MOLECULAR BIOLOGY

DAVID CLARK
Southern Illinois University

ELSEVIER
ACADEMIC
PRESS

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

MOLECULAR BIOLOGY

Senior Acquisitions Editor	Jeremy Hayhurst
Project Manager	Kyle Sarofeen
Editorial Assistant	Desiree Marr
Marketing Manager	Linda Beattie
Cover Design	Eric DeCicco
Composition	SNP Best-set Typesetter Ltd., Hong Kong
Cover Printer	Hing Yip Printing Co., Ltd.
Interior Printer	Hing Yip Printing Co., Ltd.

Elsevier Academic Press
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper. ∞

Library of Congress Cataloging-in-Publication Data
Application submitted.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 0-12-175551-7

For all information on all Elsevier Academic Press Publications
visit our Web site at www.books.elsevier.com

Printed in China
05 06 07 08 09 10 9 8 7 6 5 4 3 2 1

Dedication

This book is dedicated to Lonnie Russell who was to have been my coauthor. A few months after we started this project together, in early July 2001, Lonnie drowned in the Atlantic Ocean off the coast of Brazil in a tragic accident.

Preface

This book's subtitle, *Understanding the Genetic Revolution,* reflects the massive surge in our understanding of the molecular foundations of genetics in the last fifty years. In the next half century our understanding of how living organisms function at the molecular level, together with our ability to intervene, will expand in ways we are only just beginning to perceive.

Today we now know that genes are much more than the abstract entities proposed over a century ago by Mendel. Genes are segments of DNA molecules, carrying encoded information. Indeed, genes have now become chemical reagents, to be manipulated in the test tube. In the days of classical genetics genes represented inherited characteristics but were themselves inviolate, rather like atoms before the twentieth century. Today both genes and atoms have sub-components to be tinkered with.

A full understanding of how living organisms function includes an appreciation of how cells operate at the molecular level. This is of vital importance to all of us as it becomes ever more clear that molecular factors underlie many health problems and diseases. While cancer is the "classic" case of a disease that only became understandable when its genetic basis was revealed, it is not the only one by any means. Today the molecular aspects of medicine are expanding rapidly and it will soon be possible to personally tailor clinical treatment by taking into account the genetic make-up of individual patients.

Rather than attempting to summarize my view of modern molecular biology (the book itself, I hope, accomplishes that) in a short preface, I'd like to briefly address what this book is not. It is not intended as a reference work for faculty or researchers but rather as a survey-oriented textbook for upper division students in a variety of biological sub-disciplines. In particular it is intended for final year undergraduates and beginning graduate students.

This book does not attempt to be exhaustive in its coverage, even as a textbook. There is a second book in this series (Biotechnology: *Applying the Genetic Revolution*, 2006) co-authored with Nanette Pazdernik, which essentially picks up where this book ends. Both books, I hope, effectively survey the foundations and applications of modern molecular genetics.

Many, perhaps most, of the students using this book will be well versed in the basics of modern genetics and cell biology and so can pick and choose from the topics covered as needed. However, others will not be so well prepared, due in part to the continuing influx into molecular biology of students from related disciplines. For them I've tried to create a book whose early chapters cover the basics, before launching out into the depths.

Because of the continuing interest in applying molecular biology to an ever widening array of topics, I have tried to avoid overdoing detail (depth) in favour of breadth. This in no way minimizes the importance of the subject matter for cell biologists but instead emphasizes that molecular biology is applicable to more than just human medicine and health. The genetic revolution has also greatly impacted other important areas such as agriculture, veterinary medicine, animal behaviour, evolution, and microbiology. Students of these, and related disciplines, all need to understand molecular biology at some level.

Finally there are no references or extra reading at the ends of the chapters, for two reasons. My own cross-questioning has revealed that neither myself nor most of my colleagues and students have ever actually used such textbook references just as we rarely watch the extra material on DVDs providing actors' insights, extra scenes, and outtakes. The student has enough to deal with in the core material.

Secondly, anyone who wants up-to-date reference material is far better advised to run a web search. PubMed, Google Scholar and Scirus.com are good choices.

Feedback (hopefully positive!) is welcome.

David Clark, Carbondale, Illinois, January 2005

Acknowledgements

I would like to thank the following individuals for their help in providing information, suggestions for improvement and encouragement: Laurie Achenbach, Rubina Ahsan, Phil Cunningham, Michelle McGehee, Donna Mueller, Dan Nickrent, Joan Slonczewski. Especial thanks go to Nanette Pazdernik for help in editing many of the chapters and to Karen Fiorino for creating most of the artwork.

Introduction

Molecular Genetics Is Driving the Biotechnology Revolution

Although the breeding of plants and animals goes back thousands of years, only in the last couple of centuries has genetics emerged as a field of scientific study. Classical genetics emerged in the 1800s when the inheritance patterns of such things as hair or eye color were examined and when Gregor Mendel performed his famous experiments on pea plants. Techniques revealing how the inherited characteristics that we observe daily are linked to their underlying biochemical causes have only been developed since World War II. The resulting revelation of the molecular basis of inheritance has resulted in the increasing use of the term "molecular." Often the term "molecular biology" refers to the biology of those molecules related to genes, gene products and heredity—in other words, the term molecular biology is often substituted for the perhaps more appropriate term, **molecular genetics**. A more broad-minded definition of molecular biology includes all aspects of the study of life from a molecular perspective. Although the molecular details of muscle operation or plant pigment synthesis could be included under this definition, in practice, textbooks are limited in length. In consequence, this book is largely devoted to the molecular aspects of the storage and transmission of biological (i.e., genetic) information.

Although there is great diversity in the structures and lifestyles of living organisms, viewing life at the molecular level emphasizes the inherent unity of life processes. Perhaps it is this emergent unity, rather than the use of sophisticated molecular techniques, that justifies molecular biology as a discipline in its own right. Instead of an ever-expanding hodge-podge of methods for analyzing different organisms in more and more detail, what has been emerged from molecular analysis is an underlying theme of information transmission that applies to all life forms despite their outward differences.

Society is in the midst of two scientific revolutions. One is in the realm of technology of information, or computers, and the other in molecular biology. Both are related to the handling of large amounts of encoded information. In one case the information is man made, or at any rate man-encoded, and the mechanisms are artificial; the other case deals with the genetic information that underlies life. Biology has reached the point where the genes that control the makeup and functioning of all living creatures are being analyzed at the molecular level and can be altered by genetic engineering. In fact, managing and analyzing the vast mass of genetic information constantly emerging from experimentation requires the use of sophisticated software and powerful computers. The emerging information revolution rivals the industrial revolution in its importance, and the consequences of today's findings are already changing human lives and will continue to alter the lives of future generations. Data is accumulating about the molecules of inheritance and how they are controlled and expressed at an ever faster and faster pace. This is largely due to improved techniques, such as **PCR** (*p*olymerase *c*hain *r*eaction; see Ch. 23) and **DNA** (deoxyribonucleic acid) arrays (see Ch. 25). In particular, methods have recently been developed for the rapid, simultaneous and automated analysis of multiple samples and/or multiple genes.

One major impact of molecular biology is in the realm of human health. The almost complete sequence of the DNA molecules comprising the human genome was revealed in the year 2003. So, in theory, science has available all of the genetic information needed to make a human being. However, the function of most of a human's approximately 35,000 genes remains a mystery. Still more complex is the way in which the expression of these genes is controlled and coordinated. Inherited diseases are due to defective versions of certain genes or to chromosomal abnormalities. To understand why defective genes cause problems, it is important to investigate the normal roles of these genes. As all disease has a genetic component, the present trend is to redefine physical and mental health from a genetic perspective. Even the course of an infectious disease depends to a significant extent on built-in host responses, which are determined by host genes. For example, humans with certain genetic constitutions are at much greater risk than others of getting SARS, even though this is an emerging disease that only entered the human population in the last few years. The potential is present to improve health and to increase human and animal life spans by preventing disease and slowing the aging process. Clinical medicine is changing rapidly to incorporate these new findings.

The other main arena where biotechnology will have a massive impact is agriculture. New varieties of genetically engineered plants and animals have already been made and some are in agricultural use. Animals and plants used as human food sources are being engineered to adapt them to conditions which were previously unfavorable. Farm animals that are resistant to disease and crop plants that are resistant to pests are being developed in order to increase yields and reduce costs. The impact of these genetically modified organisms on other species and on the environment is presently a controversial issue.

Table of Contents

Detailed Contents

CHAPTER 4 *Genes, Genomes and DNA* 75

CHAPTER 5 *Cell Division and DNA Replication* 103

CHAPTER 6 *Transcription of Genes* 132

CHAPTER 7 *Protein Structure and Function* 154

CHAPTER 16 *Plasmids* 425

CHAPTER 17 *Viruses* 453

CHAPTER 18 *Bacterial Genetics* 484

CHAPTER 19 *Diversity of Lower Eukaryotes* 508

CHAPTER 20 *Molecular Evolution* 533

CHAPTER 21 *Nucleic Acids: Isolation, Purification, Detection, and Hybridization* 567

CHAPTER 22 *Recombinant DNA Technology* 599

CHAPTER ONE

Basic Genetics

Gregor Mendel Was the Father of Classical Genetics

From very ancient times, people have vaguely realized the basic premise of heredity. It was always a presumption that children looked like their fathers and mothers, and that the offspring of animals and plants generally resemble their ancestors. During the 19th century, there was great interest in how closely offspring resembled their parents. Some early investigators measured such quantitative characters as height, weight, or crop yield and analyzed the data statistically. However, they failed to produce any clear-cut theory of inheritance. It is now known that certain properties of higher organisms, such as height or skin color, are due to the combined action of many genes. Consequently, there is a gradation or quantitative variation in such properties. Such multi-gene characteristics caused much confusion for the early geneticists and they are still difficult to analyze, especially if more than two or three genes are involved.

A century before the discovery of the DNA double helix, Mendel realized that inheritance was quantized into discrete units we now call genes.

The birth of modern genetics was due to the discoveries of **Gregor Mendel** (1823–1884), an Augustinian monk who taught natural science to high school students in the town of Brno in Moravia (now part of the Czech Republic). Mendel's greatest insight was to focus on discrete, clear-cut characters rather than measuring continuously variable properties, such as height or weight. Mendel used pea plants and studied characteristics such as whether the seeds were smooth or wrinkled, whether the flowers were red or white, and whether the pods were yellow or green, etc. When asked if any particular individual inherited these characteristics from its parents, Mendel could respond with a simple "yes" or "no," rather than "maybe" or "partly." Such clear-cut, discrete characteristics are known as **Mendelian characters** (Fig. 1.01).

Today, scientists would attribute each of the characteristics examined by Mendel to a single **gene**. Genes are units of genetic information and each gene provides the instructions for some property of the organism in question. In addition to those genes that affect the characteristics of the organism more or less directly, there are also many regulatory genes. These control other genes, hence their effects on the organism are less direct and more complex. Each gene may exist in alternative forms known as **alleles**, which code for different versions of a particular inherited character (such as red versus white flower color). The different alleles of the same gene are closely related, but have minor chemical variations that may produce significantly different outcomes.

The overall nature of an organism is due to the sum of the effects of all of its genes as expressed in a particular environment. The total genetic make-up of an organism is referred to as its **genome**. In lower organisms such as bacteria, the genome may consist of approximately 2,000 to 6,000 genes, whereas in higher organisms such as plants and animals, there may be up to 50,000 genes.

Etymological Note

Mendel did not use the word "gene." This term entered the English language in 1911 and was derived from the German "Gen," short for "Pangen." This in turn came via French and Latin from the original ancient Greek "genos," which means birth. "Gene" is related to such modern words as genus, origin, generate, and genesis. In Roman times, a "genius" was a spirit representing the *inborn power* of individuals.

allele One particular version of a gene
gene A unit of genetic information
genome The entire genetic information of an individual organism
Gregor Mendel Discovered the basic laws of genetics by crossing pea plants
Mendelian character Trait that is clear cut and discrete and can be unambiguously assigned to one category or another

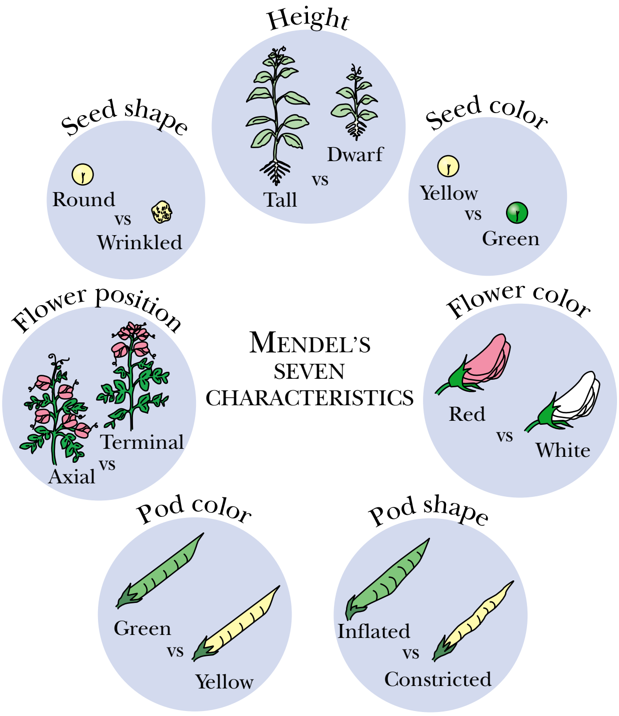


FIGURE 1.01 ***Mendelian Characters in Peas***

Mendel chose specific characteristics, such as those shown. As a result he obtained definitive answers to whether or not a particular characteristic is inherited.

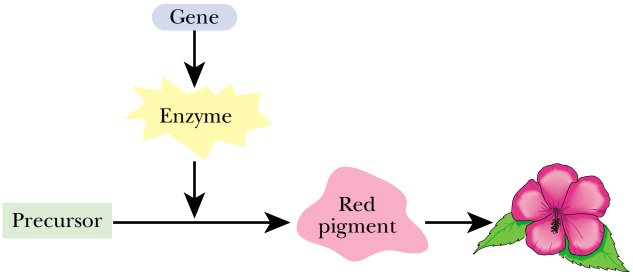


FIGURE 1.02 ***One Gene—One Enzyme***

A single gene determines the presence of an enzyme which, in turn, results in a biological characteristic such as a red flower.

Genes Determine Each Step in Biochemical Pathways

Beadle and Tatum linked genes to biochemistry by proposing there was one gene for each enzyme.

Mendelian genetics was a rather abstract subject, since no one knew what genes were actually made of, or how they operated. The first great leap forward came when biochemists demonstrated that each step in a biochemical pathway was determined by a single gene. Each biosynthetic reaction is carried out by a specific **protein** known as an **enzyme**. Each enzyme has the ability to mediate one particular chemical reaction and so the *one gene—one enzyme* model of genetics (Fig. 1.02) was put forward by G. W. Beadle and E. L. Tatum, who won a Nobel prize for this scheme in 1958. Since then, a variety of exceptions to this simple scheme have been found. For example, some complex enzymes consist of multiple subunits, each of which requires a separate gene.

A gene determining whether flowers are red or white would be responsible for a step in the biosynthetic pathway for red pigment. If this gene were defective, no red pigment would be made and the flowers would take the default coloration—white. It is easy to visualize characters such as the color of flowers, pea pods or seeds in terms of a biosynthetic pathway that makes a pigment. But what about tall versus dwarf plants and round versus wrinkled seeds? It is difficult to interpret these in terms of a single pathway and gene product. Indeed, these properties are affected by the action

Much of modern molecular biology deals with how genes are regulated. (See Chapters 9, 10 and 11.)

enzyme A protein that carries out a chemical reaction
protein A polymer made from amino acids; proteins make up most of the structures in the cell and also do most of the work

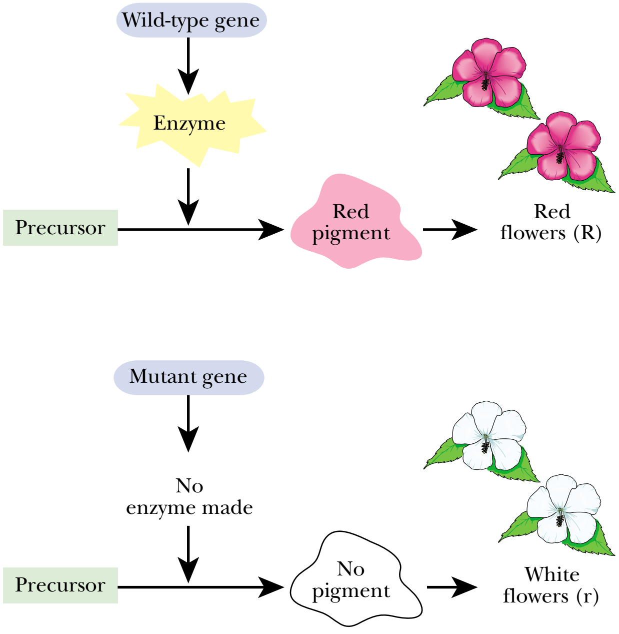


FIGURE 1.03 ***Wild-type and Mutant Genes***

If red flowers are found normally in the wild, the "red" version of the gene is called the wild-type allele. Mutation of the wild-type gene may alter the function of the enzyme so ultimately affecting a visible characteristic. Here, no pigment is made and the flower is no longer red.

of many proteins. However, as will be discussed in detail later, certain proteins control the expression of genes rather than acting as enzymes. Some of these **regulatory proteins** control just one or a few genes whereas others control large numbers of genes. Thus a defective regulatory protein may affect the levels of many other proteins. Modern analysis has shown that some types of dwarfism are due to defects in a single regulatory protein that controls many genes affecting growth. If the concept of "one gene—one enzyme" is broadened to "one gene—one protein," it still applies in most cases. [There are of course exceptions. Perhaps the most important is that in higher organisms multiple related proteins may sometimes be made from the same gene by alternative patterns of splicing at the RNA level, as discussed in Chapter 12.]

Mutants Result from Alterations in Genes

Genetics has been culturally influenced by idealized notions of a perfect "natural" or "original" state. Mutations tend to be viewed as defects relative to this.

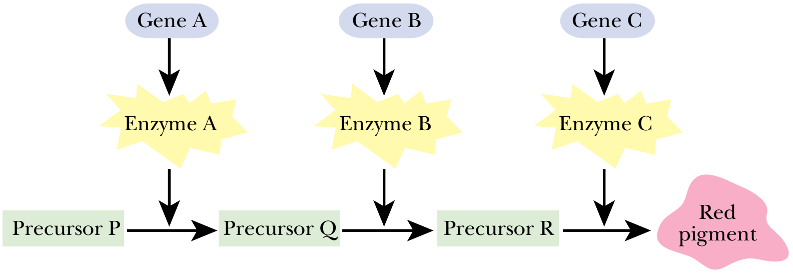
Consider a simple pathway in which red pigment is made from its precursor in a single step. When everything is working properly, the flowers shown in Figure 1.02 will be red and will match thousands of other red flowers growing in the wild. If the gene for flower color is altered so as to prevent the gene from functioning properly, one may find a plant with white flowers. Such genetic alterations are known as **mutations**. The white version of the flower color gene is defective and is a mutant allele. The properly functioning red version of this gene is referred to as the **wild-type** allele (Fig. 1.03). As the name implies, the wild-type is supposedly the original version as found in the wild, before domestication and/or mutation altered the beauties of nature. In fact, there are frequent genetic variants in wild populations and it is not always obvious which version of a gene should be regarded as the true wild-type. Generally, the wild-type is taken as the form that is common and shows adaptation to the environment.

Geneticists often refer to the red allele as "R" and the white allele as "r" (not "W"). Although this may seem a strange way to designate the color white, the idea is

mutation An alteration in the genetic information carried by a gene
regulatory protein A protein that regulates the expression of a gene or the activity of another protein
wild-type The original or "natural" version of a gene or organism

FIGURE 1.04 ***Three Step Biochemical Pathway***

In this scenario, genes A, B, and C are all needed to make the red pigment required to produce a red flower. If any precursor is missing due to a defective gene, the pigment will not be made and the flower will be white.



that the r-allele is merely a defective version of the gene for red pigment. The r-allele is NOT a separate gene for making white color. In our hypothetical example, there is no enzyme that makes white pigment; there is simply a failure to make red pigment. Originally it was thought that each enzyme was either present or absent; that is, there were two alleles corresponding to Mendel's "yes" and "no" situations. In fact, things are often more complicated. An enzyme may be only partially active or even be hyperactive or have an altered activity and genes may actually have dozens of alleles, matters to be discussed later. A mutant allele that results in the complete absence of the protein is known as a **null allele**. [More strictly, a null allele is one that results in complete absence of the gene product. This includes the absence of RNA (rather than protein) in the case of those genes where RNA is the final gene product (e.g. ribosomal RNA, transfer RNA etc)—see Chapter 3].

Phenotypes and Genotypes

Classical genetic analysis involves deducing the state of the genes by observing the outward properties of the organism.

In real life, most biochemical pathways have several steps, not just one. To illustrate this, extend the pathway that makes red pigment so it has three steps and three genes, called A, B, and C. If any of these three genes is defective, the corresponding enzyme will be missing, the red pigment will not be made, and the flowers will be white. Thus mutations in any of the three genes will have the same effect on the outward appearance of the flowers. Only if all three genes are intact will the pathway succeed in making its final product (Fig. 1.04).

Outward characteristics—the flower color—are referred to as the **phenotype** and the genetic make-up as the **genotype**. Obviously, the phenotype "white flowers" may be due to several possible genotypes, including defects in gene A, B, or C, or in genes not mentioned here that are responsible for producing precursor P in the first place. If white flowers are seen, only further analysis will show which gene or genes are defective. This might involve assaying the biochemical reactions, measuring the build-up of pathway intermediates (such as P or Q in the example) or mapping the genetic defects to locate them in a particular gene(s).

If gene A is defective, it no longer matters whether gene B or gene C are functional or not (at least as far as production of our red pigment is concerned; some genes affect multiple pathways, a possibility not considered in this analysis). A defect near the beginning of a pathway will make the later reactions irrelevant. This is known in genetic terminology as **epistasis**. Gene A is epistatic on gene B and gene C; that is, it masks the effects of these genes. Similarly, gene B is epistatic on gene C. From a practical viewpoint, this means that a researcher cannot tell if genes B or C are defective or not, when there is already a defect in gene A.

epistasis When a mutation in one gene masks the effect of alterations in another gene
genotype The genetic make-up of an organism
null allele Mutant version of a gene which completely lacks any activity
phenotype The visible or measurable effect of the genotype

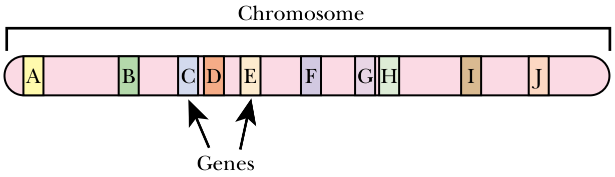


FIGURE 1.05 ***Genes Arranged along a Chromosome***

Although a chromosome is a complex three-dimensional structure, the genes on a chromosome are in linear order and can be represented by segments of a bar, as shown here. Genes are often given alphabetical designations in genetic diagrams.

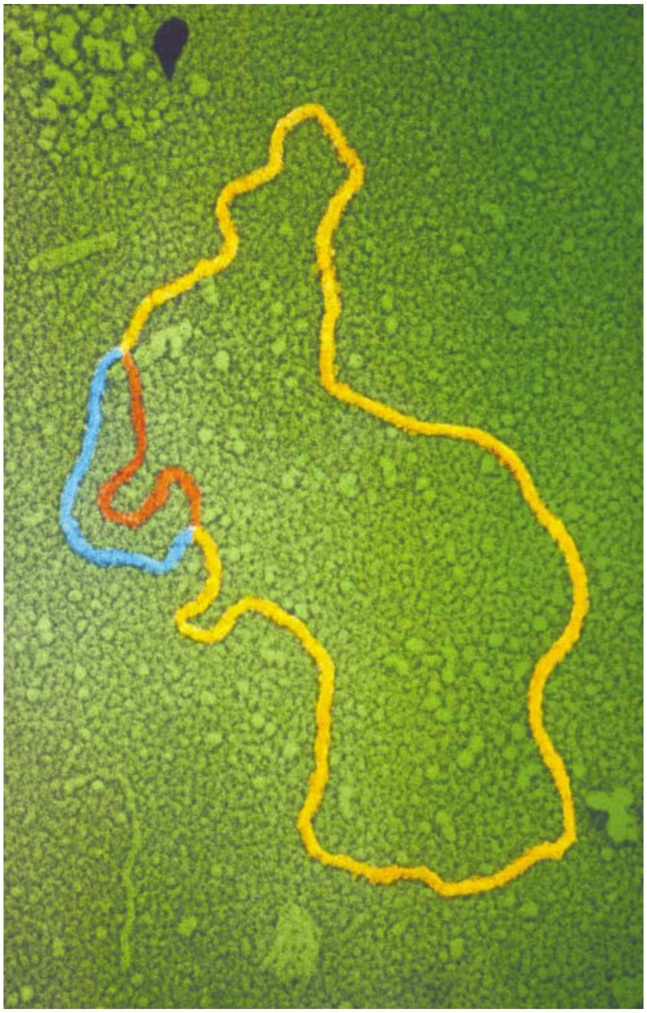


FIGURE 1.06 ***Circular DNA from a Bacterium***

Hand-tinted transmission electron micrograph (TEM) of circular bacterial DNA. This figure actually shows a small plasmid, rather than a full-size chromosome. The double-stranded DNA is yellow. An individual gene has been mapped by using an RNA copy of the gene. The RNA base pairs to one strand of the DNA forming a DNA/RNA hybrid (red). The other strand of the DNA forms a single-stranded loop, known as an "R-loop" (blue). Magnification: ×28,600. Courtesy of: P. A. McTurk and David Parker, Science Photo Library.

Chromosomes Are Long, Thin Molecules That Carry Genes

Genes are not mere abstractions. They are segments of DNA molecules carrying encoded information.

Genes are aligned along very long, string-like molecules called **chromosomes** (Fig. 1.05). Organisms such as **bacteria** usually fit all their genes onto a single circular chromosome (Fig. 1.06); whereas, higher, eukaryotic organisms have several chromosomes that accommodate their much greater number of genes. Genes are often drawn on a bar representing a chromosome (or a section of one), as shown in Figure 1.05.

One entire chromosome strand consists of a molecule of deoxyribonucleic acid, called simply DNA (see Ch. 3). The genes of living cells are made of DNA, as are the

bacteria Primitive single-celled organisms without a nucleus and with one copy of each gene
chromosome Structure containing the genes of a cell and made of a single molecule of DNA

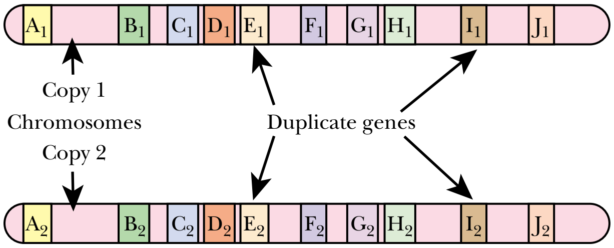


FIGURE 1.07 ***Genes Match on Each of a Pair of Homologous Chromosomes***

Higher organisms possess two copies of each gene arranged on pairs of homologous chromosomes. The genes of the paired chromosomes are matched along their length. Although corresponding genes match, there may be molecular variation between the two members of each pair of genes.

regions of the chromosome between the genes. In bacteria, the genes are closely packed together, but in higher organisms such as plants and animals, the DNA between genes comprises up to 96% of the chromosome and the functional genes only make up around 4 to 5% of the length. [Viruses also contain genetic information and some have genes made of DNA. Other viruses have genes made of the related molecule, **ribonucleic acid, RNA**.]

In addition to the DNA, the genetic material itself, chromosomes carry a variety of proteins that are bound to the DNA. This is especially true of the larger chromosomes of higher organisms where histone proteins are important in maintaining chromosome structure (see Ch. 4). [Bacteria also have histone-like proteins. However, these differ significantly in both structure and function from the true histones of higher organisms—see Chapter 9.]

Different Organisms may Have Different Numbers of Chromosomes

The cells of higher organisms usually contain two copies of each chromosome. Each pair of identical chromosomes possesses copies of the same genes, arranged in the same linear order. In Figure 1.07, identical capital letters indicate sites where alleles of the same gene can be located on a pair of chromosomes. In fact, identical chromosomes are not usually truly identical, as the two members of the pair often carry different alleles of the same gene. The term **homologous** refers to chromosomes that carry the same set of genes in the same sequence, although they may not necessarily carry identical alleles of each gene.

Different organisms differ greatly in the number of genes, the number of copies of each gene, and the arrangement of the genes on the DNA.

A cell or organism that possesses two homologous copies of each of its chromosomes is said to be **diploid** (or "2n", where "n" refers to the number of chromosomes in one complete set). Those that possess only a single copy of each chromosome are **haploid** (or "n"). Thus humans have 2 × 23 chromosomes (n = 23 and 2n = 46). Although the X and Y sex chromosomes of animals form a pair they are not actually identical (see below). Thus, strictly, a male mammal is not fully diploid. Even in a diploid organism, the reproductive cells, known as gametes, possess only a single copy of each chromosome and are thus haploid. Such a single, though complete, set of chromosomes carrying one copy of each gene from a normally diploid organism is known as its "**haploid genome**."

Bacteria possess only one copy of each chromosome and are therefore haploid. (In fact, most bacteria have only a single copy of a single chromosome, so that n = 1). If one of the genes of a haploid organism is defective, the organism may be seriously endangered since the damaged gene no longer contains the correct information that the cell needs. Higher organisms generally avoid this predicament by being diploid and having duplicate copies of each chromosome and therefore of each gene. If one copy of the gene is defective, the other copy may produce the correct product required by the cell. Another advantage of diploidy is that it allows recombination between two copies of the same gene (see Ch. 14). Recombination is important in promoting the genetic variation needed for evolution.

diploid Having two copies of each gene
haploid Having one copy of each gene
haploid genome A complete set containing a single copy of all the genes (generally used of organisms that have two or more sets of each gene)
homologous Related in sequence to an extent that implies common genetic ancestry
ribonucleic acid (RNA) Nucleic acid that differs from DNA in having ribose in place of deoxyribose and having uracil in place of thymine

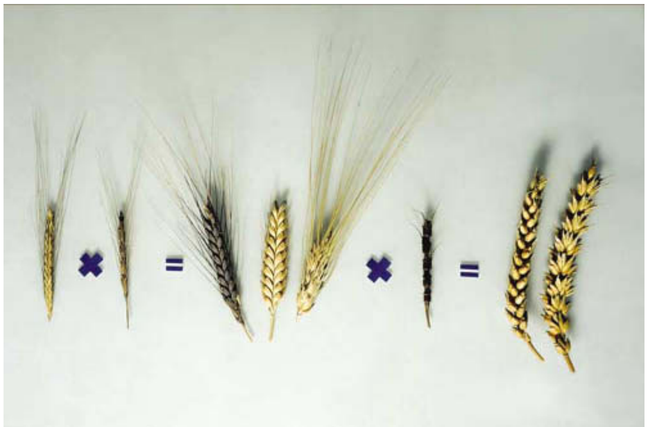


FIGURE 1.08 ***Diploid, Tetraploid and Hexaploid Wheats***

The origin of modern hexaploid bread wheat is illustrated. Einkorn wheat hybridized with goat grass to give tetraploid wheat. This in turn hybridized with the weed *Triticum tauschii* to give hexaploid bread wheat. The increase in grain yield is obvious. Courtesy of Dr. Wolfgang Schuchert Max-Planck Institute for Plant Breeding Research, Köln, Germany.

Note that haploid cells may contain more than a single copy of certain genes. For example, the single chromosome of *E. coli* carries two copies of the gene for the elongation factor EF-Tu and seven copies of the genes for ribosomal RNA. In haploid cells of the yeast *Saccharomyces cerevisiae* as many as 40% of the genes are duplicate copies. Strictly speaking, duplicate copies of genes are only regarded as genuine alleles if they occupy the same location on **homologous chromosomes**. Thus these other duplicate copies do not count as true alleles.

Occasionally, living cells with more than two copies of each chromosome can be found. **Triploid** means possessing three copies, **tetraploid** means having four copies, and so on. Animal and plant geneticists refer to the "**ploidy**" of an organism, whereas bacterial geneticists tend to use the term "**copy number**." Many modern crop plants are polyploids, often derived from hybridization between multiple ancestors. Such polyploids are often larger and give better yields. The ancestral varieties of wheat originally grown in the ancient Middle East were diploid. These were then displaced by tetraploids, which in turn gave way to modern bread wheat (*Triticum aestivum*) which is hexaploid (6n = 42) (Fig 1.08). Hexaploid bread wheat is actually a hybrid that contains four sets of genes from emmer wheat and two sets from the wild weed, *Triticum tauschii* (= *Aegilops squarrosa*). Emmer wheat is a tetraploid (4n = 28) derived from two diploid ancestors—einkorn wheat (*Triticum monococcum*) and a weed similar to modern goat grass (*Triticum speltoides* = *Aegilops speltoides*). A small amount of tetraploid wheat (*Triticum turgidum* and relatives) is still grown for specialized uses, such as making pasta.

Cases are known where there are fewer or more copies of just a single chromosome. Cells that have irregular numbers of chromosomes are said to be **aneuploid**. In higher animals, aneuploidy is often lethal for the organism as a whole, although certain aneuploid cells may survive in culture under some conditions. Although aneuploidy is usually lethal in animals, it is tolerated to a greater extent in plants. Nonetheless, in rare cases, aneuploid animals may survive. Thus, partial triploidy is the cause of certain human conditions such as Down syndrome, where individuals have an extra copy of chromosome #21. The presence of three copies of one particular chromosome is known as **trisomy**.

Dominant and Recessive Alleles

Consider a diploid plant that has two copies of a gene involved in making red pigment for flowers. From a genetic viewpoint, there are four possible types of individual plant;

aneuploid Having irregular numbers of different chromosomes
copy number The number of copies of a gene that are present
homologous chromosomes Two chromosomes are homologous when they carry the same sequence of genes in the same linear order
ploidy The number of sets of chromosomes possessed by an organism
tetraploid Having four copies of each gene
triploid Having three copies of each gene
trisomy Having three copies of a particular chromosome

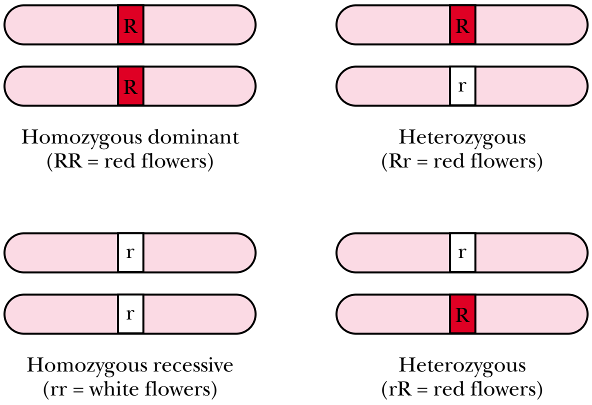


FIGURE 1.09 ***Two Different Alleles Produce Four Genotypes***

The genotypes R and r can be combined in four ways.

Genes and their alleles may interact with each other in a variety of ways. Sometimes one copy of a gene may predominate. In other cases both copies share influence.

that is, there are four possible genotypes: RR, Rr, rR and rr. The genotypes Rr and rR differ only depending on which of the pair of chromosomes carries r or R (see Fig. 1.09). When two identical alleles are present the organism is said to be **homozygous** for that gene (either RR or rr), but if two different alleles are present the organism is **heterozygous** (Rr or rR). Apart from a few exceptional cases there is no phenotypic difference between rR and Rr individuals, as it does not usually matter which of a pair of homologous chromosomes carries the r allele and which carries the R allele.

If both copies of the gene are wild-type, R-alleles (genotype, RR), then the flowers will be red. If both copies are mutant r-alleles (genotype, rr) then the flowers will be white. But what if the flower is heterozygous, with one copy "red" and the other copy "white" (genotype, Rr or rR)? The enzyme model presented above predicts that one copy of the gene produces enzyme and the other does not. Overall, there should be half as much of the enzyme, so red flowers will still be the result. Most of the time this turns out to be true, as many enzymes are present in levels that exceed minimum requirements. [In addition, many genes are regulated by complex feedback mechanisms. These may increase or decrease gene expression so that the same final level of enzyme is made whether there are two functional alleles or only one.]

From the outside, a flower that is Rr will therefore look red, just like the RR version. When two different alleles are present, one may dominate the situation and is then known as the **dominant** allele. The other one, whose properties are masked (or perhaps just function at a lower level), is the **recessive** allele. In this case, the R allele is dominant and the r allele recessive. Overall, three of the genotypes, RR (homozygous dominant), Rr (heterozygous) and rR (heterozygous), share the same phenotype and have red flowers, while only rr (homozygous recessive) plants have white flowers.

Partial Dominance, Co-Dominance, Penetrance and Modifier Genes

The assumption thus far is that one wild-type allele of the flower color gene will produce sufficient red pigment to give red flowers; in other words, the R-allele is dominant. Although one good copy of a gene is usually sufficient, this is not always the case. For example, the possession of only one functional copy of a gene for red pigment may result in half the normal amount of pigment being produced. The result may then be pale red or pink flowers. The phenotype resulting from Rr is then not the same as that seen with RR. This sort of situation, where a single good copy of a gene gives results that are recognizable but not the same as for two good copies, is known as **partial dominance**.

dominant allele Allele whose properties are expressed in the phenotype whether present as a single or double copy
heterozygous Having two different alleles of the same gene
homozygous Having two identical alleles of the same gene
partial dominance When a functional allele only partly masks a defective allele
recessive allele The allele whose properties are not observed because they are masked by the dominant allele

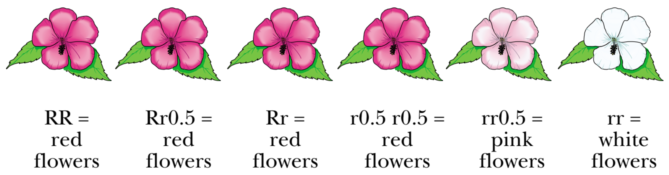


FIGURE 1.10 ***The Possible Phenotypes from Three Different Alleles***

There are six possible pairs of three different alleles. Here the r0.5 allele is a partly functional allele that makes only 50% of the normal pigment level. R is wild-type and r is null. The R R, R r0.5, R r and r0.5 r0.5 combinations will all make 100% or more of the wild type level of red pigment and so are red. The r r0.5 combination will make 50% as much pigment and so has pink flowers. The r r combination makes no pigment and so has white flowers.

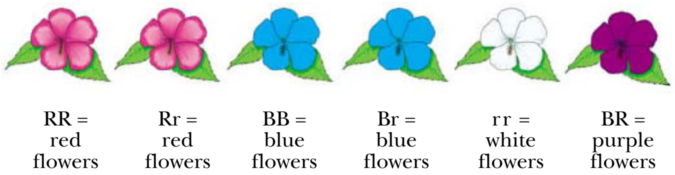


FIGURE 1.11 ***Phenotypes Resulting from Co-dominance***

Here the B allele makes an altered, blue, pigment. R is wild-type and r is null. The R R and R r combinations will make red pigment. The B r combination will make only blue pigment and the R B combination makes both red and blue pigments so has purple flowers.

