnia (FFI)

<sup>1</sup>Satellite RNAs, such as RNA2 in the lucerne transient steak virus, have nothing whatsoever to do with *satellite DNA* (q.v.), a highly repetitive form of DNA found in eukaryote chromosomes. Nor should viral satellites be confused with chromosome *satellite regions* (q.v.), which are found distal to pale-staining *nucleolar organizer regions* (q.v.).



**Figure 30.4:** The refolding model for prion replication.  $\text{PrP}^{\text{C}}$  is the normal protein and  $\text{PrP}^{\text{Sc}}$  is the pathological form. Initially,  $\text{PrP}^{\text{Sc}}$  may be supplied exogenously, e.g. by horizontal transmission through the food chain, or by vertical transmission during pregnancy (1). Alternatively,  $\text{PrP}^{\text{Sc}}$  may arise spontaneously from  $\text{PrP}^{\text{C}}$  if favored by a particular germline or somatic mutation in the  $\text{PrP}^{\text{C}}$ -encoding gene, or by other environmental causes. The disease is propagated by a chain reaction in which  $\text{PrP}^{\text{Sc}}$  molecules coming into contact with  $\text{PrP}^{\text{C}}$  (3) cause  $\text{PrP}^{\text{C}}$  to be converted into  $\text{PrP}^{\text{Sc}}$  (4), providing more  $\text{PrP}^{\text{Sc}}$  to feedback and interact with more  $\text{PrP}^{\text{C}}$  (5). Initially, the build-up of prions would be slow, but would accelerate and eventually become exponential.

It is thought that a host-encoded protein called **PrP<sup>C</sup> (cellular prion-related protein)** is an important, if not the only, component of the infectious particle. The **protein-only hypothesis** maintains that the prion contains no nucleic acid, whereas the **unconventional virus hypothesis** maintains that the agent, a **virino**, is an unusual virus which requires  $\text{PrP}^{\text{C}}$  for infection. A unified model has also been presented, involving aspects of both hypotheses.  $\text{PrP}^{\text{C}}$  is a neuron-specific membrane-associated glycoprotein present in all mammals. Normal  $\text{PrP}^{\text{C}}$  is degraded by protease treatment, but in cases of TSE, fibrils composed of highly protease-resistant aggregates of PrP (**PrP amyloids**) appear in neurons. This protease-resistant form is called **PrP<sup>Sc</sup> (scrapie prion-related protein)** or **PrP\***, and when isolated from diseased cells it is enriched for the TSE agent (although the ratio of infectious agent to protease-resistant PrP is only 1 in  $10^5$ ).

$\text{PrP}^{\text{C}}$  and  $\text{PrP}^{\text{Sc}}$  appear identical in primary structure, suggesting that the change from normal  $\text{PrP}^{\text{C}}$  to pathogenic and infectious  $\text{PrP}^{\text{Sc}}$  results from a change in conformation.  $\text{PrP}^{\text{C}}$  has a structured C-terminal domain but an N-terminal region of unstructured coil; this region adopts a predominantly  $\beta$ -sheet organization in  $\text{PrP}^{\text{Sc}}$ . It is not known how the conformational change occurs, but a model for the 'replication' of the agent involves interaction between the misfolded pathological form and its normal cellular counterpart resulting in induced refolding so that the  $\text{PrP}^{\text{C}}$  is converted to  $\text{PrP}^{\text{Sc}}$  (Figure 30.4). This is supported by the observation that prion diseases take on the characteristics of the  $\text{PrP}^{\text{Sc}}$  encoded by the host rather than the infectious agent itself, i.e. the *endogenous* PrP is being converted into a pathogenic conformer. The existence of many different strains of prion diseases is consistent with the unconventional virus hypothesis, but can be explained in terms of the protein-only hypothesis if each strain represented a different conformational form of PrP, and could autocatalytically convert normal  $\text{PrP}^{\text{C}}$  isomers into copies of itself. There are often considerable delays in cross-species infections, suggesting that the conversion of host  $\text{PrP}^{\text{C}}$  by a 'foreign' prion is initially a slow process.

Genetic studies show that about 10% of cases of Creutzfeldt–Jakob Disease, CJD (and most cases of the related diseases Gerstmann–Straussler–Scheinker syndrome and fatal familial insomnia), can be traced to germline mutations in the *PRNP* gene, which encodes  $\text{PrP}^{\text{C}}$ . Mice with mutations in the homologous *Prn-p* locus have reduced incubation times for scrapie, suggesting that mutation can make the normal prion protein more susceptible to conformational conversion by  $\text{PrP}^{\text{Sc}}$  from a different source. *Transgenic mice* (q.v.) carrying the hamster PrP-encoding gene are more susceptible to the hamster scrapie agent than wild-type mice. Perhaps most importantly, *Prn-p* gene *knockout mice* (q.v.) are resistant to scrapie infection because the TSE agent does not replicate. Additionally, these mutant mice show only a mild phenotype (altered sleep patterns, impaired long-term potentiation), and thus the endogenous function of the  $\text{PrP}^{\text{C}}$  molecule remains unknown. It is possible that a mutation in the PrP-encoding gene may predispose  $\text{PrP}^{\text{C}}$  to undergo a spontaneous conformational

change which can initiate a chain reaction of propagation in infected cells. In spontaneous cases of CJD, a somatic mutation could initiate the infection, and spreading to surrounding cells could be mediated by protein-protein contact across membranes. There is much evidence that prion-related agents can be transmitted horizontally through the food chain to humans, either by consumption of infected human brain tissue (kuru) or of beef infected with bovine spongiform encephalopathy (new variant CJD). There is also increasing evidence that somatic prion diseases can be transmitted vertically from mother to offspring.

### Box 30.1: Bacteriophage $\lambda$

**Early events: expression of immediate early and delayed early genes.** Bacteriophage  $\lambda$  is a temperate phage of *E. coli*. When  $\lambda$  infects the cell, it has the choice to replicate and eventually lyse the cell (**lytic cycle**), or to integrate into the genome and become latent (**lysogeny**). Regardless of whether  $\lambda$  follows the pathway to lysis or lysogeny, the early events of infection are the same. After entry and genome circularization, transcription is initiated at promoters  $p_L$  and  $p_R$  by host RNA polymerase. These promoters lie either side of the *cI* gene and transcription proceeds outwards (i.e. away from *cI*) terminating at  $\rho$ -dependent transcriptional terminator sites (q.v.)  $t_L$  and  $t_R$  just beyond the *N* gene on the left and the *cro* gene on the right (see figure below). *N* and *cro* are thus known as **immediate early genes**: they are expressed immediately following infection. Occasionally, right ward transcription proceeds through  $t_R$  to  $t_{R2}$ , allowing transcription of *cII*, which encodes a regulator protein, and genes *O* and *P*, which initiate  $\lambda$  replication.

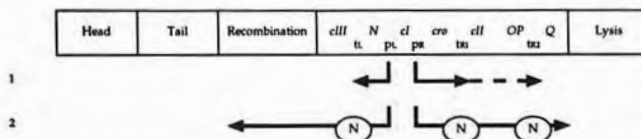
*N* is an **antiterminator protein** (q.v.) which allows transcription from the  $p_L$  and  $p_R$  promoters to proceed beyond the terminator sites. Therefore, once *N* has been synthesized, transcription left ward from  $p_L$  allows the expression not only of *N*, but also of *cIII* and the block of genes involved in recombination functions, whereas transcription right ward from  $p_R$  allows the expression not only of *cro*, *cII*, *O* and *P*, but also *Q*, which regulates late gene expression.

*cII*, *cIII*, *O*, *P*, *Q* and the recombination genes are thus known as the **delayed early genes**.

**The lytic cycle.** The lytic cycle is characterized by phage replication and the expression of the **late genes**. These encode phage particle components and proteins required for phage assembly, chromosome packaging and host-cell lysis (see figure below).

Replication functions are encoded by the delayed early genes *O* and *P*, although a number of host proteins are also required. The late genes are encoded in a single operon whose transcription initiates at promoter  $p_{R'}$ . Transcription runs off the right-hand edge of the linear  $\lambda$  map, through the *cos* site into the unassigned reading frames separating the tail genes from the *att* site. In the early phase of infection, right ward transcription from promoter  $p_{R'}$  may be initiated, but proceeds for only ~100 nucleotides before reaching a termination site. The product of the delayed early gene *Q* is an antitermination protein which allows readthrough of this terminator into the late gene operon.

Successful entry into the lytic cycle depends on the expression of *cro*, which encodes a transcriptional repressor that binds to the operator sequences  $o_L$  and  $o_R$ , overlapping the early promoters. *Cro* activity prevents expression of the regulatory proteins *C1* and *CII*, whose function is to establish lysogeny, as discussed below.



Early gene expression following infection by bacteriophage  $\lambda$ . (1) Initially, expression left ward from promoter  $p_L$  terminates at  $t_L$  and expression right wards from  $p_R$  terminates at  $t_{R1}$  (occasionally  $t_{R2}$ ), allowing expression of the immediate early genes *N* and *cro*. (2) *N* is an antiterminator protein which allows readthrough of the terminator sites, and hence expression of the delayed early genes, including the regulator genes *cII* and *Q*. In the  $\lambda$  map, genes are shown on the upper row and regulatory elements on the lower row. Transcription is shown as thick arrows, regulatory factors as circles.

DNA replication, which initially proceeds bidirectionally, switches to rolling-circle replication later in the lytic phase (see Replication). The molecular basis of this switch is not understood, but both types of replication initiate at the same origin. Rolling-circle replication produces long concatemers of the  $\lambda$  genome which are cleaved at the *cos* sites by terminase (an enzyme comprising the Nu1 and A proteins), generating the 12 nucleotide 5' overhangs characteristic of the linear genome. The left cohesive end is packaged first, and headstuffing continues until another *cos* site is encountered. Lysis releases about 100 progeny phage from the cell as well as unpackaged genomes and phage components.

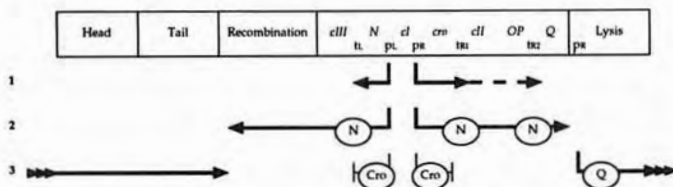
**Lysogeny and immunity to superinfection.** Lysogeny is characterized by the repression transcription and the integration of the  $\lambda$  genome into the bacterial chromosome. This is controlled by the combined activities of two transcriptional regulators, CI and CII (see figure below).

CI is the  $\lambda$  **repressor** which maintains the phage in a latent (transcriptionally inactive) state. CI binds to the operator regions  $o_L$  and  $o_R$  adjacent to the early promoters  $p_L$  and  $p_R$ , and thus prevents outward transcription. This blocks the expression of all genes, notably *N* and hence *Q*, ensuring that late gene transcription is repressed. An additional consequence of CI binding to  $o_R$  is that it activates left ward transcription from the adjacent promoter  $p_M$ . This facilitates transcription of the *cI* gene itself. Thus CI is able to maintain its own synthesis in a positive feedback loop called the **maintenance circuit**, which incidentally prevents the expression of *cro* by *countertranscription* (q.v.) through the *cro* gene. The maintenance circuit explains **immunity to superinfection** (whereby bacteria lysogenic for  $\lambda$  cannot undergo lytic infection by  $\lambda$ ): the production of surplus CI ensures that any incoming phage genomes are repressed as soon as they enter the cell. CI therefore acts as both a transcriptional acti-

vator and a transcriptional repressor, the former by recruiting RNA polymerase to the promoter, the latter by steric hindrance.

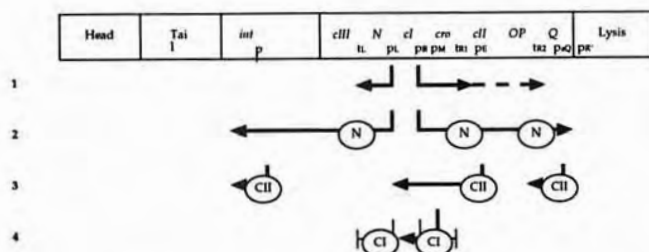
Although the positive feedback loop demonstrates how CI expression is maintained, it does not explain how it is initiated. This requires the transcriptional regulator CII, which is a the product of a delayed early gene. CII binds to three promoters:  $p_E$ ,  $p_{aQ}$  and  $p_I$ . The  $p_E$  promoter allows left ward transcription of *cI* and establishes synthesis of CI while repressing *cro* by countertranscription. Once CI is synthesized it blocks transcription from  $p_R$  and thus shuts down synthesis of CII, but by this time the CI maintenance circuit is running. The  $p_E$  promoter is stronger than  $p_M$ , and provides a burst of repressor synthesis to drive the phage into lysogeny, whereas  $p_M$  provides a low level of constitutive expression to maintain it. Promoter  $p_E$  has a poor consensus sequence, however, which explains the requirement for CII. Transcription from  $p_{aQ}$  produces an *antisense RNA* (q.v.) from the region of the *Q* gene. This interferes with the translation of any *Q* mRNA which has already been synthesized and provides a second mechanism to block expression of the late genes. Finally, transcription from  $p_I$ , which is located within the *xis* gene, facilitates expression of *int* which encodes the integrase enzyme required to insert the  $\lambda$  genome into the host chromosome (for the mechanism of  $\lambda$  integration q.v. *site-specific recombination*).

**The choice between lysis and lysogeny.** Upon infection,  $\lambda$  is committed to neither lysis nor lysogeny. Lysogeny occurs when the *cI* gene is expressed. CII blocks late gene expression by antisense repression of *Q*, facilitates synthesis of the integrase protein allowing prophage insertion, and establishes *cI* expression. Once CI is synthesized, it regulates its own synthesis through the maintenance circuit and, by binding to  $o_L$  and  $o_R$ , shuts down the expression of all other phage genes. Lysis occurs when *cro* is expressed. Cro blocks *cI* main-



The genetic cascade of the lytic cycle. The early events shown in (1) and (2) are common to both lytic and lysogenic pathways. These early events facilitate the synthesis of the antiterminator protein *Q* and the transcriptional repressor *Cro*. (3) *Q* allows transcription from  $p_R$  to proceed through the nearby terminator site and therefore facilitates expression of the late genes. Note that the late operon runs off the right-hand side of the linear map and through the head and tail genes on the left side because  $\lambda$  is circular at this stage. Meanwhile *Cro* binds to the operators adjacent to promoters  $p_L$  and  $p_R$  and prevents further expression of the early genes, which would establish lysogeny.





The genetic cascade which establishes lysogeny. The early events shown in (1) and (2) are common to both lytic and lysogenic pathways. These early events facilitate the synthesis of the transcriptional regulator CII. (3) CII binds to three promoters. At  $p_L$  it facilitates expression of the *int* gene which encodes integrase and allows integration of  $\lambda$  into the host chromosome. At  $p_E$  it establishes the expression of *cI* and represses the expression of *cro* by countertranscription. At  $p_{AQ}$  it transcribes an antisense RNA which represses Q protein synthesis, thus blocking the expression of late genes. (4) CI binds to operators adjacent to promoters  $p_L$  and  $p_R$  and prevents further expression of the early genes, but facilitates maintenance of *cI* gene expression itself from promoter  $p_M$ , thus establishing an autoregulatory loop which maintains lysogeny.

tenance expression by binding to  $O_L$  and  $O_R$ , and prevents *cI* establishment transcription, integrase synthesis and antisense repression of Q by reducing transcription from  $p_L$  and  $p_R$  and thus abolishing the expression of *cII*.

The choice between lysis and lysogeny thus breaks down to whether CI or Cro binds to the  $O_L$  and  $O_R$  operator sites, each of which consists of three binding motifs. The different properties of CI and Cro, and the distinct ways in which they bind to the operator motifs, dictates how they influence transcription and hence controls ensuing events. The predominance of each regulator in turn reflects whether CII is present or absent: in the presence of CII, *cI* expression is favored and *cro* is repressed, whereas in its absence, *cI* expression is never established and *cro* is expressed. *cII* is an early gene and is expressed soon after infection. However, the product is very unstable and is degraded rapidly by cellular proteases. A further protein, CIII, protects CII from these affects and makes it more stable, as does the host-encoded regulator CAT-cAMP (q.v. *catabolite repression*). The decision between lysis and lysogeny thus reflects the abundance of CIII and CAT-cAMP in the cell. These factors control the stability of CII and thus determine whether CI is synthesized or not. Lysogeny is favored when either CIII or CAT-cAMP is present at a high concentration (high multiplicity of infection or poor growth conditions, i.e. glucose starvation), whereas lysis is favored under most other conditions. The regulatory steps in the control of the  $\lambda$  infection cycle are summarized in the figure below.

**Induction.** Induction describes the excision of the  $\lambda$  prophage from the host chromosome and entry into the lytic cycle. Induction occurs spontaneously

at a low frequency, probably reflecting rare events where the CI repressor disengages from  $O_R$ , allowing synthesis of Cro. Induction can be stimulated by treating cells with agents that damage DNA, because this involves activation of the SOS response (q.v.). The SOS genes are usually silenced by the transcriptional repressor LexA. Damage to DNA activates the *RecA* protein (q.v.), which cleaves LexA and induces the SOS genes. The CI repressor shares some structural features with LexA and is also cleaved by activated RecA. This causes the repressor to disassociate from the operator sites, shutting down *cI* transcription from  $p_M$ , and activating transcription from  $p_L$  and  $p_R$  promoters, so heralding the lytic cycle.

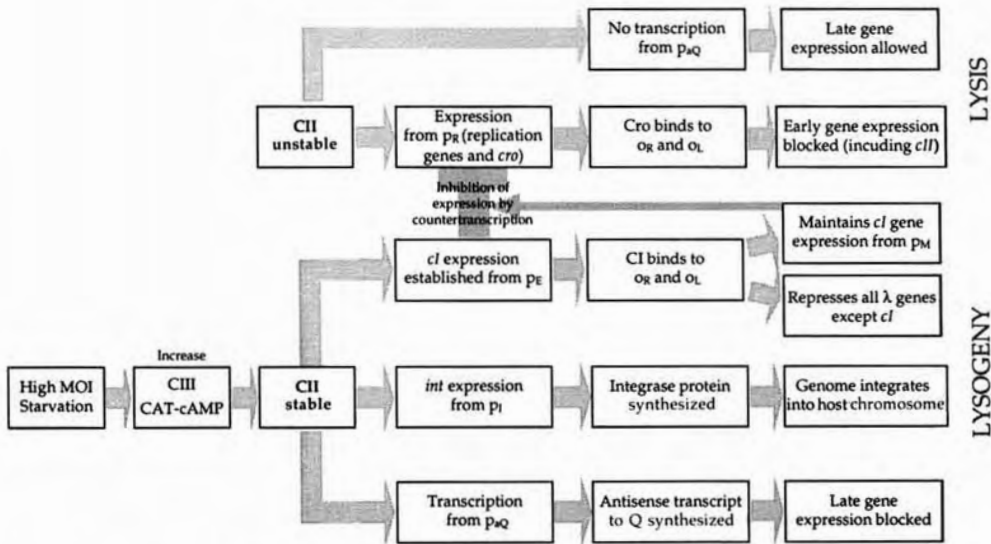
**Control of integration and excision.** Integration of  $\lambda$  into the host chromosome occurs in situations favoring lysogeny, and excision occurs in situations favoring induction. Both processes require the integrase protein, encoded by the *int* gene, whereas only excision requires excisionase, the *xis* gene product. The control of integration and excision is mediated transcriptionally and posttranscriptionally.

During the lytic cycle, transcription from  $p_L$  produces a transcript encoding both *int* and *xis*. Neither protein is synthesized, however, because transcription terminates at a region called *sib* which enables the mRNA to form a distinct secondary structure. This is cleaved by host RNase III and the distal part of the transcript, including the *int* and *xis* loci, is degraded, a regulatory mechanism known as *retroregulation* (q.v.).

When lysogeny is favored, CII activates *int* transcription from promoter  $p_I$  which is within the *xis* gene. Thus *int* is expressed but not *xis* and only integrase is synthesized.

Upon induction, transcription from  $p_L$  is derepressed, but this time both integrase and excisionase are synthesized, allowing prophage excision. The *int* and *xis* genes are not subject to retroregulation this time because integration has separated

the two genes from *sib* (the *att* site lies between *int/xis* and *sib*), and termination before *sib* generates a secondary structure which is not cleaved by RNase III.