As indicated above, there may be more than two alleles. In addition to the wild-type and null alleles, there may be alleles with partial function. Assume that a single gene dosage of enzyme is sufficient to make enough red pigment to give red flowers. Suppose there is an allele that is 50% functional, or "r0.5." Any combination of alleles that gives a total of 100% (= one gene dosage) or greater will yield red flowers. If there are three alleles, R = wild-type, r = null and r0.5 = 50% active, then the following genotypes and resulting phenotypes are possible (Fig. 1.10). In such a scenario, there are three different phenotypes resulting from six possible allele combinations.

Another possibility is alleles with altered function. For example, there may be a mutant allele that gives rise to an altered protein that still makes pigment but which carries out a slightly altered biochemical reaction. Instead of making red pigment, the altered protein could produce a pigment whose altered chemical structure results in a different color, say blue. Let's name this allele "B." Both R and B are able to make pigment and so both are dominant over r (absence of pigment). The combination of R with B gives both red and blue pigment in the same flower, which will look purple, and so they are said to be **co-dominant**. There are six possible genotypes and four possible phenotypes (colors, in this case) of flowers, as shown in Fig. 1.11.

A comlex and largely unresolved issue is that different versions of certain genes may behave differently in different individuals. Such individualized responses, especially to medication, have become a hot research topic.

As the above example shows, mutant alleles need not be recessive. There are even cases where the wild-type is recessive to a dominant mutation. Note also that a characteristic that is due to a dominant allele in one organism may be due, in another organism, to an allele that is recessive. For example, the allele for black fur is dominant in guinea pigs, but recessive in sheep. Note that a dominant allele receives a capital letter, even if it is a mutant rather than a wild-type allele. Sometimes a "+" is used for the wild-type allele, irrespective of whether the wild-type allele is dominant or recessive. A "–" is frequently used to designate a defective or mutant allele.

Does any particular allele always behave the same in each individual that carries it? Usually it does, but not always. Certain alleles show major effects in some individuals and only minor or undetectable effects in others. The term **penetrance** refers to the

co-dominance When two different alleles both contribute to the observed properties
penetrance Variability in the phenotypic expression of an allele

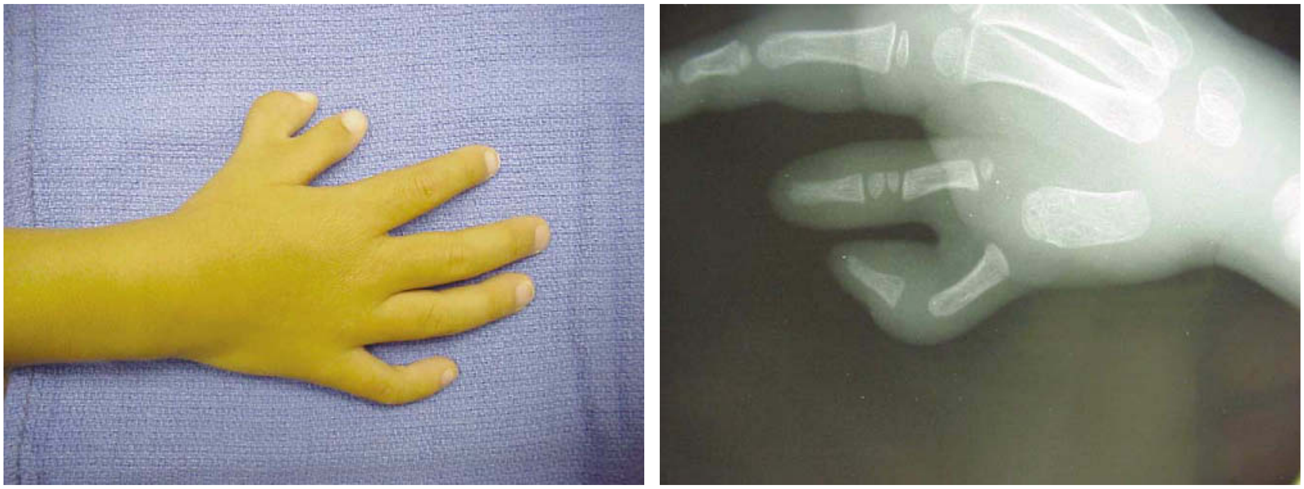


FIGURE 1.12 ***Polydactyly***

A dominant mutation may cause the appearance of extra fingers and/or toes.

relative extent to which an allele affects the phenotypic in a particular individual. Penetrance effects are often due to variation in other genes in the population under study.

In humans, there is a dominant mutation (allele = P) that causes polydactyly, a condition in which extra fingers and toes appear on the hands and feet (Fig. 1.12). This may well be the oldest human genetic defect to be noticed as the Bible mentions Philistine warriors with six fingers on each hand and six toes on each foot (II Samuel, Chapter 21, verse 20). About 1 in 500 newborn American babies shows this trait, although nowadays the extra fingers or toes are usually removed surgically, leaving little trace. Detailed investigation has shown that heterozygotes (Pp or P+) carrying this dominant allele do not always show the trait. Furthermore, the extra digits may be fully formed or only partially developed. The P allele is thus said to have variable penetrance.

Such variation in the expression of one gene is often due to its interaction with other genes. For example, the presence of white spots on the coat of mice is due to a recessive mutation, and in this case, the homozygote with two such recessive alleles is expected to show white spots. However, the size of the spots varies enormously, depending on the state of several other genes. These are consequently termed **modifier genes**. Variation in the modifier genes among different individuals will result in variation in expression of the major gene for a particular character. Environmental effects may also affect penetrance. In the fruit fly, alterations in temperature may change the penetrance of many alleles from 100% down to as low as 0%.

Genes from Both Parents Are Mixed by Sexual Reproduction

How are alleles distributed at mating? If both copies of both parents' genes were passed on to all their descendants, the offspring would have four copies of each gene, two from their mother and two from their father. The next generation would end up with eight copies and so on. Clearly, a mechanism is needed to ensure that the number of copies of each gene remains stable from generation to generation!

How does nature ensure that the correct copy number of genes is transferred? When diploid organisms such as animals or plants reproduce sexually, the parents both make sex cells, or **gametes**. These are specialized cells that pass on genetic information

To geneticists, sex is merely a mechanism for reshuffling genes to promote evolution. From the gene's perspective, an organism is just a machine for making more copies of the gene.

gametes Cells specialized for sexual reproduction that are haploid (have one set of genes)
modifier gene Gene that modifies the expression of another gene

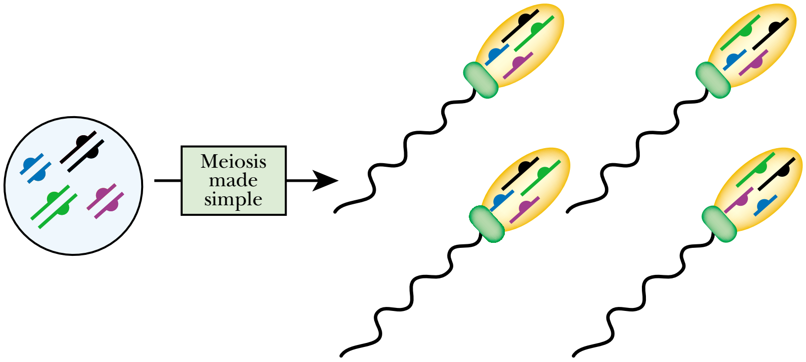


FIGURE 1.13 ***Meiosis—the Principle***

Diploid organisms distribute their chromosomes among their gametes by the process of meiosis. The principle is illustrated, but the detailed mechanism of meiosis is not shown. Chromosome reduction means that the gametes formed contain only half of the genetic material of the diploid parental cell (i.e. each gamete has one complete haploid set of genes). Each chromosome of a pair has a 50% chance of appearing in any one gamete, a phenomenon known as random segregation. While only sperm are shown here, the same process occurs during the production of ova.

to the next generation of organisms, as opposed to the **somatic cells**, which make up the body. Female gametes are known as eggs or ova (singular = ovum) and male gametes as sperm. When a male gamete combines with a female gamete at fertilization, they form a **zygote**, the first cell of a new individual (Fig 1.13). Although the somatic cells are diploid, the egg and sperm cells only have a single copy of each gene and are haploid. During the formation of the gametes, the diploid set of chromosomes must be halved to give only a single set of chromosomes. Reduction of chromosome number is achieved by a process known as **meiosis**. Figure 1.13 bypasses the technical details of meiosis and just illustrates its genetic consequences. In addition to reducing the number of chromosomes to one of each kind, meiosis randomly distributes the members of each pair. Thus, different gametes from the same parent contain different assortments of chromosomes.

Because egg and sperm cells only have a single copy of each chromosome, each parent passes on a single allele of each gene to any particular descendent. Which of the original pair of alleles gets passed to any particular descendant is purely a matter of chance. For example, when crossing an RR parent with an rr parent, each offspring gets a single R-allele from the first parent and a single r-allele from the second parent. The offspring will therefore all be Rr (Fig. 1.14). Thus, by crossing a plant that has red flowers with a plant that has white flowers, the result is offspring that all have red flowers. Note that the offspring, while phenotypically similar, are not genetically identical to either parent; they are heterozygous. The parents are regarded as generation zero and the offspring are the first, or F_1, generation. Successive generations of descendants are labeled F_1, F_2, F_3, etc.; this stands for first **filial generation**, second filial generation, etc.

Extending the ideas presented above, Figure 1.16 shows the result of a cross between two Rr plants. Each parent randomly contributes one copy of the gene, which may be an R or an r allele, to its gametes. Sexual reproduction ensures that the offspring get one copy from each parent. The relative numbers of each type of progeny as depicted in Fig. 1.16 are often referred to as **Mendelian ratios**. The Mendelian ratio in the F_2 generation is 3 red : 1 white. Note that white flowers have reappeared after skipping a generation. This is because the parents were both heterozygous for the r allele which is recessive and so was masked by the R allele.

A similar situation exists with human eye color. In this case the allele for blue eyes (b) is recessive to brown (B). This explains how two heterozygous parents (Bb) who both have brown eyes can produce a child who has blue eyes (bb, homozygous recessive). The same scenario also explains why inherited diseases do not afflict all members of a family and often skip a generation.

filial generations Successive generations of descendants from a genetic cross which are numbered F1, F2, F3, etc., to keep track of them
meiosis Formation of haploid gametes from diploid parent cells
Mendelian ratios Whole number ratios of inherited characters found as the result of a genetic cross
somatic cell Cell making up the body, as opposed to the germline
zygote Cell formed by union of sperm and egg which develops into a new individual

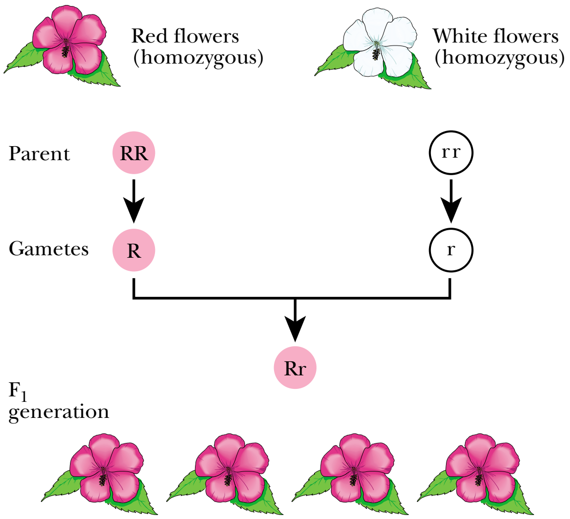
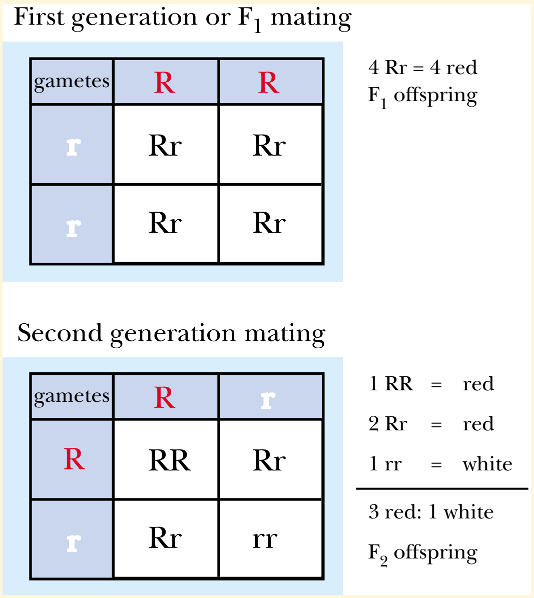


FIGURE 1.14 ***Cross between Homozygous Dominant and Recessive for Red Flower Color***

When individuals with the genotypes RR and rr are crossed, all the progeny of the cross, known as the F_1 generation, are red.



First generation or F_1 mating

gametes	R	R
r	Rr	Rr
r	Rr	Rr

4 Rr = 4 red
F_1 offspring

Second generation mating

gametes	R	r
R	RR	Rr
r	Rr	rr

1 RR	=	red
2 Rr	=	red
1 rr	=	white

3 red: 1 white
F_2 offspring

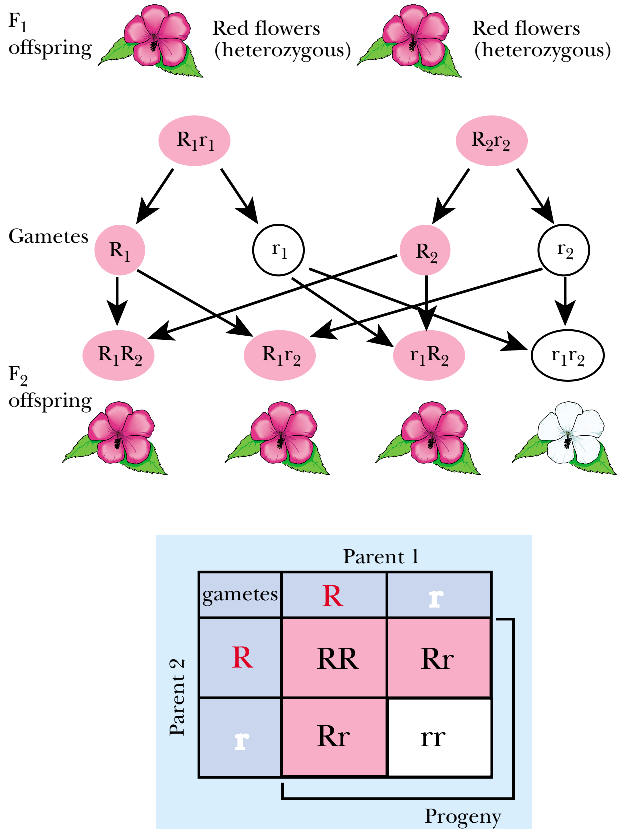
FIGURE 1.15 ***Checkerboard Determination of Genotype Ratios***

Checkerboard diagrams (also known as Punnett squares) are often used to determine the possible genotypes and their ratios that result from a genetic cross with two or more alleles. To construct a checkerboard diagram, place the possible alleles from one parent on the horizontal row and those from the other parent on the vertical row. Fill in the boxes with the combinations determined from the intersection of the vertical and horizontal rows. Then list the various phenotypes and add up the similar phenotypes. When adding similar phenotypes, Rr and rR, although genetically dissimilar, are equivalent phenotypically.

Sex Determination and Sex-Linked Characteristics

The genetic sex of many diploid organisms, including mammals and insects, is determined by which **sex chromosomes** they possess. Among mammals, possession of two **X-chromosomes** makes the organism a genetic female, whereas possession of one X-

sex chromosome A chromosome involved in determining the sex of an individual
X-chromosome Female sex chromosome; possession of two X-chromosomes causes female gender in mammals



gametes	R	r
R	RR	Rr
r	Rr	rr

FIGURE 1.16 ***The Rr × Rr Cross: Checkerboard Determination of Phenotypes***

The parents for this cross are the F_1 generation from the mating shown in Figure 1.15. At the top of the figure the possible gametes are shown for each parent flower. The arrows demonstrate how the genes distribute to give a 3 : 1 ratio of red to white flowers in the F_2 generation. At the bottom of the figure the F_2 mating is analyzed by a checkerboard to yield the same 3 : 1 ratio. Note that although no white flowers were present among the parents of this second mating, they are found in their F_2 offspring.

chromosome and one **Y-chromosome** makes the organism a genetic male. (The term "genetic" male or female is used because occasional individuals are found whose phenotypic sex does not match their genetic sex, due to a variety of complicating factors.) The checkerboard diagram for sex determination is shown in Figure 1.17.

Genes that have nothing to do with sex are also carried on the sex chromosomes. Which allele of these genes an individual will inherit correlates with the individual's sex and so they are called **sex-linked genes**. Most sex-linked genes are present in two copies in females but only one copy in males. This is because although the X- and Y-chromosomes constitute a pair, the Y-chromosome is much shorter. Thus many genes present on the X-chromosome do not have a corresponding partner on the Y-chromosome (Fig. 1.18). Conversely, a few genes, mostly involved in male fertility, are present on the Y-chromosome but missing from the X-chromosome.

If the single copy of a sex-linked gene present in a male is defective, there is no back-up copy and severe genetic consequences may result. In contrast, females with just one defective copy will usually have no problems because they usually have a good copy of the gene. However, they will be carriers and half of their male children will suffer the genetic consequences. The result is a pattern of inheritance in which the male members of a family often inherit the disease, but the females are carriers and suffer no symptoms. Figure 1.20 shows a family tree with several occurrences of an X-linked recessive disease. Males have only one X-chromosome and their Y-chromosome has no corresponding copy of the gene (symbolized by -). So any male who gets one copy of the defective allele ("a") will get the disease.

A well known example of sex-linked inheritance is red-green color blindness in humans. About 8% of men are color blind, whereas less than 1% of women show the defect. Many genes are involved in the synthesis of the three pigments for color vision, which are sensitive to red, green and blue. About 75% of color-blind people carry a sex-linked recessive mutation in the gene for the green-sensitive pigment, which is

Sex determination complicates the inheritance of a variety of other characters in many animals. Among mammals, males are more likely to suffer from certain genetic defects.

sex-linked A gene is sex-linked when it is carried on one of the sex chromosomes
Y-chromosome Male sex chromosome; possession of a Y-chromosome plus an X-chromosome causes male gender in mammals

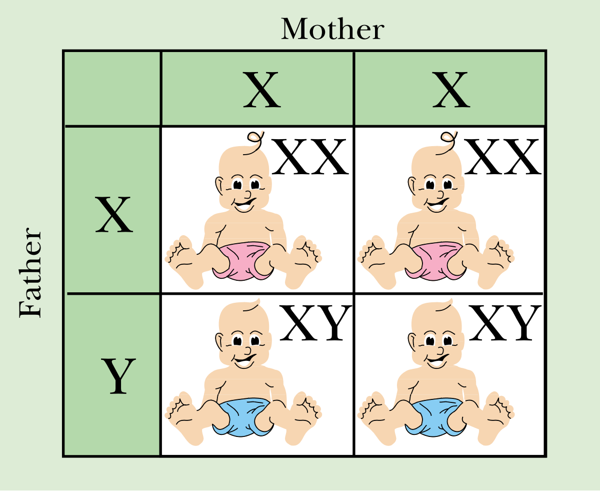


FIGURE 1.17 ***Checkerboard Diagram for Sex Determination***

The male parent contributes an X- and a Y-chromosome. The female contributes two X-chromosomes. The result is an equal proportion of males and females in each generation.

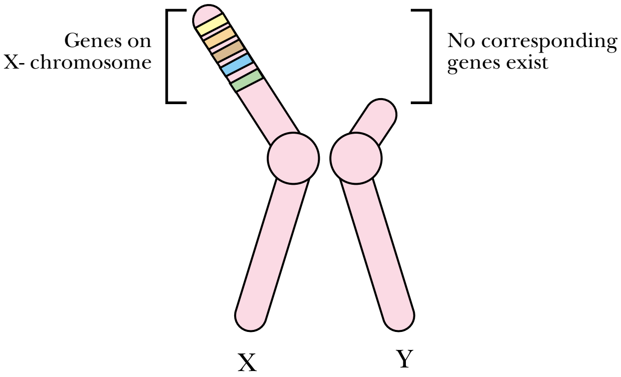


FIGURE 1.18 ***The X- and Y-Chromosomes Are Not of Equal Length***

The X- and Y-chromosome are not of equal length. The Y-chromosome lacks many genes corresponding to those on the X-chromosome. Therefore males have only one copy of these genes because they have only one X-chromosome.

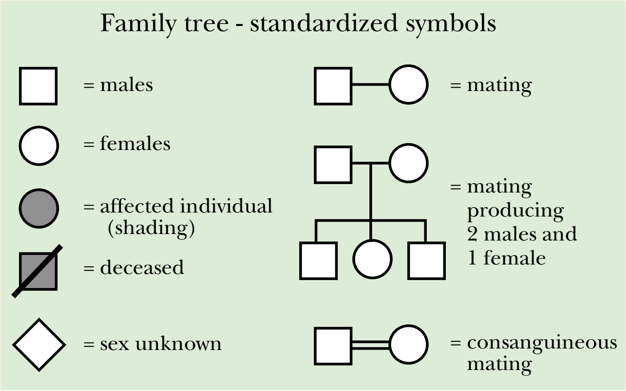

Family tree - standardized symbols
= males
= females
= affected individual (shading)
= deceased
= sex unknown
= mating
= mating producing 2 males and 1 female
= consanguineous mating

FIGURE 1.19 ***Standard Symbols for a Family Tree***

located on the X-chromosome (but absent from the Y-chromosome). More males than females are color blind, as males have only one copy of this gene. If this gene is defective, they are affected. A female has two copies of the gene and only if both are defective will she be color blind. A variety of other hereditary diseases show sex linkage and their detrimental effects are therefore more commonly observed in males than females.

Neighboring Genes Are Linked during Inheritance

The chemical nature of genes—as segments of DNA—has major effects on their inheritance.

Although there is a random distribution of strands of DNA (chromosomes) during sexual reproduction, there is not always a random distribution of alleles. To illustrate this point we must remember that most higher organisms have tens of thousands of

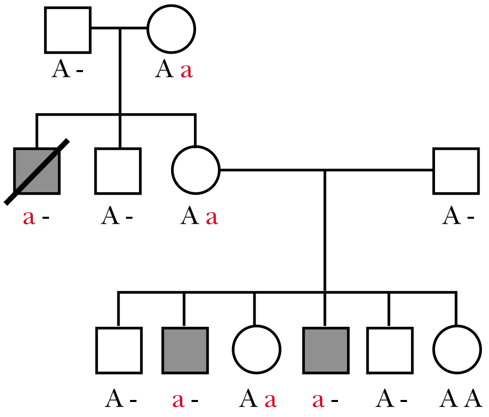


FIGURE 1.20 ***Inheritance of a Sex-linked Gene***

This family tree shows the inheritance of the wild-type ("A") and deleterious ("a") alleles of a gene that is carried on the X-chromosome. Since males have only one X-chromosome, they have only a single allele of this gene. The symbol "–" is used to indicate the absence of a gene. When the defective allele "a" is passed on to males, they will suffer its deleterious effects.

genes carried on multiple pairs of homologous chromosomes. Consider just a few of these genes—call them A, B, C, D, E, etc.—which have corresponding mutant alleles—a, b, c, d, e, etc. These genes may be on the same chromosome or they may be on different chromosomes. Let's assume that genes A, B and C are on one pair of homologous chromosomes and D and E are on a separate pair. Organisms that are heterozygous for all of these genes will have the genotype Aa, Bb, Cc, Dd, Ee. Consequently, A, B and C will be on one of a pair of homologous chromosomes and a, b, and c will be on the other member of the pair. A similar situation applies to D and E and d and e.

Alleles carried on different chromosomes are distributed at random among the offspring of a mating. For example, there is as much chance of allele d accompanying allele A during inheritance as allele D. In contrast, when genes are carried on the same chromosome, their alleles will not be distributed at random among the offspring. For example, because the three alleles A, B, and C are on the same chromosome, that is, the same molecule of DNA, they will tend to stay together. The same applies to a, b, and c. Such genes are said to be linked and the phenomenon is known as **linkage** (Fig. 1.21). Note that if two genes are very far apart they will not be linked in practice even if they reside on the same chromosome.

Recombination during Meiosis Ensures Genetic Diversity

However, the alleles A, B, and C (or a, b, and c) do not *always* stay together during reproduction. **Crossing over** occurs during the process of meiosis when the gametes are formed. First, the chromosome carrying a specific sequence of genes lines up next to the homologous chromosome with allele sites matching. Second, swapping of segments of the chromosomes can now occur by breaking and rejoining of the neighboring DNA strands. Note that the breaking and joining occurs in equivalent regions of the two chromosomes and neither chromosome gains or loses any genes overall. The genetic result of such crossing over, the shuffling of different alleles between the two members of a chromosomal pair, is called **recombination** or **crossing over** (Fig. 1.21). The farther apart two genes are on the chromosome, the more likely a crossover will form between them and the higher will be their frequency of recombination.

Genetic linkage is often defined, from a molecular viewpoint, as the tendency of alleles carried by the same DNA molecule to be inherited together. However, if two genes are very far apart on a very long DNA molecule, linkage may not be observed in practice. For example, consider a long chromosome, carrying genes A, B, C, D and

crossing over When two different strands of DNA are broken and are then joined to one another
linkage Two alleles are linked when they are inherited together more often than would be expected by chance, usually this is because they reside on the same DNA molecule (that is, on the same chromosome)
recombination Mixing of genetic information from two chromosomes as a result of crossing over

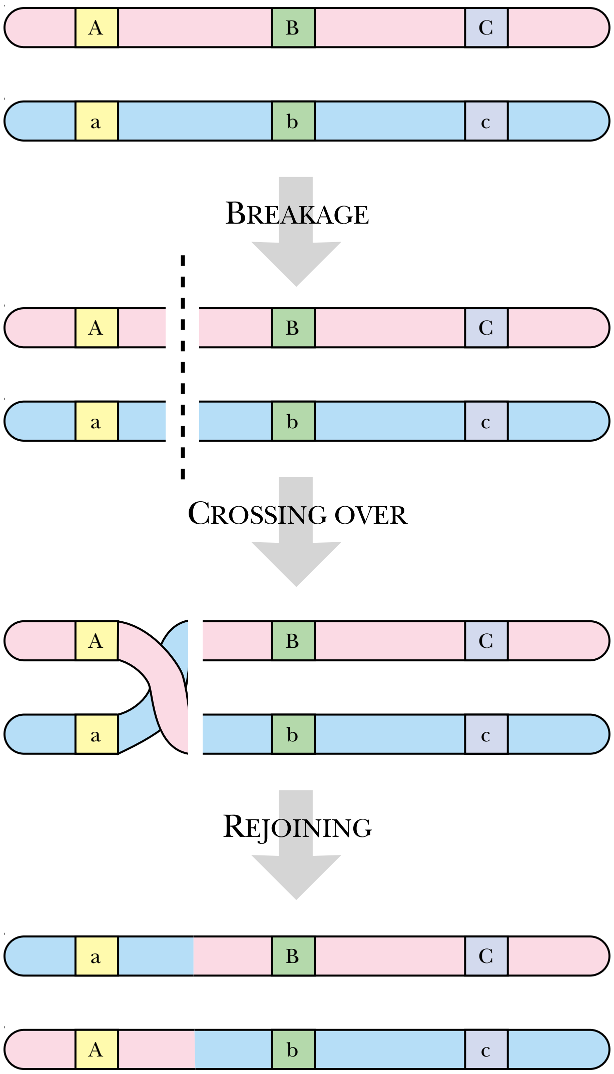


FIGURE 1.21 ***Linkage of Genes and Recombination during Meiosis***

At the top, the two members of a chromosome pair are shown, each carrying different alleles. Because the three alleles A, B, and C are on the same molecule of DNA, they will tend to stay together. So if the offspring inherits allele A from one parent, it will usually get alleles B and C, rather than b and c. The genes A, B and C are linked and the phenomenon is termed linkage. During meiosis the DNA is broken and the chromosomes are rejoined such that part of one chromosome is exchanged with the homologous partner. This exchange of genetic information is known as recombination, and it occurs at many sites along a pair of chromosomes.

E. It can be observed that A is linked to B and C and that C and D are linked to E, but that no linkage is observed between A and E in breeding experiments (Fig. 1.22). Given that A is on the same DNA molecule as B and that B is on the same DNA molecule as C etc., it can be deduced that A, B, C, D and E must all be on the same chromosome. In genetic terminology, it is said that A, B, C, D and E are all in the same **linkage group**. Even though the most distant members of a linkage group may not directly show linkage to each other, their relationship can be deduced from their mutual linkage to intervening genes.

It is often important to know the precise location of a gene. For example, this is true of the genes responsible for hereditary defects. In the past, geneticists measured the recombination frequencies of genes in order to estimate how far apart they were on the chromosome. Nowadays, physical methods are used to measure the distances between genes in terms of the length of the DNA molecule upon which they are carried.

Escherichia coli is a Model for Bacterial Genetics

Using simpler organisms has allowed more detailed analysis of gene structure and function.

Bacteria are smaller and multiply faster than flies. Bacterial cultures contain many millions of individuals for analysis. Indeed a typical culture of a vigorous bacterium such as *E. coli* may contain as many as 5 × 10^9 cells per ml—roughly the same number as

linkage group A group of alleles carried on the same DNA molecule (that is, on the same chromosome)

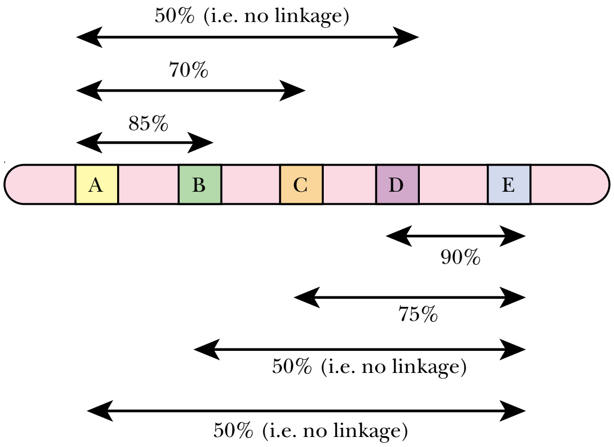


FIGURE 1.22 ***Linkage Groups***

If genes A, B, C, D and E are all on the same chromosome, they will show linkage. The extent of linkage depends primarily on their distance from each other on the chromosome. For example, the alleles of two genes close to each other may be inherited together 90% of the time, whereas the alleles of more distant genes will stay together less often. These percentages are somewhat deceptive since alleles on different chromosomes will accompany each other 50% of the time due to random segregation. Thus 50% is the lowest possible numerical value for "linkage" and does not in fact imply either the presence or absence of linkage.

Morgan Used the Fruit Fly, *Drosophila*, to Study Genetics

In the early 20th century, the fruit fly, *Drosophila melanogaster*, achieved fame as the model organism for genetic analysis. Fruit flies are small, easy to feed, and yield a new generation in a few days—a big time saving advantage compared with flowering plants. Larger numbers of individuals can be examined and, most important, many generations can be observed in a single year of experimentation.

Mendel did not actually discover linkage, as the characters he worked with were mostly on different chromosomes. Two were actually on the same chromosome but far enough apart for their linkage to go unnoticed. He was lucky in being able to lay the foundations of genetics without the complications that linkage introduces. T. H. Morgan pioneered work on *Drosophila*, beginning in 1909. He was largely responsible for discovering and analyzing the phenomena of linkage, sex-determination and sex-linked genes that have been described above.

FIGURE 1.23 ***Fruit Flies Used for Genetics***

Fruit flies of the species *Drosophila melanogaster* are raised in milk bottles for genetic research. The milk bottles are sterilized in an autoclave, then partly filled with nutritious growth medium plus a sheet of filter paper. The fly larvae eat the pale brown medium and pupate on the paper. When the flies need to be examined, they are knocked unconscious with ether, from which they recover in about 10 minutes. Courtesy of: Dr Jeremy Burgess, Science Photo Library.

the total human world population. After World War II, the use of bacteria and their viruses took genetic analysis down to the level of the DNA molecule and allowed the mapping of different mutations within the same gene. Bacteria are not merely ideal for high-powered genetic analysis; they are also convenient for biochemical investigations. It was at this point that ***Escherichia coli*** (or, commonly, ***E. coli***), a bacterium found as a harmless inhabitant of the large intestine of man and other animals, came to the forefront of genetic research.

Bacterial genetics gave rise to a standard terminology for naming genes. Consider the biochemical pathway for synthesis of the amino acid threonine (Fig. 1.24). This pathway consists of three steps, catalyzed by three enzymes that are encoded by three genes, *thrA*, *thrB*, and *thrC*.

Note that related genes are all designated by a three-letter abbreviation, which, hopefully, indicates their function. Each separate gene of such a related group is additionally followed by a capital letter of the alphabet. The gene designation is printed in *italics*, or if written by hand it may be underlined. The wild-type allele is indicated with a "+" sign; e.g., *thrA*$^+$. A defective allele may have a "–" sign; e.g. *thrA*$^-$. Different mutations in the same gene receive allele numbers, for example, *thrB1, thrB2*, *thrB57*, etc. [This convention does not apply to eukaryotes, in part because the number of genes is much larger and their nomenclature has never been properly standardized. Nonetheless, eukaryotic genes are generally italicized.]

The bacterium *E. coli* has approximately 4,000 genes (about one tenth as many as a human) arranged within a single circular chromosome (Fig. 1.25). The *E. coli* chro-

***Escherichia coli* (*E. coli*)** A species of bacterium commonly used in genetics and molecular biology

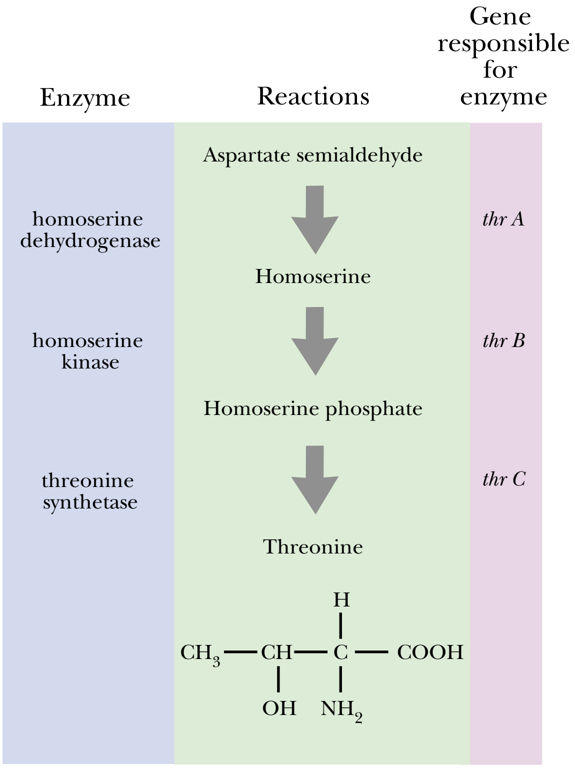


FIGURE 1.24 ***The Threonine Pathway***

The genes that code for enzymes necessary to convert aspartate semialdehyde to threonine are designated in italics.

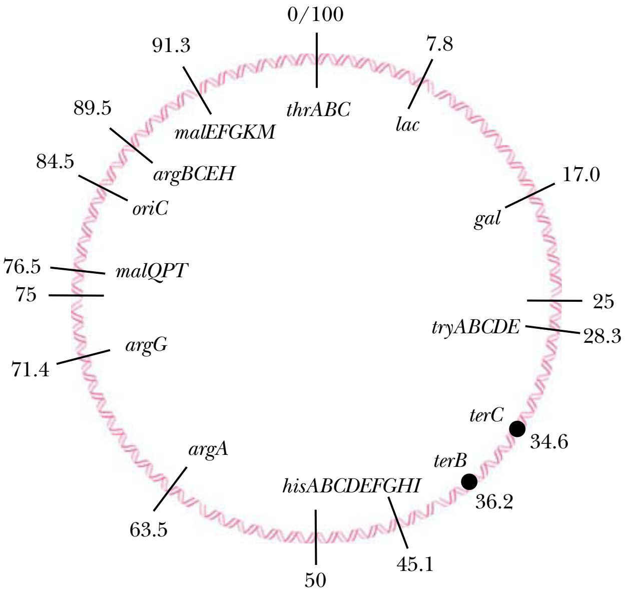


FIGURE 1.25 ***The* E. coli *Chromosome***

The circular *E. coli* chromosome has been divided into 100 map units. Starting with zero at *thrABC*, the units are numbered clockwise from 0 to 100. Various genes are indicated with numbers corresponding to their position on the map. The replication origin (*oriC*) and termini (*ter*) of replication are also indicated. Note that chromosome replication does not start at zero map units—the zero point was an arbitrary designation.

> Plasmids, the extra circles of DNA found in bacteria, have become vital tools in modern genetic technology. See especially Chapters 16, 18 and 22.

mosome is divided into 100 **map units**. The position of the *thrABC* genes was arbitrarily chosen as the zero/100 position (i.e., start and end). The **origin of replication** (***oriC***) is the point at which the chromosome starts to divide and the **terminus** (***ter***) is where replication terminates (see Ch 5).

Note that bacteria only have a single chromosome and therefore only contain a single copy of most of their genes. Therefore, dominant and recessive alleles of the same gene in the same bacterial cell are not normally found. However, bacterial cells

map unit A subdivision that is one hundredth of the length of the bacterial chromosome
origin of chromosome (*oriC*) Origin of replication of a chromosome
terminus of replication (*ter*) The place on any DNA molecule where replication ends

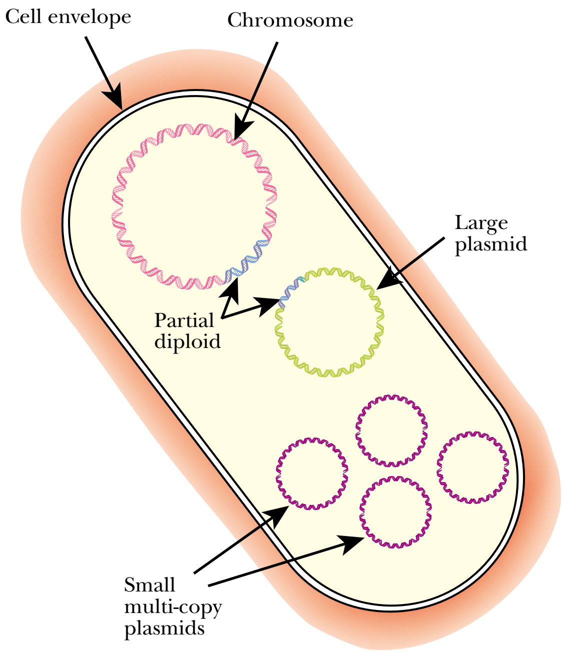


FIGURE 1.26 ***Plasmid within an* E. coli *Cell Shows Partial Diploidy***

The bacterial chromosome is indicated in red; the large single-copy plasmid is indicated in green and the small multiple-copy plasmid in purple. Note that a segment of the bacterial chromosome, colored blue, has been duplicated and is carried by the larger plasmid, making the cell a partial diploid.

By convention, when bacterial geneticists describe the genotype of a bacterial strain they list only those genes with mutations. A gene that is not mentioned is assumed to be wild-type. Furthermore, the "–" sign that indicates a gene is mutated is also usually omitted. Thus merely listing a gene implies that it is mutated.