The choice between lysis and lysogeny.

**Box 30.2: HIV and the molecular biology of AIDS**

**The HIV infection cycle.** The **human immunodeficiency viruses** (HIV-1 and HIV-2) are closely related retroviruses of the lentivirus group, which also includes the simian immunodeficiency viruses (SIV). The general replication strategy of the retroviruses is discussed in the main text, and HIV follows the same cycle. However, lentiviruses have more complex genomic organization than other retroviruses and encode a number of regulatory gene products in addition to the standard *gag*, *pol* and *env* genes. These enable them to switch between latent and lytic infection strategies.

HIV is a lymphotropic virus which infects T-cells displaying the surface receptor CD4. HIV infection usually causes a short-term disease with relatively mild syndromes. However, CD4 cells are depleted during HIV infection, due to both the lytic infection of these cells and due to their targeting by the immune response. Since T-cells are the cells which regulate the immune response, a secondary and long-term result of HIV infection is an **acquired immunodeficiency syndrome (AIDS)**. HIV infection is also associated with neurological disorders which may result from HIV infection of CD4-carrying neurons, or interaction of the virus with brain cells via an alternative receptor. The retroviral replication cycle, which allows the virus to lie dormant within the host genome for protracted periods, combined with its tropism for CD4-producing T-cells, means it is difficult to develop an effective strategy for treatment of HIV infections. In addition, the viral-encoded reverse transcriptase is highly error-prone, resulting in genetic diversity within the population of viruses infecting a given individual, and allowing immune system evasion — for this reason an effective vaccine has been difficult to develop.

**HIV gene expression and regulation.** The HIV genome contains a number of genes not found in other retroviruses. Both HIV-1 and HIV-2 carry the genes *tat*, *rev*, *nef*, *vpr* and *vif*; in addition, HIV-1 carries a unique gene *vpu*, and HIV-2 a unique gene *vpx*. The *tat*, *rev* and *nef* genes encode proteins with regulatory functions in the HIV infection cycle. The Tat protein is essential for replication and binds to a *cis*-acting element, *Tar*, found in the long-terminal repeats. *Tar* forms a complex secondary structure which inhibits transcription and

protein synthesis, the latter possibly by targeting the RNA for degradation. Tat binds both in the proviral DNA and the transcript arising from it, and alleviates this repression.

The Rev protein is also a posttranscriptional regulator and acts by controlling the nuclear export of splice intermediates. Normally, the presence of splicing components at the splice junctions of cellular pre-mRNA prevents export to the nucleus, so that only fully spliced transcripts are exported for translation. In HIV infections, fully spliced transcripts encode regulatory proteins only, and thus in the absence of Rev, no structural proteins are made. Rev binds to a Rev response element, the **antirepression sequence**, allowing the export of unspliced and partially spliced RNA, and hence the synthesis of structural proteins.

Nef is a negative-acting regulator and may prevent uncontrolled HIV proliferation, which would lead to an enhanced host immune response. Vif, Vpu and Vpr stimulate the production of infectious progeny virus, by either regulating viral gene expression or the assembly of mature virions. Their precise mechanisms of action are unclear.

**Lysis or latency?** In the same way that bacteriophage  $\lambda$  can enter the lytic or lysogenic pathways depending on the environment, HIV can also respond to exterior signals. Transcription of the HIV prophage leads to the production of mature progeny virions, but also in elevated levels of Rev, which shuts down the synthesis of Rev and Tat proteins by stimulating the export of unspliced transcripts encoding structural proteins. Tat is required for transcription and protein synthesis, and thus reduction in Tat levels shuts down viral gene expression, resulting in latency. The virus can enter the lytic cycle once again in response to host-encoded transcriptional regulators recognizing LTR promoter elements. It is thought that the signal transduction pathway initiated by antigen binding to T-cell receptors may be involved in this reactivation process. Thus, the HIV virus is released in short bursts from immune-activated T-cells before transcription is repressed and the virus becomes latent. It may be these short bursts, as well as the genetic variability of the virus, which allows it to evade the host immune response.

**Box 30.3:** The infection cycle of herpes simplex virus 1 (HSV1)

**The lytic cycle.** Herpes simplex virus 1 is an  $\alpha$ -type herpesvirus with a 150 kbp double-stranded linear DNA genome. It infects many different types of cell in many species, due to its interaction with heparan sulfate molecules on cell surfaces. Uptake involves interaction with an FGF receptor.

Once inside the cell, the linear genome is released from the nucleocapsid into the nucleus, where it is immediately circularized. The nucleocapsid also contains two proteins: an RNase termed **virion host shut-off protein (VHS)**, and **VP16**, a protein which cooperates with the host-encoded Oct-1 transcription factor to transcribe the five **immediate early genes** or  **$\alpha$ -genes** of the HSV-1 genome. Most of the products of the  $\alpha$ -genes are genetic regulators, either of their own genes or of the downstream  **$\beta$ -genes**, whose function concerns DNA metabolism and replication. Once these stages are underway, a large set of **~40 late genes** ( **$\gamma$ -genes**) becomes active, producing proteins concerned with DNA packaging and virion assembly. Progeny virions are transported to the cell surface via the endoplasmic reticulum.

**Latent infection.** Lytic HSV-1 infection occurs in most cells, but in neurons, infection is latent. The basis of this cell-specific latency is the transcription of a set of **latency-associated transcripts (LATs)** and their subsequent splicing. The transcription unit for the LATs overlap one of the immediate early genes,  $\alpha 0$ , but in the antisense direction. The LATs are produced during lytic infection, but are unspliced and exported to the cytoplasm. In the latent infection, the spliced LATs are restricted to the nucleus, and may in some way down regulate lytic gene expression. However, no LAT-associated polypeptides have been detected, so it is likely that the LATs function at the RNA level. The LATs are not required for the *establishment* of latent infection, but are critical for reactivation of the lytic cycle. During latent infection, HSV-1 is maintained episomally in the cell without replicating. The virus is thus a useful vector for gene transfer to neurons in live animals, and is used for *gene therapy* (q.v.). Recombinant HSV-1 vectors can be constructed where inserted foreign genes are driven from the LAT promoters.

**References**

- Dimmock, N.J. and Primrose, S.B. (1993) *An Introduction to Modern Virology*. 4th edn. Blackwell Science, Oxford.
- Murphy, F.A., Fauquet, C.M., Bishop D.H.L., Ghabrial, S.A., Jarvis, A.W., Martelli, G.P., Mayo, M.A. and Summers, M.D. (eds) (1995) *Virus Taxonomy. Sixth Report of the International Committee on Taxonomy of Viruses*. Springer, New York.
- Further reading**
- Aguzzi, A. and Weissmann, C. (1997) Prion research: the next frontiers. *Nature* 389: 795–798.
- Banerjee, A.K. and Barik, S. (1992) Gene expression of vesicular stomatitis virus genome. *Virology* 188: 417–429.
- Berns, K.I. (1990) Parvovirus replication. *Microbiol. Rev.* 54: 316–329.
- Caughey, B. and Chesebro, B. (1997) Prion protein and the transmissible spongiform encephalopathies. *Trends Cell Biol.* 7: 56–62.
- Coffin, J.M. (1992) Structure and classification of retroviruses. In: *The Retroviridae* (ed. J. Levy), vol. 1, pp. 1437–1500. Plenum Press, New York.
- Collmer, C. and Howell, S. (1992) Role of satellite RNA in the expression of symptoms caused by plant viruses. *Annu. Rev. Phytopathol.* 30: 419–442.
- Diener, T.O. (1991) The frontiers of life: The viroids and viroid-like satellite RNAs. In: *Viroids and Satellites: Molecular Parasites at the Frontiers of Life* (ed. K. Maramorosch), pp. 1–20. CRC Press, Boca Raton, FL.
- Eckhart, W. (1991) Polyomavirinae and their replication. In: *Fundamental Virology* (eds B.N. Fields and D.M. Knipe), 2nd edn., pp. 727–741. Raven Press, New York.
- Estes, M.K. (1991) Rotaviruses and their replication. In: *Fundamental Virology* (eds B.N. Fields and D.M. Knipe), 2nd edn., pp. 619–642. Raven Press, New York.
- Gabizon, R. and Taraboulos, A. (1997) Of mice and (mad) cows: Transgenic mice help to understand prions. *Trends Genet.* 13: 264–269.
- Hendrix, R.W., Roberts, J.W., Stahl, F.W. and Weisberg, R.A. (eds) (1983) *Lambda II*. Cold Spring Harbor Press, Cold Spring Harbor, New York.
- Horwitz, M.S. (1990) Adenoviruses and their replication. In: *Virology* (eds B.N. Fields and D.M. Knipe), 2nd edn., pp. 1679–1722. Raven Press, New York.
- Joshi, S. and Joshi, R.L. (1996) Molecular biology of Human Immunodeficiency Virus type 1. *Transfusion Sci.* 17: 351–378.



- Keppel, F., Fayewt, O. and Georgopoulos K. (1988) Strategies for bacteriophage DNA replication. In: *The Bacteriophages* (ed. R. Calendar), Vol. 2, pp. 145–262. Plenum Press, New York.
- Kingsbury, D. (ed.) (1991) *The Paramyxoviruses*. Plenum Press, New York.
- Luong, G. and Palese, P. (1992) Genetic analysis of influenza virus. *Curr. Opin. Genet. Dev.* 2: 77–81.
- Roizman, B. and Sears, A.E. (1993) The replication of Herpes simplex viruses In: *The Human Herpesvirus* (eds B. Roizman, C. Lopez and R.J. Whitley), pp. 11–68. Raven Press, New York.

### Websites

- All the virology on the web: an index of virology resources and information available on the internet — <http://www-micro.msb.le.ac.uk/garryfavweb/garryfavweb.html>

**This Page Intentionally Left Blank**

# Index

t, table; f, figure; bx, box.

$\alpha$ -complementation,  
109–110bx, 330

$\alpha$ -helices, 289–90f

A-DNA, 229, 230f

A-rule, 185

A-site (amino acyl-tRNA site),  
of ribosome, 314

*ABL/BCR* gene fusion, 255–6,  
257–8

Abzymes, 345

*Ac-Ds* elements, 176, 178, 344

Acute transforming  
retroviruses, 254, 256

Adaptive response, 190

Adaptors, 326

Additive effects, 10t

AIDS, 485

Alkaline phosphatase, 324,  
326

Allele frequency, 212

Allele replacement, 352

Allele-specific hybridization,  
358

Allele-specific PCR, 280

Alleles, 2, 107, 201

Allelic complementation,  
109–110bx

Allelic exclusion, 5, 381

Allopolyploidy, 47

Allostery, 295  
in transcriptional control,  
455–7f

Allotype, 379

Alternative splicing, 418–20  
in *Drosophila* sex  
determination, 420f

*Alu* element, 96, 142, 144

Amber codon, 127, 204

Amber suppressor, 220

Ambisense genome, 470

Amino acid sequence, 287–9

Amino acids, 305–6bx

Aminoacyl-tRNA, 128

Aminoacyl-tRNA synthetase,  
129

*Amphioxus*, 300

Amplicon, 219, 342

Anaphase lag, 49

Anaphase-promoting  
complex, 32

Aneuploidy, 47–49

Anisotropic bending, 230

Antibodies, 345, 379, 386–7  
for protein analysis, 307,  
345

Anticipation, 16bx, 218

Anticodon, 128

Anticodon loop, 321

Antigen switching,  
trypanosomes, 378

Antiparallel strands, 225

Antipodal effects, 42

Antisense RNA, 224  
in control of gene  
expression, 320, 365  
in gene therapy, 365  
in plasmid replication, 276–7  
in protein synthesis, 320

Antisense strand, 444

Antitermination, 460–1, 481–2

AP endonuclease, 192

AP-site, 185

Apoptosis, 33–34

Aptamers, 366

Artificial chromosome  
vectors, 154t, 329

Asymmetric PCR, 284

Attached chromosome, 52

Attenuation, attenuator  
control, 461

Autopolyploidy, 47

Autosomal inheritance, 15bx

Autosomes, 57

Autozygosity, 149

Auxotroph, 215

$\beta$ -sheets, 289–90f

$\beta$ -strands, 289–90

$\beta$ -turn, 289–90

B-chromosomes, 58

B-DNA, 229, 230f

Backcross, 3f

Bacterial artificial  
chromosome (BAC), 153,  
329

Bacteriophage, 467

Bacteriophage  $\lambda$ , 481–84bx  
cloning vectors, 327–9

gene regulation, 475–6,  
481–84bx  
integration and excision,  
385, 484

Bacteriophage M13, 473–4,  
237–9

Bacteriophage Mu, 468  
host range variation, 385–6  
replication, 473

Baculovirus expression  
system, 342

Balance mechanism, 76

Balancer chromosome, 51

Balbani rings, 59

Baltimore classification, 470

Bands and interbands, 59

Barr body, 100

Barrel motifs in proteins,  
291–2

Base analog, 185

Base composition, 134–5

Base excision repair, 191–2

Base loss, 185

Base pairs  
alternative, 227  
Hoogsteen, 227  
Watson-Crick, 226–7, 228f

Base ratio, 134

Base stacking, 226

Bases, in nucleic acids, 223,  
225f

Basic DNA-binding domain,  
241

Beckwith-Wiedemann  
syndrome, 98–99

Biotin-streptavidin system,  
360

Biphasic genome  
organization, 58, 143

Blue-white selection, 330

Bone morphogenetic proteins,  
430–1

Boundary element, 41

Box (DNA or protein  
sequence motif), 301

*BRCA1*, *BRCA2*, 152, 260

Breakage-fusion-bridge cycle,  
174

Burkitt's lymphoma, 258

- C-gene segment, 387  
 C-value, 134  
 C-value paradox, 135–6  
 CAK, 28  
 Calcium signaling, 438–9  
 Calmodulin (CaM), 438–9  
 Cancer, 253  
   familial, 258–9  
   multiple hit hypothesis, 253, 258  
   tumor progression, 253  
 Candidate genes, 335–6  
 Cap, 413  
 Cap-binding protein, 316–7  
 Capping, 413  
 Capsnatching, 421, 474–5  
 Cassette, 165–6, 377  
 Cassette mutagenesis, 346–7f  
 Catabolite repression, 464  
*cdc* mutants, 26  
*Cdc2/CDC28* genes, 26–8  
 CDK-activating kinase, 28  
 CDK-cyclin inhibitors, 28, 33, 260  
 cDNA, 333  
   cDNA capture/cDNA selection, 155  
   cDNA libraries, 333–4  
   cDNA synthesis, 333  
 Cell cycle  
   bacterial, 21–23  
   eukaryotes, 23–34  
 Cell division, bacteria, 22–23  
 Central dogma, 111t  
 Centromere, 60–61  
 CEPH families, 150  
 Character, 1, 2  
 Checkpoints (of cell cycle), 24  
 Chemical cleavage of  
   mismatch, 337  
 Chi, 372  
 Chiasmata, 374  
 Chimera, 54, 364  
 Chloroplast DNA, properties of, 266  
 Chromosome jumping, 335  
 Chromatid, 57  
 Chromatin, 35–42  
   diminution, 77  
   domains, 40–42  
   open and repressed, 40  
   remodeling, 33, 40  
   role in gene regulation, 39–42  
   structure, 35–39  
 Chromomeres, 59, 60  
 Chromosome banding, 57–59t  
 Chromosome breakpoints, 49  
 Chromosome deletions, 50  
 Chromosome derivatives, 49–50  
 Chromosome duplications, 50  
 Chromosome imbalance (aneuploidy), 45, 47–49, 202  
 Chromosome landing, 336  
 Chromosome mutation, 45–55, 201  
   balanced, 45, 51–52  
   constitutional, 45–54  
   numerical, 45–49, 201–3  
   somatic, 45, 54–55, 253–8  
   structural, 49–55, 201–3  
   unbalanced, 45–51  
 Chromosome number, 134  
 Chromosome painting, 156  
 Chromosome puffs, 59  
 Chromosome theory of inheritance, 57  
 Chromosome walk, 153, 335  
 Chromosomes, 57–64  
   classification/  
     nomenclature, 62–64bx  
   metaphase, 39  
   molecular structure, 60–64  
   morphology, 57–60  
 Circular dichroism, 307  
 Circular permutation, 472  
*cis*-acting elements, 113, 452  
*cis*-dominance, 110, 113  
*cis*-trans test, 104, 109–110bx  
 Cistron, 104–106  
 Class switching, 381–2  
 Classical genetic analysis, 213–5  
 Clastogen, 53  
 Clone contig map, 152–3  
 Cloning vectors, 323, 327–9t  
 Coding region, 107  
 Codominance, 7  
 Codon, 127  
 Codon assignment, 130  
 Codon family, 129  
 Codon usage/choice/bias/  
   preference, 130  
 Cofactor, 292  
 Coiled coil, 291  
 Coiling in snails, 67f  
 Cointegrate, 172  
*ColE1* replicon, 276–77  
 Colony blot, 358–9  
 Comparative genomics, 156  
 Competence  
   artificially induced, 330  
   for natural transformation, 120  
 Complementary base pairs, 226  
 Complementation analysis, 104, 109–110bx  
 Complementation map, 110  
 Complementation group, 110  
 Complex mutation, 202  
 Complexity, genomes, 135  
 Compound chromosome, 52  
 Concatemer, 472  
 Concerted evolution, 300  
 Conditional mutants, 210, 214  
 Conduction (*cis*-mobilization) of DNA, 119  
 Conjugation, 117–9  
 Conservative mutation, 204  
 Constitutive mutation, 210  
 Context-dependent regulation, 456  
 Contiguous gene syndromes, 50  
 Continuous character, 11  
 Controlling elements, 178  
 Cooperative binding, 295  
 Cooperative transposition, 174  
*cop* mutant, 272, 274–5  
 Cosmid, 327–9  
 Cosuppression, 114  
 Cot analysis, 157–159bx  
 Cotranslational frameshifting (recoding), 319  
 Cotranscriptional regulation, 456, 458  
 Counterselection, 214, 330–1, 351–2  
 CpG islands, 95, 155  
*Cre-loxP* system, 353, 386  
 CREB, 441  
 Creutzfeldt-Jakob disease, 480–81  
 cRNA (complementary RNA), 359  
 Cross-feeding, 110  
 Cross-talk, 439  
 Cryptic plasmids, 271, 272  
 Cryptic satellite DNA, 142  
 Cryptic splice site, 416, 420  
 Curing, 272  
 Cyclic nucleotides  
   as second messengers, 434–6  
   structure, 225, 226f  
 Cyclin-dependent kinases, 26–28  
 Cyclins, 26–27  
 Cytogenetic maps, 146  
 Cytokine receptors, 428–9  
 Cytokines, 425–6  
 Cytoplasmic determinants, 67  
 Cytoplasmic inheritance, 263, 268–9