Consider the genotype: ser*A14 leu-6 thi*

This genotype tells us that the bacterial strain in question has no defects in the genes for making the amino acid threonine. It does have a fully identified defect in one of the genes for making the amino acid serine (ser*A14*). Its defect in leucine synthesis has been partly characterized and numbered, but which of the leucine genes is altered remains uncertain (*leu-6*). The mutation that prevents synthesis of the vitamin thiamine (*thi*) is still uncharacterized.

sometimes carry extra genetic elements known as **plasmids**. These are circular molecules of DNA that replicate in step with cell division, like the chromosome, but are generally much smaller (Fig. 1.26). They often carry genes that provide extra capabilities to the bacterial cell, but that are not essential for normal growth and division. Plasmids may be present in single or multiple copies. If a plasmid carries extra alleles corresponding to genes already on the bacterial chromosome, the bacteria are said to be **partial diploids** for those particular genes (Fig. 1.26).

Bacteria and the viruses that infect them (known as bacteriophages) were the most important organisms used when science made the transition from classical genetics to molecular genetics/biology. As will become apparent in subsequent chapters, the unveiling of the molecular basis of heredity allowed a much deeper understanding of genetic mechanisms. The next chapter will review the variety of living organisms and focus on those used most often in genetic analysis. Then, in Chapter 3, the structure of DNA will be examined and it will become apparent how DNA encodes the genetic information.

partial diploidy Situation in which a cell is diploid for only some of its genes
plasmid Circular molecule of double stranded helical DNA which replicates independently of the chromosomes of the host cell. Rare linear plasmids have been discovered

CHAPTER **TWO**

Cells and Organisms

What Is Life?

Although there is no definition of life that suits all people, everyone has an idea of what being alive means. Generally, it is accepted that something is alive if it can *grow* and *reproduce*, at least during some stage of its existence. Thus, we still regard adults who are no longer growing and those individuals beyond reproductive age as being alive. We also regard sterile individuals, such as mules or worker bees as being alive, even though they lack the ability to reproduce. Part of the difficulty in defining life is the complication introduced by multicellular organisms. Although a multicellular organism as a whole may not grow or reproduce some of its cells may still retain these abilities.

Perhaps the key factor that characterizes life is the ability to self-replicate. This includes both the **replication** of the genetic information (the genome) and of the structure carrying and protecting it (the cell). Growth and reproduction need both information and energy in order to process raw materials into new living matter, and ultimately to create new organisms identical or, at any rate very similar, to the original organism. This brings us to another characteristic of life, which is that it evolves. Descendents are not identical to their ancestors but gradually accumulate changes in their genetic information over time. Both accurate replication and occasional evolutionary change are due to the properties of the nucleic acid molecules, DNA and RNA, which carry the genetic information. Furthermore, life forms do not merely grow and divide they also respond to stimuli from the environment. Some responses involve such complex structures as the nervous system of higher animals. However, many responses operate at the genetic level and are therefore included in this book.

No satisfactory technical definition of life exists. Despite this we understand what life entails. In particular, life involves a dynamic balance between duplication and alteration.

The basic ingredients needed to sustain life include the following:

Genetic information Biological information is carried by the **nucleic acid** molecules, **deoxyribonucleic acid** (**DNA**) and **ribonucleic acid** (**RNA**). The units of genetic information are known as **genes** and each consists physically of a segment of a nucleic acid molecule. DNA is used for long-term storage of large amounts of genetic information (except by some viruses—see Ch. 17). Whenever genetic information is actually used, working copies of the genes are carried on RNA. The total genetic information possessed by an organism is known as its **genome**. Whenever an organism reproduces, the DNA molecules carrying the genome must be replicated so that the descendents may receive a complete copy of the genetic information.

Mechanism for energy generation By itself, information is useless. Energy is needed to put the genetic information to use. Living creatures must all obtain energy for growth and reproduction. **Metabolism** is the set of processes in which energy is acquired, liberated and used for biosynthesis of cell components.

Machinery for making more living matter Synthesis of new cell components requires chemical machinery. In particular, the **ribosomes** are needed for making proteins, the **macromolecules** that make up the bulk of all living tissue.

A characteristic outward physical form Living creatures all have a material body that is characteristic for each type of life form. This structure contains all the metabolic and biosynthetic machinery for generating energy and making new living matter. It also contains the DNA molecules that carry the genome.

Identity or self All living organisms have what one might call an identity. The term self-replication implies that an organism knows to make a copy of itself—

deoxyribonucleic acid (DNA) The nucleic acid polymer of which the genes are made
gene A unit of genetic information
genome The entire genetic information from an individual
macromolecule Large polymeric molecule; in living cells especially DNA, RNA, protein or polysaccharide
metabolism The processes by which nutrient molecules are transported and transformed within the cell to release energy and to provide new cell material
nucleic acid Polymer made of nucleotides that carries genetic information
replication Duplication of DNA prior to cell division
ribonucleic acid (RNA) Nucleic acid that differs from DNA in having ribose in place of deoxyribose and having uracil in place of thymine
ribosome The cell's machinery for making proteins

not merely to assemble random organic material. Living organisms use raw material from the environment to make more of their own selves, ultimately to make complete copies of themselves. This concept of self versus non-self reaches its most sophisticated expression in the immune systems that protect higher animals against disease. But even primitive creatures attempt to preserve their own existence.

Living Creatures Are Made of Cells

Looking around at the living creatures that inhabit this planet, one is first struck by their immense variety: squids, seagulls, sequoias, sharks, sloths, snakes, snails, spiders, strawberries, soybeans, and so forth. Although highly diverse to the eye, the biodiversity represented by these creatures is actually somewhat superficial. The most fascinating thing about life is not its superficial diversity but its *fundamental unity*. All of these creatures, together with microscopic organisms too small to see with the naked eye, are made up of **cells,** structural units or compartments that have more or less the same components.

Matter is divided into atoms. Genetic information is divided into genes. Living organisms are divided into cells.

The idea that living cells are the structural units of life was first proposed by Schleiden and Schwann in the 1830s. Cells are microscopic structures that vary considerably in shape. Many are spherical, cylindrical or roughly cuboidal but many other shapes are found, such as the long branched filaments of nerve cells. Many microscopic life forms consist of a single cell, whereas creatures large enough to see usually contain thousands of millions. Each cell is enclosed by a cell membrane composed of **proteins** and **phospholipids** and contains a complete copy of the genome (at least at the start of its life). Living cells possess the machinery to carry out metabolic reactions and generate energy and are usually able to grow and divide. Moreover, living cells always result from the division of pre-existing cells; they are never assembled from their component parts. This implies that living organisms too can only arise from pre-existing organisms. In the 1860s, Louis Pasteur confirmed experimentally that life cannot arise spontaneously from organic matter. Sterilized nutrient broth did not "spoil" or "go bad" unless it was exposed to microorganisms in the air.

In most multicellular organisms, the cells are specialized in a variety of ways (Fig. 2.01). The development of specialized roles by particular cells or whole tissues is referred to as **differentiation**. For example, the red blood cells of mammals lose their nucleus and the enclosed DNA during development. Once these cells are fully differentiated, they can perform only their specialized role as oxygen carriers and can no longer grow and divide. Some specialized cells remain functional for the life span of the individual organism, whereas others have limited life spans, sometimes lasting only a few days or hours. For multicellular organisms to grow and reproduce, some cells clearly need to keep a complete copy of the genome and retain the ability to grow and divide. In single-celled organisms, such as **bacteria** or protozoa, each individual cell has a complete genome and can grow and reproduce; hence, the complications of having multiple types of cell are largely absent.

Essential Properties of a Living Cell

At least in the case of unicellular organisms, each cell must possess the characteristics of life as discussed above (Fig. 2.02). Each living cell must generate its own energy and

bacteria Primitive, relatively simple, single-celled organisms that lack a cell nucleus
cell The cell is the basic unit of life. Each cell is surrounded by a membrane and usually has a full set of genes that provide it with the genetic information necessary to operate
differentiation Progressive changes in the structure and gene expression of cells belonging to a single organism that leads to the formation of different types of cell
phospholipid A hydrophobic molecule found making up cell membranes and consisting of a soluble head group and two fatty acids both linked to glycerol phosphate
protein Polymer made from amino acids that does most of the work in the cell

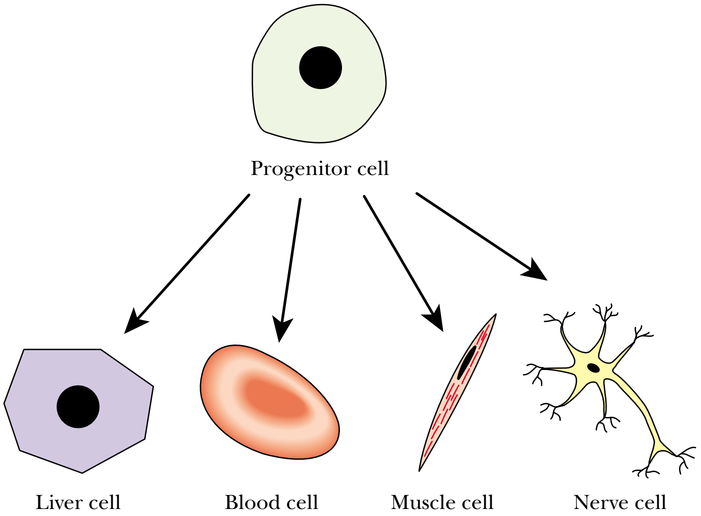


FIGURE 2.01 ***Some Cells Differentiate***

In multicellular organisms, cells differentiate from unspecialized precursor cells. Differentiation allows cells to specialize functionally. Their form is related to their function.

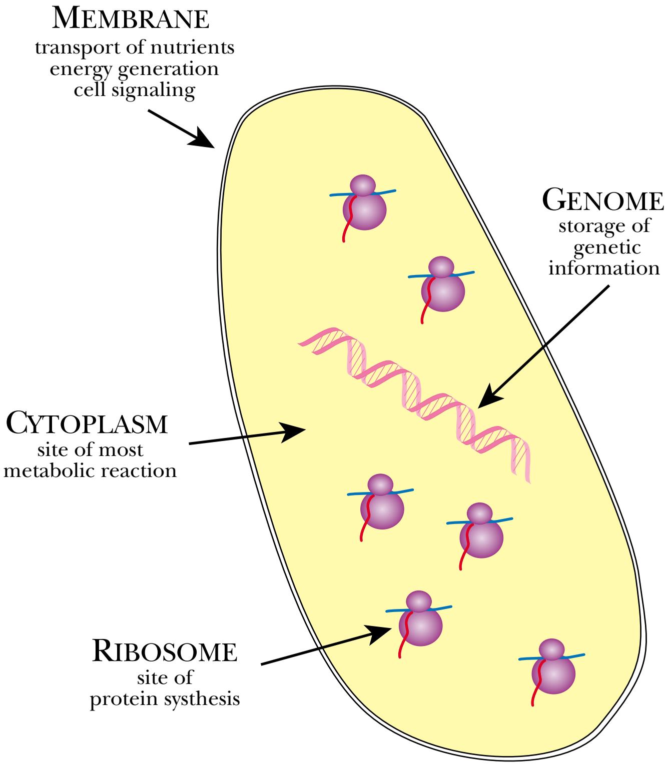


FIGURE 2.02 ***Essential Features of a Living Cell***

Simple cells possess certain elements considered essential to support life. Fundamentals for life include a membrane to separate the inside of the cell (the cytoplasm) from the environment; a means to store genetic information (the genome); and an apparatus (ribosome) to synthesize proteins.

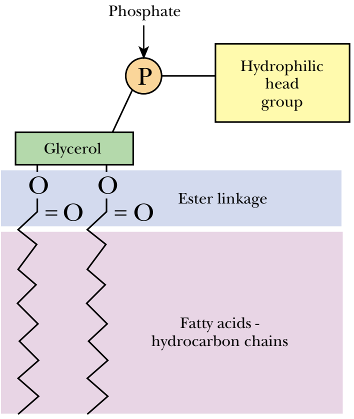


FIGURE 2.04 ***Phospholipid Molecule***

Phospholipid molecules of the kind found in membranes have a hydrophilic head group attached via a phosphate group to glycerol. Two fatty acids are also attached to the glycerol via ester linkages.

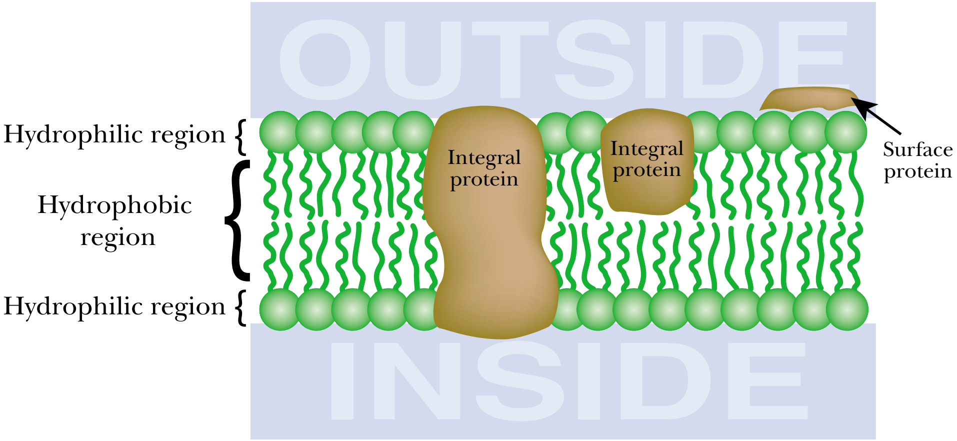


FIGURE 2.03 ***A Biological Membrane***

A biological membrane is formed by phospholipid and protein. The phospholipid layers are oriented with their hydrophobic tails inward and their hydrophilic heads outward. Proteins may be within the membrane (integral) or lying on the membrane surfaces.

Membranes do not merely separate living tissue from the non-living exterior. They are also the site of many biosynthetic and energy yielding reactions.

synthesize its own macromolecules. Each must have a genome, a set of genes carried on molecules of DNA. [Partial exceptions occur in the case of multicellular organisms, where responsibilities may be distributed among specialized cells and some cells may lack a complete genome.]

A cell must also have a surrounding **membrane** that separates the cell interior, the **cytoplasm**, from the outside world. The cell membrane, or cytoplasmic membrane, is made from a double layer of phospholipids together with proteins (Fig. 2.03). [Some single-celled protozoa, such as *Paramecium*, have multiple nuclei within each single cell. In addition, in certain tissues of some multi-cellular organisms several nuclei may share the same cytoplasm and be surrounded by only a single cytoplasmic membrane. Such an arrangement is known as a syncytium when it is derived from multiple fused cells.] Phospholipid molecules consist of a water-soluble head group, including phosphate, found at the surface of the membrane, and a lipid portion consisting of two hydrophobic chains that form the body of the membrane (Fig. 2.04). The phospholipids form a hydrophobic layer that greatly retards the entry and exit of water-soluble molecules. For the cell to grow, it must take up nutrients. For this, transport proteins, which penetrate through the membrane, are necessary. Many of the metabolic reactions involved in the breakdown of nutrients to release energy are catalyzed by soluble enzymes located in the cytoplasm. Other energy-yielding series of reactions, such as the respiratory chain or the photosynthetic system, are located in membranes. The proteins may be within or attached to the membrane surfaces (Fig. 2.03).

The cytoplasmic membrane is physically weak and flexible. Many cells therefore have a tough structural layer, the cell wall, outside the cell membrane. Most bacterial and plant cells have hard cell walls, though animal cells usually do not. Thus a cell wall is not an essential part of a living cell.

Soluble enzymes located in the cytoplasm catalyze biosynthesis of the low molecular weight precursors to protein and nucleic acids. However, assembly of proteins requires a special organelle, the ribosome (Fig. 2.05). This is a subcellular machine that consists of several molecules of RNA and around 50 proteins. It uses information that is carried from the genome into the cytoplasm by special RNA molecules, known as **messenger RNA**. The ribosome decodes the nucleic acid-encoded genetic information on the messenger RNA to make protein molecules.

cytoplasm The portion of a cell that is inside the cell membrane but outside the nucleus
membrane A thin flexible structural layer made of protein and phospholipid that is found surrounding all living cells
messenger RNA (mRNA) The class of RNA molecule that carries genetic information from the genes to the rest of the cell

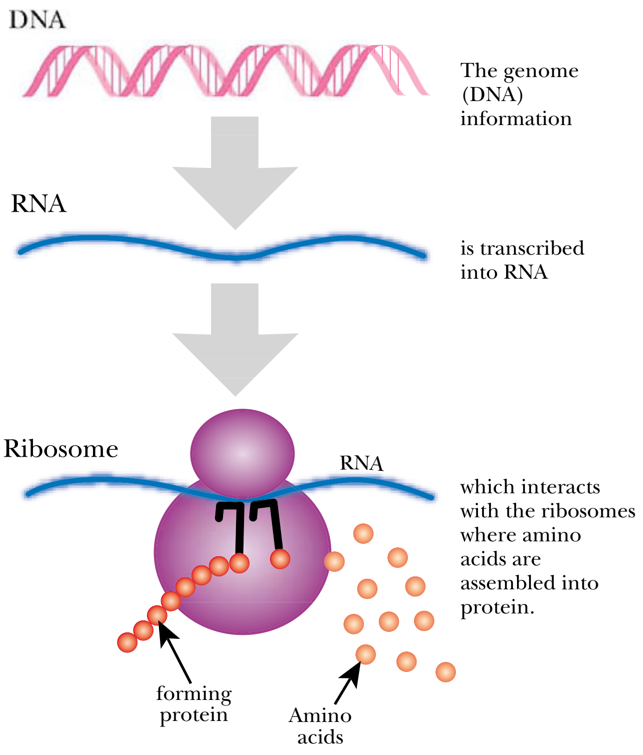

DNA
The genome (DNA) information
RNA
is transcribed into RNA
Ribosome
RNA
which interacts with the ribosomes where amino acids are assembled into protein.
forming protein
Amino acids

FIGURE 2.05 ***Ribosomes Make Protein***

The information stored in DNA is transported to the ribosome where proteins are synthesized from components known as amino acids.

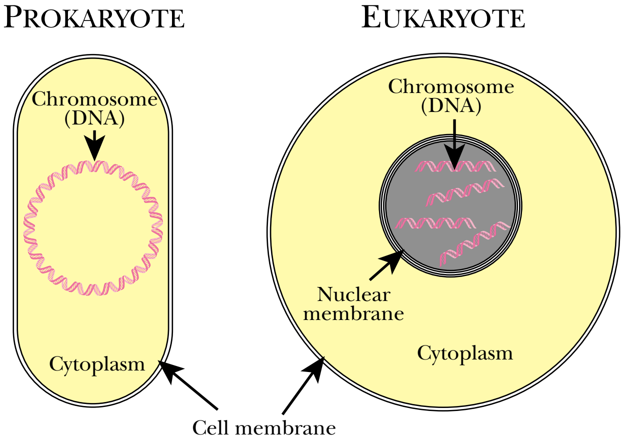


FIGURE 2.06 ***Prokaryotic and Eukaryotic Cells***

A comparison of prokaryotic and eukaryotic cells shows that the eukaryotes have a separate compartment called the nucleus that contains their DNA.

Based on differences in compartmentalization, living cells may be divided into two types, the simpler **prokaryotic** cell and the more complex **eukaryotic** cell. By definition, prokaryotes are those organisms whose cells are not subdivided by membranes into a separate **nucleus** and cytoplasm. All prokaryote cell components are located together in the same compartment. In contrast, the larger and more complicated cells of higher organisms (animals, fungi, plants and protists) are subdivided into separate compartments and are called eukaryotic cells. Figure 2.06 compares the design of prokaryotic and eukaryotic cells.

eukaryote Higher organism with advanced cells, which have more than one chromosome within a compartment called the nucleus
nucleus An internal compartment surrounded by the nuclear membrane and containing the chromosomes. Only the cells of higher organisms have nuclei.
prokaryote Lower organism, such as a bacterium, with a primitive type of cell containing a single chromosome and having no nucleus

Prokaryotic Cells Lack a Nucleus

Cells are separated from their environments by membranes. In the more complex cells of eukaryotes, the genome is separated from the rest of the cell by another set of membranes.

Bacteria (singular, bacterium) are the simplest living cells and are classified as prokaryotes. By definition, prokaryotes lack a nucleus and their DNA is therefore in the same compartment as the cytoplasm. Bacterial cells (Fig. 2.07) are always surrounded by a membrane (the cell or cytoplasmic membrane) and usually also by a cell wall. Like all cells, they contain all the essential chemical and structural components necessary for life. Typically, each bacterial cell has a single **chromosome** carrying a full set of genes providing it with the genetic information necessary to operate as a living organism. [Occasional bacteria are known that have more than one chromosome, however this is relatively uncommon.] Typically, bacteria have 3,000–4,000 genes, although some have as few as 500. The minimum number of genes to allow the survival of a living cell is uncertain. Experiments are presently in progress to successively delete genes from certain very small bacterial genomes in an attempt to create a truly minimal cell.

A typical bacterial cell, such as ***Escherichia coli***, is rod shaped and about two or three micrometers long and a micrometer wide. A micrometer (μm), also known as a micron, is a millionth of a meter (i.e., 10^{-6} meter). Bacteria are not limited to a rod shape (Fig. 2.08); spherical, filamentous or spirally twisted bacteria are also found. Occasional giant bacteria occur, such as *Epulopiscium fishelsoni*, which inhabits the surgeonfish and measures a colossal 50 microns by 500 microns—an organism visible to the naked eye. Typical eukaryotic cells are 10 to 100 microns in diameter.

A smaller cell has a larger surface-to-volume ratio. Smaller cells transport nutrients relatively faster, per unit mass of cytoplasm (i.e., cell contents), and so can grow more rapidly than larger cells. Because bacteria are less structurally complex than animals and plants, they are often referred to as "lower organisms." However, it is important to remember that present-day bacteria are at least as well adapted to modern conditions as animals and plants, and are just as highly evolved as so-called "higher organisms." In many ways, bacteria are not so much "primitive" as specialized for growing more efficiently in many environments than larger and more complex organisms.

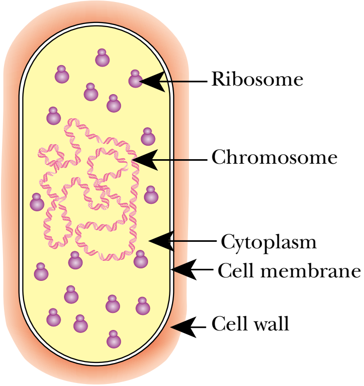


FIGURE 2.07 ***Typical Bacterium***

The components of a bacterium are depicted.

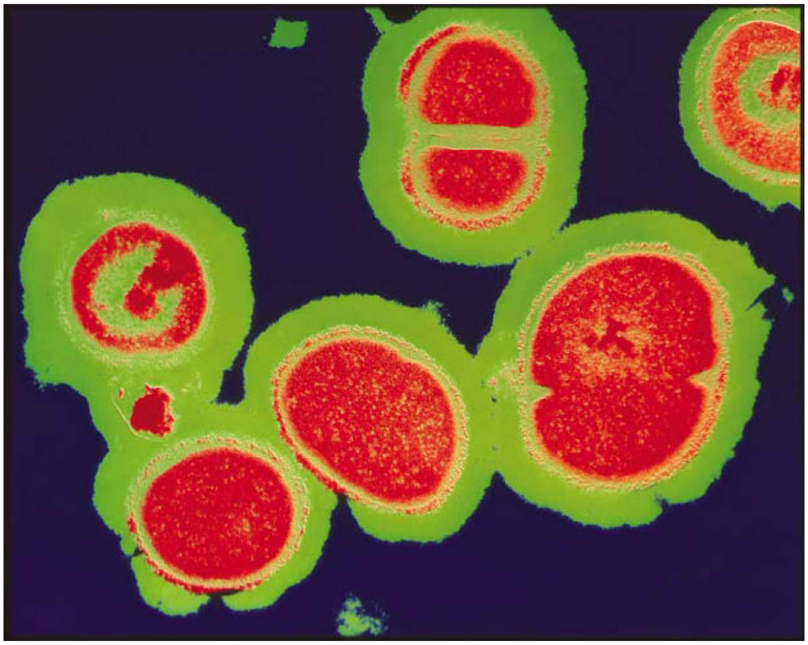


FIGURE 2.08 ***False Color TEM of* Staphylococcus aureus**

Colored transmission electron micrograph (TEM) of a cluster of *Staphylococcus aureus* seen dividing. *S. aureus* may cause boils, usually by entering the skin through a hair follicle or a cut. They are also responsible for internal abscesses and most types of acute suppurative infection. Magnification: ×24,000. Provided by Dr Kari Lounatmaa, Science Photo Library.

chromosome Structure containing the genes of a cell and made of a single molecule of DNA
Escherichia coli A bacterium commonly used in molecular biology

FIGURE 2.09 ***Hot spring in Ethiopia***

Hot springs are good sites to find archaebacteria. These springs are in the Dallol area of the Danakil Depression, 120 metres below sea level. The Danakil Depression of Ethiopia is part of the East African Rift Valley. Hot water flows from underground to form these pools. The water is heated by volcanic activity and is at high pressure, causing minerals in the rock to dissolve in the water. The minerals precipitate out as the water cools at the surface, forming the deposits seen here. Provided by Bernhard Edmaier, Science Photo Library.

Eubacteria and Archaebacteria Are Genetically Distinct

At a fundamental level, three domains of life, eubacteria, archaebacteria and eukaryotes, have replaced the old-fashioned division into animal and vegetable.

There are two distinct types of prokaryotes, the **eubacteria** and **archaebacteria**, which are no more genetically related to each other than either group is to the eukaryotes. Both eubacteria and archaebacteria show the typical prokaryotic structure—in other words, they both lack a nucleus and other internal membranes. Thus, cell structure is of little use for distinguishing these two groups. The eubacteria include most well known bacteria, including all those that cause disease. When first discovered, the archaebacteria were regarded as strange and primitive. This was largely because most are found in extreme environments (Fig. 2.09) and/or possessed unusual metabolic pathways. Some grow at very high temperatures, others in very acidic conditions and others in very high salt. The only major group of archaebacteria found under "normal" conditions are the methane bacteria, which, however, have a very strange metabolism. They contain unique enzymes and cofactors that allow the formation of methane by a pathway found in no other group of organisms. Despite this, the **transcription** and **translation** machinery of archaebacteria resembles that of eukaryotes, so they turned out to be neither fundamentally strange nor truly primitive when further analyzed.

Biochemically, there are major differences between the eubacterial and archaebacterial cells. In all cells, the cell membrane is made of phospholipids, but the nature and linkage of the lipid portion is quite different in the eubacteria and archaebacteria (Fig. 2.10). The cell wall of eubacteria is always made of peptidoglycan, a molecule unique to this group of organisms. Archaebacteria often have cell walls, but these are made of a variety of materials in different **species**, but peptidoglycan is never

Archaebacteria (or Archaea) Type of bacteria forming a genetically distinct domain of life. Includes many bacteria growing under extreme conditions
Eubacteria Bacteria of the normal kind as opposed to the genetically distinct Archaebacteria
species A group of closely related organisms with a relatively recent common ancestor. Among animals, species are populations that breed among themselves but not with individuals of other populations. No satisfactory definition exists for bacteria or other organisms that do not practice sexual reproduction.
transcription Process by which information from DNA is converted into its RNA equivalent
translation Making a protein using the information provided by messenger RNA

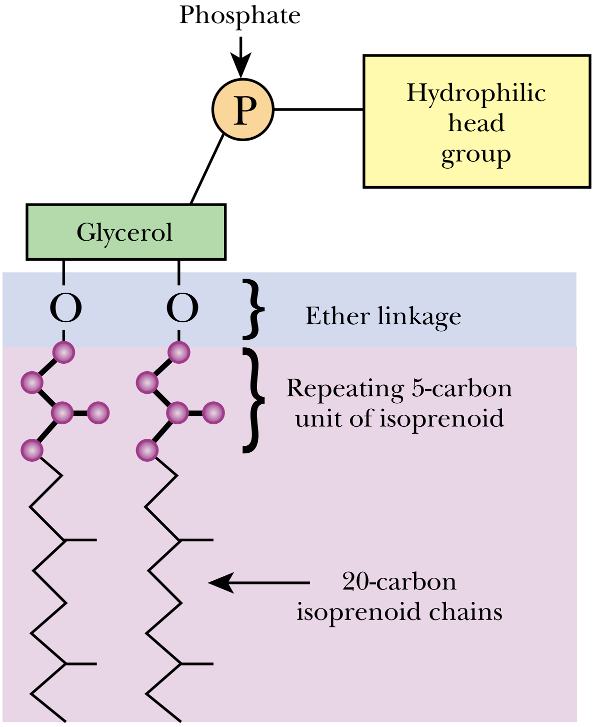

Phosphate
P
Hydrophilic head group
Glycerol
O
O
Ether linkage
Repeating 5-carbon unit of isoprenoid
20-carbon isoprenoid chains

FIGURE 2.10 ***Lipids of Archaebacteria***

In eubacteria and eukaryotes, the fatty acids of phospholipids are esterified to the glycerol. In archaebacteria, the lipid portion consists of branched isoprenoid hydrocarbon chains joined to the glycerol by ether linkages (as shown here). Such lipids are much more resistant to extremes of pH, temperature and ionic composition.

present. Thus the only real cellular structures possessed by prokaryotes, the cell membrane and cell wall, are in fact chemically different in these two groups of prokaryotes. The genetic differences will be discussed later when molecular evolution is considered (see Ch. 20).

Bacteria Were Used for Fundamental Studies of Cell Function

Most of the early experiments providing the basis for modern day molecular biology were performed using bacteria such as *Escherichia coli* (see below), because they are relatively simple to analyze. Some advantages of using bacteria to study cell function are:

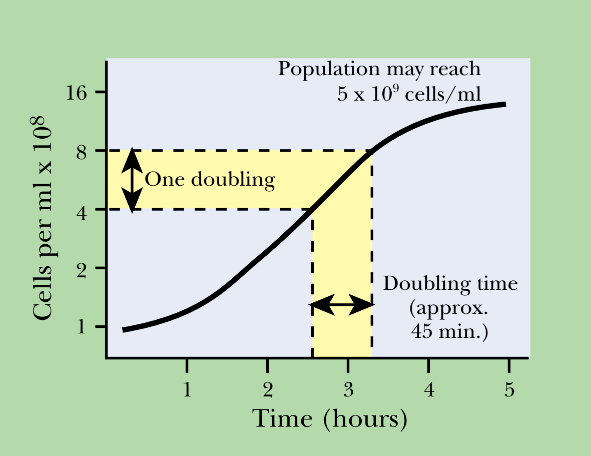
1. Bacteria are single-celled microorganisms. Furthermore, a bacterial culture consists of many identical cells due to lack of sexual recombination during cell division. In contrast, in multi-cellular organisms, even an individual tissue or organ contains many different cell types. All the cells in a bacterial culture respond in a reasonably similar way, whereas those from a higher organism will give a variety of responses, making analysis much more difficult.
2. The most commonly used bacteria have about 4,000 genes as opposed to higher organisms, which have up to 50,000. Furthermore, different selections of genes are expressed in the different cell types of a single multicellular organism.
3. Bacteria are **haploid**, having only a single copy of most genes, whereas higher organisms are **diploid**, possessing at least two copies of each gene. As discussed in Ch. 1, the multiple gene copies may differ in a variety of ways, making research results more complex.

Biologists have always been pulled in two directions. Studying simple creatures allows basic principles to be investigated more easily. And yet we also want to know about ourselves.

diploid Possessing two copies of each gene
haploid Possessing only a single copy of each gene

FIGURE 2.11 ***Graph of Exponential Growth of Bacterial Culture***

The number of bacteria in this culture is doubling approximately every 45 minutes. This is typical for fast growing bacteria such as *Escherichia coli* that are widely used in laboratory research. The bacterial population may reach 5 × 10^9 cells per ml or more in only a few hours under ideal conditions.



E. *coli* is normally harmless, although occasional rogue strains occur. Even these few **pathogenic** *E. coli* strains mostly just cause diarrhea, by secreting a mild form of a toxin related to that found in cholera and dysentery bacteria. However, the notorious *E. coli* O157:H7 carries two extra toxins and causes bloody diarrhea that may be fatal, especially in children or the elderly. In outbreaks of *E. coli*, the bacteria typically contaminate ground meat used in making hamburgers. Several massive recalls of frozen meat harboring *E. coli* O157:H7 have occurred in the late 1990's. For example, in 1997 the Hudson Foods plant in Columbus, Nebraska was forced to shut down and 25 million pounds of ground beef were recalled.

4. Bacteria can be grown under strictly controlled conditions and many will grow in a chemically defined culture medium containing mineral salts and a simple organic nutrient such as glucose.
5. Bacteria grow fast and may divide in as little as 20 minutes, whereas higher organisms often take days or years for each generation (Fig. 2.11).
6. A bacterial culture contains around 10^9 cells per ml. Consequently genetic experiments that need to analyze large numbers of cells can be done conveniently.
7. Bacteria can be conveniently stored for short periods (a couple of weeks) by placing them in the refrigerator and for longer periods (20 years or more) in low temperature freezers at –70°C. Upon thawing, the bacteria resume growth. Thus it is not necessary to keep hundreds of cultures of bacterial mutants constantly growing just to keep them alive.

In practice, bacteria are usually cultured by growing them as a suspension in liquid inside tubes, flasks or bottles. They can also be grown as colonies (visible clusters of cells) on the surface of an agar layer in flat dishes, known as Petri dishes (Fig. 2.12). Agar is a carbohydrate polymer extracted from seaweed that sets, or solidifies, like gelatin.

It should be noted that the convenient properties noted above apply to commonly grown laboratory bacteria. In contrast, many bacterial species found in the wild are difficult or, by present techniques impossible, to culture in the laboratory. Many others have specialized growth requirements and most rarely grow to the density observed with the bacteria favored by laboratory researchers.

pathogenic Disease causing

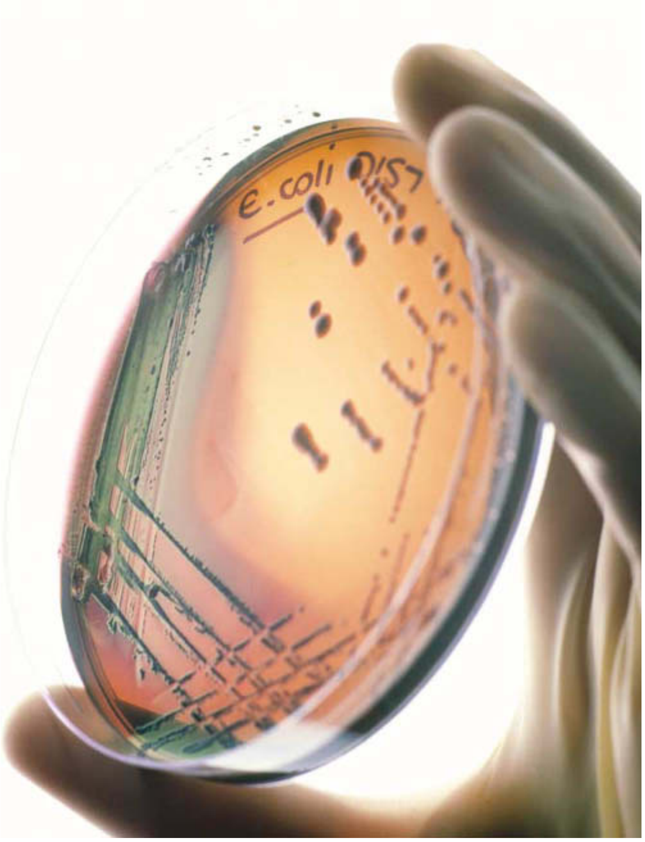


FIGURE 2.12 ***Bacterial Colonies in a Petri Dish***

A Petri dish showing colonies of *Escherichia coli* 0157:H7 growing on nutrient agar medium. This strain sometimes causes foodborne illness. It may cause bloody diarrhoea and occasionally kidney failure, particularly in the elderly or very young. This *E. coli* strain originates from the intestines of cattle and spreads to contaminate beef and milk. Provided by TEK Image, Science Photo Library.

The famous K-12 laboratory strain of *E. coli* was chosen as a research tool because of its fertility. In 1946, Joshua Lederberg was attempting to carry out genetic crosses with bacteria. Until then, no mechanisms for gene transfer had been demonstrated in bacteria, and genetic crosses were therefore thought to be restricted to higher organisms. Lederberg was lucky, as most bacterial strains, including most strains of *E. coli*, do not mate. But among those he tested was one strain (K-12) of *E. coli* that happened to give positive results. Mating in *E. coli* K-12 is actually due to a **plasmid**, an extra circular molecule of DNA within the bacterium that is separate from the chromosome. Because the plasmid carries the genes for fertility, it was named the **F-plasmid**.

According to Jaques Monod, who discovered the operon (see Ch. 9): "What applies to E. coli applies to E. lephant."

Escherichia coli (E. coli) Is a Model Bacterium

Although many different types of bacteria are used in laboratory investigations, the bacterium used most often in molecular biology research is *Escherichia coli*. *E. coli* is a rod-shaped bacterium of approximately 1 by 2.5 microns. Its natural habitat is the colon (hence "coli"), the lower part of the large intestine of mammals, including humans. The knowledge derived by examining *E. coli* has been used to untangle the genetic operation of other organisms. In addition, bacteria, together with their viruses and plasmids, have been used experimentally during the genetic analysis of higher organisms.

F-plasmid A particular plasmid which confers ability to mate on its bacterial host, *Escherichia coli*
plasmid Circular molecule of double stranded helical DNA which replicates independently of the host cell's chromosomes. Rare linear plasmids have been discovered

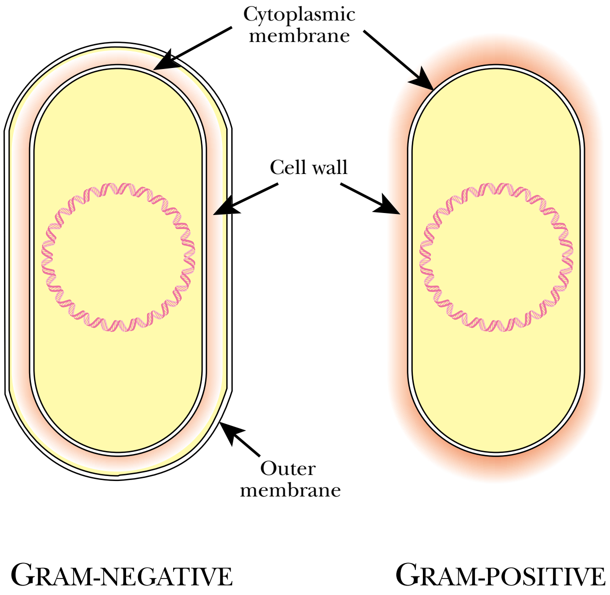


FIGURE 2.13 ***Gram-Negative and Gram-Positive Bacteria***

Gram-negative bacteria have an extra membrane surrounding the cell wall.

E. coli is a **gram-negative bacterium**, which means that it possesses two membranes. Outside the cytoplasmic membrane possessed by all cells are the cell wall and a second, outer membrane (Fig. 2.13). (Although gram-negative bacteria do have two compartments, they are nonetheless genuine prokaryotes, as their chromosome is in the same compartment as the ribosomes and other metabolic machinery. They do not have a nucleus, the key characteristic of a eukaryote). The presence of an outer membrane provides an extra layer of protection to the bacteria. However, it can be inconvenient to the biotechnologist who wishes to manufacture genetically engineered proteins from genes cloned into *E. coli*. The outer membrane hinders protein secretion. Consequently there has been a recent upsurge of interest in **gram-positive** bacteria, such as *Bacillus*, which lack the outer membrane.

Where Are Bacteria Found in Nature?