- Cytotype, 175, 265
- D-gene segments, 387
- D-loop (displacement loop), 232
- D-loop (part of transfer RNA), 321
- Dam methylation, 94
- Dcm methylation, 94
- Deamination, 185–6t  
of 5-methylcytosine, 94, 185, 211
- Death domain, 33–34, 431
- Degeneracy, genetic code, 129
- Deletion loop, 50
- Densely methylated island, 404
- Determination, 71
- Development, 65–92  
cell-cell interactions in, 68  
early *Xenopus laevis*, 80–82bx  
gene regulation in, 65–66  
genomic equivalence in, 65–66  
genomic nonequivalence, 77bx, 219  
mosaic, 68–69  
regulative, 68–69  
role of environment, 75
- Developmental noise, 17–18bx
- Diacylglycerol, 436
- Dictyostelium discoideum*, life cycle, 79–80bx
- Difference cloning, 336
- Differential display, 283
- Differentiation, 65  
maintenance, 71–72  
simple models, 66–67, 78–80bx
- Digoxigenin system, 360
- Dihybrid cross, 8–9f
- Direct reversal repair, 187–190
- Directed mutation, 212, 218–220bx
- Directional cloning, 326
- DIRVISH, 152
- Dislodgment, 272
- Disulfide bonds, 293
- DMS protection assay, 250t, 251f
- DNA, 224  
breathing, 223  
classified by function, 137t  
conformational polymorphism, 223, 230  
helical morphology, 228–30, 233bx  
sequence polymorphism, 150, 211  
structure, double helix, 228–30  
structure, recognition by proteins, 236  
topology, 232, 233–4bx  
DNA-binding motifs in proteins, 237–43, 237t  
DNA damage, 185–6t  
DNA fingerprinting, 158–9bx  
DNA glycosylases, 191t  
DNA gyrase, 397  
DNA library, 323, 331–4  
gridded, 153  
screening, 323, 334–6  
DNA ligase, 323, 324, 326, 400  
DNA methylase (methyltransferase), 93  
DNA methylation, 93–101  
bacteria, 93–94  
control of transposable elements, 96, 170–1  
CpG/CpNpG in eukaryotes, 94–101  
general role in gene regulation, 95–97  
imprinting and X-chromosome inactivation, 97–101  
DNA microarrays, 304  
DNA modification, 93  
DNA polymerases, 394, 407  
*E. coli*, 394–5t  
eukaryote, 396–7t  
in molecular cloning, 324t  
thermostable, 396, ch21  
DNA primase, 397  
DNA profiling, 158–9bx  
DNA repair, 187–199  
error free, 187–197  
mutagenic, 197–199  
DNA replication, 389–409  
by displacement, 393  
models, 389  
priming, 404, 405t  
semiconservative, 389–90  
semidiscontinuous, 390–91  
DNA sequencing, 161–3bx  
DNA tumor viruses, 259–60  
DNA viruses, 470–3  
DNase I footprinting, 250t, 251f  
DNase I hypersensitive site, 40, 454, 462  
Domain swap, 247  
Domains, of protein structure, 290–2  
Dominance, 1–2, 6–7, 209  
Dominance relationships, 7–8t  
Dominant negative (*trans*-dominant), 110, 114, 210  
Dominant positive, 210  
Donation (*trans*-mobilization) of DNA, 119  
Donor conjugal DNA synthesis, 118–9f  
DOP-PCR, 281  
Dosage compensation, 75  
Dosage effects, 45, 207  
Dosage repetition, 301  
Dot blot, 357  
Double minute chromosome, 219  
Down's syndrome, 48  
*Drosophila melanogaster*  
eye development, 86–87  
homeotic genes, 87–89bx  
model organism, 145–6  
pattern formation in syncytial embryo, 82–85bx  
segmentation, 87–89bx  
dsRNA-binding domain, 243–4  
E-site (exit site), of ribosome, 314  
Ectopic expression, 210  
Editosome, 422  
Edman degradation, 307  
Electrophoresis, 355  
Electroporation, 330, 349  
Elongation factors  
protein synthesis, 316, 317–8f  
transcription, 458–9  
Embryonic stem cells, 364–5  
Endosymbiont theory, 268  
Enhancement, 10t, 221  
Enhanceosome, 454  
Enhancer, 451–2  
Enhancer trap, 345  
Entrapment vectors, 344–5t  
Environment, 13–15, 17–18bx  
Enzyme cascades, intracellular signaling, 431  
Enzyme mismatch cleavage, 337  
Epigenesis, 65  
Epigenetic gene regulation, 95–101  
Epigenetic information, 93, 127  
Epimutation, 95  
Episomes, 165  
Epistasis, 10t  
Euchromatin, 38  
Exclusion mapping, 148  
Exon duplication, 287, 301

- Exon phase, role in evolution, 303
- Exon shuffling, 287, 301
- Exon skipping, 416
- Exon sliding, 303
- Exon trapping, 155
- Exons, 137
- Expressed sequence tag, 155–6, 281
- Expression cloning, 339–342
- E. coli*, 340–1
- eukaryote hosts, 341–2
- native and fusion proteins, 340–1
- Expression linked copy, 378
- Expression vectors, 340
- Expressivity, 8
- F transfer region (*tra* operon, *tra* region), 124bx
- F' plasmid, 119
- F-factor, in bacterial gene mapping, 125bx
- F-factor/F-plasmid, 117–9, 124bx
- Fate maps, 68
- Fiber diffraction, 308
- Fiber FISH, 152
- Fidelity of replication, 183–5
- Field inversion gel electrophoresis, 355–6
- Filter hybridization, 357–9
- FISH, 152
- 5-methylcytosine, 94, 211
- 5'→3', 225
- Flow sorting, 333
- FLP-FRP system, 386
- Fluctuation test, 218
- Fluorescence activated chromosome sorting (FACS), 333
- Foldback elements, 176, 178
- Foldon, 94
- Footprints
- of precisely excised transposons, 173
- of proteins, 250–1
- Forward mutation, 209
- Fos, *c-fos* gene, 254–8, 441
- Fragile sites, 52–53
- Fragile-X syndrome, 53, 218
- Frameshift fidelity, 184–5
- Frameshift mutation, 205
- fts* mutants, 22
- Functional cloning, 334–5
- Functional genomics, 304–5
- Fusion protein, 45, 207, 255–8, 340–1
- G-banding, 58–59
- G-protein coupled receptors, 426–7
- G-proteins, classes and activities, 427t
- G<sub>0</sub>, 23
- G<sub>1</sub> and G<sub>2</sub> gap phases, 23
- Gain of function, 209–211, 350–2
- Gap genes, 83–84
- GC content, 134–5
- Gel electrophoresis, 323, 355–6bx
- large DNA molecules, 355–6bx
- Gel retardation assay, 250t, 251f
- Gene amplification
- drug resistance, 219
- expression vectors, 342
- in cancer, 255–8
- programmed, 77, 219
- Gene augmentation therapy, 365
- Gene batteries, 458
- Gene conversion, 208
- allelic, 376
- in DNA repair, 371
- in passive transposition, 169–70
- information cassette switching, 377
- Gene density, 138
- Gene disruption, 45, 207
- Gene dosage, 45, 207
- Gene duplication and divergence, 287, 298–300
- Gene duplication, mechanisms, 299–300
- Gene expression, 111–3
- prokaryote and eukaryote strategies compared, 115–6
- Gene families, 297–304
- chimeric, 300–303
- conventional, 297–300
- Gene fusion, 45, 207
- Gene knock-in, 352
- Gene knockout, 352
- Gene locus, 2, 107
- Gene mapping
- bacteria, 125
- eukaryotes, 144–156
- Gene mutation, 201
- Gene number, 138
- Gene product, 111
- Gene regulation, 113–114
- global and narrow domain, 114
- induction and repression, 114
- Gene size, 137
- Gene structure, 106–107, 137
- Gene superfamily, 297
- Gene targeting, 305, 349–53
- Gene therapy, 349, 352, 365–6bx
- Gene tracking, 150
- Gene transfer, bacterial, 117–126
- Gene trap, 305, 345
- General transcription factors, 447
- Genes, 2, 103–110 (see also individual genes by name) and cistrons, structure/function relationship, 104–106 history of term, 103 nomenclature, 107–108
- Genetic code, 127–131
- properties, 128t
- universal, 128f
- variations in codon assignment, 130t
- Genetic information, 93, 127
- Genetic linkage, 9, 373
- Genetic mapping, 144, 146–151
- quantitative trait loci, 150–151
- Genetic markers, 149–50t
- Genetic networks, 215, 305
- Genetic screens, 214
- Genome, 133
- Genomes
- physico-chemical properties, 134–135
- structure and organization, 133–144
- Genomic libraries, 332–3
- Genomics, 144–156
- Genotype, 2
- Genotoxic agents, 185, 187t
- Germinal mutation, 201
- Globin genes
- mutations, 216–7
- transcriptional regulation, 462–3
- Greek key motif, 291
- Growth factors, 425–6
- Growth transformation, 253
- Guide RNA (gRNA), 422
- Gypsy transposon, 41, 180
- Hairpin, 230–1
- Haploid number, 134
- Haploinsufficiency, 210
- Haplotype, 374

- HAPPY mapping, 146–7f  
 Harlequin staining, 59  
 Headful mechanism, 472, 474f  
 Hedgehog family proteins  
   *Drosophila* Hedgehog, 87–89bx, 138  
   in vertebrate limb development, 91–92bx  
 Helicase, 397, 398t  
 Helix-loop-helix, 241, 246, 291  
 Helix-turn-helix, 237–9, 291  
 Helmstetter-Cooper model, 21–22  
 Hemoglobin disorders, molecular basis, 216–7  
 Hemoglobinopathies, 216–7  
 Hepatitis B virus, 472  
 Hereditary persistence of fetal hemoglobin, 217  
 Heredity, 2  
 Heritability, 18bx  
 Herpes simplex virus, infection cycle, 486  
 Heterochromatin, 38  
   role in gene silencing, 42  
 Heterochronic mutations, 74  
 Heterogeneous nuclear RNA (hnRNA), 411  
 Heterozygous, 2  
 Hfr strains, 119  
 High mobility group (HMG) proteins, 37, 239  
 Histones  
   core, 37, 241–2  
   linker, 36–37, 242  
   histone fold, 36, 241–2  
   modification, 33, 39–42  
 HIV infection cycle, 485  
 HNF3/Fork head family, 239  
 HO endonuclease, 377  
 Holliday model, 370–2  
 Homeodomain, 238–9  
 Homeologous chromosomes, 47  
 Homeotic genes, 73, 87–89bx, 138  
 Homing introns and inteins, 179  
 Homogeneously staining region, 219  
 Homologous chromosome pairs, 57  
 Homologous recombination  
   hotspots and coldspots, 161  
   mechanism, 369–372  
   molecular basis, 372–3  
   regional frequency variation, 161  
   role in genetic mapping, 373–6  
 Homology, 297  
 Homology-dependent gene silencing, 114  
 Homopolymer tailing, 327  
 Homozygous, 2  
 Horizontal gene transfer, 298  
 Host controlled restriction-modification systems, 93–94  
*Hox* genes, 298  
   in vertebrate limb development, 91–92bx  
   vertebrate, 89–90bx, 138  
 Human Genome Project, 145  
 Hybrid dysgenesis, 175  
 Hydatidiform moles, 49, 97  
 Hydrogen bonds  
   in nucleic acids, 226–8  
   in protein-nucleic acid binding, 244–5  
   in proteins, 293–4  
 Hydrophobic core, 293  
 Hypervariable minisatellite DNA, 142  
 Hypochromic effect, 357  
 Hypostasis, 10t  
 Illegitimate recombination, 370, 382–3t  
 Immediate early genes  
   cellular, 33, 440–1  
   viral, 481, 486  
 Immunoglobulin genes, 386–7  
 Immunoglobulins, 379, 386–7  
   diversity, 379  
 Imprinting boxes, 97  
*In situ* hybridization, 152, 339, 359  
*In situ* PCR, 284  
*In vitro* mutagenesis, 346–8  
   systematic, 347  
*In vitro* packaging, 330  
*In vivo* footprinting, 250  
 Incompatibility groups, 276, 277–8bx  
 Incompatibility, of plasmids, 272, 275–6  
 Indel, 205  
 Induced fit, 244  
 Induction  
   of gene expression, 114, 463–4bx  
   bacteriophage  $\lambda$ , 483  
   community effect, 70  
   homeogenetic, 69  
   in development, 69–71f  
   instructive, 69  
   lateral inhibition, 69–70f  
   morphogens in, 69  
   of latent virus, 468  
   permissive, 69  
 Informational suppression, 220–1  
 Informosome, 412–3  
 Inhibition domains (of transcriptional repressors), 465  
 Initiation codon, 127, 130  
 Initiation factors, protein synthesis, 316–7  
 Initiator tRNA, role in protein synthesis, 316–7  
 Inositol phospholipids, 436–8  
 Inositol-1,4,5-trisphosphate, 436–7  
 Integrase  
   retroviral, 165  
   bacteriophage  $\lambda$ , 385, 484  
 Integrins, 179–80  
 Interaction trap, 345  
 Interference, 160–1  
 Internal ribosome entry site, 319  
 Intrabodies, 366  
 Intrinsic termination, of transcription, 460  
 Intron-encoded proteins, 417–8  
 Introns, 104, 137  
   origins and evolution, 303  
   phase, 302–3  
   splicing, 416–7  
 Inverse PCR, 284  
 Inverse transposition, 174  
 Inversions, 50–51f  
 Ion channels, 426  
 Iron response element, 320, 423–4  
 IS elements, 176–7  
 Isoaccepting tRNAs, 129, 314–5  
 Isoallele, 6, 209, 211  
 Isochore model, 58, 143–4  
 Isochromosomes, 51  
 Isoelectric focusing, 307, ch25  
 Isotropic bending, 230  
 Isotypes, 379  
 Isotypic exclusion, 381  
 Iterons, 276  
 J-gene segments, 387  
 Jak-STAT pathway, 429  
 Jelly roll motif, 292  
 Jun, *c-jun* gene, 254–8, 441  
 Junk DNA, 137  
 K-homology domain, 244  
 Karyogram, 134