Bacteria are found almost everywhere. Bacteria have been found 40 miles high in the atmosphere and seven miles deep beneath the ocean floor. Some bacteria live in the sea, others live in fresh water, and others are found growing happily in sewage. Some bacteria live in the soil, some are found living in the roots of plants, and some live inside animals. Most of the bacteria that live inside animals are harmless, and some are even of positive value in aiding digestion or synthesizing vitamins that are absorbed by their host animal.

The total number of bacteria on our planet is estimated at an unbelievable 5×10^{30}. Over 90% are in the soil and subsurface layers below the oceans. The total amount of bacterial carbon is 5×10^{17} grams, nearly equal to the total amount of carbon found in plants. Probably over half of the living matter on Earth is microbial.

Familiar animals and plants are vastly outnumbered by microorganisms, in every natural habitat.

In addition to the "normal" habitats, some bacteria live in extreme environments where most other life forms cannot survive. Some bacteria can live in very concentrated salt solutions, such as the Dead Sea and the Great Salt Lake. Antarctic lakes that only thaw for a short period of each year contain bacteria. Other bacteria inhabit hot sulfur springs, where temperatures approach boiling point and the pH is close to 1. Bacteria even grow in some thermal deep sea vents where the temperature is above 100°C and the high pressure keeps the water liquid. Bacteria from these habitats may

gram-negative bacterium Type of bacterium that has both an inner (cytoplasmic) membrane plus an outer membrane which is located outside the cell wall
gram-positive bacterium Type of bacterium that has only an inner (cytoplasmic) membrane and lacks an outer membrane

Patients are usually given antibiotics to treat bacterial infections. These are chemical substances capable of killing most bacteria by inhibiting specific biochemical processes, but which are relatively harmless to people. The most commonly used antibiotics, the penicillins and cephalosporins, are synthesized by a kind of fungus known as mold (see Fig. 2.14). However, many antibiotics are made by one kind of bacteria in order to kill other types of bacteria. The *Streptomyces* group of soil bacteria produces a wide range of antibiotics including streptomycin, kanamycin and neomycin. Some antibiotics, like chloramphenicol, were originally made by molds but nowadays can be chemically synthesized. Finally, some antibiotics, such as sulfonamides, are entirely artificial and are only synthesized by chemical corporations.

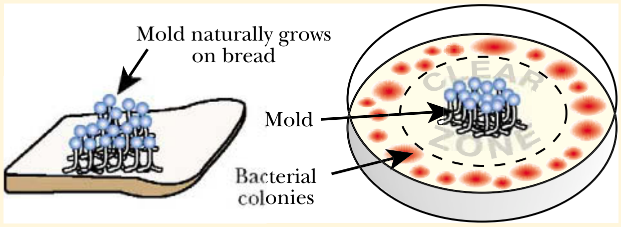


FIGURE 2.14 ***Bacterial Growth Is Suppressed by Bread Mold***

The blue mold that often grows on bread makes **penicillin**. When penicillin is produced by molds grown on agar in a Petri dish, it will diffuse outwards and suppress the growth of bacteria in a circle around it.

provide products that are useful because of their resistance to extreme conditions. *Thermus aquaticus*, a bacterium from hot springs, has provided the heat stable **DNA polymerase** needed for the polymerase chain reaction (PCR), a widely used technique (see Ch. 23).

When different bacteria compete to live in the same habitat, they often resort to biological warfare. Some bacterial strains secrete toxic chemicals in order to kill off others that are competing for the same resources. Certain bacteria synthesize toxic proteins, known as **bacteriocins**. These proteins are designed to kill closely related bacterial strains, yet are harmless to the producer strain. Nisin, a bacteriocin produced by some strains of *Lactococcus lactis* acts as a food preservative and kills food-borne pathogens including *Listeria monocytogenes* and *Staphylococcus aureus*. Nisin and related bacteriocins are relatively short peptides of molecular weight 3.5 kDa. They are formed naturally by the strains of *Lactococcus* that are used to make silages and fermented foods such as wara, a Nigerian cheese product, and kimchi (Korean traditional fermented vegetables). Although scientists have found relatively few practical applications for bacteriocins, the plasmids which carry the genes for bacteriocins have provided the most widely used **vectors** for carrying genes in genetic engineering (described in Ch. 22).

Streptomycin and related **antibiotics** are also made by bacteria, especially those of the *Streptomyces* group, to kill competing bacteria in the soil environment. These antibiotics are not proteins (unlike the colicins) and have been widely used clinically.

antibiotics Chemical substances that inhibit specific biochemical processes and thereby stop bacterial growth selectively; that is, without killing the patient too.
bacteriocin A toxic protein made by bacteria to kill other, closely related, bacteria
DNA polymerase An enzyme that elongates strands of DNA, especially when chromosomes are being replicated
penicillin An antibiotic made by a mold called *Penicillium*, which grows on bread producing a blue layer of fungus
PCR See polymerase chain reaction
vector (a) In molecular biology a vector is molecule of DNA which can replicate and is used to carry cloned genes or DNA fragments; (b) in general biology a vector is an organism (such as a mosquito) that carries and distributes a disease-causing microorganisms (such as yellow fever or malaria)

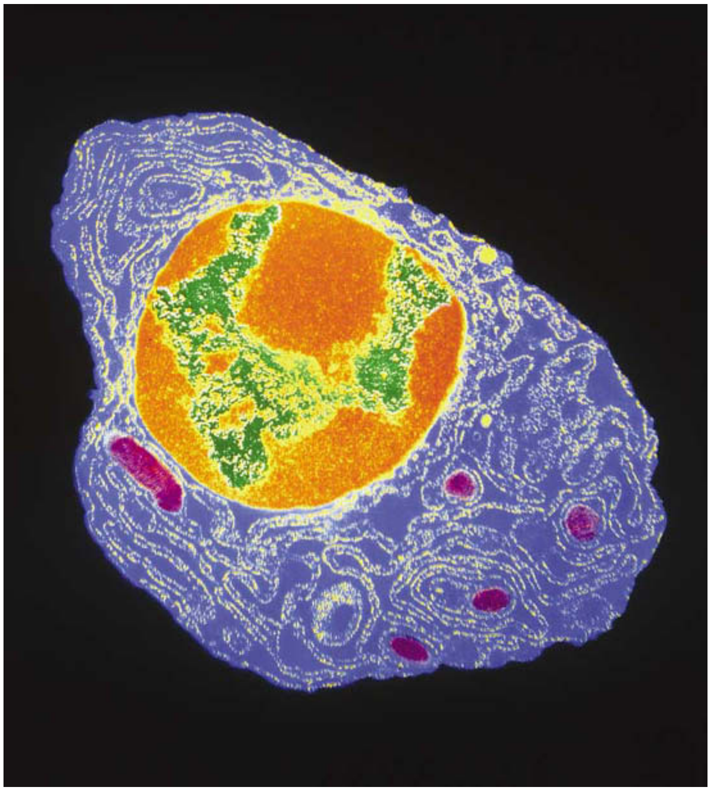


FIGURE 2.15 ***A Eukaryote Has Multiple Cell Compartments***

False color transmission electron micrograph of a plasma cell from bone marrow. Multiple compartments surrounded by membranes, including a nucleus, are found in eukaryotic cells. Characteristic of plasma cells is the arrangement of heterochromatin (orange) in the nucleus, where it adheres to the inner nuclear membrane. Also typical is the network of rough endoplasmic reticulum (yellow dotted lines) in the cytoplasm. The oval or rounded crimson structures in the cytoplasm are mitochondria. Magnification ×4,500. Provided by Dr. Gopal Murti, Science Photo Library.

Some Bacteria Cause Infectious Disease, but Most Are Beneficial

Bacteria are best known to the layman for causing infectious disease. Cholera, tuberculosis, bubonic plague (“Black Death”), anthrax, syphilis, gonorrhea, whooping cough, diphtheria and a variety of other diseases are caused by bacteria. These diseases were widespread before modern technology and hygiene largely eliminated them from advanced societies. This was mostly due to clean water, sewers, flush toilets and soap, rather than specifically “medical” advances such as the use of antibiotics or vaccinations.

Only a small proportion of bacteria causes disease. Many bacteria help maintain the ecosystem by degrading waste materials. For example, soil bacteria degrade the remains of dead plants and animals and take part in the breakdown of animal waste. Bacteria also degrade many man-made chemicals and pollutants. If “good” bacteria did not maintain the environment, higher life-forms could not survive.

If higher organisms disappeared from the Earth, the prokaryotes would survive and evolve. They do not need us although we need them.

Very occasionally bacteria which are even tinier than usual infect other, larger bacteria. This results in a bacterial disease of bacteria! The best known example of this is *Bdellovibrio bacterivorus*. This penetrates the outer membrane of a wide range of gram-negative bacteria, including *E. coli, Pseudomonas*, etc., and takes up residence in the space between the inner and outer membranes. *Bdellovibrio* lives on nutrients it steals from the host cell. After a few hours, the host cell bursts and releases half a dozen new *Bdellovibrio* cells.

Eukaryotic Cells Are Sub-Divided into Compartments

A eukaryotic cell has its genome inside a separate compartment, the nucleus. In fact, eukaryotic cells have multiple internal cell compartments surrounded by membranes (Fig 2.15). The nucleus itself is surrounded by a double membrane, the **nuclear enve-**

nuclear envelope Envelope consisting of two concentric membranes that surrounds the nucleus of eukaryotic cells

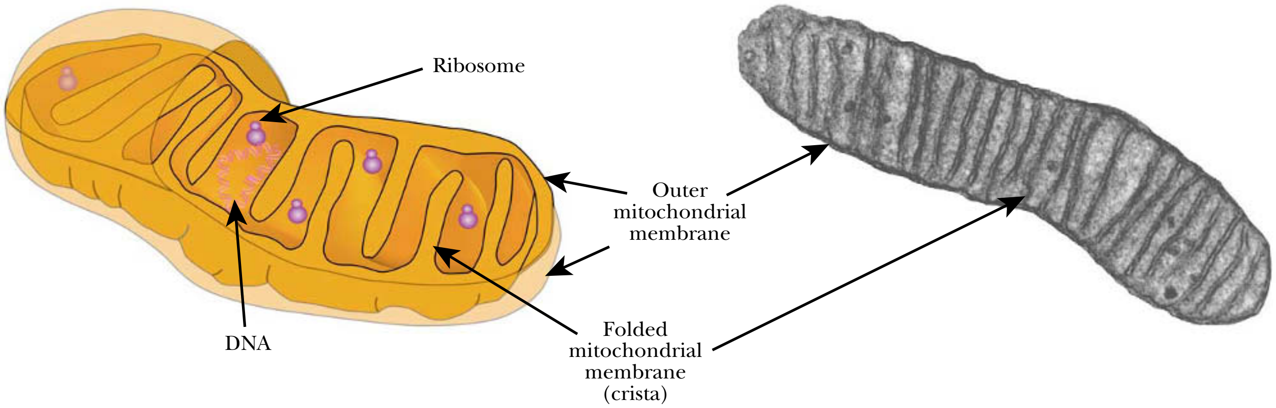

Ribosome
Outer mitochondrial membrane
DNA
Folded mitochondrial membrane (crista)

FIGURE 2.16 ***Mitochondrion***

A mitochondrion is surrounded by two concentric membranes. The inner membrane is folded inward to form **cristae**. These are the site of the respiratory chain that generates energy for the cell.

lope, which separates the nucleus from the cytoplasm, but allows some communication with the cytoplasm via **nuclear pores** (Fig 2.15). The genome of eukaryotes consists of 10,000–50,000 genes carried on several chromosomes. Eukaryotic chromosomes are linear, unlike the circular chromosomes of bacteria. Most eukaryotes are diploid, with two copies of each chromosome. Consequently, they possess at least two copies of each gene. In fact, eukaryotic cells often have multiple copies of certain genes as the result of gene duplication.

Eukaryotes possess a variety of other membranes and **organelles**. Organelles are subcellular structures that carry out specific tasks. Some are separated from the rest of the cell by membranes (so-called **membrane-bound organelles**) but others (e.g., the ribosome) are not. The **endoplasmic reticulum** is a membrane system that is continuous with the nuclear envelope and permeates the cytoplasm. The **Golgi apparatus** is a stack of flattened membrane sacs and associated vesicles that is involved in secretion of proteins, or other materials, to the outside of the cell. **Lysosomes** are membrane-bound structures specialized for digestion, containing degradative enzymes.

Life is modular. Complex organisms are subdivided into organs. Large and complex cells are divided into organelles.

All except a very few eukaryotes contain **mitochondria** (singular, mitochondrion; Fig. 2.16). These are generally rod-shaped organelles, bounded by a double membrane. They resemble bacteria in their overall size and shape. As will be discussed in more detail (see Ch. 20), it is thought that mitochondria are indeed evolved from bacteria that took up residence in the primeval ancestor of eukaryotic cells. Like bacteria, mitochondria each contain a circular molecule of DNA. The mitochondrial genome is similar to a bacterial chromosome, though much smaller. The mitochondrial DNA has some genes needed for mitochondrial function.

Mitochondria are specialized for generating energy by respiration and are found in all eukaryotes. (A few eukaryotes are known that cannot respire; nonetheless these retain remnant mitochondrial organelles—see below.) In eukaryotes, the enzymes of respiration are located on the inner mitochondrial membrane, which has numerous infoldings to create more membrane area. This contrasts with bacteria, where the respiratory chain is located in the cytoplasmic membrane, as no mitochondria are present.

crista (plural cristae) Infolding of the photosynthetic membrane in chloroplast
endoplasmic reticulum Internal system of membranes found in eukaryotic cells
Golgi apparatus A membrane bound organelle that takes part in export of materials from eukaryotic cells
lysosome A membrane bound organelle of eukaryotic cells that contains degradative enzymes
membrane-bound organelles Organelles that are separated from the rest of the cytoplasm by membranes
mitochondrion Membrane-bound organelle found in eukaryotic cells that produces energy by respiration
nuclear pore Pore in the nuclear membrane through which the nucleus communicates with the cytoplasm
organelle Subcellular structure that carries out a specific task. Membrane-bound organelles are separated from the rest of the cytoplasm by membranes but other organelles such as the ribosome are not.

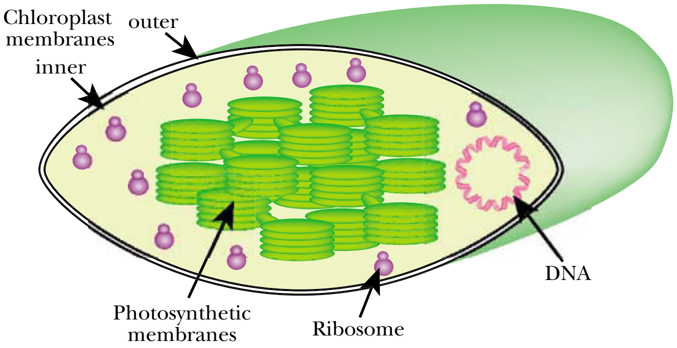


FIGURE 2.17 ***Chloroplast***

The chloroplast is bound by a double membrane and contains infolded stacks of membrane specialized for photosynthesis. The chloroplast also contains ribosomes and DNA.

Chloroplasts are membrane-bound organelles specialized for photosynthesis (Fig. 2.17). They are found only in plants and some single-celled eukaryotes. They are oval to rod shaped and contain complex stacks of internal membranes that contain the green, light-absorbing pigment **chlorophyll** and other components needed for trapping light energy. Like mitochondria, chloroplasts contain a circular DNA molecule and are thought to have evolved from a photosynthetic bacterium.

The Diversity of Eukaryotes

Unlike prokaryotes that fall into two distinct genetic lineages (the eubacteria and archaebacteria), all eukaryotes are genetically related, in the sense of being ultimately derived from the same ancestor. Perhaps this is not surprising since all eukaryotes share many advanced features that the prokaryotes lack. When it is said that all eukaryotes are genetically related, it is in reference to the nuclear part of the eukaryotic genome, not the mitochondrial or chloroplast DNA molecules that have become part of the modern eukaryotic cell.

A wide variety of eukaryotes live as microscopic single cells. However, most eukaryotes are larger multicellular organisms that are visible to the naked eye. Traditionally, these higher organisms have been divided into the plant, fungus and animal kingdoms. This classification still holds, provided one remembers to include several new groups to account for the single-celled eukaryotes. Some single-celled eukaryotes may be viewed as plants, fungi or animals. Others are intermediate or possess a mixture of properties and need their own miniature kingdoms.

Eukaryotes Possess Two Basic Cell Lineages

The most primitive multicellular organisms are merely aggregates of more or less identical cells. However, most multicellular organisms consist of distinct tissues and organs containing a variety of specialized cells. Furthermore, most cells in higher organisms do not contribute to the next generation, but die when the multicellular individual of whom they are part dies. These are known as **somatic cells** (Fig. 2.20). Only the **germ line cells** take part in forming a new individual. This, of course, complicates genetic analysis. Although all cells in any multicellular organism start with an identical copy of the genome, they differentiate to give quite different structures that perform different functions. Understanding development is a major challenge facing molecular biology today. In animals there is a sharp division between somatic cells and germ line cells that persists throughout the life cycle. However, plants do not set aside special germ cells until close to the time that gametes are made.

germ line cells Reproductive cells producing eggs or sperm that take part in forming the next generation
chlorophyll Green pigment that absorbs light during photosynthesis
somatic cells Cells making up the body but which are not part of the germ cell line.

The Symbiotic Theory of Organelle Origins

A well accepted theory of mitochondrial (and chloroplast) origin is that certain bacteria were ingested by ancestral eukaryotes and have lived in a symbiotic relationship with their descendents ever since. Figure 2.18 suggests how this could have occurred. The mitochondrion contains DNA and ribosomes. The DNA of the mitochondria more closely resembles that of bacteria than of eukaryotes.

Certain primitive single-celled eukaryotes, such as ***Entamoeba*** and ***Giardia***, lack the ability to respire and instead live by **fermentation** (Fig. 2.19). It was once believed that they lacked mitochondria and had branched off from the ancestral eukaryote before it had captured the bacterium that gave rise to the mitochondrion. More recently, it was suggested that the ancestors to these organisms did originally possess mitochondria, but lost them secondarily during the course of evolution. However, recent work has shown that even *Entamoeba* and *Giardia* retain small remnant organelles ("mitosomes") corresponding to mitochondria. Although the capability for respiration has indeed been completely lost, the remnant organelles function in assembling the iron sulfur clusters found in several essential proteins.

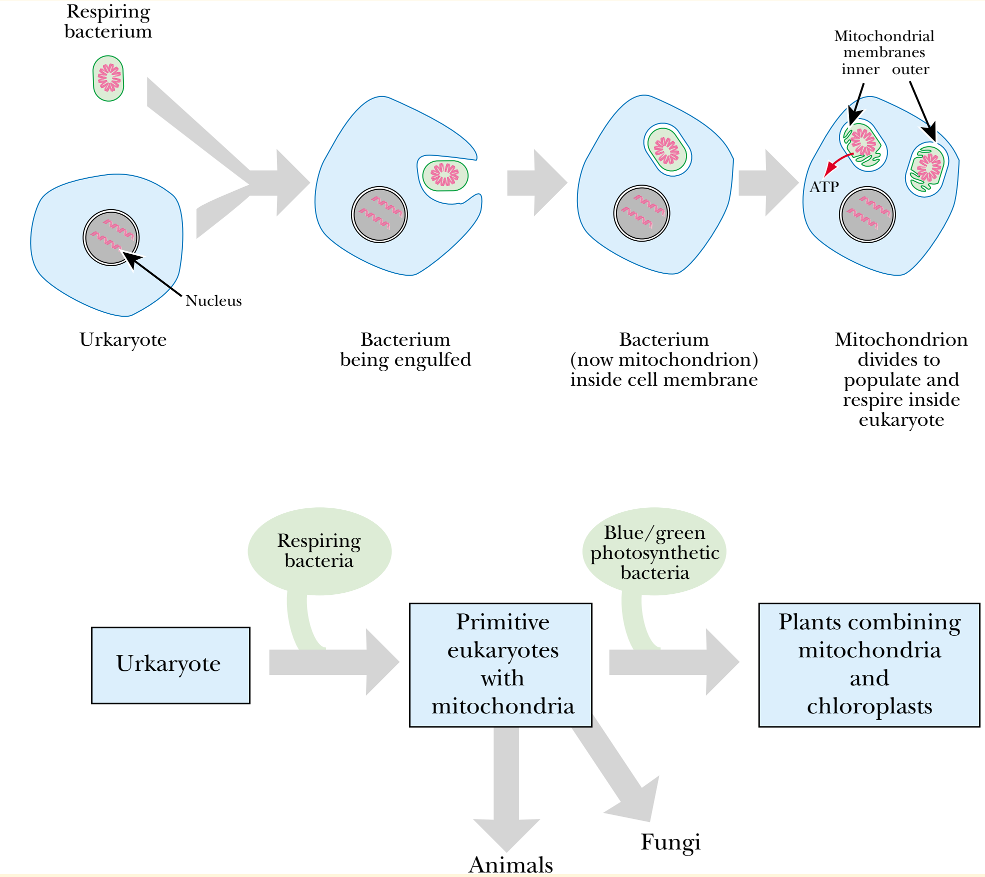

Respiring bacterium
Mitochondrial membranes
inner outer
ATP
Nucleus
Urkaryote
Bacterium being engulfed
Bacterium (now mitochondrion) inside cell membrane
Mitochondrion divides to populate and respire inside eukaryote
Respiring bacteria
Blue/green photosynthetic bacteria
Urkaryote
Primitive eukaryotes with mitochondria
Plants combining mitochondria and chloroplasts
Animals
Fungi

FIGURE 2.18 ***Symbiosis with Respiring Bacteria Gives Rise to the Primitive Eukaryote***

The ancestor to the eukaryote, or "urkaryote" engulfs a respiring bacterium by surrounding it with an infolding of the cell membrane. Consequently there is now a double membrane around the newly enveloped bacterium. The symbiont, now called a "mitochondrion", divides by fission like a bacterium and provides energy for the primitive eukaryote. The mitochondrion develops infoldings of the inner membrane that increase its energy producing capacity.

Entamoeba A very primitive single-celled eukaryote that lacks mitochondria
fermentation A biochemical process that releases energy without oxygen or light
Giardia A very primitive single-celled eukaryote that lacks mitochondria

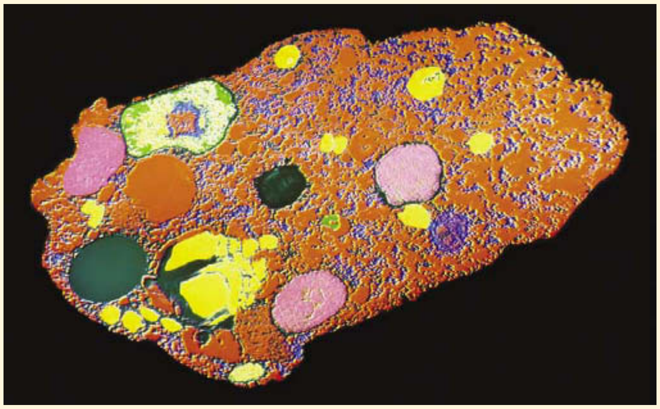


FIGURE 2.19 Entamoeba: *an Anaerobic Eukaryote*

Some single-celled eukaryotes lack true respiratory mitochondria and must grow by fermentation. Shown here is a false-color transmission electron micrograph of *Entamoeba histolytica*, a parasitic amoeba, which is ingesting human red blood cells (green ovals). The white/green oval (at left) with a blue and pink central circular area is the nucleus. *Entamoeba* invades and destroys the tissues of the intestines, causing amoebic dysentery. It may spread to the liver causing abscesses to develop. The infection is acquired through contamination of food or water or through the agency of flies. Magnification: ×830. Courtesy of: London School of Hygiene & Tropical Medicine, Science Photo Library.

Organisms Are Classified

Living organisms have two names, both printed in italics; for example, *Escherichia coli* or *Saccharomyces cerevisiae*. The first name refers to the **genus** (plural, genera), a group of closely related species. After its first use in a publication, the genus name is often abbreviated to a single letter, as in "*E. coli*." Next comes the species, or individual, name. The genus and species are the smallest subdivision of the system of biological classification. Classification of living organisms facilitates the understanding of their origins and the relationships of their structure and function. The highest level of classification is the **domain**. There are considered to be three domains:

Biological classification attempts to impose a convenient filing system upon organisms related by continuous evolutionary branching.

1. ***Eubacteria*** These are prokaryotic cells (traditional bacteria). Interestingly, this group includes the genomes of mitochondria and chloroplasts that have been symbiotically related to eukaryotes.
2. ***Archaebacteria:*** From a structural viewpoint, these are prokaryotes like eubacteria in that they lack a nucleus. However, their gene sequences and other biochemical features indicate they are, if anything, slightly more closely related genetically to eukaryotes than to eubacteria.
3. ***Eukaryotes:*** Higher organisms whose DNA is carried on several chromosomes which are found inside the nucleus. Their cells are divided into separate compartments and usually contain other organelles in addition to the nucleus. Eukaryotes are divided into four **kingdoms**:

   Protoctista—An accumulation of primitive, mostly single-celled eukaryotes often referred to as protists that don't belong to the other three main kingdoms. There are several groups that are distinct enough that some scientists would elevate them in rank to miniature kingdoms.

domain (of life) Highest ranking group into which living creatures are divided, based on the most fundamental genetic properties
genus A group of closely related species
kingdom Major subdivision of eukaryotic organisms, in particular the plant, fungus and animal kingdoms

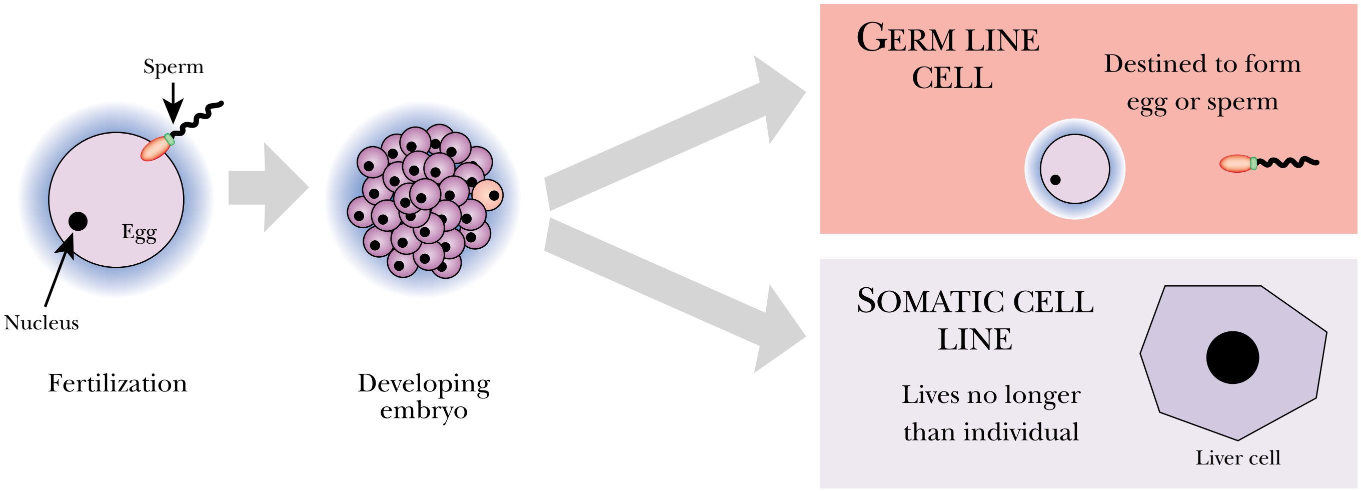


FIGURE 2.20 ***Somatic Cells versus Germ Line***

After an egg is fertilized and begins its development into an animal embryo, cells have two fates. A small number of cells form the germ line, which gives rise to the gametes (eggs or sperm) that give rise to future generations. However, most cells are part of the somatic cell line, which forms the remainder of the organism. These somatic cells will die either before the organism as a whole, or with it, as part of the natural life cycle.

Domain
Kingdom
Phylum
Class
Order
Family
Genus
species

Plants—Possess both mitochondria and chloroplasts and are photosynthetic. Typically they are non-mobile and have rigid cell walls made of cellulose.

Fungi—Possess mitochondria but lack chloroplasts. Once thought to be plants that had lost their chloroplasts, it is now thought they never had them. Their nourishment comes from decaying biomatter. Although fungi are non-mobile, they lack cellulose and their cell walls are made of chitin. They may be more closely related to animals than plants.

Animals—Lack chloroplasts but possess mitochondria. Differ from fungi and plants in lacking a rigid cell wall. Typically mobile. They are divided into 20 to 30 **phyla** (singular, phylum), depending somewhat on personal taste. Some phyla include:

Porifera—sponges

Cnidaria—sea anemones and jellyfish

Platyhelminthes—flatworms

Nematoda—roundworms

Arthropoda—insects, crustaceans, etc.

Annelida—segmented worms, such as earthworms

Mollusca—snails, squids, etc.

phylum (plural phyla) Major groups into which animals are divided, roughly equivalent in rank to the divisions of plants or bacteria

Echinodermata—starfish, sea urchins

Chordata—vertebrates and their relatives.

Phyla are divided into classes, such as mammals.

Classes are divided into orders, such as primates.

Orders are divided into families, such as hominids.

Families are divided into genera, such as *Homo*.

Genera are divided into species, such as *Homo sapiens*

Some Widely Studied Organisms Serve as Models

Biologists have always concentrated their attention on certain living organisms, either because they are convenient to study or are of practical importance. Inevitably, model organisms are atypical in some respects. For example, few bacteria grow as fast as *E. coli* and few mammals breed as fast as mice. Nonetheless, information discovered in such model systems is assumed to apply also to related organisms. In practice this often proves to be true, at least to a first approximation. As discussed above, the basic principles of molecular biology have been investigated in simple single-celled prokaryotes. However, to obtain knowledge that is useful in medicine and agriculture, researchers need model organisms that are much more closely related to humans and to crop plants, respectively. Even these models have their limitations; ultimately, human cells and agriculturally useful animals and plants have to be studied directly.

Yeast Is a Widely Studied Single-Celled Eukaryote

Biotechnology is a new word but not a new occupation. Brewing and baking both use yeast and date back to the earliest human civilizations.

Yeast is widely used in molecular biology for many of the same reasons as bacteria. It is the eukaryote about which most is known and the first whose genome was sequenced—in 1996. Yeasts are members of the fungus kingdom and are about equally related to animals and plants. A variety of yeasts are found in nature, but the one normally used in the laboratory is brewer's yeast, *Saccharomyces cerevisiae* (Fig. 2.21). This is a single-celled eukaryote that is easy to grow in culture. Even before the age of molecular biology, yeast was widely used as a source of material for biochemical analysis. The first enzymatic reactions were characterized in extracts of yeast and the word enzyme is derived from the Greek for "in yeast".

Although it is a "higher organism", yeast measures up quite well to the list of useful properties that make bacteria easy to study. In addition, it is less complex genetically than many other eukaryotes:

a. Yeast is single-celled microorganism. Like bacteria, a yeast culture consists of many identical cells. Although larger than bacteria, yeast cells are only about a tenth the size of the cells of higher animals.

b. Yeast has a haploid genome of about 12 Mb of DNA with about 6,000 genes, as compared to *E. coli*, which has 4,000 genes, and humans, who have approximately 25,000.

c. The natural life cycle of yeast alternates between a diploid phase and a haploid phase. Thus it is possible to grow haploid cultures of yeast, which, like bacteria, have only a single copy of each gene, making research interpretations easy.

d. Unlike many higher organisms, yeast has relatively few of its genes—about 5%—interrupted by intervening sequences, or introns.

e. Yeast can be grown under controlled conditions in chemically defined culture medium and forms colonies on agar like bacteria.

f. Yeast grows fast, though not as fast as bacteria. The cell cycle takes approximately 90 minutes (compared to around 20 minutes for fast growing bacteria).

g. Yeast cultures can contain around 10^9 cells per ml of culture media, like bacteria.

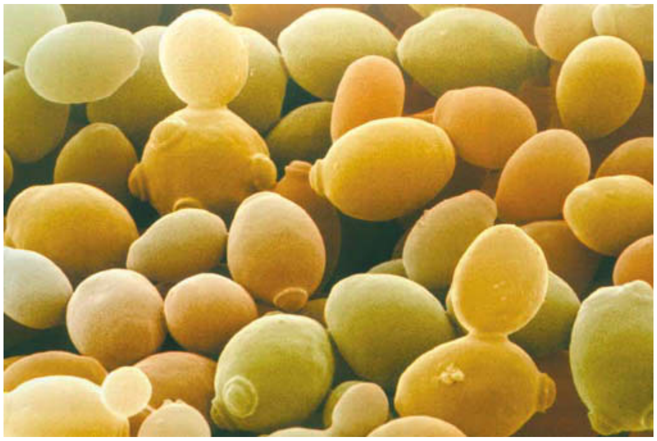


FIGURE 2.21 ***Yeast Cells***

Colored scanning electron micrograph (SEM) of budding yeast cells (*Saccharomyces cerevisiae*). The larger mother cells are budding off smaller daughter cells. Magnification: ×4,000. Courtesy of: Andrew Syred, Science Photo Library.

h. Yeast can be readily stored at low temperatures.

i. Genetic analysis using recombination is much more powerful in yeast than in higher eukaryotes. Consequently, collections of yeast strains that each have one yeast gene deleted are available.

Yeast illustrates the genetic characteristics of higher organisms in a simplified manner.

Yeast may grow as diploid or haploid cells (Fig. 2.22). Both haploid and diploid yeast cells grow by **budding**, rather than symmetrical cell division. In budding, a bulge, referred to as a bud, forms on the side of the mother cell. The bud gets larger and one of the nuclei resulting from nuclear division moves into the bud. Finally, the cross wall develops and the new cell buds off from the mother. Especially under conditions of nutritional deprivation, diploid yeast cells may divide by meiosis to form haploid cells, each with a different genetic constitution. This process is analogous to the formation of egg and sperm cells in higher eukaryotes. However, in yeast, the haploid cells appear identical and there is no way to tell the sexes apart and so we refer to mating types. In contrast to the haploid gametes of animals and plants, the haploid cells of yeast may grow and divide indefinitely in culture. Two haploid cells, of opposite mating types, may fuse to form a zygote.

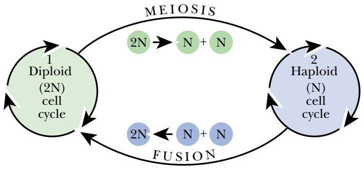


FIGURE 2.22 ***Yeast Life Cycle***

The yeast cell alternates between haploid and diploid phases and is capable of growth and cell division in either phase.

In its haploid phase, *Saccharomyces cerevisiae* has 16 chromosomes and nearly three times as much DNA as *E. coli*. Despite this, it only has 1.5 times as many genes as *E. coli*. Thus a substantial portion of yeast DNA apparently lacks genetic information and so is **non-coding DNA**. It is easier to use the haploid phase of yeast for isolating mutations and analyzing their effects. Nonetheless, the diploid phase is also useful for studying how two alleles of the same gene interact in the same cell. Thus, yeast can be used as a model to study the diploid state and yet take advantage of its haploid phase for most of the genetic analysis.

A Roundworm and a Fly Are Model Multicellular Animals

> "If all the matter in the universe except the nematodes were swept away, our world would still be dimly recognizable. . ."
>
> —N.A. Cobb, 1914

Nematodes in oceanic mud or inland soils may all look the same. Nonetheless, they harbor colossal genetic diversity.

Ultimately, researchers have to study multicellular creatures. The most primitive of these that is widely used is the roundworm, *Caenorhabditis elegans*. Nematodes, or roundworms, are best known as parasites both of animals and plants. Although it is related to the "eelworms"—nematodes that attack the roots of crop plants—*C. elegans*, is a free-living and harmless soil inhabitant that lives by eating bacteria. A single acre

budding Type of cell division seen in yeasts in which a new cell forms as a bulge on the mother cell, enlarges, and finally separates
non-coding DNA DNA sequences that do not code for proteins or functional RNA molecules

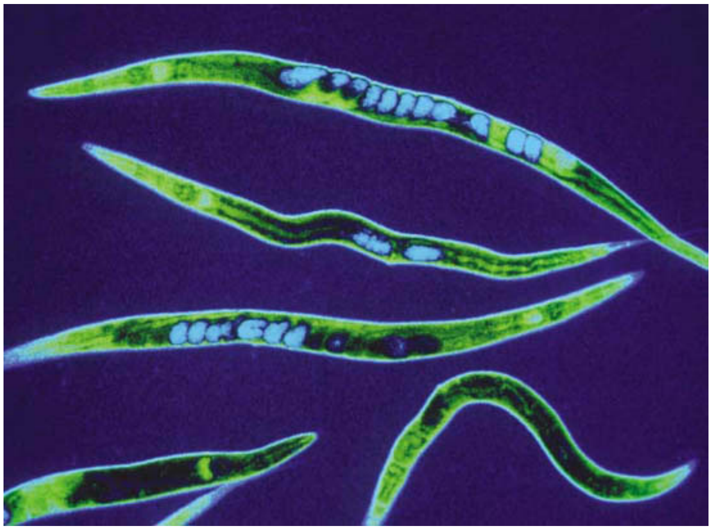


FIGURE 2.23
Caenorhabditis elegans

False-color scanning optical micrograph of the soil-dwelling bisexual nematode *Caenorhabditis elegans*. The round internal structures are eggs. *C. elegans* is convenient for genetic analysis because of its tendency to reproduce by self-fertilization. This results in offspring that are all identical to the parent. It takes only three days to reach maturity and thousands of worms can be kept on a culture plate. Approximate magnification: ×80C. Courtesy of: James King-Holmes, Science Photo Library.

of soil in arable land may contain as many as 3,000 million nematodes belonging to dozens of different species.

The haploid genome of *Caenorhabditis elegans* consists of 97 Mb of DNA carried on six chromosomes. This is about seven times as much total DNA as in a typical yeast genome. *C. elegans* has an estimated 20,000 genes and so contains a much greater proportion of non-coding DNA than lower eukaryotes such as yeast. Its genes contain an average of 4 intervening sequences each.

The adult *C. elegans* is about 1 mm long and has 959 cells and the lineage of each has been completely traced from the fertilized egg (i.e., the zygote). It is thus a useful model for the study of animal development. In particular, **apoptosis**, or programmed cell death, was first discovered and has since been analyzed genetically using *C. elegans*. Although very convenient in the special case of *C. elegans*, such a fixed number of cells in an adult multicellular animal is extremely rare. *C. elegans*, which lives about 2–3 weeks, is also used to study life span and the aging process. RNA interference, a gene-silencing technique that relies on double-stranded RNA, was discovered in *C. elegans* in 1998 and is now used to study gene function during development in worms and other higher animals. RNA interference is discussed in Ch. 11.

As noted in Chapter 1, the fruit fly, *Drosophila melanogaster* (usually called *Drosophila*) was chosen for genetic analysis in the early part of the 20th century. Fruit flies live on rotten fruit and have a 2 week life cycle, during which the female lays several hundred eggs. The adults are about 3 mm long and the eggs about 0.5 mm. Once molecular biology came into vogue it became worthwhile to investigate *Drosophila* at the molecular level, in order to take advantage of the wealth of genetic information already available. The haploid genome has 180 Mb of DNA carried on 4 chromosomes. Although we normally think of *Drosophila* as more advanced than a primitive roundworm, it has an estimated 14,000 genes—6,000 fewer than the roundworm, *C. elegans*. Genes from *Drosophila* contain approximately 3 intervening sequences each on average. Research on *Drosophila* has concentrated on cell differentiation, development, signal transduction and behavior.