- Karyotype**  
   abnormal, 46t  
   normal, 134  
**Killer factors**, 273  
**Killer plasmids**, 273, 274  
**Killer system**, 272  
**Kinetochore**, 60–1  
**Kinetoplast DNA**, 268  
**Kissing complex**, 277  
**Klenow fragment**, 394  
**Klinefelter's syndrome**, 48  
**Knudson, two hit hypothesis**, 258–9  
**Kozak consensus**, 317
- Lac operon**, 463–4bx  
**Lampbrush chromosomes**, 60  
**Lariat intermediate**, 415  
**Leading strand — lagging strand model**, 390  
**Leaky allele**, 210  
**Lesions, of DNA**, 185–6  
**Lethal allele**, 8  
**Leucine zipper**, 241, 246  
**Licensing factor**, 29–30  
**Ligase chain reaction**, 285  
**LINEs**, 96, 142, 144  
**Linkage**, 9, 373  
**Linkage equilibrium/disequilibrium**, 374  
**Linkage mapping**, 375, 147–151  
   limitations to accuracy, 159–161bx  
**Linkers**, 326  
**Linking number paradox**, 37  
**Lipid second messengers**, 436  
**Lipofection**, 349  
**Locus control region**, 41, 451, 462–3  
**Lod scores**, 148  
**Long range restriction map**, 338  
**Loops, in polypeptides**, 290  
**Loss of function**, 210  
**Loss of heterozygosity**, 259  
**Lysogeny**, 468  
**Lytic infection**, 468
- M phase**, 23  
**M phase/maturation-promoting factor**, 25–26  
**Macromolecules**  
   determining mass, 320bx  
   separation, 320, 355  
**Macromutation**, 202  
**MADS box**, 73  
**Maintenance methylase**, 93
- Major groove**, 228  
**MAP kinase**, 432–3t  
**Mapping function**, 160  
**MAT locus**, 377  
**Maternal effect**, 5–6, 67f  
**Maternal genes, *Drosophila***, 82–85bx  
**Maternal inheritance**, 263, 264f  
**Mating type switching**, 377  
**Matrix-associated region**, 41  
**Mediator**, 453  
**Megaplasmsids**, 271  
**Meiosis**, 373  
**Meiotic drive**, 6–8  
**Meiotic map**, 146  
**Mendel's First Law**, 3  
**Mendel's Second Law**, 8–9  
**Mendelian inheritance**, 1–11  
**Meristic character**, 11  
**Merozygote**, 117  
**Meselson–Stahl experiment**, 389–90  
**Mesoderm induction, *Xenopus laevis***, 81  
**Messenger RNA (mRNA)**, 224, 313  
   life cycle in bacteria and eukaryotes, 313–4  
   processing, 412–421  
**Metabolic block**, 215  
**Methylation interference assay**, 250t, 251f  
**Microballistics**, 349  
**Microsatellite DNA**, 143  
**Minisatellite DNA**, 142–3  
**Minor groove**, 228  
**Misinstructional lesion**, 185  
**Mismatch**, 194  
**Mismatch repair**  
   long patch, 194–5  
   short patch, 196–8  
**Missense mutation**, 204  
**Mitochondrial DNA**  
   mutants, 264–5t  
   organization and gene expression, 266–7  
   replication, 267–8  
**Mitochondriopathies**, 264  
**Mitosis**, 32  
   control of, 30–32  
   mitotic recombination, 375  
   mitotic segregation, 376  
**Mixoploidy**, 54  
**Mobile genetic elements**, 165–182  
**Mobilization, of plasmids for transfer**, 118  
**Model organisms, genomics**, 133, 145–146t
- Modules, of proteins**, 287, 300–301  
**Molecular chaperones**, 39, 294–5  
**Molecular clock**, 213  
**Molecular cloning**, 323  
   analysis of cloned DNA, 336–339  
   general principles, 323–331  
   recovery of cloned DNA, 331  
   strategies, 331–336, 332t  
**Molecular markers**, 149–50t  
**Monoallelic expression**, 5, 97–100, 381  
**Monocistronic mRNA**, 104  
**Monohybrid cross**, 3–4f  
**Monoploid number**, 134  
**Morbid map**, 50  
**Morphogenesis**, 65, 73–75  
**Mosaic**, 54  
**Motif, in DNA or polypeptide sequence**, 301  
**Motif, in protein structure**, 290–1t  
**mRNA guanylyltransferase**, 408, 413  
**Muller classification**, 209  
**Multienzyme proteins**, evolution by gene fusion, 302  
**Multigene families**, 136, 139, 298  
**Multiple cloning site**, 327  
**Multiple crossovers**, 160  
**Multiple start sites**, 448  
**Multiplex PCR**, 281  
**Mutagen**, 187, 188t  
**Mutagenesis**  
   gene targeting, 349–53  
   induced, 183–8  
   *in vitro*, 346–8  
   natural mechanisms, 183–8, 207–8, 172–5  
**Mutant**, 201  
**Mutant alleles**, 209–211  
**Mutation**, 183, 201  
   functional consequences, 203–7  
   structural categories, 201–3t  
**Mutation frequency**, 212  
**Mutation hotspots**, 211  
**Mutation pressure**, 212  
**Mutation rate**, 183, 212  
**Mutation screening**, 337t  
**Mutator/antimutator genes**, 183, 393  
**Myc, c-myc gene**, 254–8  
**MyoD gene family**, 140



- N*-formylmethionine, 130, 305–6  
*N*-glycosidic bond, 224  
*N*-region diversity, 380  
 Natural selection, 212–3  
 Neomorph, 210  
 Nested genes, 106  
 Nested PCR, 280  
 Neutral mutation, 203  
 Neutron scattering, 308  
 Niewkoop center, 80–82  
 NOESY, 308  
 Nonallelic interactions, 9–11ft  
 Noncoding region, 106  
 Nonconjunction, 49  
 Nonconservative mutation, 204  
 Nondisjunction, 49f  
 Nonhistone proteins, 37, 239  
 Noninstructional lesion, 185  
 Nonreciprocal recombination, 376–8  
 Nonsense codon, 127  
 Nonsense mutation, 204  
 Nonsynonymous mutation, 204  
 Norm of reaction, 13  
 Northern blot/hybridization, 338, 357–8  
 Northwestern screen, 250, 358  
 Nuclear localization sequences, 310  
 Nuclear magnetic resonance spectroscopy, 307–8  
 Nuclear matrix, 38  
 Nuclear receptor family, 241, 431  
 Nuclear scaffold, 38  
 Nucleases  
   biological roles, 399  
   in molecular cloning, 324t  
   nuclease mapping, 338–9t, 340f  
 Nucleic acids, 223  
   backbone, 225  
   conformational polymorphism, 223, 230  
   helical morphology, 228–30, 233bx  
   primary structure, 223–6, 228f  
   secondary structure, 226–31  
   tertiary structure, 231–2  
 Nucleic acid hybridization, 323, 356–9bx  
   in solution, 357  
   parameters, 356–7  
 Nucleic acid probes, 359–60bx  
   and nonisotopic labeling, 360  
 Nucleic acid synthesis, 407–8bx  
 Nucleic acid-binding proteins, 235–52  
   recognition elements in, 235–7  
 Nucleic acid-protein interaction (*see* Protein-nucleic acid interaction)  
 Nucleoid, structure and organization, 42–3  
 Nucleolar organizer region, 58  
 Nucleosides, 224  
 Nucleosome, 35–36f  
   core particle, 35  
   displacement, 39  
   phasing, 37–339  
   structure during replication/transcription, 39  
 Nucleotide excision repair, 192–4  
 Nucleotides, 225, 226f, 227t  
 Null allele, 210  
  
 Ochre codon, 127, 204  
 Ochre suppressor, 220  
 Okazaki fragments, 391  
 Oligonucleotide mutagenesis, 356–7f  
 Oligonucleotides, 225  
 Oncogene activation, mechanisms, 255–258tf  
 Oncogenes, 253–8, 259, 261  
 One gene one enzyme model, 103  
 Opal codon, 127, 204  
 Opal suppressor, 220  
 Open reading frame, 106, 127, 313  
 Operator, 452, 463  
 Operon, 107, 456  
 Organelle genetics, 263–5  
 Organelle genomes, 263–9  
 Organelle plasmids, 268  
 Organizer, 80–82  
 Origin of replication, 401–4  
 Orthologous genes, 298  
 Overlapping genes, 106  
  
 P-DNA, 174  
 P-element, 178–9  
 P-element mutagenesis, 344  
 P-site (peptidyl-tRNA site), of ribosome, 314  
 P1 artificial chromosome (PAC), 153, 329  
 p53, 33, 260, 441, 456  
 Packaging ratio, 35  
 Packaging site, 122  
 Padlock probes, 360  
 Pair rule genes, 83, 84  
 Paired family, 239  
 Paralogous chromosome segments, 300  
 Paralogous genes, 298  
 Paramutation, 7, 96  
 Paranemic joint, 231  
 Parasexual exchange, bacteria, 117  
 Parental imprinting, 5, 97–100f  
   enhancer competition, 98  
   role in mammalian development, 99–100  
 Partial redundancy, 140  
 Parvovirus, 474  
 Passive transposition, 169, 170f  
 Pattern formation, 65, 72–75  
 PCR, 279–85  
 PCR mutagenesis, 282  
 PCR products, cloning, 283  
 Pedigree analysis, 15–16bx  
 Penetrance, 8  
 Peptide nucleic acid (PNA), 366  
 Peptidyl-tRNA, 318  
 PEST domain, 310  
 Phage display, 305, 346  
 Phagemids, 327–9  
 Phasmids, 327  
 Phenocopy, 17  
 Phenotype, 2  
 Phenotypic variance, 17–18bx  
 Philadelphia chromosome, 258  
 Phosphodiester bond, 225  
 Phosphoinositide 3-kinase, 347f  
 Photoreactivation, 190  
 Physical mapping, low resolution, 151–2t  
 Physical maps, 146  
 Pilus, 118  
 Plaque lift, 358–9  
 Plasmid cloning vectors, 327–9  
 Plasmid maintenance, 272, 275–6t  
 Plasmid partition, 22, 272  
 Plasmid phenotypes, 273  
 Plasmid replication, 276–7  
 Plasmid rescue, 176, 344  
 Plasmids, 271–8  
   classification, 271–3  
   conjugative and nonconjugative, 272, 273, 277–8bx

- copy number, 273-5
- Plectonemic joint, 232
- Pleiotropy, 2, 103
- Ploidy, 45, 133-4
- Point mutation, 202
- Polar mutation, 110, 319
- Poly(A) tail, 413
- Poly(A)+ RNA, 334, 414
- Polyadenylate polymerase, 408, 413
- Polyadenylation, 413
- Polycistronic mRNA, 104, 266-8
- Polygenic theory, 12
- Polymerase chain reaction (see PCR)
- Polymorphism, 201, 212
- Polypeptides, 287-9
  - chemical bonds in, 287-9f
  - N→C polarity, 287-9
- Polyploidy, 46-7
- Polyproteins, 105
- Polyteny, polytene
  - chromosomes, 47, 58-59
- Position effect variegation, 42
- Position effects, 45, 207
- Positional cloning, 335-6
- Positional information, 65, 72
- Positive-negative dual selection, 351-2f
- Post-replicative mismatch repair, 94, 194-5
- Potency, 72
- POU domain, 238-9, 246
- Precise excision, of
  - transposons, 172
- Preformation, 65
- Prime plasmid, 272
- Primer extension, 38-9t, 340f
- Primers
  - arbitrary, 281
  - for PCR, 279
- Priming strategies, 405t
- Primosome, 397
- Prion hypothesis, 479-81
- Programmed misreading, 319
- Programmed mutation, 212, 218-220bx
- Promiscuous DNA, 268
- Promoter, 443
  - structure of bacterial, 446f
  - structure of eukaryote, 448-9, 446f
- Promoter clearance, 443
- Promoter trap, 345
- Proofreading
  - in replication, 183-5
  - in translation, 129
- Prophage, 468
- Prosthetic group, 292
- Protein degradation, 310
- Protein families, 297-304
  - chimeric, 287, 300-303
  - conventional, 297-300
  - selective expansion, 303-4
- Protein folding, 293-4
- Protein modification, 295-7t
- Protein polymorphisms, 150, 211
- Protein secretion, 309
- Protein sequencing, 307
- Protein synthesis, 313-322
  - elongation cycle, 317-8
  - initiation, 130, 316-7f
  - overview of mechanism, 315-316
  - regulation, 318-20
  - termination, 318f
- Protein targeting, 309-10bx
  - to organelles, 310
  - signal sequences, 308-9
  - to membranes, 310
- Protein truncation test, 337
- Protein-nucleic acid
  - interaction, 232, 235-252, 343
  - characterization of, 249-52
  - direct binding, 244
  - modulation of tertiary structure, 245
  - ordered water molecules, 245
  - protein dimerization, 244
  - sequence specificity, 246-9
- Proteins, 287-310
  - primary structure, 287-9
  - quaternary structure, 292
  - secondary structure, 289-90
  - structural determination, 307-8bx
  - tertiary structure, 290-5
- Proteome, 288, 304
- Proto-oncogenes, 254-5
- Prototroph, 215
- Provirus, 468
- Pseudoautosomal inheritance, 4, 15-16
- Pseudoautosomal region, 4, 57
- Pseudogene, 107, 139
  - nonprocessed, 140
  - processed, 140, 166, 182
- Pseudoknot, 232
- Puffer fish, use in genome analysis, 155
- Pulsed field gel electrophoresis, 355-6
- Quantitative inheritance, 11-15
- Quantitative PCR, 284
- Quantitative trait loci, 150-1
- Quenching, 455
- R-banding, 59
- RACE, 284
- Radiation hybrid mapping, 146-7f
- Raf, *raf* gene, 254-8, 432
- RAG proteins, 381
- Ramachandran plot, 288-9f
- Random genetic drift, 212-3
- Random mutagenesis, 348
- RAPD markers, 150, 281
- Ras, *ras* gene, 254-8, 432
- RB1 gene, 29, 33, 258-60
- Reading frame, 127-8f
- Readout, of DNA sequences
  - by proteins, 246-9
- Readthrough (protein synthesis), 319
- Readthrough mutation, 204
- Reassociation kinetics, 135, 157-159bx
- RecA
  - in recombination, 384bx
  - in SOS mutagenesis, 197-8
- RecBCD nuclease, 372
- RecBCD pathway, 372
- Receptor serine/threonine kinases, 430-1t
- Receptor tyrosine kinases, 427-8f
- Recessive, 1-2, 209
- Reciprocal cross, 4
- Recombinant selection, 330-331
- Recombinant DNA, 323, 370
- Recombination fraction, 148
- Recombination frequency, 147
- Recombination
  - hotspots and coldspots, 161
  - as a repair mechanism, 197, 371
  - different classes, 369-387, 370t
  - molecular basis, 384
- Recombination signal sequence, 380
- Redundancy, structural and functional, 140
- Refolding model of prion replication, 480f
- Regional specification, 65
- Rel family, 242-3, 256
- Relaxed control, relaxed plasmid, 272, 277

- Release factor, 318
- Reovirus, 475
- Rep protein  
bacterial helicase, 398  
in plasmid replication, 276-7
- Repair deficiency syndromes, 189t
- Repairosome, 193
- Repetitive DNA, 133, 136, 137, 141f  
mutations involving, 207-8  
transposable elements as, 141-2
- Replication, 389-409  
fidelity of, 183-5  
initiation, 400-404  
linear genomes, 405-6t  
modes of, 401t  
regulation, 29-30, 406-7  
termination, 404-6
- Replication banding, 59
- Replication fork, 390
- Replication intermediates, 409
- Replication slipping, 184, 208f
- Replication time zone, 406-7
- Replication-deficient mutants, 392-3t
- Replicative form, 392, 473
- Replicative intermediates, 392
- Replication, 118-9f
- Replisome, 392-400  
prokaryote and eukaryote compared, 401t
- Reporter genes, 342-3t
- Reporter vectors, 343
- Representational difference analysis, 283
- Resolution, of cointegrate, 169, 386
- Response elements, 458
- Restriction enzymes/  
endonucleases, 93, 323-6, 353-5bx
- Restriction fragment length polymorphism, 150, 211
- Restriction map, 153-4, 336-8f
- Restriction point, 25
- Retinoblastoma, 258-9
- Retinoblastoma protein, 29, 33, 258-60
- Retroelements, 180-182
- Retrons, 182
- Retroposition, 169-70, 171f
- Retroposons, 182
- Retroregulation, 423, 484
- Retrotransposition, 169-170, 171f
- Retrotransposons, 180-1
- Retroviruses  
generation of transgenic animals, 364  
life cycle, 169-170, 171f, 468, 475, 476f  
mechanism of integration, 169-170, 171f
- Rev protein, 421, 485
- Reverse genetics, 213
- Reverse hybridization, 358
- Reverse transcriptase, 407, 169-70, 283-4, 333, 475-6
- Reverse transcriptase PCR, 283-4, 339
- Reversion, 209
- $\rho$ -dependent termination, 460
- Ribbon-helix-helix motif, 242
- Ribosomal proteins, 314
- Ribosomal RNA (rRNA), 224, 314  
genes, 58, 136, 139, 300  
processing, 412
- Ribosome binding site, 313
- Ribosomes, 314
- Ribozymes, 224, 232
- Ring chromosome, 51
- RNA, 223-4
- RNA editing, 266-8, 421, 422t
- RNA export, 421
- RNA polymerases, 407, 443  
bacterial, 445-6  
eukaryotic, 447t
- RNA polymerase II, C-terminal domain, 449
- RNA processing, 411-24, 411t
- RNA splicing, 414-421  
nuclear introns, 414-6
- RNA stability, regulation of, 422-4
- RNA structure, 228f, 230-1  
recognition by proteins, 236-7
- RNA targeting, 421
- RNA tumor viruses, 254, 256, 258
- RNA viruses, 470, 474-5  
protein synthesis in eukaryotes, 477t
- RNA-binding motifs in proteins, 237t, 243-4
- RNP domain, 243
- Rot analysis, 157-159bx
- RT-PCR, 283-4, 339
- Runaway replication, 274-5
- S phase, 23
- SAP kinase, 433-4
- Satellite association, 58
- Satellite DNA, 137, 142
- Satellite nucleic acids, 479
- Satellite region, 58
- Satellite viruses, 479
- Scanning hypothesis, 317, 319
- Scanning mutagenesis, 347
- SCID, 381
- Second messengers, 434-440
- Second site mutations, 220-1bx
- Segment polarity genes, 73, 83
- Segregation, equal, 3
- Selection pressure, 213
- Selection  
in genetic screens, 214  
in molecular cloning, 330-1  
natural and artificial, 17-18bx, 212-3
- Selenocysteine, 130
- Selenocysteine insertion sequence, 130
- Selenoproteins, 130
- Self-splicing introns, 266-8, 416-8
- Selfish DNA, 137
- Sense codon, 127
- Sense strand, 444
- Sequence homogenization, 287, 299, 300
- Sequence tagged sites, 153
- Serial analysis of gene expression (SAGE), 304
- Sex determination, 75-6t
- Sex-chromosomes, 57
- Sex-linkage, sex-linked inheritance, 4-5f, 15-16bx
- Sexduction, 119, 125bx
- SH2 domain, 432, 435, 441
- Shapiro intermediate, 168f
- Shine-Dalgarno sequence, 316
- Short tandem repeat polymorphism, 150
- Shotgun sequencing, 163
- Shuttle vector, 327, 341-2, 360-1
- Sib pair analysis, 149
- Sickle cell anemia/trait, 212-3, 216-7bx
- $\sigma$ -factor, 445
- Signal joint, 380
- Signal transduction, 425-42  
immune system, 429-30
- Silencers, 377, 454, 452
- Silent mutation, 204
- Simple sequence DNA, 142
- Simple sequence length polymorphisms, 150, 211
- SINEs, 96, 142, 144
- Single stranded (DNA)  
binding protein, 295, 399