Zebrafish are used to Study Vertebrate Development

Danio rerio, (previously *Brachydanio rerio*) the zebrafish, is increasingly being used as a model for studying genetic effects in vertebrate development. Zebrafish are native to the slow freshwater streams and rice paddies of East India and Burma, including the Ganges River. They are small, hardy fish, about an inch long that have been bred

apoptosis Programmed suicide of unwanted cells during development or to fight infection

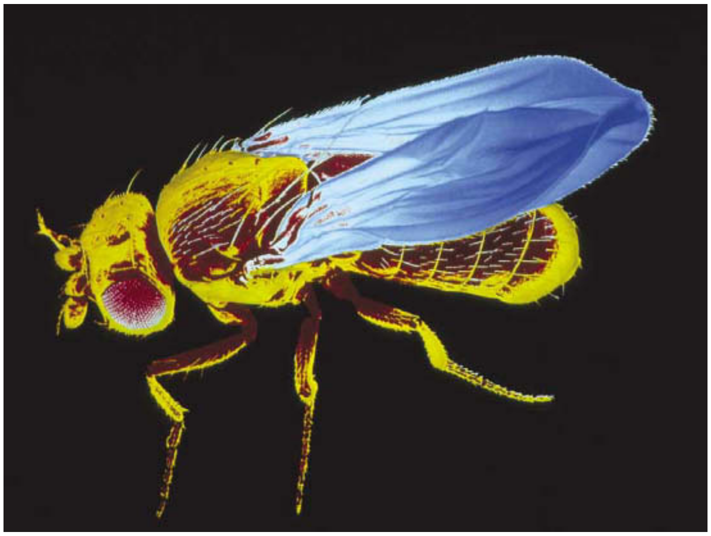


FIGURE 2.24 Drosophila melanogaster*, the Fruit Fly*

False-color scanning electron micrograph of the fruit fly *Drosophila melanogaster*. The fruit fly has many external characteristics that can reveal mutation events. This specimen is of the wild type, known as *Oregon R*. Magnification: ×18. Courtesy of: Dr Jeremy Burgess, Science Photo Library.

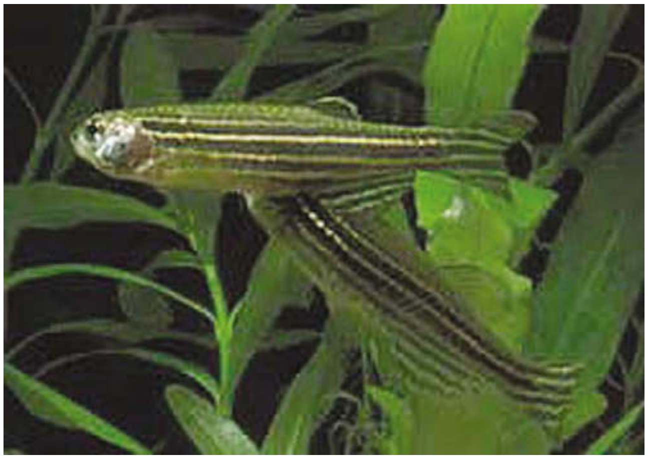


FIGURE 2.25 *Danio rerio*

Danio rerio, the zebrafish, has recently been adopted as a model for the genetic study of embryonic development in higher animals.

Late in 2003, zebrafish became the first commercially available genetically engineered pets. Fluorescent red zebrafish are marketed in the USA by Yorktown Technologies as GloFish^{TM}. They fluoresce red when illuminated with white light, or better, black light (i.e. near UV) due to the presence of a gene for a red fluorescent protein taken from a sea coral. The principle is similar to that of the widely used green fluorescent protein taken from jellyfish (see Ch. 25 for use of GFP in genetic analysis). The price of about $5 per fish makes GloFish^{TM} about five times as expensive as normal zebrafish. The fish were developed at the National University of Singapore by researcher Zhiyuan Gong with the ultimate objective of monitoring pollution. A second generation of more specialized red fluorescent zebrafish will fluoresce in response to toxins or pollutants in the environment.

for many years by fish hobbyists in home aquariums where they may survive for about five years. The standard "wild-type" is clear-colored with black stripes that run lengthwise down its body (Fig. 2.25). Its eggs are laid in clutches of about 200. They are clear and develop outside the mother's body, so it is possible to watch a zebrafish egg grow into a newly formed fish under a microscope. Development from egg to adult takes about three months. Zebrafish are unusual in being nearly transparent so it is possible to observe the development of the internal organs.

Zebrafish have about 1,700 Mb of DNA on 25 chromosomes and show about 75% homology with the human genome. Genetic tagging is relatively easy and microinjecting the eggs with DNA is straightforward. Consequently, the zebrafish has become a favorite model organism for studying the molecular genetics of embryonic development.

Mouse and Man

Only a few animals have been investigated intensively. The rest are assumed to be similar except for minor details.

The ultimate aim of molecular medicine is to understand human physiology at the molecular level and to apply this knowledge in curing disease. As discussed in Chapter 24, science now has available the complete sequence of the human genome, but researchers have little idea of what the products of most of these genes actually do. Since direct experimentation with humans is greatly restricted, animal models are necessary. Although a range of animals has been used to investigate various topics, the rat and the mouse are the most widespread laboratory animals. Rats were favored in the early days of biochemistry when metabolic reactions were being characterized. Mice are smaller and breed faster than rats, and are easier to modify genetically. Consequently, the mouse is used more often for experiments involving genetics and molecular biology. Mice live from one to three years and become sexually mature after about 4 weeks. Pregnancy lasts about three weeks and may result in up to 10 offspring per birth.

Humans have two copies each of approximately 30,000 genes scattered over 23 pairs of chromosomes. Mice have a similar genome, of 2,600 Mb of DNA carried on 20 pairs of chromosomes. Less than 1% of mouse genes lack a homolog in the human genome. The average mouse (or human) gene extends over 40 kilobases of DNA that consists mostly of non-coding intervening sequences (approximately 7 per gene). Nowadays there are many strains of mutant mice in which one or more particular genes have been altered or disrupted. These are used to investigate gene function (Fig. 2.26).

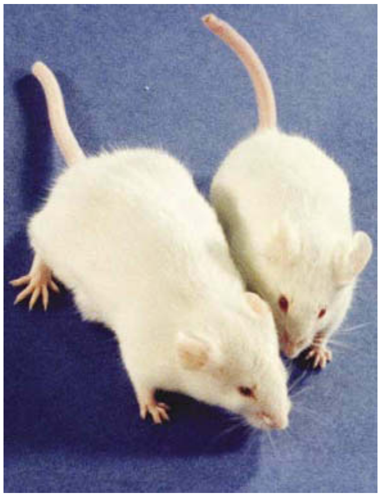


FIGURE 2.26 ***Transgenic Mice***

The larger mouse contains an artificially introduced human gene, which causes a difference in growth. Mice with the human growth hormone gene grow larger compared to mice without this gene.

Intact humans cannot be used for routine experiments for ethical reasons. However, it is possible to grow cells from both humans and other mammals in culture. Many cell lines from humans and monkeys are now available. Such cells are much more difficult to culture than genuine single-celled organisms. Cell lines from multicellular organisms allow fundamental investigations into the genome and other cell components. Historically, the most commonly used cell lines, e.g. HeLa cells, are actually cancer cells. Unlike cells that retain normal growth regulation, cancer cells are "immortalized", that is they are not limited to a fixed number of generations. In addition, cancer cell lines can often divide in culture in the absence of the complex growth factors needed to permit the division of normal cells.

Arabidopsis Serves as a Model for Plants

Flowering plants have more genes than any other type of organism. The function of most of these genes is still a mystery.

Historically, the molecular biology of plants has lagged behind other groups of organisms. Ironically, plants now hold the record for the highest number of genes (40,000 to 50,000 genes for rice—some 10,000 more than humans). If our criterion for superiority is gene number, then it is the plants who represent the height of evolution, not mammals. Why do plants have so many genes? One suggestion is that because plants are immobile they cannot avoid danger by moving. Instead they must stand and face it like a man—or rather like a vegetable. This means that plants have accumulated many genes involved in defense against predators and pests as well as for adapting to changing environmental conditions. One of the most active areas in biotechnology today is the further genetic improvement of crop plants. Genetic manipulation of plants is not hindered by the ethical considerations that apply to research on animals or humans. Moreover, crop farming is big business.

Arabidopsis thaliana, the mouse-ear cress, has become the model for the molecular genetics of higher plants. It is structurally simple and also has the smallest genome of any flowering plant, 125 Mb of DNA—just over 10 times as much DNA as yeast, yet carried on only five pairs of chromosomes. *Arabidopsis* has an estimated 25,000 genes with an average of 4 intervening sequences per gene. *Arabidopsis* can be grown indoors and takes about 6–10 weeks to produce several thousand offspring from a single original plant. Though slow by bacterial standards, this is much faster than waiting a year for a new crop of peas, as Mendel did.

Arabidopsis shares with yeast the ability to grow in the haploid state, which greatly facilitates genetic analysis. Pollen grains are the male germ line cells of plants and are therefore haploid. When pollen from some plants, including *Arabidopsis*, is grown in tissue culture, the haploid cells grow and divide and may eventually develop into normal looking plants. These are haploid, and therefore sterile. Diploid plants may be reconstituted by fusion of cells from two haploid cell lines. Alternatively, diploidy can be artificially induced by agents such as colchicine that interfere with mitosis to cause a doubling of the chromosome number. In the latter case, the new diploid line will be homozygous for all genes.

FIGURE 2.27 Arabidopsis thaliana, *the Mouse-ear Cress*

The plant most heavily used as a model for molecular biology research is *Arabidopsis thaliana*, a member of mustard family (Brassicaceae). Common names include Mouse-ear cress, Thale cress and Mustard weed. Courtesy of: Dr Jeremy Burgess, Science Photo Library.

Haploidy, Diploidy and the Eukaryote Cell Cycle

Eukaryotes are normally regarded as diploid, having two copies of each gene carried on pairs of homologous chromosomes. While this is true of the majority of multicellular animals and many single-celled eukaryotes, there are significant exceptions. Many plants are polyploid, especially angiosperms (flowering plants). About half of the present-day angiosperms are thought to be polyploid, especially tetraploid or hexaploid. For example, coffee (ancestral haploid number = 11) exists as variants with 22, 44, 66, or 88 chromosomes (i.e. 2n, 4n, 6n and 8n). Polyploid plants have larger cells and the plants themselves are often larger. In particular, polyploids have often been selected among domesticated crop plants, since they tend to give bigger plants with higher yields.

Polyploidy is unusual in animals, being found in occasional insects and reptiles. So far the only polyploid mammal known is a rat from Argentina that was discovered to be tetraploid in 1999. It actually has only 102 chromosomes, having lost several from the original tetraploid set of 4n = 112. The tetraploid rat has larger cells than its diploid relatives. The only haploid animal known is an arthropod, a mite, *Brevipalpus phoenicis*, which was discovered in 2001. Infection of these mites by an endosymbiotic bacterium causes feminization of the males. The genetic females of this species reproduce by parthenogenesis (i.e. development of unfertilized eggs into new individuals).

Many eukaryotes alternate between haploid and diploid phases. However, the properties and relative importance of the two phases varies greatly with the organism.

In most animals, only the gametes, the egg and sperm cells, are haploid. After mating two haploid gametes fuse to give a diploid zygote that develops into a new animal. However, in plants and fungi, haploid cells often grow and divide for several generations before producing the actual gametes. It seems likely that in the ancestral eukaryote a phase consisting of haploid cells alternated with a diploid phase. In yeasts, both haploid and diploid cells may be found and both types grow and divide in essentially the same manner (see above). In lower plants, such as mosses and liverworts, the haploid phase, or **gametophyte**, may even form a distinct multicellular plant body.

The concept of germline versus somatic cells applies to animals but not to other higher organisms.

During animal development, there is an early division into **germline** and **somatic** cells. Only cells from the germline can form gametes and contribute to the next generation of animals. Somatic cells have no long-term future but grow and divide only as long as the individual animal continues to live. Hence, genetic defects arising in somatic cells cannot be passed on through the gametes to the next generation of animals. However, they may be passed on to other somatic cells. Such somatic inheritance is of great importance as it provides the mechanism for cancer. In plants and fungi there is no rigid division into germline and somatic cells. The cells of many higher plants are **totipotent**. In other words, a single cell from any part of the plant has the potential to develop into a complete new plant, which can develop reproductive tissues

gametophyte Haploid phase of a plant, especially of lower plants such as mosses and liverworts, where it forms a distinct multicellular body
germline cell Cell capable of forming gametes and so contributing to the next generation of animals
somatic cell Cell making up the body, as opposed to the germline
totipotent Capable of giving rise to a complete multicellular organism

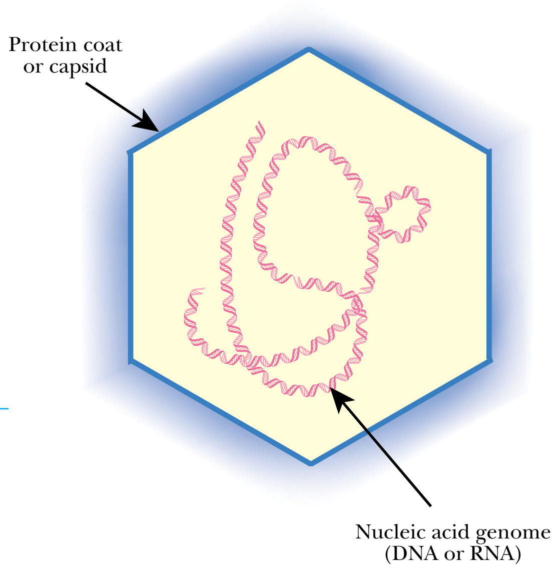


FIGURE 2.28 ***Structural Components of a Virus***

A virus is composed of a protein coat and nucleic acid. Note that there are no ribosomes or phospholipid membranes and only one type of nucleic acid is present.

TABLE 2.01 Polyploidy in Crop Plants

Plant	Ancestral haploid number	Chromosome number	Ploidy level
wheat	7	42	6n
domestic oat	7	42	6n
peanut	10	40	4n
sugar cane	10	80	8n
white potato	12	48	4n
tobacco	12	48	4n
cotton	13	52	4n

and produce gametes. This is not normally possible for animal cells. [The experimental cloning of animals, such as Dolly the sheep, is an artificial exception to this rule.]

Viruses Are Not Living Cells

The characteristics of living cells were outlined early in this chapter. A common, but somewhat technical, definition of a living cell is as follows: Living cells contain both DNA and RNA and can use the genetic information encoded in these to synthesize proteins by using energy that they generate themselves. This definition is designed not so much to explain, positively, how a cell works as to exclude **viruses** from the realm of living cells. The essential features of a virus are shown in Figure 2.28. Viruses are packages of genes in protein coats and are much smaller than bacteria. Viruses are obligate **parasites** that must infect a host cell in order to replicate themselves. Whether viruses are alive or not is a matter of opinion; however, viruses are certainly not living cells. Virus particles (**virions**) do contain genetic information in the form of DNA or RNA, but are incapable of growth or division by themselves. A virus may have its

parasite An organism or genetic entity that replicates at the expense of another creature
virion A virus particle
virus Subcellular parasite with genes of DNA or RNA which replicates inside the host cell upon which it relies for energy and protein synthesis. In addition, it has an extracellular form, in which the virus genes are contained inside a protective coat

Viruses are packages of genes that are not alive by themselves but may take over living cells. Once in control the virus uses the cell's resources to manufacture more viruses.

genome made of DNA or RNA, but only one type of nucleic acid is present in the virion of any given type of virus.

Viruses lack the machinery to generate their own energy or to synthesize protein. After invading a host cell, the virus does not grow and divide like a cell itself. The virion disassembles and the virus genes are expressed using the machinery of the host cell. In particular, viral proteins are made by the host cell ribosomes, using virus genetic information. In many cases, only the virus DNA or RNA enters the host cell and the other components are abandoned outside. After infection, virus components are manufactured by the infected cell, as directed by the virus, and are assembled into new virus particles. Usually the host cell is killed and disintegrates. Typically, several hundred viruses may be released from a single infected cell. The viruses then abandon the cell and look for another host. [Note that some viruses cause "chronic" or "persistent" infections where virus particles are made slowly and released intermittently rather than as a single burst. In this case the host cell may survive for a long time despite infection. In addition, many viruses may persist inside the host cell for a long time in a latent, non-replicating, state and only change to replicative mode under certain conditions—see Ch. 17.]

Some scientists regard viruses as being alive based on the viral possession of genetic information. The majority, however, do not accept that viruses are truly alive, since viruses are unable to generate energy or to synthesize protein. Viruses are thus on the borderline between living and non-living. Virus particles are in suspended animation, waiting for a genuine living cell to come along so they can infect it and replicate themselves. Nonetheless, a host cell whose life processes have been subverted by a virus does duplicate the viral genetic information and produces more virus particles. Thus viruses possess some of the properties of living creatures. Viruses are very important from a practical viewpoint. Firstly, many serious diseases are due to virus infection. Secondly, many genetic manipulations that are now used in genetic engineering are carried out using viruses.

Merely being a parasite does not prevent an organism from being a living organism. For example, **rickettsias** are degenerate bacteria that cause typhus fever and related diseases. They cannot grow and divide unless they infect a suitable host cell. However, rickettsias can generate energy and make their own proteins, provided they obtain sufficient complex nutrients from the animal cell they invade. Furthermore, rickettsias reproduce by growing and dividing like other bacteria. Viruses are subcellular parasites, totally dependent on other life forms for their energy, materials and even the equipment to manufacture their own components.

Bacterial Viruses Infect Bacteria

Even bacteria can get sick, usually as the result of infection by a virus. Bacterial viruses are sometimes referred to as **bacteriophages**, or phages for short. Phage comes from a Greek word meaning to eat. When bacteria catch a virus, they do not merely get a mild infection, like a cold, as humans usually do. They are doomed. The bacteriophage takes over the bacterial cell and fills it up by manufacturing more bacteriophages, as shown in Figure 2.29. Then the bacterial cell bursts and liberates the new crop of bacteriophages to infect more bacteria. This takes only about an hour or so. In a matter of hours, a bacteriophage epidemic could wipe out a culture of bacteria numbering several times the earth's human population.

Bacterial viruses infect only bacteria. Some have relatively broad host ranges, whereas others infect only a single species or even just a few particular strains of bacteria. Generally speaking, any particular disease, whether caused by bacteria or by viruses, infects only a closely related group of organisms.

bacteriophage A virus that infects bacteria
rickettsia Type of degenerate bacterium that is an obligate parasite and infects the cells of higher organisms

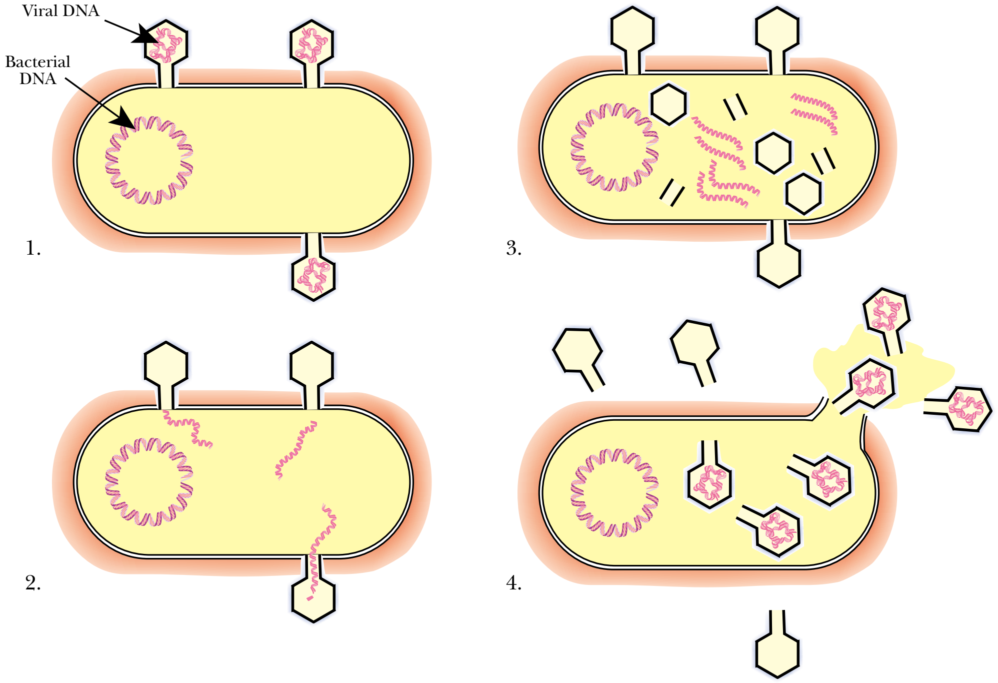


FIGURE 2.29 ***Virus Entry into a Cell***

Components of a new virus are synthesized under the direction of viral DNA but using the synthetic machinery of the host cell. First (1) a virus binds to the host cell and then (2) inserts its nucleic acid into the host cell. The synthetic machinery of the host cell then manufactures the viral proteins and nucleic acids according to the genetic information carried by the viral DNA (3). Finally, the virus causes the cell to burst, releasing the newly synthesized viruses (4) that seek a new host. The host cell dies as the result of the viral infection.

Human Viral Diseases Are Common

An immense variety of viruses exists (see Chapter 17 for more details). Viruses infect every other life-form, from bacteria to eukaryotes, including humans.

Many common childhood diseases such as measles, mumps and chickenpox are caused by viruses, as are the common cold and flu. More dangerous viral diseases include polio, smallpox, herpes, Lassa fever, Ebola and AIDS. Do viruses ever do anything useful? Yes; infection by a mild virus can provide resistance against a related but more dangerous virus (see Ch. 17). Viruses may carry genes from one host organism to another, in a process known as transduction (see Ch. 18), and have thereby played a major role in molecular evolution (see Ch. 20). The ability of viruses to carry genes between organisms may be put to good use by genetic engineers. All the same, about the best that can be said for the natural role of viruses is that most of them do relatively little damage and only a few cause highly virulent diseases.

Viral diseases usually cannot be cured once they have been caught. Either the victim's body fights off the infection or it does not, although some antiviral drugs can help the host in the fight. However, viral diseases can often be prevented by **immunization**, if a potential victim is **vaccinated** before catching the virus. In this case, the invading virus will be killed by the immune system, which has been put on alert by the vaccine, and the disease will be prevented.

Antibiotics are of no use against viruses; they only kill bacteria. So why do doctors often prescribe antibiotics for viral diseases like flu or colds? There are two main reasons. The valid reason is that giving antibiotics may help combat secondary or opportunistic infections caused by bacteria, especially in virally-infected patients who are in poor health. However, massive over-prescription of antibiotics occurs because many patients would be upset if faced with the truth. They would rather be given medicine, even if it is of no use, than face the fact that there is no cure. This abuse has in turn contributed to the spread of antibiotic resistance among many infectious bacteria (see Ch. 16)—thus creating a major health problem.

immunization Process of preparing the immune system for future infection by treating the patient with weak or killed versions of an infectious agent
vaccination Artificial induction of the immune response by injecting foreign proteins or other antigens

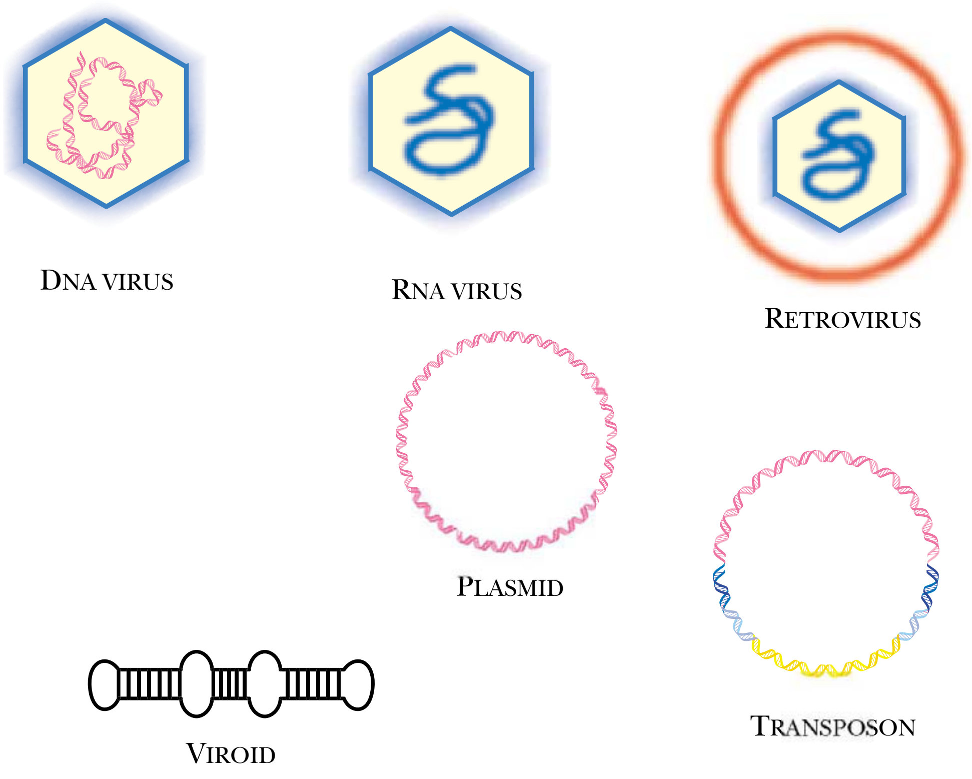


FIGURE 2.30 ***The Variety of Subcellular Genetic Elements—"Gene Creatures"***

These structures possess some of the characteristics of life. However, they use their host's machinery to replicate. The plasmid and the viroid lack a protein shell. The transposon is merely a segment of DNA (yellow) with special ends (blue) inserted into another DNA molecule.

A Variety of Subcellular Genetic Entities Exist

A whole range of entities exist that have genetic information, but do not themselves possess the machinery of life and cannot exist without a host cell to parasitize (Fig. 2.30). Viruses are the most complex of these subcellular genetic elements. In this book, these elements will sometimes be collectively referred to as "gene creatures" to emphasize that they possess genetic information, but have no cell structure or metabolism of their own. Gene creatures may be thought of as inhabiting cells, much as living cells live in their own, larger-scale environment. The term gene creatures is intended to focus attention on the properties of these genetic elements in contrast to the traditional viewpoint, which regards them merely as parasites or accessories to "real cells". These assorted genetic elements will be dealt with in subsequent chapters. Here they will just be introduced, to give some idea of the range of gene creatures that share the biosphere with the more traditional life forms (Fig. 2.31).

As discussed above, viruses carry their genes inside a protective shell of protein. **DNA viruses** have their genes in the form of DNA, and **RNA viruses** contain genes as RNA. **Retroviruses** have RNA copies of their genes inside the virus particle, but once inside the host cell, they make a DNA copy of their genome (see Ch. 17).

Viroids and **plasmids** are self-replicating molecules of nucleic acid that lack the protein coat characteristic of a virus. Viroids are naked molecules of RNA that infect plants and trick the infected plant cell into replicating more viroid RNA (see Ch. 17). Like a virus, they are released into the environment and must find a new cell to infect.

An amazing variety of quasi-independent genetic elements are widespread in the biosphere. They range from those causing major diseases of cellular organisms to those whose existence is scarcely noticeable without sophisticated molecular analysis.

DNA virus A virus whose genome consists of DNA
plasmid Self-replicating genetic elements that are sometimes found in both prokaryotic and eukaryotic cells. They are not chromosomes nor part of the host cell's permanent genome. Most plasmids are circular molecules of double stranded DNA although rare linear plasmids and RNA plasmids are known
retrovirus Type of virus which has its genes as RNA in the virus particle but converts this to a DNA copy inside the host cell by using reverse transcriptase
RNA virus A virus whose genome consists of RNA
viroid Naked single-stranded circular RNA that forms a stable highly base-paired rod-like structure and replicates inside infected plant cells. Viroids do not encode any proteins but possess self-cleaving ribozyme activity

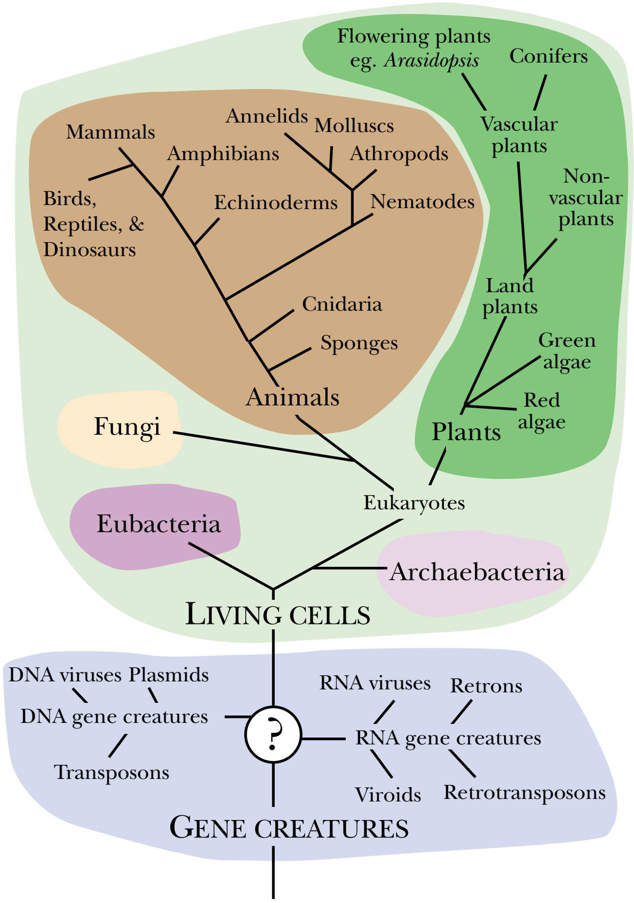


FIGURE 2.31 ***The Molecular Biologist's "Tree of Life"***

This tree of life includes both the traditional living creatures, such as plants and animals, as well as the two genetically distinct types of prokaryotic cell (eubacteria and archaebacteria). At the bottom are shown a variety of gene creatures, whose relationships are still mostly uncertain.

Unlike a virus, their extra-cellular phase lacks a protective protein shell. Plasmids are self-replicating molecules of DNA that live permanently inside host cells (see Ch. 16). Although some plasmids can be transferred from one host cell to another, they have no extra-cellular phase and so unlike viruses or viroids, they do not destroy their host cell. Plasmids are widely used to carry genes during many genetic engineering procedures.

Transposable elements, or **transposons**, are simpler still. They are nucleic acid molecules, usually DNA, that lack the ability to self-replicate. In order to get replicated, they must insert themselves into other molecules of DNA that are capable of replicating themselves. Thus transposable elements require a host DNA molecule, such as the chromosome of a cell, a virus genome or a plasmid. Transposable refers to the fact that these elements possess the ability of jumping from one host DNA molecule to another, a property that is essential for their survival and distribution (see Ch. 15).

Prions are infectious protein molecules, the ultimate parasites. They contain no nucleic acid and possess genetic information only in the sense of being gene products. Prions infect cells in the nervous systems of animals and cause diseases, the most famous of which is bovine spongiform encephalopathy, better known as mad cow disease. The prion protein is actually a misfolded version of a normal protein found in nerve cells, especially in the brain. When the prion infects a nerve cell, it promotes the misfolding of the corresponding normal proteins, which causes the cell to die. The prion protein is actually encoded by a gene belonging to the host animal that it infects.

prion Distorted, disease-causing form of a normal brain protein which can transmit an infection
transposable element or transposon Segment of DNA that can move as a unit from one location to another, but which always remains part of another DNA molecule

CHAPTER THREE

DNA, RNA and Protein

Nucleic Acid Molecules Carry Genetic Information

Chemical Structure of Nucleic Acids

DNA and RNA Each Have Four Bases

Nucleosides Are Bases Plus Sugars; Nucleotides Are Nucleosides Plus Phosphate

Double Stranded DNA Forms a Double Helix

Base Pairs are Held Together by Hydrogen Bonds

Complementary Strands Reveal the Secret of Heredity

Constituents of Chromosomes

The Central Dogma Outlines the Flow of Genetic Information

Ribosomes Read the Genetic Code

The Genetic Code Dictates the Amino Acid Sequence of Proteins

Various Classes of RNA Have Different Functions

Proteins, Made of Amino Acids, Carry Out Many Cell Functions

The Structure of Proteins Has Four Levels of Organization

Proteins Vary in Their Biological Roles

Nucleic Acid Molecules Carry Genetic Information

Chapter 1 discussed how the fundamentals of modern genetics were laid when Mendel found that hereditary information consists of discrete fundamental units now called genes. Each gene is responsible for a single inherited property or characteristic of the organism. Just as the discovery that atoms are made of subatomic particles ushered in the nuclear age, so the realization that genes are made up of **DNA** molecules opened the way both to a deeper understanding of life and to its artificial alteration by genetic engineering.

Genetic information is encoded by molecules named **nucleic acids** because they were originally isolated from the nucleus of eukaryotic cells. There are two related types of nucleic acid, **deoxyribonucleic acid** (**DNA**) and **ribonucleic acid** (**RNA**). The master copy of each cell's genome is stored on long molecules of DNA, which may each contain many thousands of genes. Each gene is thus a linear segment of a long DNA molecule. In contrast, RNA molecules are much shorter, are used to transmit the genetic information to the cell machinery, and carry only one or a few genes. [Certain viruses use RNA to encode their genomes as well as transmitting genetic information to the cell machinery. These RNA viruses have short genomes, rarely more than a dozen genes, as opposed to the hundreds or thousands of genes carried on the DNA genomes of cells.]

Genetic information is carried on long linear polymers, the nucleic acids. Two classes of nucleic acid, DNA and RNA, divide up the responsibility of storing and deploying the genetic information.

Chemical Structure of Nucleic Acids

DNA and RNA are linear polymers made of subunits known as **nucleotides**. The information in each gene is determined by the order of the different nucleotides, just as the information in this sentence is due to the order of the 26 possible letters of the alphabet. There are four different nucleotides in each type of nucleic acid and their order determines the genetic information (Fig. 3.01).

Each nucleotide has three components: a **phosphate group**, a five-carbon sugar, and a nitrogen-containing **base** (Fig. 3.02). The phosphate groups and the sugars form the backbone of each strand of DNA or RNA. The bases are joined to the sugars and stick out sideways.

In DNA, the sugar is always **deoxyribose**; whereas, in RNA, it is **ribose**. Both sugars are **pentoses**, or five-carbon sugars. Deoxyribose has one less oxygen than ribose (Fig. 3.03). It is this chemical difference that gave rise to the names deoxyribonucleic acid and ribonucleic acid. Both sugars have five-membered rings consisting of four carbon atoms and an oxygen. The fifth carbon forms a side chain to the ring. The five carbon atoms of the sugar are numbered 1′, 2′, 3′, 4′ and 5′ as shown in Fig. 3.02. By convention, in nucleic acids, numbers with prime marks refer to the sugars and numbers without prime marks refer to the positions around the rings of the bases.

Nucleotides are joined by linking the phosphate on the 5′-carbon of the (deoxy) ribose of one nucleotide to the 3′-position of the next as shown in Fig. 3.04. The phosphate group is joined to the sugar on either side by ester linkages, and the overall structure is therefore a **phosphodiester** linkage. The phosphate group linking the sugars has a negative charge.

base Alkaline chemical substance, in molecular biology especially refers to the cyclic nitrogen compounds found in DNA and RNA
deoxyribonucleic acid (DNA) Nucleic acid polymer of which the genes are made
deoxyribose The sugar with five carbon atoms that is found in DNA
DNA Deoxyribonucleic acid, nucleic acid polymer of which the genes are made
nucleic acid Class of polymer molecule consisting of nucleotides that carries genetic information
nucleotide Monomer or subunit of a nucleic acid, consisting of a pentose sugar plus a base plus a phosphate group
pentose A five carbon sugar, such as ribose or deoxyribose
phosphate group Group of four oxygen atoms surrounding a central phosphorus atom found in the backbone of DNA and RNA
phosphodiester The linkage between nucleotides in a nucleic acid that consists of a central phosphate group esterified to sugar hydroxyl groups on either side
ribonucleic acid (RNA) Nucleic acid that differs from DNA in having ribose in place of deoxyribose
ribose The 5-carbon sugar found in RNA

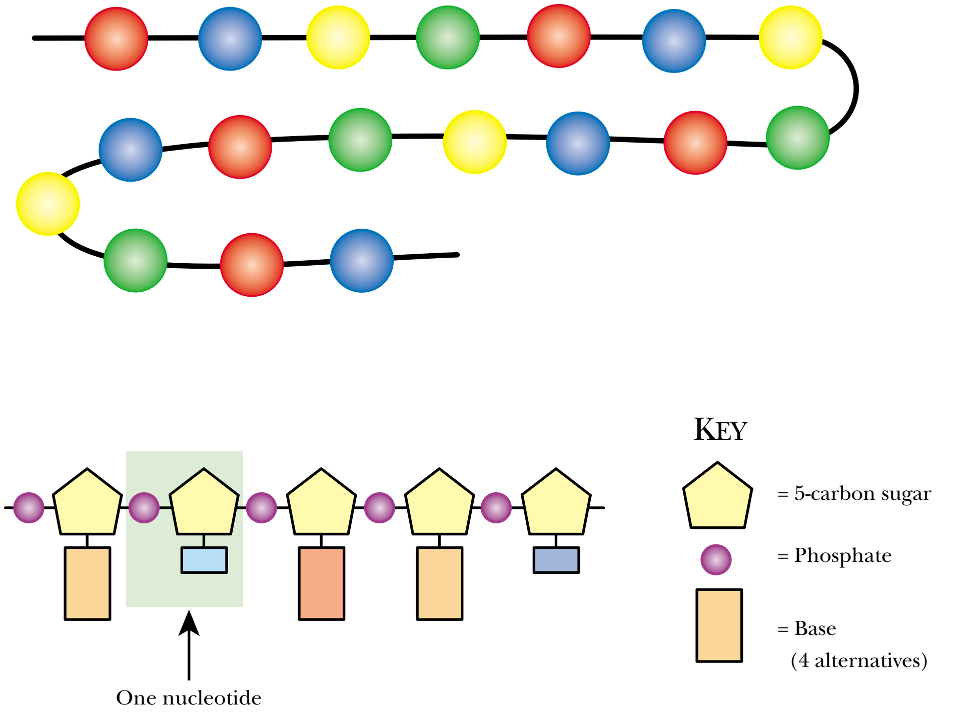


FIGURE 3.01 ***The Order of the Nucleotides Encodes the Genetic Information***

Nucleotides are ordered along a string of DNA or RNA. It is the ordering of the different nucleotides that dictates the nature of the information within the nucleic acid.