- Single stranded DNA viruses, 473-4
- SIR proteins, 377, 454
- Sister chromatid exchange, 54, 376
- Site-directed mutagenesis, 346-7f
- Site-specific recombination, 370, 378-9, 385-6  
in transgenic organisms, 353
- Slow transforming retroviruses, 258
- SMAD family, 431
- Small nuclear RNA (snRNA), 224  
role in splicing, 415-6
- Somatic cell hybrids, 152
- Somatic hypermutation, 219-20
- Somatic mutation, 201, 253-8
- Somatic recombination, 379
- Sonic hedgehog, 91-92, 138
- SOS mutagenesis, 197-8
- SOS response, 197  
induction of bacteriophage, 198-9
- Southern blot/hybridization, 357-8
- Southwestern screen, 249-50, 358
- Specification  
autonomous, 68f  
conditional, 68f  
in development, 71
- Spheroplasts, 361
- Spliced leader RNA, 420
- Spliceosome, 415-6
- Splicing factors, 416
- Sporulation in *Bacillus subtilis*, 78-79bx
- Squelching, 453
- SR proteins, 416
- SRY, 75
- SSCP analysis, 37
- START, 25, 28-29
- STATs, 429, 4421
- Stem cell, 72
- Stem-loop, 230-1
- Steroid receptors, 431
- Stop codon, 127
- Strand slipping (*see* Replication slipping)
- Stringency, 357
- Stringent control, stringent plasmids, 272, 277
- Substitution fidelity, 183-4
- Subtractive hybridization, 336
- Subviral agents, 467, 477-81
- Suicide enzyme, 190
- Supercoiling, 232, 233-4bx
- Superfamilies, 297
- Suppression, 10t, 220-1bx
- SV40 472
- Svedberg units, 320
- Switch regions, 382
- Synonymous mutation, 204
- Synteny, 374
- T-cell receptor genes, 386-7
- T-cell receptors, 386-7
- T-vector, 283
- Target site duplication, 167f
- Targeted mutation  
by homologous recombination, 352  
spontaneous, 185
- Targeting vectors, 351-2, 360
- Tat-*Tar*, 421, 485
- TATA box, 448
- TATA-binding protein (TBP), 242, 447
- Tautomeric shifts, 184
- Tautomers, 184, 223
- TBP-associated factors (TAFs), 242, 449
- Telomerase, 61-62f
- Telomeres, 61-62, 405
- Temperate phage, 468
- Terminal deoxynucleotidyl transferase, 324, 408
- Terminal redundancy, 472
- Termination codon, 127
- Tetrad analysis, 160
- TFIID, 449
- TFIIH, 193, 449
- Thalassemias, 216-7
- Thermal melting profile of DNA, 357
- 30 nm fiber, 38
- 3T3 cell assay, 255
- Threshold character, 11
- Ti plasmid, 273, 349, 362-3
- Ti vectors, 362-3
- Tn3 resolvase, 386
- Topoisomerase, 397, 398t
- TP53 gene (*see* p53)
- Trans-acting factors, 113, 452
- Trans-sensing, 7, 114
- Trans-splicing, 266-8, 420
- Transcribed spacer sequences, 106-7
- Transcript analysis, 338-9t
- Transcription, 443  
initiation in bacteria, 445-7  
initiation in eukaryotes, 447-50  
principles, 444-5
- Transcription factors,  
activation by signaling proteins, 440-1, 455-6  
activation domains, 464-5bx  
combinatorial activity, 456  
constitutive, 448  
dimerization domains, 244  
DNA-binding domains, 237-43  
examples and binding sites, 459t  
mechanism of action, 451-5  
regulation of activity, 455-6
- Transcription unit, 106
- Transcription-coupled repair, 192-4
- Transcriptional arrest, 458
- Transcriptional elongation,  
regulation of, 458-9
- Transcriptional initiation  
bacteria, 445-7  
RNA polymerase I, 448  
RNA polymerase II, 448-9  
RNA polymerase III, 450
- Transcriptional map, 155
- Transcriptional regulation  
bacteria, 456-8  
 $\beta$ -globin gene, 462-3bx  
eukaryotes, 458-9  
*lac* operon, 463-4bx
- Transcriptional termination, 460
- Transduction, 120-123  
generalized, 120, 122  
specialized, 122-3
- Transfection  
of eukaryotic cells, 348, 349t, 361  
of bacteria, 118
- Transfer RNA (tRNA), 224, 314-315  
adaptor role in protein synthesis, 314-5  
charging, 127-129, 315  
processing, 411-2  
structure, 314-5, 321-2bx
- Transfer RNA introns, 418
- Transformation  
of bacteria, natural, 118, 119-20  
artificial of *E. coli*, 330
- Transforming growth factor- $\beta$  superfamily, 430-1
- Transgenesis, 348-353  
inducible transgene activity, 353  
random integration of DNA, 350-2



- site specific recombination systems, 353
- Transgenic animals and plants, 348–9
- Transition, 202
- Translation, 127–129
- Translation, regulation (*see* Protein synthesis)
- Translesion synthesis, 198, 394–5
- Translocation
  - chromosome mutation, 51–53f
  - in protein synthesis, 318
  - reciprocal, 52–53f
  - Robertsonian/whole arm, 52–53
- Transmissible spongiform encephalopathy, 479–81
- Transplacement, 350–1f
- Transposable elements, 165
  - class I (*see* Retroelements)
  - class II (*see* Transposons)
  - classification, 175–182
  - uses, 175, 176t
- Transposase, 165
- Transposition, 165, 370
  - aberrant, 174
  - consequences of, 172–5
  - conservative and replicative, 167–169
  - mechanisms, 166–70
  - regulation, 170–172
- Transposon tagging, 176, 344
- Transposons, 175–180
- Transvection, 114
- Transversion, 202
- Triple helix, 231
- Triple helix therapy, 365–6
- Triplet, 127
- Triplet repeat syndromes, 218bx
- Tumor suppressor genes, 253, 258–60t
- Tumor viruses, 469
- Turner's syndrome, 48
  - 12–23 rule, 380
- 2–D electrophoresis, 307, 356
- Two hybrid system, 305, 345
- Ubiquitin, in protein degradation, 310
- Unassigned reading frame, 127
- Unequal crossover, 207
- Unequal exchange, 207, 208f
- Unequal sister chromatid exchange, 207
- Unique sequence DNA, 133, 136
- Unstable mutant alleles, 172
- Untranslated region, 106
- UV-induced DNA lesions, 185–90
- UvrABC nuclease, 185–90
- V-gene segments, 387
- Variation, phenotypic, 1, 2
- Variegation, 54
- V(D)J recombination, 380
- Vectorette PCR, 284
- Very short patch mismatch repair, 196–8
- Viral oncogenes, 254, 256t
- Virion, 467
- Viroids, 273, 478
- Virus, 467
- Viruses
  - as gene transfer vectors, 349
  - infection strategy, 468–9
  - regulation of gene expression, 475–7
  - replication strategy, 469–75
- Virusoid, 479
- VNTR sequences, 142, 150
- Vulval specification, *Caenorhabditis elegans*, 85–86bx
- Wee mutants, 26
- Western blot, 307
- Whole genome duplication, 299–300
- Wild type, 201
- Winged helix, 239
- Wingless, 87–89bx
- Wnt signaling proteins
  - Drosophila* Wingless, 87–89bx
  - in vertebrate limb development, 91–92bx
- Wobble hypothesis, wobble rules, 129, 227
- X-chromosome inactivation, 5, 100–101
- X-inactivation center, 100
- X-linked inheritance, 4–5f, 15–16bx
- X-ray crystallography, 307–8
- Xeroderma pigmentosum, 189
- XIST gene, 99–101
- Y-linked inheritance, 5, 15–16bx
- YAC transgenic mice, 351–2
- Yeast artificial chromosome (YAC), 153, 327–9, 361
- Yeast cloning vectors, 360–1bx
- Z-DNA, 229, 230f
- Zinc binuclear cluster, 241
- Zinc finger
  - Cys<sub>2</sub>-His<sub>2</sub>, 239–40f
  - multicysteine, 240–1
  - DNA recognition code, 248–9t
- Zoo blot, 155, 358
- Zygotic genes, *Drosophila*, 83–85