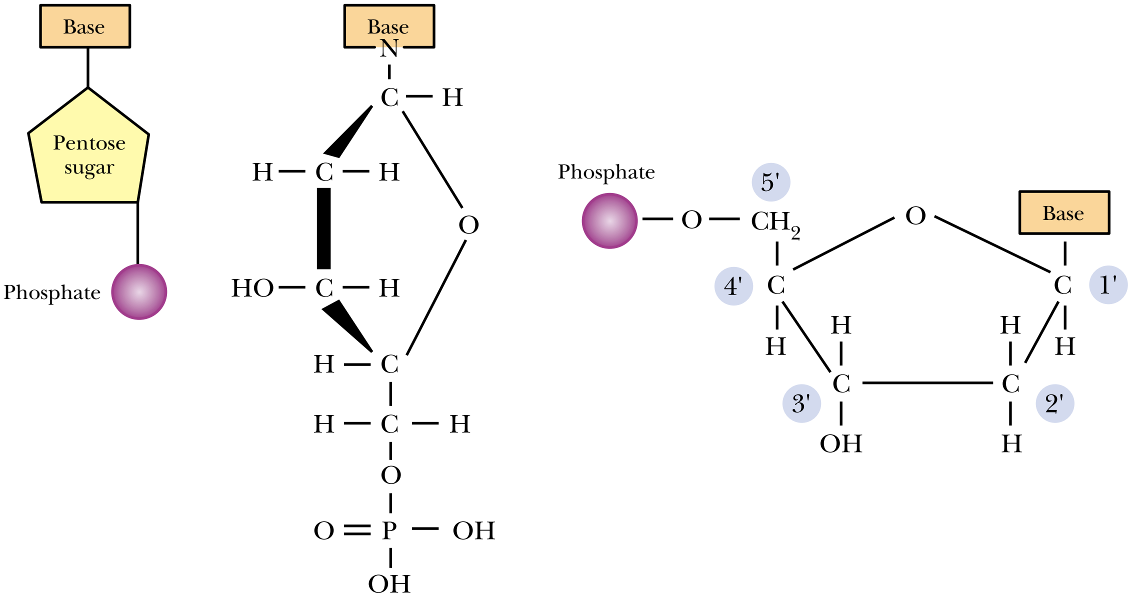


FIGURE 3.02 ***Three Views of a Nucleotide***

The three components of a nucleotide are shown to the left. The structures on the right show the pentose sugar (deoxyribose) connected to the phosphate and the base.

FIGURE 3.03 ***The Sugars Composing RNA and DNA***

Ribose is the five-carbon sugar (pentose) found in RNA. Deoxyribose is the pentose of DNA. It has one less oxygen than ribose as it has a hydrogen in place of the hydroxyl group on position 2′ of the ribose ring.

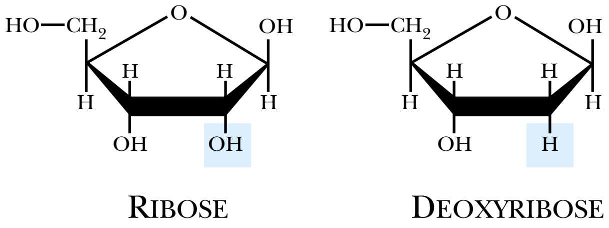
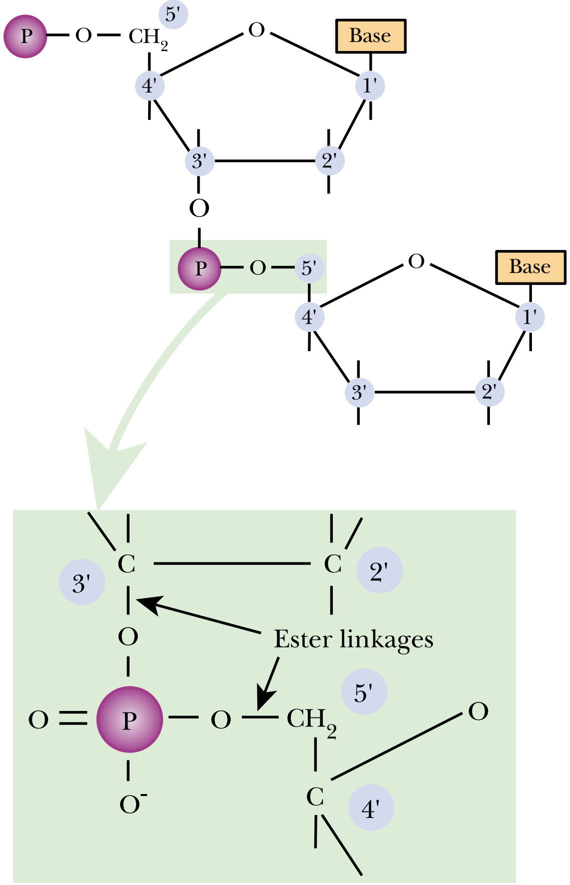


FIGURE 3.04 ***Nucleotides Are Joined by Phosphodiester Linkages***

The nucleotides that form the backbone of DNA and RNA are joined together by linkages involving their phosphate groups. One nucleotide is linked via its 5′-carbon to the oxygen of the phosphate group and another nucleotide is linked via its 3′-carbon to the other side of the central phosphate. These linkages are termed phosphodiester groups.



DNA and RNA Each Have Four Bases

There are five different types of nitrogenous bases associated with nucleotides. DNA contains the bases **adenine**, **guanine**, **cytosine** and **thymine**. These are often abbreviated to A, G, C and T, respectively. RNA contains A, G and C, but T is replaced by **uracil** (U). From the viewpoint of genetic information, T in DNA and U in RNA are equivalent.

The bases found in nucleic acids are of two types, **pyrimidines** and **purines**. The smaller pyrimidine bases contain a single ring whereas the purines have a fused double ring. Adenine and guanine are purines; and thymine, uracil and cytosine are pyrimidines. The purine and pyrimidine ring systems and their derivatives are shown in Figure 3.05.

adenine (A) A purine base that pairs with thymine, found in DNA or RNA
cytosine (C) One of the pyrimidine bases found in DNA or RNA and which pairs with guanine
guanine (G) A purine base found in DNA or RNA that pairs with cytosine
purine Type of nitrogenous base with a double ring found in DNA and RNA
pyrimidine Type of nitrogenous base with a single ring found in DNA and RNA
thymine (T) A pyrimidine base found in DNA that pairs with adenine
uracil (U) A pyrimidine base found in RNA that may pair with adenine

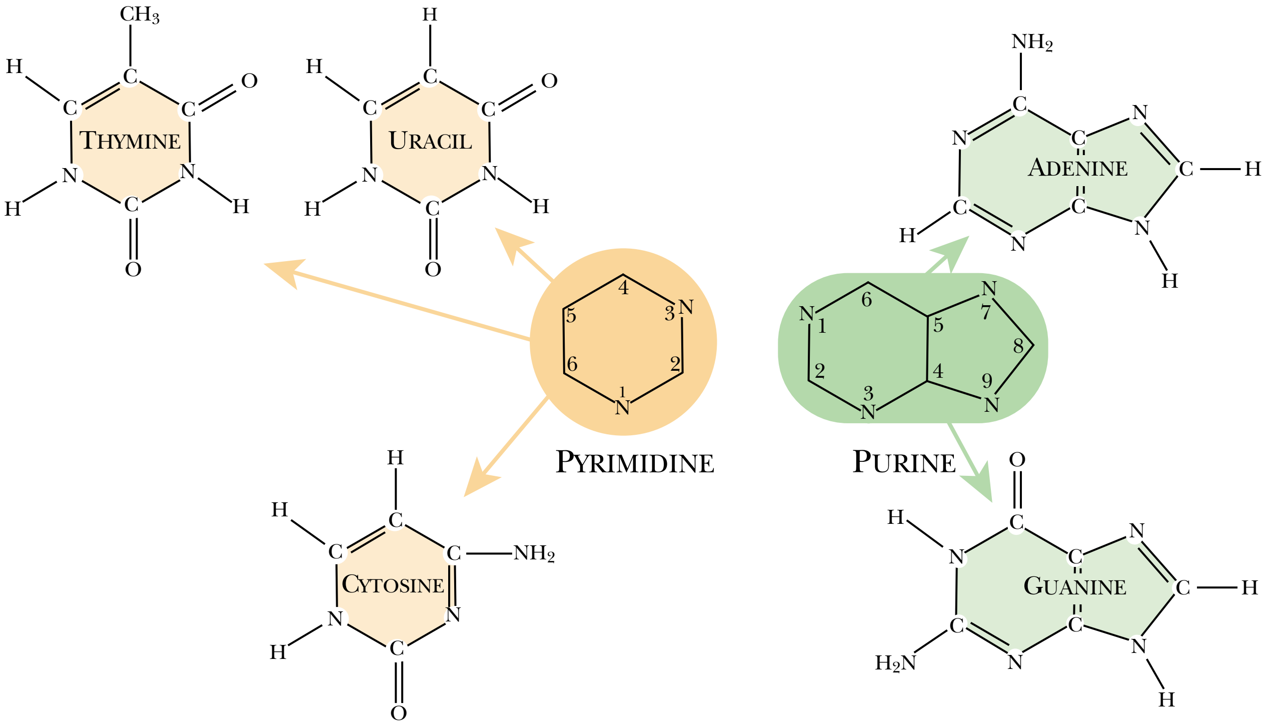


FIGURE 3.05 ***The Bases of the Nucleic Acids***

The four bases of DNA are adenine, guanine, cytosine and thymine. In RNA, uracil replaces thymine. Pyrimidine bases contain one-ring structures, whereas purine bases contain two-ring structures.

TABLE 3.01 Naming Bases, Nucleosides and Nucleotides

Base	Abbreviations		Nucleoside	Nucleotide
Adenine	ade	A	adenosine	adenosine monophosphate (AMP)
Guanine	gua	G	guanosine	guanosine monophosphate (GMP)
Cytosine	cyt	C	cytidine	cytidine monophosphate (CMP)
Thymine	thy	T	thymidine	thymidine monophosphate (TMP)
Uracil	ura	U	uridine	uridine monophosphate (UMP)

Nucleosides Are Bases Plus Sugars; Nucleotides Are Nucleosides Plus Phosphate

A base plus a sugar is known as a **nucleoside**. A base plus a sugar plus phosphate is known as a **nucleotide**. If necessary, one may distinguish between **deoxynucleosides** or **deoxynucleotides** where the sugar is deoxyribose, and **ribonucleosides** or **ribonucleotides** that contain ribose. The names of the nucleosides are similar to the names of the corresponding bases (see Table 3.01). The nucleotides do not have names of their own but are referred to as phosphate derivatives of the corresponding nucleoside. For example, the nucleotide of adenine is **adenosine** monophosphate, or AMP.

Three-letter abbreviations for the bases such as ade, gua, etc., are sometimes used when writing biochemical pathways or for the names of genes involved in nucleotide metabolism. When writing the sequence of a nucleic acid, the single letter abbrevia-

adenosine The nucleoside consisting of adenine plus (deoxy)ribose
deoxynucleoside A nucleoside containing deoxyribose as the sugar
deoxynucleotide A nucleotide containing deoxyribose as the sugar
nucleoside The union of a purine or pyrimidine base with a pentose sugar
nucleotide Monomer or subunit of a nucleic acid, consisting of a pentose sugar plus a base plus a phosphate group
ribonucleoside A nucleoside whose sugar is ribose (not deoxyribose)
ribonucleotide A nucleotide whose sugar is ribose (not deoxyribose)

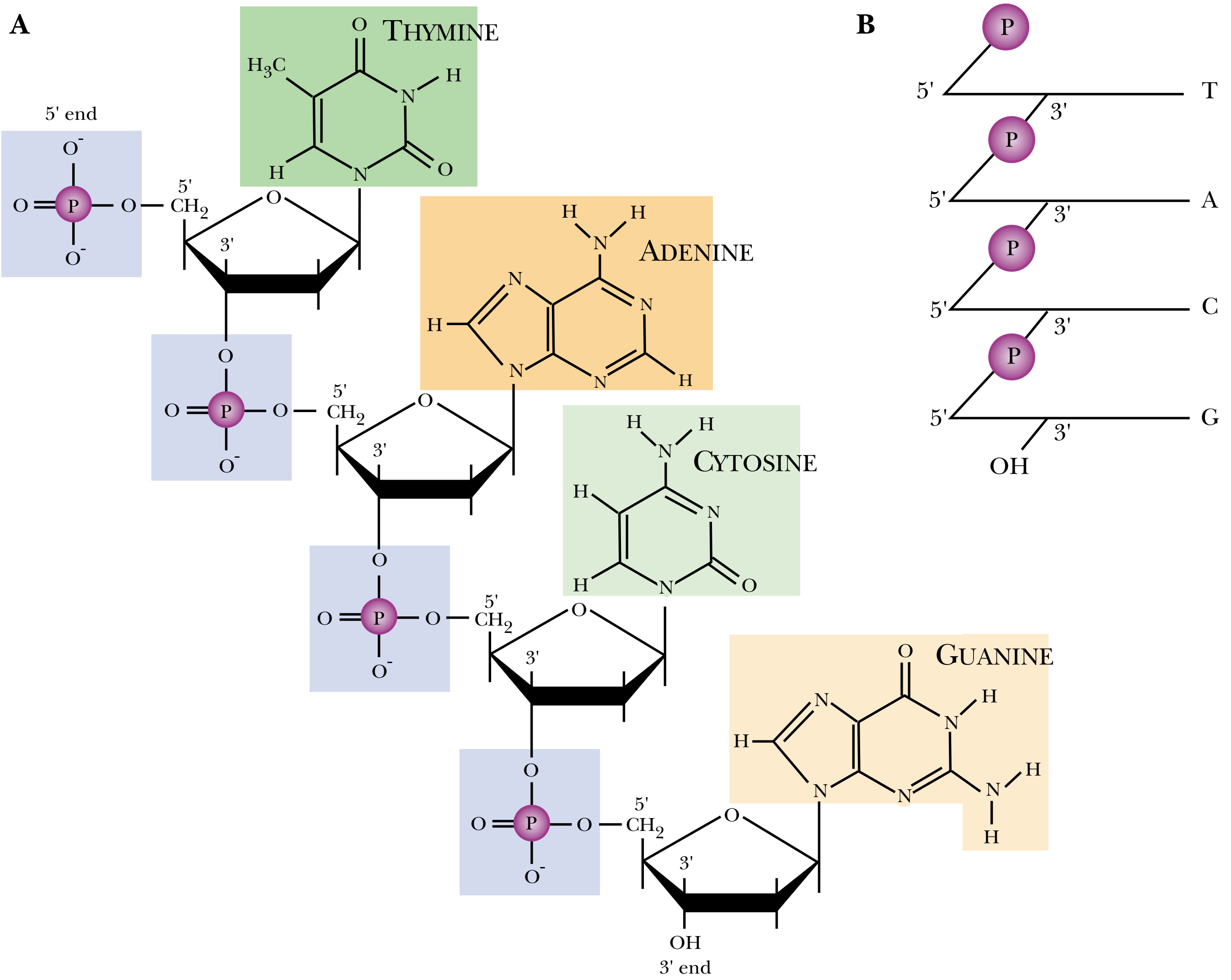


FIGURE 3.06 ***Some Variations in the Ways Nucleic Acids are Represented***

(A) More elaborate drawings show the chemical structures of the nucleic acid components, including the pentose sugar, phosphate groups and bases. (B) Simple line drawings may be used to summarize the linkage of sugars by the 5′ and 3′ phosphodiester bonds. Here, the protruding bases have been abbreviated to a single letter.

tions are used (A, T, G and C for DNA or A, U, G and C for RNA). The letter N is often used to refer to an unspecified base.

Double Stranded DNA Forms a Double Helix

A strand of nucleic acid may be represented in various ways, either in full or abbreviated to illustrate the linkages (Fig. 3.06). As remarked above, nucleotides are linked by joining the 5′- phosphate of one to the 3′-hydroxyl group of the next. Typically, there is a free phosphate group at the 5′-end of the chain and a free hydroxyl group at the 3′-end of a nucleic acid strand. Consequently, a strand of nucleic acid has polarity and it matters in which direction the bases are read off. The 5′-end is regarded as the beginning of a DNA or RNA strand. This is because genetic information is read starting at the 5′-end. [In addition, when genes are replicated, nucleic acids are synthesized starting at the 5′-end as described in Ch. 5.]

The structure of the DNA double helix is critical to replication of the genes, as described in more detail in Chapter 5.

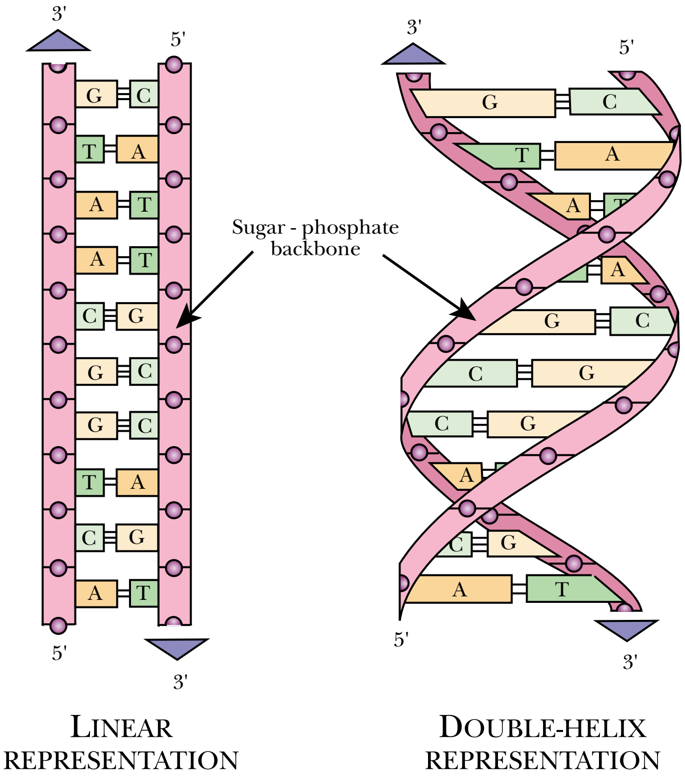


FIGURE 3.07
Representations of Double Stranded DNA

On the left DNA is represented as a double line consisting of two complementary strands. Actually DNA forms a double helix, as shown to the right.

RNA is normally found as a single-stranded molecule, whereas DNA is double-stranded. Note that the two strands of a DNA molecule are **antiparallel**, as they point in opposite directions. This means that the 5′-end of one strand is opposite the 3′-end of the other strand (Fig. 3.07). Not only is DNA double-stranded, but the two separate strands are wound around each other in a helical arrangement. This is the famous **double helix** first proposed by Francis Crick and James Watson in 1953 (Fig. 3.08). The DNA double helix is stabilized both by hydrogen bonds between the bases (see below) and by stacking of the aromatic rings of the bases in the center of the helix.

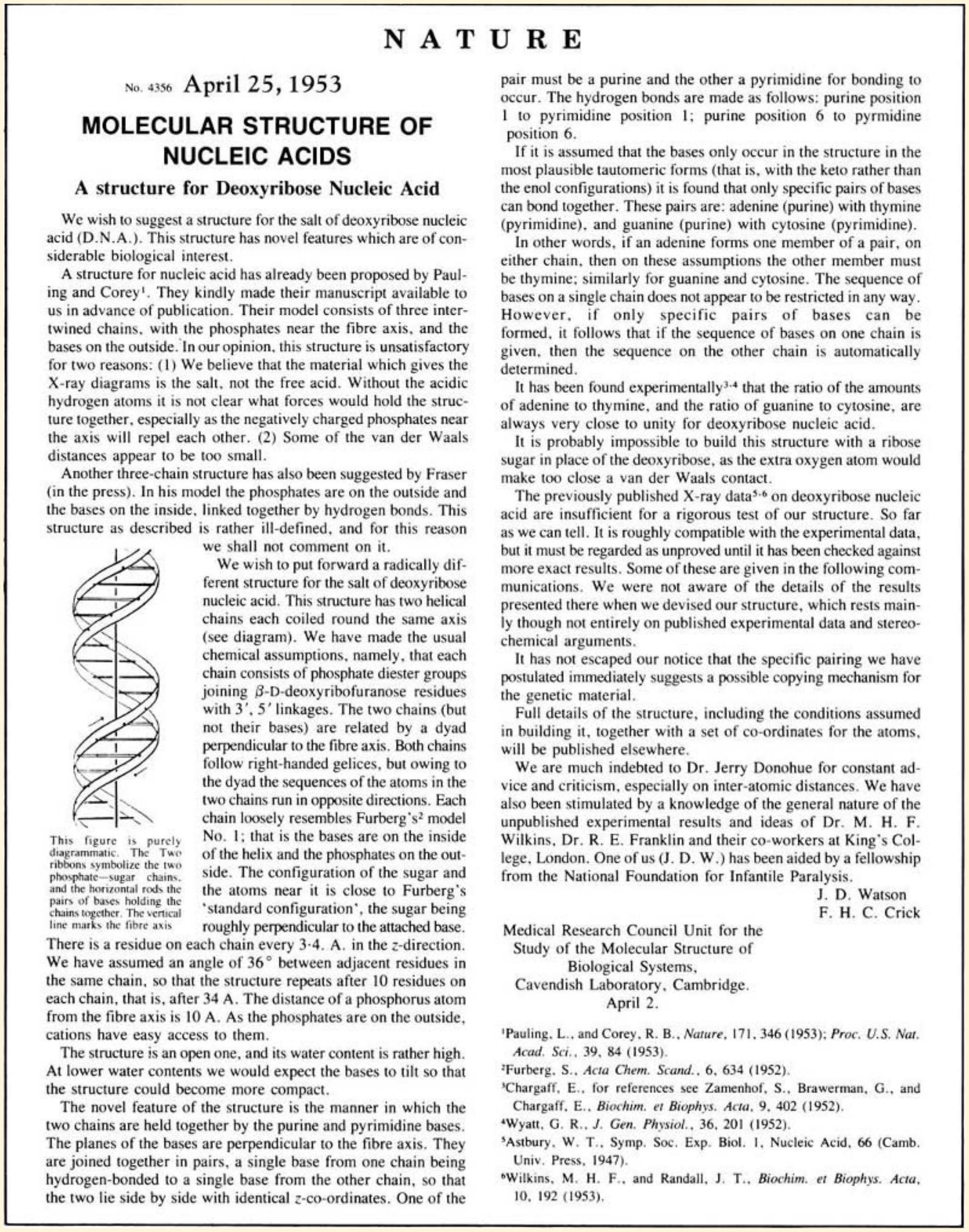
DNA forms a **right-handed** double helix. To tell a right-handed helix from a left-handed helix, the observer must look down the helix axis (in either direction). In a right-handed helix, each strand turns clockwise as it moves away from the observer (in a left-handed helix it would turn counterclockwise).

Base Pairs are Held Together by Hydrogen Bonds

In double stranded DNA, the bases on each strand protrude into the center of the double helix where they are paired with the bases in the other strand by means of **hydrogen bonds**. Adenine (A) in one strand is always paired with thymine (T) in the other, and guanine (G) is always paired with cytosine (C) (Fig. 3.10). Consequently, the number of adenines in DNA is equal to the number of thymines, and similarly the numbers of guanine and cytosine are equal. Note that the nucleic acid bases have amino or oxygen side-groups attached to the ring. It is these chemical groups, along with the nitrogen atoms that are part of the rings themselves, that allow the formation of hydrogen bonds. The hydrogen bonding in DNA **base pairs** involves either oxygen

antiparallel Parallel, but running in opposite directions
base pair Two bases held together by hydrogen bonds
double helix Structure formed by twisting two strands of DNA spirally around each other
hydrogen bond Bond resulting from the attraction of a positive hydrogen atom to both of two other atoms with negative charges
right-handed helix In a right-handed helix, as the observer looks down the helix axis (in either direction), each strand turns clockwise as it moves away from the observer

Working in Cambridge, England, James Watson and Francis Crick based their model of the double helix partly on the interpretation of data from X-ray crystallography by Rosalind Franklin and Maurice Wilkins, which suggested a helical molecule. Chemical analysis by Erwin Chargaff showed that DNA contained equimolar amounts of A and T and also of G and C. This, and chemical modeling, led Watson and Crick to propose that DNA was double stranded and that A in one strand is always paired with T in the other. Similarly, G is always paired with C. Watson and Crick published their landmark paper in *Nature* (Fig. 3.08) in 1953.



NATURE

No. 4356 April 25, 1953

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′, 5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed gelices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4. A. in the *z*-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

This figure is purely diagrammatic. The Two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical *z*-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrmidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell. It is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on inter-atomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at King's College, London. One of us (J. D. W.) has been aided by a fellowship from the National Foundation for Infantile Paralysis.

J. D. Watson
F. H. C. Crick

Medical Research Council Unit for the
Study of the Molecular Structure of
Biological Systems,
Cavendish Laboratory, Cambridge.
April 2.

[1] Pauling, L., and Corey, R. B., *Nature*, 171, 346 (1953); *Proc. U.S. Nat. Acad. Sci.*, 39, 84 (1953).

[2] Furberg, S., *Acta Chem. Scand.*, 6, 634 (1952).

[3] Chargaff, E., for references see Zamenhof, S., Brawerman, G., and Chargaff, E., *Biochim. et Biophys. Acta*, 9, 402 (1952).

[4] Wyatt, G. R., *J. Gen. Physiol.*, 36, 201 (1952).

[5] Astbury, W. T., Symp. Soc. Exp. Biol. 1, Nucleic Acid, 66 (Camb. Univ. Press, 1947).

[6] Wilkins, M. H. F., and Randall, J. T., *Biochim. et Biophys. Acta*, 10, 192 (1953).

FIGURE 3.08 ***DNA Is a Double Helix***

This one-page paper published in *Nature* described the now-famous double helix. *J. D. Watson & F. H. C. Crick*, Molecular Structure of Nucleic Acids, A Structure for Deoxyribose Nucleic Acid, Nature 171 (1953) 737.

In 2003 the Double Helix celebrated its 50th anniversary. In Great Britain, the Royal Mail issued a set of five commemorative stamps illustrating the double helix together with some of the technological advances that followed, such as comparative genomics and genetic engineering. In addition, the Royal Mint issued a £2 coin depicting the DNA double helix itself (Fig. 3.09).

FIGURE 3.09 ***Double Helix—50th Anniversary Coin***

A £2 coin commemorating the discovery of the double helix was issued in 2003 by Great Britain.

or nitrogen as the atoms that carry the hydrogen, giving three alternative arrangements: O–H–O, N–H–N and O–H–N.

Each base pair consists of one larger purine base paired with a smaller pyrimidine base. So, although the bases themselves differ in size, all of the allowed base pairs are the same width, providing for a uniform width of the helix. The A-T base pair has two hydrogen bonds and the G-C base pair is held together by three, as shown in Figure 3.10. Before the hydrogen bonds form and the bases pair off, the shared hydrogen atom is found attached to one or the other of the two bases (shown by the complete lines in Fig. 3.10). During **base pairing**, this hydrogen also bonds to an atom of the second base (shown by the dashed lines).

Although RNA is normally single-stranded, many RNA molecules fold up, giving double-stranded regions. In addition, a strand of RNA may be found paired with one of DNA under some circumstances. Furthermore, the genome of certain viruses consists of double-stranded RNA (see Ch. 17). In all of these cases, the uracil in RNA will base pair with adenine. Thus the base-pairing properties of the uracil found in RNA are identical to those of the thymine of DNA.

Complementary Strands Reveal the Secret of Heredity

Due to the rules for base-pairing, the sequence of a DNA strand can be deduced if the sequence of its partner is known.

If one of the bases in a base pair of double stranded DNA is known, then the other can be deduced. If one strand has an A, then the other will have a T, and vice versa. Similarly, G is always paired with C. This is termed complementary base pairing. The significance is that if the base sequence of either one of the strands of a DNA molecule is known, the sequence of the other strand can be deduced. Such mutually deducible sequences are known as **complementary sequences**. It is this complementary nature of a DNA double helix that allows genetic information to be inherited. Upon cell division, each daughter cell must receive a copy of the parental genome.

base pairing A pair of two complementary bases (A with T or G with C) held together by hydrogen bonds

complementary sequences Two nucleic acid sequences whose bases pair with each other because A, T, G, C in one sequence correspond to T, A, C, G, respectively, in the other

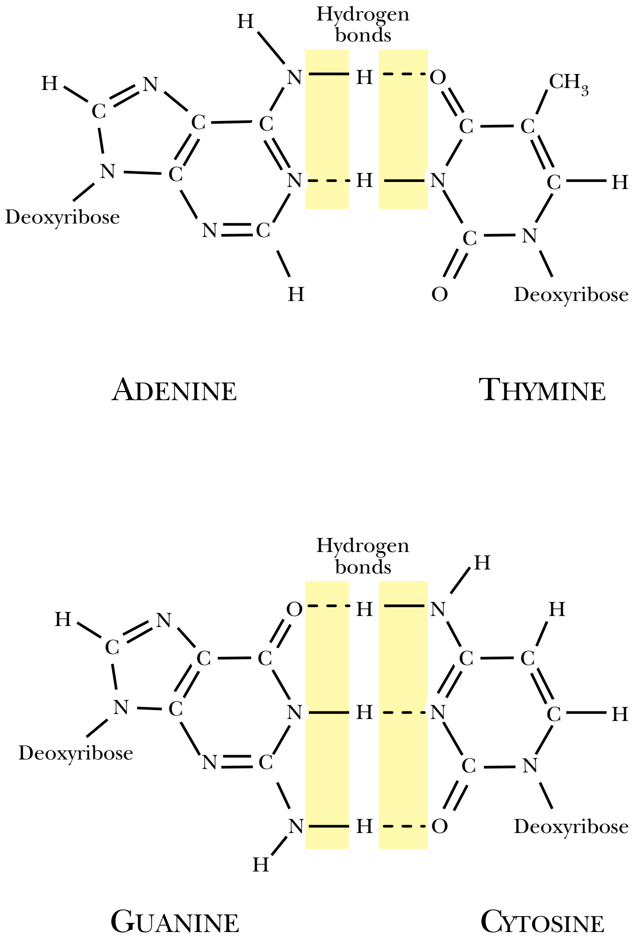


FIGURE 3.10 ***Base Pairing by Hydrogen Bond Formation***

Purines (adenine and guanine) pair with pyrimidines (thymine and cytosine) by hydrogen bonding (colored regions).When the purines and pyrimidines first come together, they form the bonds indicated by the dotted lines.

This requires accurate duplication or replication of the DNA (Fig. 3.11). This is achieved by separating the two strands of DNA and using complementary base pairing to make a new partner for each original strand (see Ch. 5 for details).

Constituents of Chromosomes

Genes are segments of large DNA molecules known as **chromosomes** (Fig. 3.12). Each chromosome is thus an exceedingly long single molecule of DNA. In addition to the DNA, which comprises the genes themselves, the chromosome has some accessory protein molecules, which help maintain its structure. The term **chromatin** refers to this mixture of DNA and protein, especially as observed with the microscope in the nuclei of eukaryotic cells. The genes are arranged in linear order. In front of each gene is a **regulatory region** of DNA involved in switching the gene on or off. Between genes are spacer regions of DNA often referred to as **intergenic regions**. In prokaryotes, groups of genes may be clustered close together with no intergenic regions. Such clusters are

Genetic information includes both the genes themselves and regions of DNA involved in controlling gene expression.

chromatin Complex of DNA plus protein which constitutes eukaryotic chromosomes
chromosome Structure containing the genes of a cell and made of a single molecule of DNA
intergenic region DNA sequence between genes
regulatory region DNA sequence in front of a gene, used for regulation rather than to encode a protein

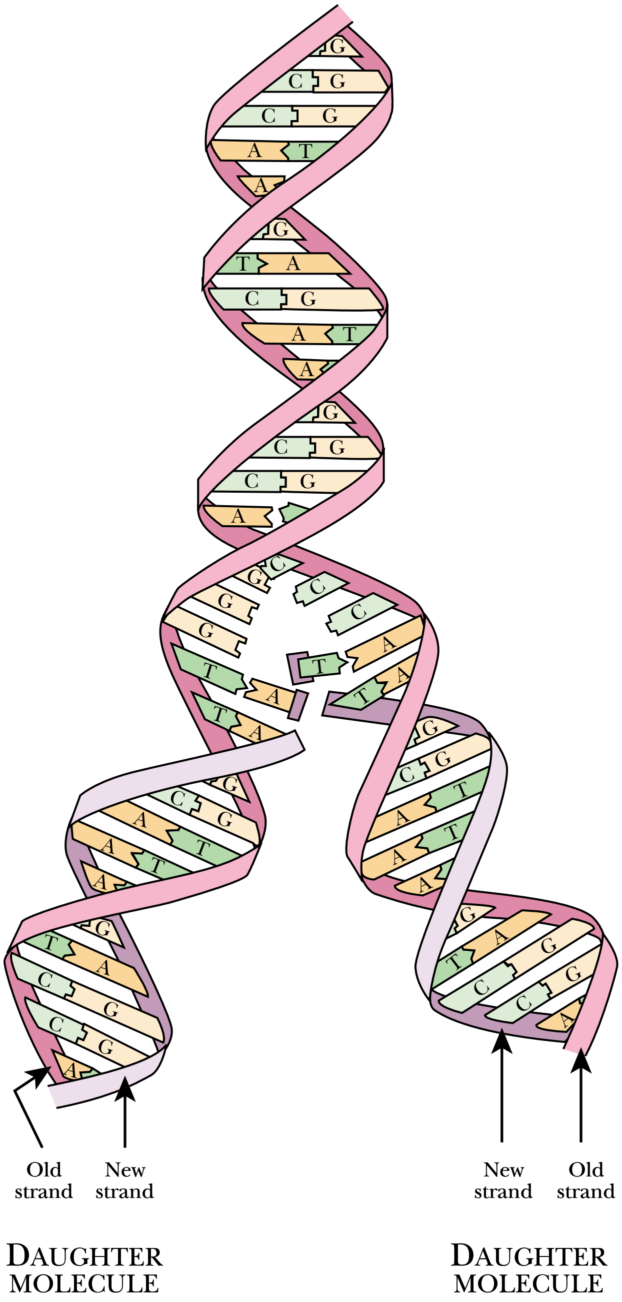


FIGURE 3.11 ***Complementary Strands Allow Duplication***

Because DNA strands are complementary, double-stranded DNA can be split into single strands each carrying sufficient information to recreate the original molecule. Complementary base pairing allows the synthesis of two new strands so restoring double-stranded DNA.

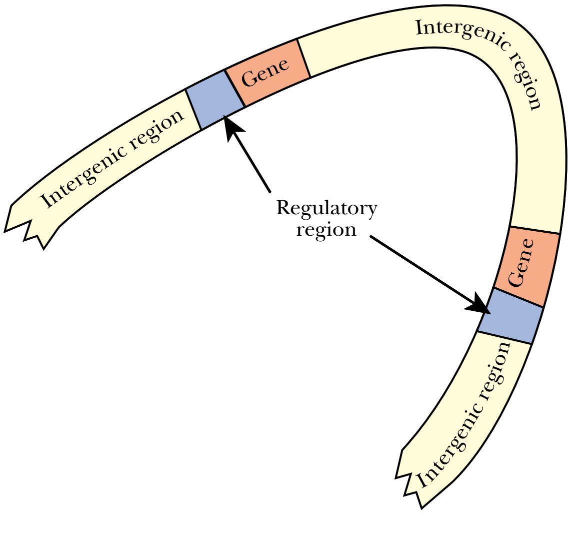


FIGURE 3.12 ***The General Pattern of Information on a Chromosome***

Genes are normally preceded by regions of DNA involved in regulation. Between the genes are regions of DNA that apparently do not carry useful genetic information. These are called intergenic regions and vary greatly in size.

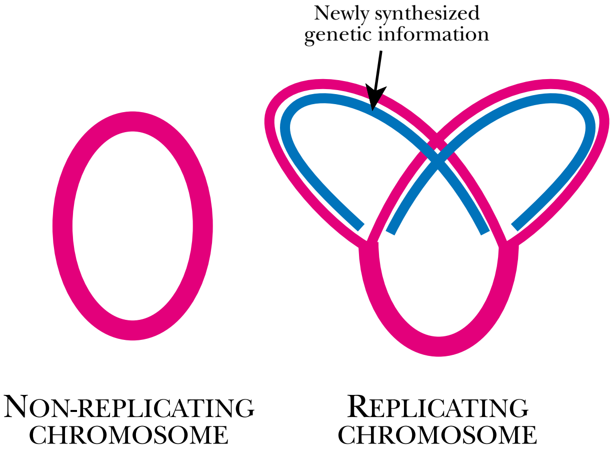

Newly synthesized genetic information
NON-REPLICATING CHROMOSOME
REPLICATING CHROMOSOME

FIGURE 3.13 ***The Circular Bacterial Chromosome and Its Replication***

The bacterial chromosome is circular and not linear. When the double stranded DNA is duplicated, the chromosome is opened forming loops that allow replication of each DNA strand.

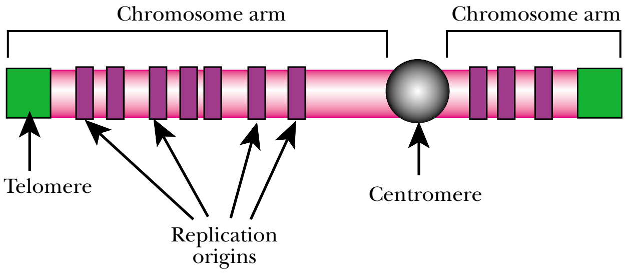


FIGURE 3.14 ***Structural Components of a Eukaryotic Chromosome***

The eukaryotic chromosome is a linear molecule with specific DNA sequences called telomeres at each end. More or less in the center is an organized region called the centromere that is involved in chromosome division. Along the chromosome are multiple regions where replication is initiated.

Details of replication mechanism and structure vary between the linear chromosomes of eukaryotes and the circular chromosomes of most bacteria.

called **operons** and each is under the control of a single regulatory region. Operons are transcribed to give single mRNA molecules, each consisting of several genes.

Chromosomes from bacteria are circular molecules of double-stranded DNA. Since bacteria generally have only around 3,000–4,000 genes, and the intergenic regions are very short, one chromosome is sufficient to accommodate all of their genes. When bacteria divide, the chromosome opens up at the origin of replication and replication proceeds around the circle in both directions (Fig. 3.13).

Chromosomes from higher organisms such as animals and plants are linear molecules of double stranded DNA. They have a **centromere**, usually located more or less in the middle, and structures known as **telomeres** at the two ends (see Fig. 3.14). Both centromeres and telomeres contain special repetitive DNA sequences allowing their recognition by particular proteins. [One exception to this rule is that the yeast, *Saccharomyces*, lacks repetitive sequences at its centromere. However, this is not a general property of fungi, as other fungi do have repetitive centromere sequences.] The centromere is used at cell division when the chromosomes replicate. The newly divided daughter chromosomes are pulled apart by spindle fibers (or microtubules) attached to the centromeres via protein structures known as **kinetochores**. Due to the mechanism of initiating DNA synthesis (see Ch. 5), the far ends of linear DNA molecules are shortened by a few bases each round of replication. In those cells that are permitted to continue growing and dividing, the end sequences are repaired by the enzyme **telomerase**. Telomeres are critically important in cell differentiation, cancer and aging.

Higher organisms have much more DNA than bacteria. This is partly because they possess more genes–higher eukaryotes may have up to 50,000 genes. However, the major reason is that eukaryotes have much longer intergenic regions and other **non-coding DNA**. In fact, as shall be discussed later (Ch. 4), in higher eukaryotes, the genes are only a small proportion of the total DNA. Consequently, higher organisms need

centromere Region of eukaryotic chromosome, usually more or less central, where the microtubules attach during mitosis and meiosis
kinetochore Protein structure that attaches to the DNA of the centromere during cell division and also binds the microtubules
non-coding DNA DNA sequences that do not code for proteins or functional RNA molecules
operon A cluster of prokaryotic genes that are transcribed together to give a single mRNA (i.e. polycistronic mRNA)
telomerase Enzyme that adds DNA to the end, or telomere, of a chromosome
telomere Specific sequence of DNA found at the end of linear eukaryotic chromosomes

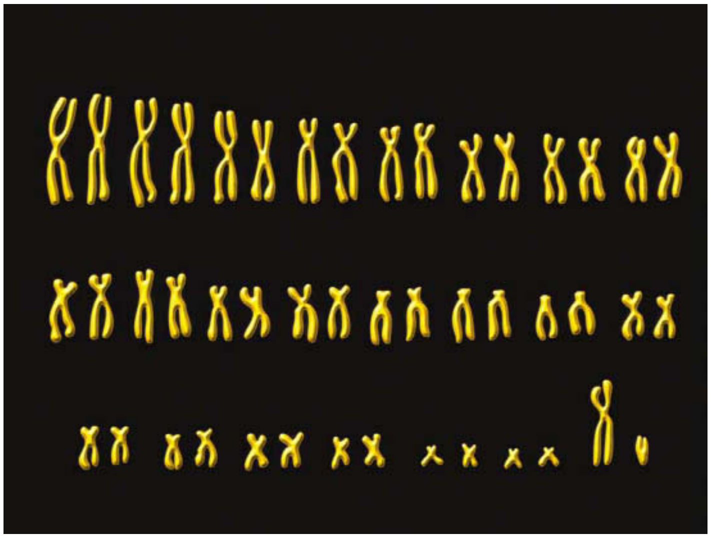


FIGURE 3.15 ***A Set of Human Chromosomes***

A human karyotype is a complete set of chromosomes containing 22 pairs plus one "X" and one "Y" chromosome (lower right) if the individual is male (as shown here). Females possess two "X" chromosomes. Courtesy of Alfred Pasieka, Science Photo Library.

Humans have vast amounts of DNA making up 46 linear chromosomes. However, most of this is non-coding DNA as discussed further in Chapters 4 and 24.

several chromosomes to accommodate all their DNA. Since eukaryotes are usually diploid, their chromosomes come in pairs. For example, humans have two duplicate sets of 23 different chromosomes, making a total of 46 chromosomes. The chromosomes are only visible under the light microscope during cell division and it is then that a complete set of chromosomes can be visualized (Fig. 3.15). The complete set of chromosomes found in the cells of a particular individual is known as the **karyotype**.

Chromosomes and specific regions of chromosomes may be identified by their staining patterns after using specific stains that emphasize regions lacking genes. This **chromosome banding technique** has been used to identify major chromosome abnormalities (Fig. 3.16).

As shown in Chapter 2, chromosomes of higher organisms are found in a separate membranous compartment of the cell, the nucleus. The nucleus is divided off from the rest of the cell by the nuclear envelope, consisting of two concentric membranes. The genes control the rest of the cell by dispatching genetic information in the form of special messenger molecules, the **messenger RNA**, through pores in the nuclear envelope.

The Central Dogma Outlines the Flow of Genetic Information

Under normal circumstances, genetic information flows from DNA to RNA to protein. As a result, proteins are often referred to as "gene products". Some RNA molecules are also "gene products" as they act without being translated into protein.

Genetic information flows from DNA to RNA to protein during cell growth. In addition, all living cells must replicate their DNA when they divide. The **central dogma** of molecular biology is a scheme showing the flow of genetic information during both the growth and division of a living cell (Fig. 3.17). During cell division each daughter cell receives a copy of the genome of the parent cell. As the genome is present in the form of DNA, cell division involves the duplication of this DNA. **Replication** is the process by which two identical copies of DNA are made from an original molecule of DNA. Replication occurs prior to cell division. An important point is that information does not flow from protein to RNA or DNA. However, information flow from RNA "backwards" to DNA is possible in certain special circumstances due to the operation of reverse transcriptase. In addition, replication of RNA occurs in viruses with an RNA genome (neither complication is shown in Fig. 3.17).

The genetic information stored as DNA is not used directly to make protein. During cell growth and metabolism, temporary, working copies of the genes known as

central dogma Basic plan of genetic information flow in living cells which relates genes (DNA), message (RNA) and proteins
chromosome banding technique Visualization of chromosome bands by using specific stains that emphasize regions lacking genes
karyotype The complete set of chromosomes found in the cells of a particular individual
messenger RNA (mRNA) The molecule that carries genetic information from the genes to the rest of the cell
replication Duplication of DNA prior to cell division

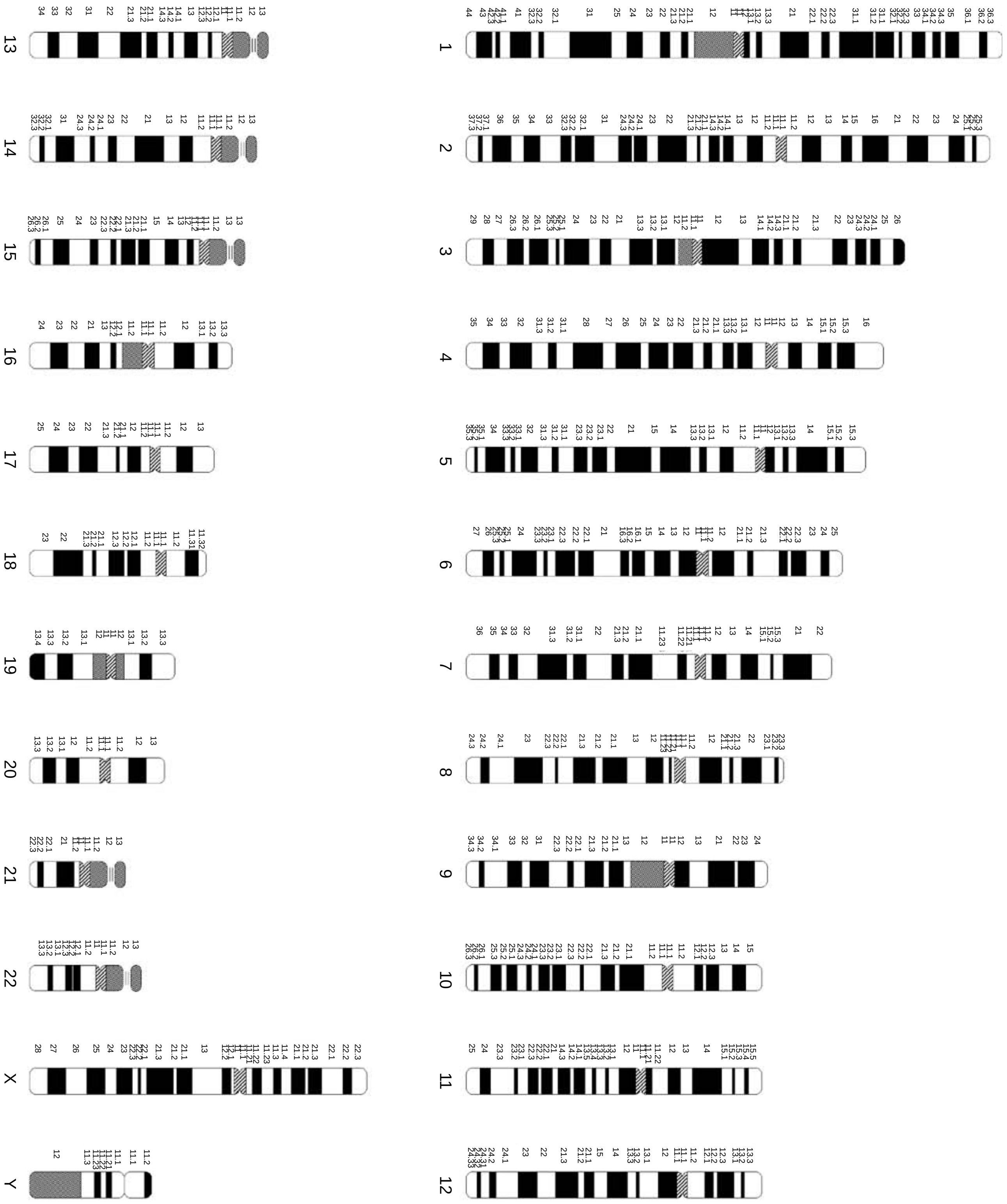


FIGURE 3.16 ***Banding Patterns of Human Chromosomes***

Representation of the banding patterns seen in metaphase chromosomes during meiosis. The bands are originally visualized by dyes. The relative distances between these bands are the same for an individual chromosome, so this is a useful way of identifying a particular chromosome. Courtesy of Dept. of Clinical Cytogenetics, Addenbrookes Hospital, Cambridge, UK, Science Photo Library.

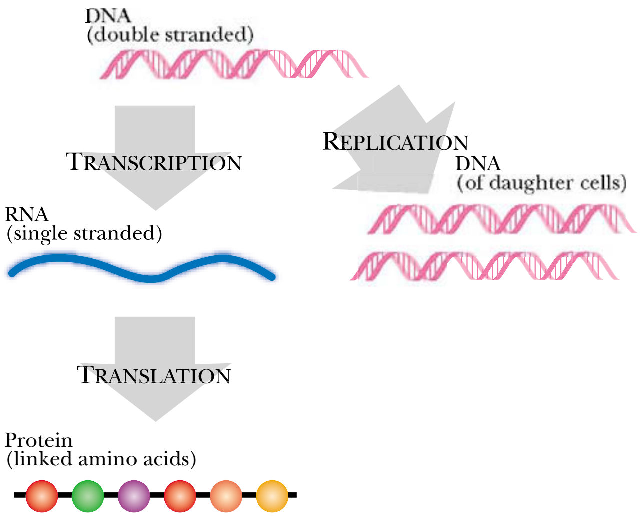


FIGURE 3.17 ***The Central Dogma (Simple Version)***

The information flow in cells begins with DNA, which may either be replicated, giving a duplicate molecule of DNA, or be transcribed to give RNA. The RNA is read (translated) as a protein is built.

messenger RNA (mRNA) are used. These are RNA copies of genetic information stored by the DNA and are made by a process called **transcription**. The messenger RNA molecules carry information from the genome to the cytoplasm, where the information is used by the **ribosomes** to synthesize **proteins**. In eukaryotes, mRNA is not made directly. Instead, transcription yields precursor RNA molecules (pre-mRNA) that must be processed, to produce the actual mRNA as detailed in Ch. 12.

The DNA that carries the primary copy of the genes is present as gigantic molecules, each carrying hundreds or thousands of genes. In contrast, any individual messenger RNA molecule carries only one or a few genes' worth of information. Thus, in practice, multiple short segments of DNA are transcribed simultaneously to give many different messenger RNA molecules. In eukaryotes, each mRNA normally carries only a single gene, whereas in prokaryotes, anywhere from one to a dozen genes may be transcribed as a block to give an mRNA molecule carrying several genes, usually with related functions (Fig. 3.18).

Translation is the synthesis of proteins using genetic information carried by messenger RNA. Proteins consist of one or more polymer chains known as **polypeptides**. These are made from subunits called **amino acids**. Translation thus involves transfer of information from nucleic acids to an entirely different type of macromolecule. This decoding process is carried out by ribosomes. These submicroscopic machines read the messenger RNA and use the information to make a polypeptide chain. Proteins, which make up about two-thirds of the organic matter in a typical cell, are directly responsible for most of the processes of metabolism. Proteins perform most of the enzyme reactions and transport functions of the cell. They also provide many structural components and some act as regulatory molecules, as described below.

Proteins are made by a subcellular machine, the ribosome, that uses the genetic code to read information encoded by nucleic acids.

Ribosomes Read the Genetic Code

This introductory section will summarize protein synthesis as it occurs in bacteria. It should be noted that the details of protein synthesis differ between bacteria and higher organisms (see Ch. 8). The bacterial ribosome, as described in more detail below, consists of two subunits, small (30S) and large (50S). **S-values** tell how fast a particle

amino acid Monomer from which the polypeptide chains of proteins are built
polypeptide chain A polymer that consists of amino acids
protein Polymer made from amino acids; may consist of several polypeptide chains
ribosome The cell's machinery for making proteins
S-value The sedimentation coefficient is the velocity of sedimentation divided by the centrifugal field. It is dependent on mass and is measured in Svedberg units
transcription Conversion of information from DNA into its RNA equivalent
translation Making a protein using the information provided by messenger RNA

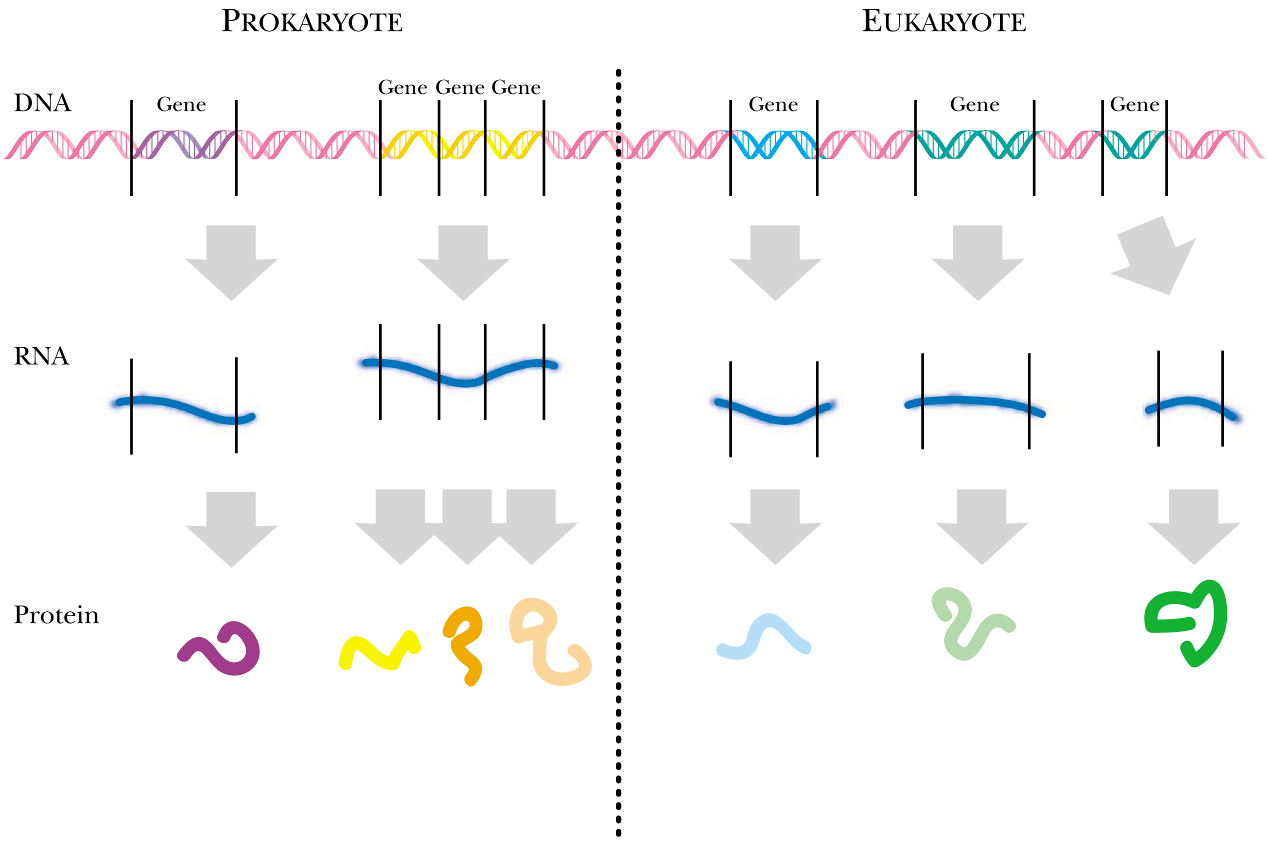


FIGURE 3.18 ***Differing Patterns of Transcription***

In eukaryotes, each gene is transcribed to give a separate mRNA that encodes only a single protein. In prokaryotes, an mRNA molecule may carry information from a single gene or from several genes that are next to each other on the chromosome.

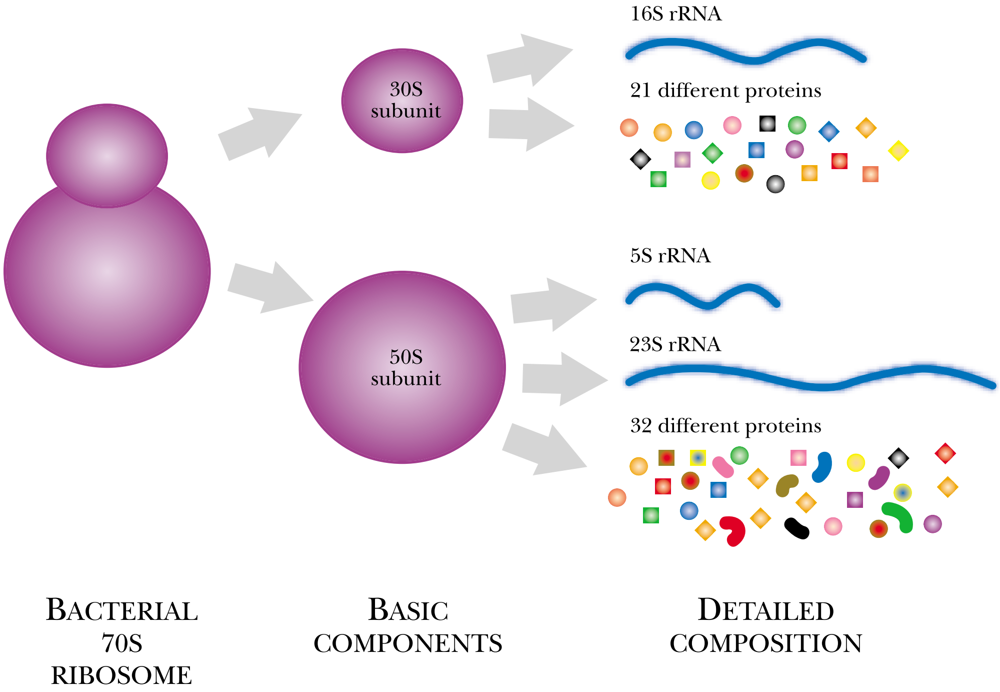


FIGURE 3.19 ***Structural Components of a Ribosome***

The bacterial ribosome can be broken down into two smaller subunits and finally into RNA molecules and proteins.

FIGURE 3.20 ***The Genetic Code***

A codon consisting of three base pairs determines each amino acid to be added to a growing polypeptide chain. The codon table shows the 64 different codons, in RNA language, alongside the amino acids they encode. Three of the codons act as stop signals. AUG (methionine) and GUG (valine) act as start codons.

	2nd (middle) base				
1st base	U	C	A	G	3rd base
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA stop	UGA stop	A
	UUG Leu	UCG Ser	UAG stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

sediments in an ultracentrifuge. They give a rough indication of size but are not linearly related to molecular weight. A complete ribosome with a 30S and a 50S subunit has an S-value of 70S (not 80S).

By weight, the ribosome itself consists of about two-thirds **ribosomal RNA (rRNA)** and one-third protein. In bacteria, the large subunit has two rRNA molecules, 5S rRNA and 23S rRNA, and the small subunit has just the one 16S rRNA. In addition to the rRNA, there are 52 different proteins, 31 in the large subunit and the other 21 in the small subunit (Fig. 3.19). The rRNA molecules are NOT themselves translated into protein; instead, they form part of the machinery of the ribosome that translates the mRNA.

The Genetic Code Dictates the Amino Acid Sequence of Proteins

The bases of DNA or RNA are grouped in threes for decoding.

There are 20 amino acids in proteins but only four different bases in the messenger RNA. So nature cannot simply use one base of a nucleic acid to code for a single amino acid when making a protein. During translation, the bases of mRNA are read off in groups of three, which are known as **codons**. Each codon represents a particular amino acid. Since there are four different bases, there are 64 possible groups of three bases; that is, 64 different codons in the **genetic code**. However, there are only 20 different amino acids making up proteins, so some amino acids are encoded by more than one codon. In addition, three of the codons are used for punctuation to stop the growing chain of amino acids (Fig. 3.20). In addition, the codon, AUG, encoding methionine, acts as a start codon. Thus newly made polypeptide chains start with the amino acid methionine. [Much less often, GUG encoding valine, may also act as the start codon. However, even if the start codon is GUG the first amino acid of the newly made protein is methionine (not valine).]

To read the codons a set of adapter molecules is needed. These molecules, known as **transfer RNA (tRNA)**, recognize the codon on the mRNA at one end and carry the corresponding amino acid attached to their other end (Fig. 3.21). These adapters represent a third class of RNA and were named transfer RNA since they transport amino acids to the ribosome in addition to recognizing the codons of mRNA. Since there are numerous codons, there many different tRNAs. [Actually, there are fewer different tRNA molecules than codons as some tRNA molecules can read multiple codons—see Ch. 8 for details.] At one end, the tRNA has an **anticodon** consisting of three bases that are complementary to the three bases of the codon on the messenger RNA. The

anticodon Group of three complementary bases on tRNA that recognize and bind to a codon on the mRNA
codon Group of three RNA or DNA bases that encodes a single amino acid
genetic code The code for converting the base sequence in nucleic acids, read in groups of three, into the sequence of a polypeptide chain
ribosomal RNA (rRNA) Class of RNA molecule that makes up part of the structure of a ribosome
transfer RNA (tRNA) RNA molecules that carry amino acids to a ribosome

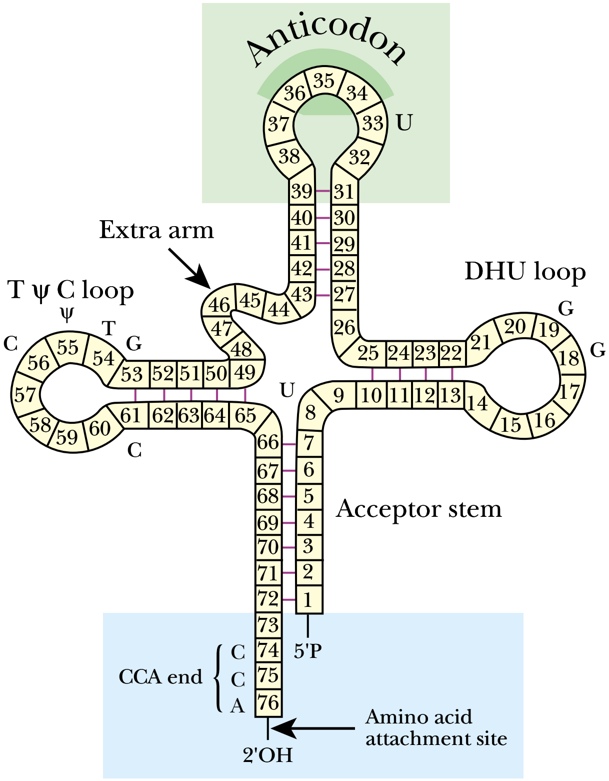


FIGURE 3.21 ***Transfer RNA Contains the Anticodon***

Each transfer RNA molecule has an anticodon that is complementary to the codon carried on the messenger RNA. The codon and anticodon bind together by base pairing. At the far end of the tRNA is the acceptor stem ending in the bases CCA (cytosine, cytosine, adenine). Here is attached the amino acid that corresponds to the codon on the mRNA.

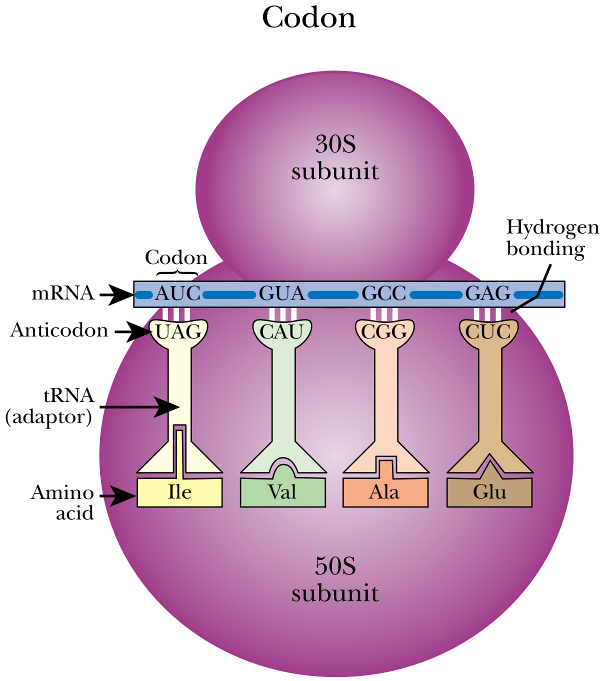


FIGURE 3.22 ***Stylized Relationship of Charged tRNA to mRNA and the Ribosome***

Note: This figure does not show the correct physical arrangement—instead it illustrates the coding relationships between the tRNA and mRNA. The mRNA binds to the 30S subunit of the ribosome. The anticodons of the tRNAs carrying amino acids bind to the corresponding codons on the mRNA. In real life only two tRNAs are present on the ribosome at any given time and the codons on the mRNA are contiguous, with no gaps between.

codon and anticodon recognize each other by base pairing and are held together by hydrogen bonds. At its other end, each tRNA carries the amino acid corresponding to the codon it recognizes.

The small (30S) subunit binds the messenger RNA and the large (50S) subunit is responsible for making the new polypeptide chain. Figure 3.22 shows the relationship between the mRNA and the tRNAs in a stylized way. In practice, only two tRNA molecules are base-paired to the messenger RNA at any given time. After binding to the mRNA, the ribosome moves along it, adding a new amino acid to the growing polypeptide chain each time it reads a codon from the message (Fig. 3.23). A more detailed account of protein synthesis is given in Chapter 8.

FIGURE 3.23 ***Ribosome Elongating a Polypeptide Chain***

A new amino acid is added to the polypeptide chain each time a new tRNA arrives at the ribosome, bringing its attached amino acid. The anticodon of the tRNA binds to the mRNA. The large subunit cross links the incoming amino acid to the growing chain, such that the incoming tRNA ends up carrying the growing polypeptide chain. The 30S subunit of the ribosome then moves one step along the mRNA. This results in ejection of the left-most tRNA and readies the mRNA to accept the next incoming tRNA. The polypeptide chain continues to grow until a "stop codon" is reached.

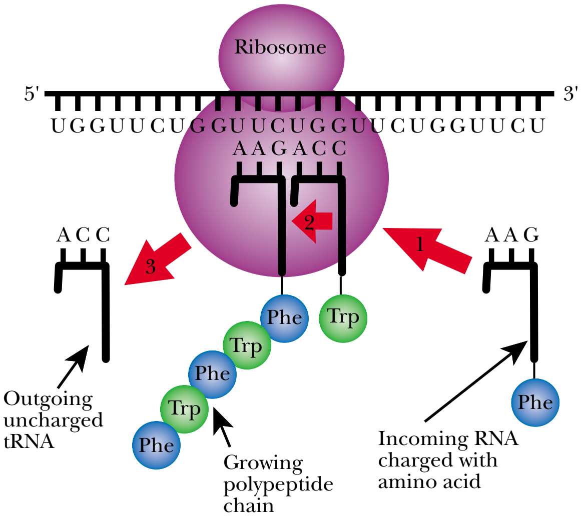

Ribosome
5'
3'
UGGUUCUGGUUCUGGUUCUGGUUCU
AAGACC
ACC
AAG
3
2
1
Phe
Trp
Trp
Phe
Trp
Phe
Phe
Outgoing uncharged tRNA
Growing polypeptide chain
Incoming RNA charged with amino acid

TABLE 3.02 Major Classes of Non-Translated RNA

Name	Function
Ribosomal RNA	comprises major portion of ribosome and is involved in synthesis of polypeptide chains
Transfer RNA	carries amino acids to ribosome and recognizes codons on mRNA
Small nuclear RNA	involved in the processing of messenger RNA molecules in the nucleus of eukaryotic cells (also called snRNA, or "snurps")
Guide RNA	involved in processing of RNA or DNA in some organisms
Regulatory RNA	functions in the regulation of gene expression by binding to proteins or DNA or to other RNA molecules
Antisense RNA	functions in regulating gene expression by base pairing to mRNA
Recognition RNA	part of a few enzymes (e.g., telomerase); enables them to recognize certain short DNA sequences
Ribozymes	enzymatically active RNA molecules

Various Classes of RNA Have Different Functions

Originally, genes were regarded as units of heredity and alleles were defined as alternative versions of a gene. However, these concepts have been broadened as knowledge of genome structure has increased. Molecular insights led first to the view of genes as segments of DNA encoding proteins—the one gene—one enzyme model of Beadle and Tatum. In this case, messenger RNA acts as an intermediary between the DNA, which is used for storage of genetic information, and the protein, which functions in running the cell. The concept of a gene was then further extended to include segments of DNA that encode RNA molecules that are not translated into protein but function as RNA. The most common examples are the ribosomal RNA and transfer RNA involved in protein synthesis. The term "gene products" therefore refers to such non-translated RNA molecules as well as proteins. For convenience, the major classes of non-translated RNA are summarized in Table 3.02.

In addition to the chemical differences discussed above (ribose instead of deoxyribose and uracil instead of thymine), RNA differs from DNA in several respects. RNA is usually single stranded, although most RNA molecules do fold up, thus producing double stranded regions. RNA molecules are usually much shorter than DNA and only

RNA is not so simple after all. Several classes of RNA exist that carry out a variety of roles in addition to carrying information for protein synthesis. See especially Chapter 11 for novel insights into the role of RNA in regulation.

antisense RNA RNA complementary in sequence to messenger RNA and which, therefore, base pairs with it and prevents translation
ribozyme RNA molecule that acts as an enzyme

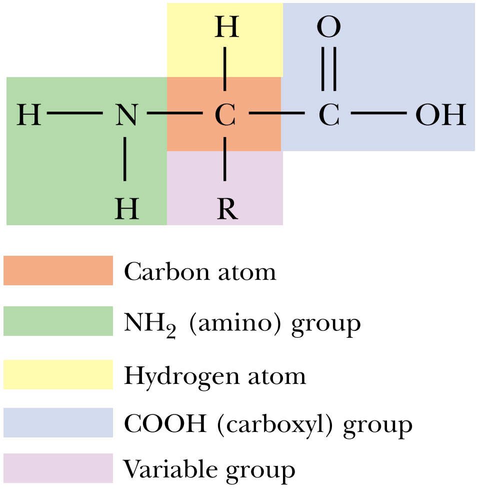


FIGURE 3.24 ***General Features of Amino Acids***

Almost all amino acids found in proteins have the features shown in common. In glycine, the simplest amino acid, the R group is a single hydrogen atom. In proline, the R group consists of a ring structure that bonds to the nitrogen atom shown. This therefore only has a single attached hydrogen and becomes an imino group (—NH—).

carry the information for one or a few genes. Moreover, RNA is usually much shorter-lived than DNA, which is used for long-term storage of the genome. Some classes of RNA molecules, especially tRNA, contain unusual, chemically modified bases that are never found in DNA (see Ch. 8).

The above differences in function between RNA and DNA apply to living cells. However, certain viruses carry their genomes as either single or double-stranded RNA. In such cases, multiple genes will obviously be present on these RNA genomes. Furthermore, double-stranded viral RNA can form a double helix, similar though not identical in structure to that of DNA. The properties of viruses and the novel aspects of their genomes are discussed more fully in Chapter 17.

Proteins, Made of Amino Acids, Carry Out Many Cell Functions

Proteins are multifunctional biological polymers that consist of one or more polypeptide chains. The information carried by messenger RNA is translated to give a polypeptide chain. The linear sequence of nucleotides in the RNA (read in groups of three—i.e. codons) corresponds to the linear sequence of the amino acids that make up the polypeptide chain. That is, the mRNA and the polypeptide chain are co-linear.

Typically, about 60% of the organic matter in a cell is protein. Most of the cell's activities and many of its structures depend on its proteins.

Some proteins act as **enzymes** to catalyze biochemical reactions including the generation of energy and the synthesis of nucleotides and their assembly into nucleic acids. Other proteins are structural, or transport nutrients or take part in cell movement (mechanical proteins). Finally, there are proteins involved in information processing. Molecules whose primary role is to carry information (nucleic acids like DNA and messenger RNA) are basically linear molecules with a regular repeating structure. Molecules that form cellular structures or have active roles carrying out reactions are normally folded into three-dimensional (3-D) structures. These include both proteins and most non-translated RNA molecules, including tRNA and rRNA.

Proteins are made from a linear chain of monomers, known as amino acids (Fig. 3.24), and are folded into a variety of complex 3-D shapes. A chain of amino acids is called a **polypeptide chain** (Fig. 3.25). There are 20 different amino acids used in making proteins. All have a central carbon atom, the **alpha carbon**, surrounded by a hydrogen atom, an amino group (NH_2), a carboxyl group (COOH), and a variable side chain, the **R-group** (Fig. 3.24). Amino acids are joined together by **peptide bonds** (Fig. 3.25). The first amino acid in the chain retains its free amino group and this end is often called the

alpha carbon The central carbon atom of an amino acid, to which the amino, carboxyl and R groups are attached
enzyme A protein or RNA molecule that catalyses a chemical reaction
peptide bond Type of chemical linkage holding amino acids together in a protein molecule
polypeptide chain Polymeric chain of amino acids
R-group Chemical group forming side chain of amino acid

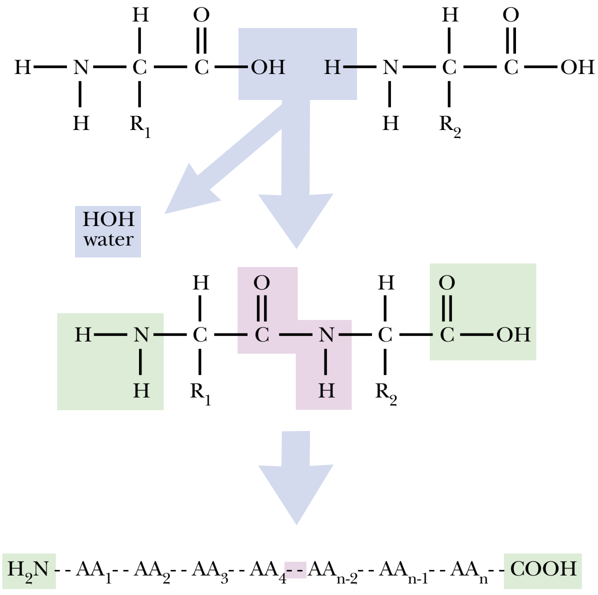

HOH
water
H_2N - - AA_1 - - AA_2 - - AA_3 - - AA_4 - - AA_{n-2} - - AA_{n-1} - - AA_n - - COOH

FIGURE 3.25 ***Formation of a Polypeptide Chain***

A polypeptide chain is formed as amino and carboxyl groups on two neighboring amino acids combine and eliminate water. The linkage formed is known as a peptide bond. No matter how many amino acids are added, the growing chain always has an N- or amino terminus, and a C- or carboxy terminus.

Two amino acids eliminate water from their amino and carboxy regions.

A peptide bond is formed.

A polypeptide is formed when an amino acid (AA) join, leaving a linear structure with a N-terminus (NH_2) an a carboxy terminus (COOH).

amino- or **N-terminus** of the polypeptide chain. The last amino acid to be added is left with a free carboxyl group and this end is often called the **carboxy-** or **C-terminus**.

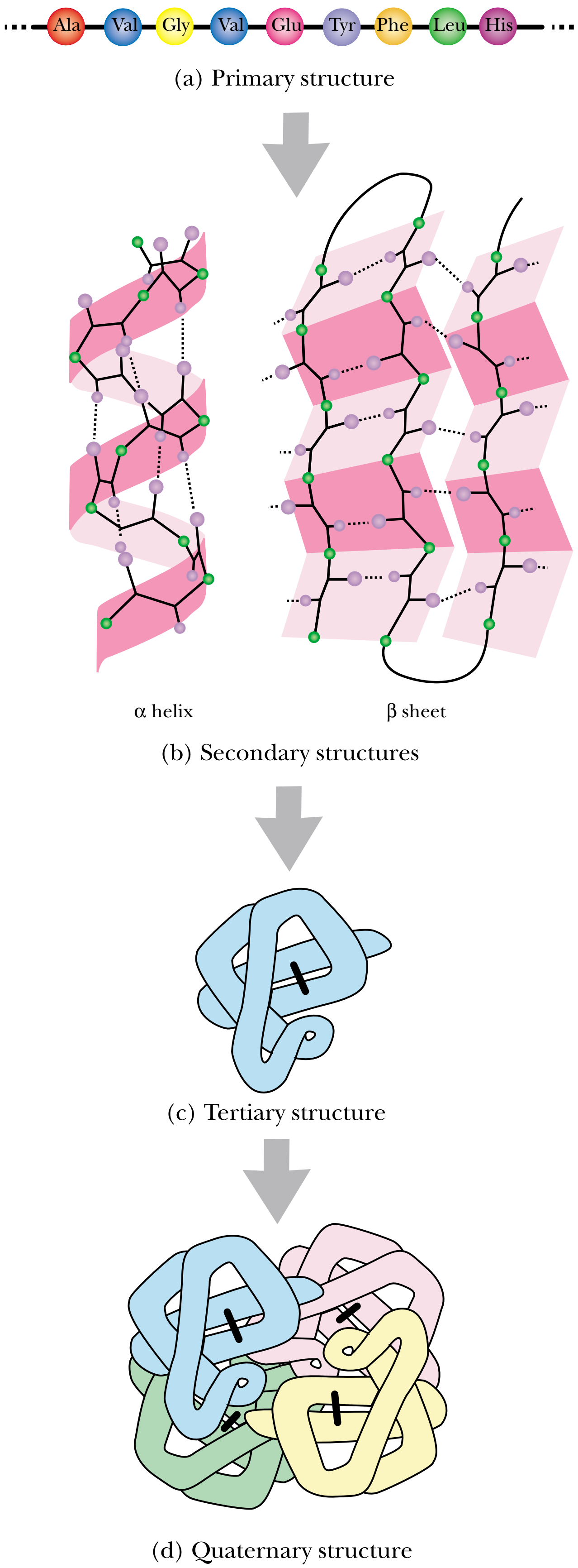
Some proteins consist of a single polypeptide chain; others contain more than one. To function properly, many proteins need extra components, called **cofactors** or **prosthetic groups**, which are not made of amino acids. Many proteins use single metal atoms as cofactors; others need more complex organic molecules.

The Structure of Proteins Has Four Levels of Organization

For a protein to be functional, the polypeptide chains must be folded into their correct 3-D structures. The structures of biological polymers, both protein and nucleic acid, are often divided into levels of organization (Fig. 3.26). The first level, or **primary structure**, is the linear order of the monomers—i.e., the sequence of the amino acids for a protein, or of the nucleotides in the case of DNA or RNA. **Secondary structure** is the folding or coiling of the original polymer chains by means of hydrogen bonding. Although DNA is not a protein, hydrogen bonding between base pairs forms the famous double helix. In proteins, hydrogen bonding between peptide groups results in several possible helical or wrinkled sheet-like structures (see Ch. 7 for details).

The next level is the **tertiary structure**. The polypeptide chain, with its preformed regions of secondary structure, is then folded to give the final 3-D structure. This level of folding depends on the side chains of the individual amino acids. In certain cases, proteins known as chaperonins help other proteins to fold correctly (see Ch. 7). As there are 20 different amino acids, a great variety of final 3-D conformations is possible. Nonetheless, many proteins are roughly spherical. Lastly, **quaternary structure** is the assembly of several individual polypeptide chains to give the final structure. Not

amino- or N-terminus The end of a polypeptide chain that is made first and that has a free amino group
carboxy- or C-terminus The end of a polypeptide chain that is made last and has a free carboxy-group
cofactor Extra chemical group non-covalently attached to a protein that is not part of the polypeptide chain
primary structure The linear order in which the subunits of a polymer are arranged
prosthetic group Extra chemical group covalently attached to a protein that is not part of the polypeptide chain
quaternary structure Aggregation of more than one polymer chain in final structure
secondary structure Initial folding up of a polymer due to hydrogen bonding
tertiary structure Final 3-D folding of a polymer chain



(a) Primary structure

(b) Secondary structures

(c) Tertiary structure

(d) Quaternary structure

FIGURE 3.26 ***Four Levels of Protein Structure***

The final protein structure is best understood by following the folding process from simple to complex. The primary structure is the specific order of the amino acids (a). The secondary structure is due to regular folding of the polypeptide chain due to hydrogen bonding (b). The tertiary structure results from further folding of the polypeptide due to interactions between the amino acid side chains (c). Finally, the quaternary structure is the assembly of multiple polypeptide chains (d).

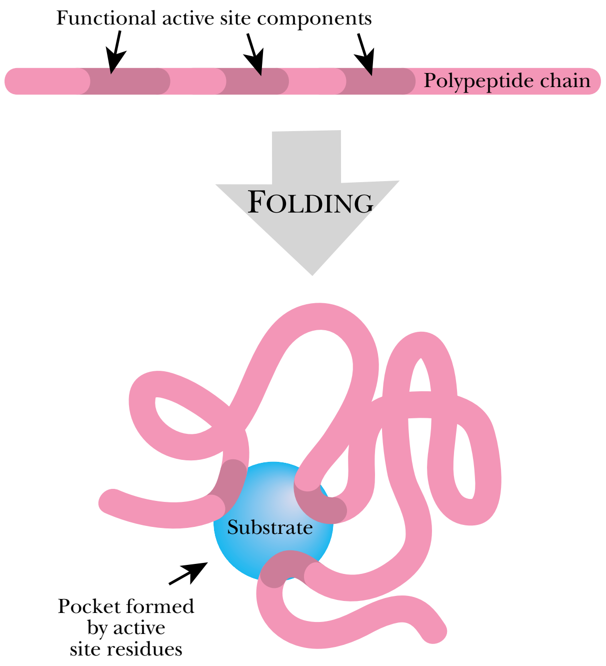

Functional active site components
Polypeptide chain
FOLDING
Substrate
Pocket formed by active site residues

FIGURE 3.27 Polypeptide Forms an Active Site after Folding

Folding of the protein brings together several regions of the polypeptide chain that are needed to perform its biological role. The active site forms a pocket for binding the substrate. Some of the amino acid residues at the active site are also involved in chemical reactions with the substrate.

all proteins have more than one polypeptide chain; some just have one, so they have no quaternary structure.

Proteins Vary in Their Biological Roles

Functionally, proteins may be divided into four main categories: **structural proteins**, enzymes, **regulatory proteins** and **transport proteins**.

1. Structural proteins make up many sub-cellular structures. The flagella with which bacteria swim around, the microtubules used to control traffic flow inside cells of higher organisms, the fibers involved in contractions of a muscle cell, and the outer coats of viruses are a few examples of structures constructed using proteins.
2. Enzymes are proteins that facilitate chemical reactions. An enzyme first binds another molecule, known as its **substrate**, and then performs some chemical operations with it. Some enzymes bind only a single substrate molecule; others may bind two or more, and react them together to make the final product. In any case, the enzyme needs an **active site**, a pocket or cleft in the protein, where the substrate binds and the reaction occurs. The active site of the protein is produced by the folding up of its polypeptide chain correctly so that amino acid residues that were spread out at great distances in the linear chain now come together and will cooperate in binding the substrate to facilitate the enzyme reaction (Fig. 3.27).
3. Although regulatory proteins are not enzymes, they do bind other molecules and so they also need active sites to accommodate these. Regulatory proteins vary enormously. Many of them can bind both small signal molecules and DNA. The presence or absence of the signal molecule determines whether or not the gene is switched on (Fig. 3.28).

active site Special site or pocket on a protein where the substrate binds and the enzyme reaction occurs
regulatory protein A protein that regulates the expression of a gene or the activity of another protein
structural protein A protein that forms part of a cellular structure
substrate The molecule altered by the action of an enzyme
transport protein A protein that carries other molecules across membranes or around the body

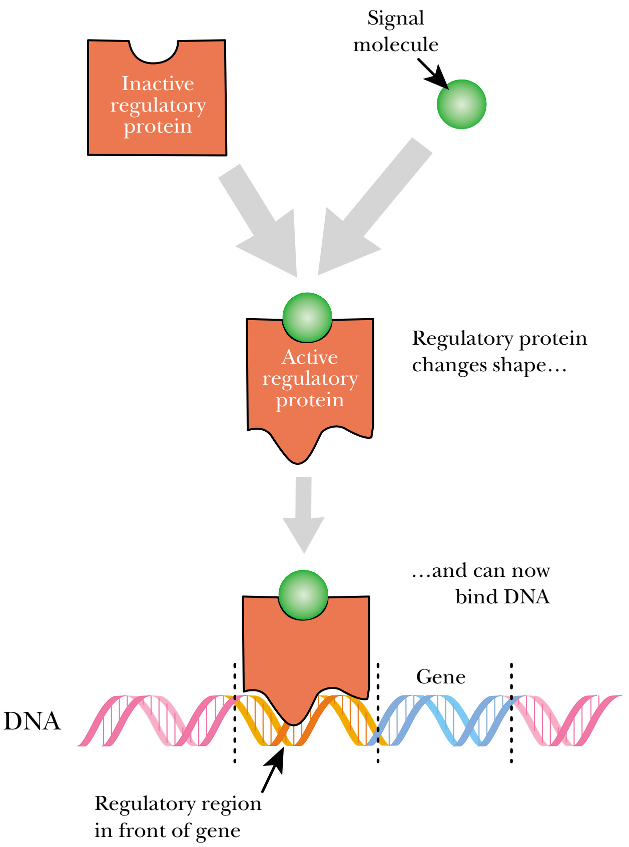

Signal molecule
Inactive regulatory protein
Active regulatory protein
Regulatory protein changes shape…
…and can now bind DNA
Gene
DNA
Regulatory region in front of gene

FIGURE 3.28 ***Regulatory Protein***

Regulatory proteins usually exist in two conformations. Receiving a signals promotes a change in shape. The regulatory protein may then bind to DNA and alter the expression of a gene.

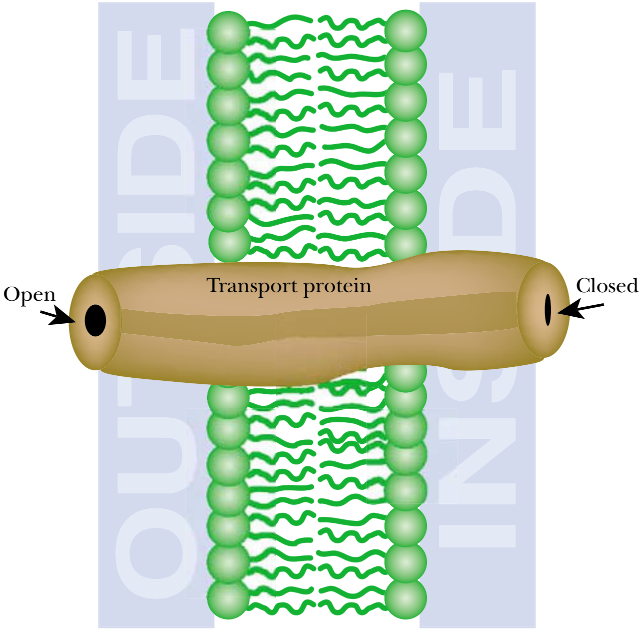


FIGURE 3.29 ***Transport Proteins***

Transport proteins are often found in cell membranes where they are responsible for the entry of nutrients or the export of waste products.

4. Transport proteins are found mostly in biological membranes, as shown in Figure 3.29, where they carry material from one side to the other. Nutrients, such as sugars, must be transported into cells of all organisms, whereas waste products are deported. Multi-cellular organisms also have transport proteins to carry materials around the body. An example is hemoglobin, which carries oxygen in blood.

CHAPTER FOUR

Genes, Genomes and DNA

History of DNA as the Genetic Material

Despite their complexity, living organisms obey the laws of chemistry.

Until early in the nineteenth century, it was believed that living matter was quite different from inanimate matter and was not subject to the normal laws of chemistry. In other words, organisms were thought to be made from chemical components unique to living creatures. Furthermore, there was supposedly a special vital force that mysteriously energized living creatures. Then, in 1828, Friedrich Wohler demonstrated the conversion in a test tube of ammonium cyanate, a laboratory chemical, to urea, a "living" molecule also generated by animals. This was the first demonstration that there was nothing magical about the chemistry of living matter.

Further experiments showed that the molecules found in living organisms were often very large and complex. Consequently, their complete chemical analysis was time consuming and is indeed, still continuing today. The de-mystification of life chemistry reached its peak in the 1930s when the Russian biochemist Alexander Oparin wrote a book outlining his proposal for the chemical origin of life. Although the nature of the genetic material was still unknown, Oparin put forward the idea that life, with its complex molecular composition, evolved from small molecules in the primeval ocean as a result of standard physical and chemical forces (see Ch. 20).

Avery found that purified DNA could carry genetic information from one strain of bacterium to another. This revealed that DNA was the genetic material.

Until the time of World War II, the chemical nature of the *inherited genetic information* remained very vague and elusive. DNA was actually discovered in 1869 by Frederich Miescher, who extracted it from the pus from infected wounds! However, it was nearly a century before its true significance was revealed by Oswald Avery. In 1944, Avery found that the virulent nature of some strains of bacteria that caused pneumonia could be transmitted to related harmless strains by a chemical extract. Avery purified the essential molecule and demonstrated that it was DNA, although he did not use the name "DNA," since its structure was then uncharacterized. When DNA from virulent strains was added to harmless strains, some took up the DNA and were "transformed" into virulent strains (see Ch. 18 for the mechanism of transformation). Avery concluded that the genes were made of DNA and that somehow genetic information was encoded in this molecule. Since DNA was known to have only half a dozen components, it had not been a leading competitor for the role of genetic material; it was viewed as too simple to encode the information for a living creature!

The question of how DNA, with only half a dozen components, could act as the genetic information was answered by James Watson and Francis Crick in 1953. Their now famous double helix provided a chemical basis for the genetic code and suggested

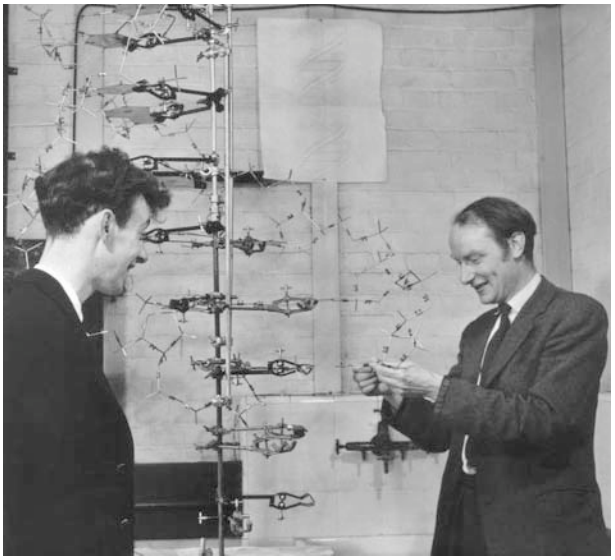


FIGURE 4.01 ***Watson and Crick in the 1950s***

James Watson (b.1928) at left and Francis Crick (b.1916), with their model of part of a DNA molecule in 1953. Courtesy of: A. Barrington Brown, Science Photo Library.

a mechanism for DNA replication. In 1950 Maurice Wilkins and his assistant Raymond Gosling took the first images of DNA using X-ray diffraction. Gosling's work was continued by Rosalind Franklin who joined Wilkins' group the following year. Watson and Crick used a X-ray diffraction picture taken by Rosalind Franklin and Raymond Gosling in 1952 as the basis for their structural model. Rosalind Franklin died in 1958 of cancer aged 37, probably due to the effects of the X-rays. Unraveling the chemical basis for inheritance won Watson, Crick and Wilkins the Nobel Prize in Physiology or Medicine for 1962 "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material".

X-ray diffraction showed that two strands of DNA are twisted together forming a double helix.

This central finding underlies our whole understanding of how living cells operate and what life means. Since the discovery of the double helix, the genetic code has been worked out, and starting in 1995 with the bacterium *Haemophilus influenzae*, the DNA of a variety of organisms has been totally sequenced. As the third millennium begins, the human genome has been sequenced, but researchers are still working to assemble the data into complete contiguous sequences for each chromosome. This chapter will discuss how much genetic information is needed to operate a living cell and how that information is arranged on the DNA. Much of the information about these processes comes from studying bacteria, but the information frequently applies also to eukaryotes.

The Double Helix by James D. Watson
Published in 1968 by Atheneum, New York

This book gives a personal account of the greatest biological advance of the 20th century—the unraveling of the structure of the DNA double helix by James Watson and Francis Crick. Like the bases of DNA, Watson and Crick formed a complementary pair. Crick, a physicist with an annoying laugh, was supposed to be working towards a Ph.D. on protein X-ray crystallography. Watson was a homeless American biologist, wandering around Europe with a post-doctoral fellowship, looking for something to do.

Despite spending much time carousing, the intrepid heroes, Crick and Watson, beat their elders to the finish line. Watson describes with relish how the great American chemist, Linus Pauling, placed the phosphate backbone of DNA down the middle, so failing to solve the structure. The data proving the phosphate backbone was on the outside of the double helix came from Rosalind Franklin, an X-ray crystallographer at London University. Of her Watson says, ". . . the best home for a feminist was in another person's lab."

The Director of the Cavendish Laboratory at Cambridge was Sir William Bragg, the august inventor of X-ray crystallography. Despite being depicted as a stuffy has-been who nearly threw Crick out for loud-mouthed insubordination, Bragg wrote the foreword to the book. After all, when younger scientists under your direction make the greatest discovery of the century, it is no time to bear a grudge!

The biographies of great scientists are usually exceedingly dull. Who cares, after all, what Darwin liked for breakfast or what size shoes Mendel wore? It is their discoveries, and how they changed the world, that are fascinating. "The Double Helix" is different. Biographers are generally minor figures, understandably hesitant to criticize major achievers. Watson, himself a big name, happily lacks such respect, and cheerfully castigates other top scientists. It is this honest portrayal of the flaws and fantasies of those involved in unraveling the DNA double helix that keeps the readers attention.

If your stomach can't stand any more sagas about caring investigators who work on into the early hours hoping that their discoveries will help sick children, this book is for you. Like most candid scientists, Watson and Crick did not work for the betterment of mankind; they did it for fun.

How Much Genetic Information Is Necessary to Maintain Life?

Bacteria typically carry all of their genes on a single circular chromosome. Occasional species of bacteria have two or more different chromosomes and a few bacteria even contain linear chromosomes (see Ch. 18). Higher organisms have linear chromosomes and the number ranges from a handful to over a thousand in a few flowering plants.

The simplest living cell probably needs around 200–300 genes.

The smallest eubacterial genome is that of *Mycoplasma genitalium*, which has one circular chromosome, consisting of 580,000 base pairs (**bp**) of DNA. Since the average gene is about 1,000 base pairs long, and there is little non-coding DNA between bacterial genes, *M. genitalium* should have approximately 500 genes. The precise number is disputed and ranges from 468 to 517. Extensive analysis of mutations suggests that around 300 of these genes are essential for the growth and reproduction of *M. genitalium*. Comparison with other small bacterial genomes suggests that around 250 genes may be the minimum essential for a living cell. (This estimate also takes into account those pathways needed for synthesis of all vital cell components, some of which are missing in *M. genitalium* because it is a parasitic bacterium.) The smallest prokaryotic genome belongs to *Nanoarchaeum equitans*, a marine archaebacterium that was discovered in 2002. *N. equitans* has about 15% less DNA than *M. genitalium* and may also be a parasite, as it cannot grow unless attached to the surface of other microorganisms. Despite having less DNA its genes are more closely spaced and in consequence *N. equitans* actually has more coding sequences than *M. genitalium*—approximately 550.

Most bacteria have a few thousand genes.

Although parasitic bacteria may have less than 1,000 genes, most free-living bacteria have 2,000 to 4,000 genes. Occasional bacteria with complex life cycles, such as *Myxococcus*, may have 9,000 to 10,000 genes. Free-living eukaryotes typically have from 6,000 to 50,000 genes (see Table 4.01). However, the parasitic eukaryote *Encephalitozoon cuniculi* (Protozoa, Microspora) has only 2.9 million base pairs (**Mbp**) of DNA, implying that it possesses no more than 3,000 genes—less than many bacteria.

Non-Coding DNA

Some regions of DNA contain useful genetic information, other regions do not.

The number of genes ranges from roughly 500 to 50,000—a 100-fold range. In contrast, the amount of DNA ranges from 0.5 Mbp to nearly 50,000 Mbp—a 100,000-fold range. This discrepancy is due to **non-coding DNA** in eukaryotes with larger DNA content. This, as its name indicates, is DNA whose base sequence is largely meaningless and does not encode useful genetic information, at least as far as we know currently.

In addition to segments of DNA that give rise to gene products (i.e., protein or RNA), DNA molecules contain many other regions, including both regulatory sites and non-coding regions. Any segment of DNA, whether coding or not, can be referred to as a **locus (plural, loci)**; that is, a location on a chromosome (or other molecule of DNA). Since any DNA sequence can occur in alternative versions, the term **allele** is used for these even if the DNA in question is non-coding.

Non-coding DNA accounts for the majority of the DNA in most higher animals and plants.

Although bacteria have relatively little non-coding DNA, eukaryotes have significant amounts. Even a relatively primitive eukaryote, such as yeast, has nearly 50 percent non-coding DNA. For example, yeast has about three times as much DNA as *E. coli* but only 1.5 times as many genes. Higher eukaryotes have even greater proportions of non-coding DNA. Mammals such as mice and men have an estimated

allele One particular version of a gene, or more broadly, a particular version of any locus on a molecule of DNA
bp Abbreviation for base pair(s)
locus (plural, loci) A place or location on a chromosome; it may be a genuine gene or just any site with variations in the DNA sequence that can be detected, like RFLPs or VNTRs
Mbp Megabase pairs or million base pairs
non-coding DNA DNA that does not code for proteins or functional RNA molecules

TABLE 4.01 Genome Sizes

Organism	Number of Genes	Amount of DNA (bp)	Number of Chromosomes
Viruses			
Bacteriophage MS2	4	3,600	1 (ssRNA)*
Tobacco Mosaic Virus	4	6,400	1 (ssRNA)*
ΦX174 bacteriophage	11	5,387	1 (ssDNA)
Influenza	12	13,500	8 (ssRNA)
T4 bacteriophage	200	165,000	1
Poxvirus	300	187,000	1
Bacteriophage G	680	498,000	1
Prokaryotes			
Mitochondrion (human)	37	16,569	1
Mitochondrion (*Arabidopsis*)	57	366,923	1
Chloroplast (*Arabidopsis*)	128	154,478	1
Nanoarchaeum equitans	550	490,000	1
Mycoplasma genitalium	480	580,000	1
Methanococcus	1,500	1.7 Mbp	1
Escherichia coli	4,000	4.6 Mbp	1
Myxococcus	9,000	9.5 Mbp	1
Eukaryotes (haploid genome)			
Encephalitozoon	2,000	2.5 Mbp	11
Saccharomyces	5,700	12.5 Mbp	16
Caenorhabditis	19,000	100 Mbp	6
Drosophila	12,000	140 Mbp	5
Homo sapiens	25,000	3,300 Mbp	23
Arabidopsis	25,000	115 Mbp	5
Oryza sativa (Rice)	45,000	430 Mbp	12

*ssRNA = single stranded RNA; ssDNA = single stranded DNA; all other genomes consist of double stranded DNA.

20,000 to 30,000 genes carried on a total of 300 Mbp of DNA. This means that just over 85 percent is non-coding. However, flowering plants, which are estimated to have roughly as many genes as mammals, possess 100-fold more DNA. Some amphibians, such as frogs and newts, possess almost as much.

In eukaryotes, the genes themselves are often interrupted by stretches of non-coding DNA.

In prokaryotes, almost all the non-coding DNA is found between genes as **intergenic DNA**. In eukaryotes, the situation is more complicated. Not only is non-coding DNA scattered throughout the eukaryotic chromosomes between the genes, but the actual genes themselves are often interrupted with non-coding DNA. These **intervening sequences** are known as **introns**, whereas the regions of the DNA that contain coding information are known as **exons**. Most eukaryotic genes consist of exons alternating with introns (Fig. 4.02).

In lower, single-celled eukaryotes, such as yeast, introns are relatively rare and often quite short. In contrast, in higher eukaryotes, most genes have introns and they

exon Segment of a gene that codes for protein and that is still present in the messenger RNA after processing is complete
intergenic DNA Non-coding DNA that lies between genes
intervening sequence An alternative name for an intron
intron Segment of a gene that does not code for protein but is transcribed and forms part of the primary transcript

FIGURE 4.02 ***Intervening Sequences Interrupt Eukaryotic Genes***

Regions of non-coding DNA between genes are called intergenic DNA. Non-coding regions that interrupt the coding regions of genes are called introns.

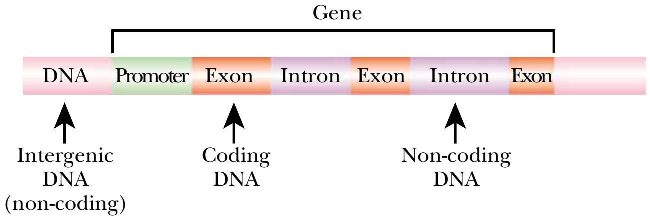
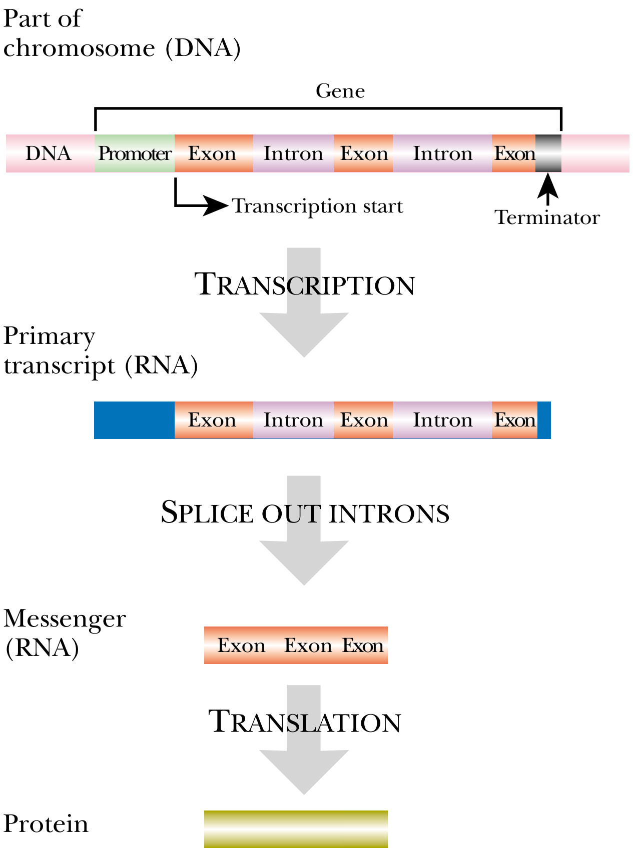


FIGURE 4.03 ***Removal of Introns Before Protein Synthesis***

A gene on the chromosomal DNA, consisting of a promoter, introns and exons, is transcribed to give the primary transcript, an RNA molecule containing both the introns and exons. Processing of the primary transcript removes the introns to leave an mRNA carrying only exons. The protein reflects only the information in the exons.



Genes that are interrupted must have the non-coding regions removed when messenger RNA is made.

are often longer than the exons. In some genes, the introns may occupy 90 percent or more of the DNA. For example, the mutated gene causing cystic fibrosis was found to occupy 250,000 base pairs and have 24 exons, which encoded a protein of 1,480 amino acids. Since 1,480 amino acids need only 4,440 base pairs to encode them, this means that scarcely 2 percent of the cystic fibrosis gene is actual coding DNA. The rest consists of intervening sequences—23 introns.

In order to synthesize the protein encoded by an interrupted gene, the introns must be removed at some stage. Splicing out of introns is accomplished after transcription at the mRNA stage within the nucleus. When a gene is expressed, the DNA is first transcribed to give a long RNA molecule, known as the **primary transcript**, that includes the introns. The primary transcript is then processed to remove the introns, yielding the mRNA (Fig. 4.03). Because this chapter is concerned with genome structure, the details of intron extraction will be deferred until Chapter 12.

Coding DNA May Be Present within Non-coding DNA

Introns are not totally absent from prokaryotes, but they are extremely rare. Moreover, there is usually only a single intron in a gene, unlike in eukaryotes where many

primary transcript The original RNA molecule obtained by transcription from a DNA template, before any processing or modification has occurred

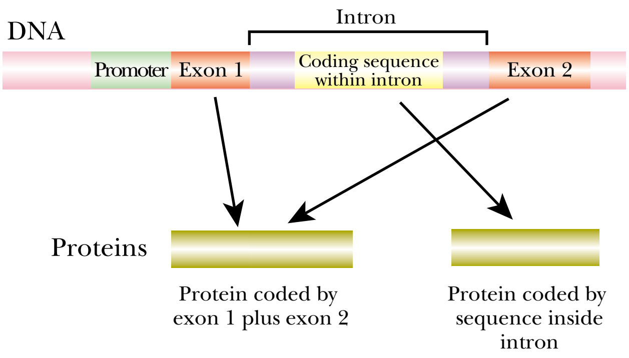

DNA
Intron
Promoter
Exon 1
Coding sequence within intron
Exon 2
Proteins
Protein coded by exon 1 plus exon 2
Protein coded by sequence inside intron

FIGURE 4.04 ***Intron Containing a Coding Sequence of Its Own***

An interesting situation can arise when a coding sequence has an intron within it that itself includes a coding sequence for another protein.

genes have multiple introns. Furthermore, most known examples are within the genes of bacterial viruses, rather than the chromosomal genes of bacteria themselves. For example, bacteriophage T4 possesses several introns, including one each in the genes encoding thymidylate synthase and ribonucleotide reductase. The T4 introns are homologous to the self-splicing introns of lower eukaryotes (for splicing mechanisms, see Ch. 12). This family of introns takes the complexity one step further, as there is a coding sequence for a separate protein located entirely within the intron (Fig. 4.04). This protein is concerned with survival of the intron.

In those rare cases where chromosomal genes of prokaryotes are interrupted, the genes often encode RNA molecules rather than proteins. Both tRNA and rRNA genes have been found with introns in both the eubacteria and archaebacteria. For example, one of the leucine tRNA genes of cyanobacteria (blue-green photosynthetic bacteria) and the corresponding gene in the DNA of chloroplasts contain self-splicing introns inserted at equivalent positions.

Repeated Sequences Are a Feature of DNA in Higher Organisms

Repeated sequences are frequently found in the DNA of higher organisms.

Most genes are present only as single copies. Such unique sequences account for almost all bacterial DNA. However, in higher organisms, unique sequences may comprise as little as 20 percent of the total DNA. For example, humans have 65 percent unique DNA, whereas frogs have only 22 percent. The rest of the DNA is made up of **repeated sequences (or repetitive sequences)** of one kind or another. Repeated sequence are what their name suggests, DNA sequences that are repeated many times throughout the genome. In some cases the repeat sequences follow each other directly—tandem repeats (see below), whereas others are spread separately around the genome—interspersed sequences. Some repeated sequences are genuine genes, but the majority consist of non-coding DNA.

Individual members of a family of repeated sequences are rarely identical in every base. Nonetheless, one may imagine an ideal, so-called **consensus sequence**, from which they are all derived by only minor alterations (Fig. 4.05). Such a consensus sequence is deduced in practice by examining many related individual sequences and including those bases most often found at each position. In other words, consensus sequences are found by comparing many sequences and taking the average.

In addition to multiply repeated sequences, eukaryotic cells also possess **pseudogenes**. These are defective duplicate copies of genuine genes, whose defects prevent them from being expressed. Pseudogenes are present in only one or two copies and may be next to the original, functional version of the gene or may be far away, even on a different chromosome. In quantitative terms, pseudogenes account for only a

consensus sequence Idealized base sequence consisting of the bases most often found at each position.
pseudogene Defective copy of a genuine gene
repeated sequences DNA sequences that exist in multiple copies
repetitive sequences Same as repeated sequences

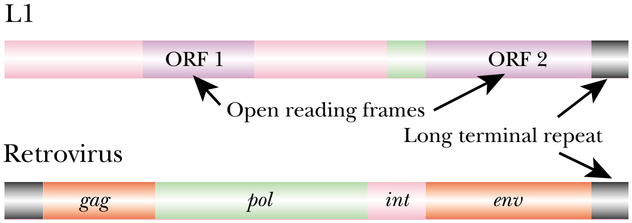
FIGURE 4.05 ***Deduction of Consensus Sequence***

The frequency of base appearances is used to derive a consensus sequence that is most representative of the series of related sequences shown.

Actual sequence observed:	A	t	C	C	G	T	A	T	G	T
(bases that differ from consensus are shown in lower case)	A	G	C	a	t	T	A	T	G	T
	A	G	g	C	G	T	t	T	G	T
	c	G	C	C	G	c	A	T	G	a
	A	a	t	C	G	T	A	T	c	T
	A	G	C	g	a	g	A	T	G	T
	A	G	C	C	G	T	A	T	G	T
	g	G	C	C	a	T	A	g	t	T
	A	G	a	C	G	c	A	a	G	T
	A	G	t	C	G	T	A	T	a	T
Number of times most common base appears at each position:	8	8	6	8	7	7	9	8	7	9
Derived *consensus sequence:*	A	G	C	C	G	T	A	T	G	T

FIGURE 4.06 ***Structure of the LINE-1 Element***

An example of a LINE-1 or L1 element is shown. L1 contains blocks of DNA that show homology with the *pol* and LTR sequences of retroviruses, as well as two coding sequences or open reading frames (ORF1 and ORF2) involved in its own replication.



Genes for ribosomal RNA are usually found in multiple copies. In higher organisms there may be thousands of copies.

About 7% of human DNA consists of repeats of the 300 bp Alu element.

tiny fraction of the DNA. However, they are believed to be of great importance in molecular evolution as the precursors to new genes (see Ch. 20). Sometimes both copies of a duplicated gene remain functional and repeated duplication may even give families of related genes. The multiple copies gradually diverge to a greater or lesser extent as they adapt to carry out similar but related roles. Thus the repeated sequences due to a gene family are closely related but not absolutely identical.

Since each prokaryotic cell contains 10,000 or more ribosomes, it is not surprising that their DNA usually contains half a dozen copies of the genes for rRNA and tRNA. In the much larger eukaryotic cell, there are hundreds or thousands of copies of the rRNA and tRNA genes. Sequences present in hundreds or thousands of copies are referred to as **moderately repetitive sequences**. About 25 percent of human DNA falls into this category. This includes multiple copies of highly used genes, like those for rRNA as well as nonfunctional stretches of DNA that are repeated many times.

Much of the moderately repetitive non-coding DNA is formed of **LINEs**, which are "**Long INterspersed Elements**." They are thought to be derived from retrovirus-like ancestors. In mammalian genomes, there are 20,000–50,000 copies of the LINE-1 (L1) family (Fig. 4.06). A complete L1 element is around 7,000 bp and contains two coding sequences. However, most individual L1 elements are shorter and many contain sequence rearrangements that disrupt the coding sequences, rendering them nonfunctional.

Another 10 percent of human DNA consists of sequences present in hundreds of thousands to millions of copies. Much of this **highly repetitive DNA** consists of **SINEs**, or **Short INterspersed Elements**. These sequences are almost all nonfunctional as far

highly repetitive DNA DNA sequences that exist in hundreds of thousands of copies
LINE Long interspersed element
long interspersed element (LINE) Long sequence found in multiple copies that makes up much of the moderately repetitive DNA of mammals
moderately repetitive sequence DNA sequences that exist in thousands of copies (but less than a hundred thousand)
short interspersed element (SINE) Short sequence found in multiple copies that makes up much of the highly or moderately repetitive DNA of mammals
SINE Short interspersed element

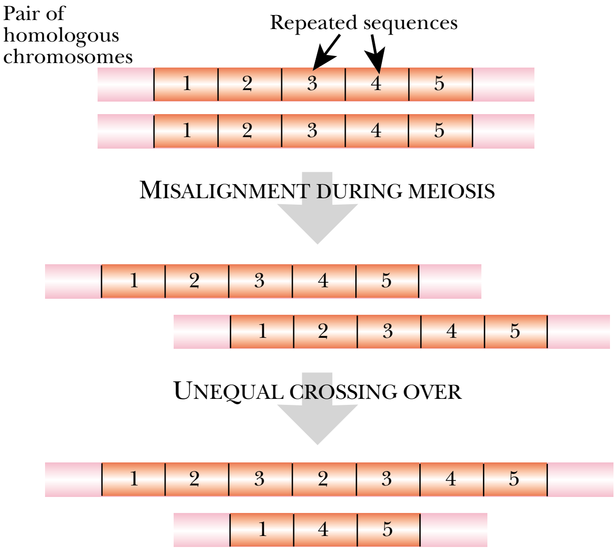


FIGURE 4.07 *Unequal Crossover due to Misalignment*

A pair of homologous chromosomes contains repeated elements. Since repeated elements may be readily misaligned during meiosis, crossing over will sometimes occur in regions that are not comparable in each chromosome. The result is one longer and one shorter DNA fragment.

as is known. The best known SINE is the 300 base pair **Alu element**. It is named after the restriction enzyme *Alu I*, which cuts it at a single site 170 bp from the front of the sequence. (See Ch. 22 for restriction enzymes.) From 300,000 to 500,000 copies (per haploid genome) of the Alu element are scattered throughout human DNA. Though apparently useless, they make up 6 to 8 percent of a human's genetic information. They occur singly or in small clusters and the majority are mutated or incomplete.

Most mammals contain SINEs topographically related to the Alu element. However, the original sequence, as found in mice, hamsters, etc., is only 130 bp long. For example, the mouse contains about 50,000 copies of the Alu-related **B1 element**. The human Alu element possesses two tandem repeats of this ancestral 130 bp B1 sequence plus an extra, unrelated 31 bp insertion of obscure origin. The mouse has less than 100,000 copies of the B1 element, so it would be classified as moderately repetitive DNA, whereas humans have more than 100,000 copies of the related Alu element, which is therefore classified as highly repetitive DNA. Clearly, the division into "moderately" repetitive and "highly" repetitive DNA is somewhat arbitrary.

Satellite DNA Is Non-coding DNA in the Form of Tandem Repeats

Tandem repeats cluster together forming regions of inert satellite DNA.

Unlike the LINEs and SINEs, which by definition are scattered throughout the genome, a significant amount of highly repetitive DNA in eukaryotic cells is found as long clusters of **tandem repeats**. This is also known as **satellite DNA**. Tandem means that the repeated sequences are next to each other in the DNA without gaps between. The amount of satellite DNA is highly variable. In mammals such as the mouse, satellite DNA accounts for about 8 percent of the DNA, whereas in the fruit fly, *Drosophila*, it comprises nearly 50 percent.

Long series of tandem repeats tend to misalign when pairs of chromosomes line up for recombination during meiosis. **Unequal crossing over** will then produce one shorter and one longer segment of repetitive DNA (Fig. 4.07). Thus, the exact number of tandem repeats varies from individual to individual within the same population.

Alu element An example of a SINE, a particular short DNA sequence found in many copies on the chromosomes of humans and other primates
B1 element An example of a SINE found in mice; the precursor sequence from which the human Alu element evolved
satellite DNA Highly repetitive DNA of eukaryotic cells that is found as long clusters of tandem repeats and is permanently coiled tightly into heterochromatin
tandem repeats Repeated sequences of DNA (or RNA) that lie next to each other
unequal crossing over Crossing over in which the two segments that cross over are of different lengths; often due to misalignment during pairing of DNA strands

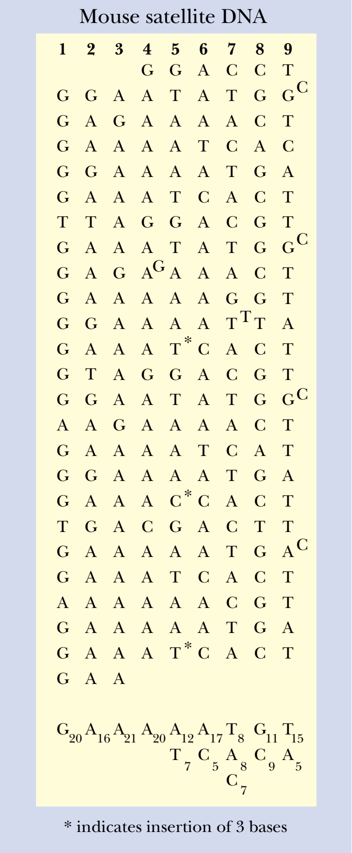


FIGURE 4.08 ***Repeating Motifs in Mouse Satellite DNA***

Variations in the consensus 9 bp satellite DNA sequence GAAAAATGT are shown.

Person to person variation in the overall length of short tandem repeats allows individual identification and is used in forensic analysis.

TABLE 4.02 Distribution of a 64 bp Human VNTR

% of population	# of repeats
7	18
11	16
43	14
36	13
4	10

In insects, the repeating sequences of satellite DNA are very short and consist of only one or very few different sequences. Thus in *Drosophila virilis* a 7 bp repeat with a consensus sequence of ACAAACT accounts for almost all of the satellite DNA. About half of the repeats have the consensus itself and the rest differ by one or, rarely, two bases. Satellite sequences vary considerably from one organism to another. The more commonly used *Drosophila melanogaster* has more complex satellite DNA that includes the 7 bp sequence just described as well as other 5, 10 and 12 bp repeats. In mammals, the satellite sequences are relatively complex. Although there is an overall 9 bp consensus in the mouse, there is much more variation among the repeats (Fig. 4.08).

Satellite DNA is inert and is permanently coiled tightly into what is known as **heterochromatin**. The heterochromatin is located around the **centromeres** of the chromosomes, suggesting that it serves some structural role. Note, however, that these satellite DNA sequences are quite distinct from the **centromere sequences**, which are needed for attachment of the spindle fibers during cell division.

Minisatellites and VNTRs

Segments of DNA consisting of short tandem repeats, but in much fewer copies than satellites, are known as **mini-satellites** or **VNTRs** (**Variable Number of Tandem Repeats**). Typically there may be from five to 50 tandem repeats in a VNTR. In mammals, VNTRs are common and are scattered over the genome, although they tend to be found close to the telomeres.

Due to unequal crossing over, the number of repeats in a given VNTR varies among individuals. Although VNTRs are non-coding DNA and not true genes, nonetheless the different versions are referred to as alleles. For example, Table 4.02 shows the distribution of one human VNTR of 64 bp among the population.

Some hyper-variable VNTRs may have as many as 1,000 different alleles and give unique patterns for almost every individual. This quantitative variation may be used for the identification of individuals by **DNA fingerprinting**.

Origin of Selfish DNA and Junk DNA

It is thought that most of the repetitive and non-coding DNA found in the chromosomes of eukaryotes is useless to the organism concerned. Such useless DNA is sometimes referred to as **junk DNA**. Where did it come from? It is thought that many

centromere Structure found on a chromosome and used to build and organize microtubules during mitosis
centromere sequence (CEN) A recognition sequence found at the centromere and needed for attachment of the spindle fibers
DNA fingerprint Individually unique pattern due to multiple bands of DNA produced using restriction enzymes, separated by electrophoresis and usually visualized by Southern blotting
heterochromatin Highly condensed form of chromatin that is genetically inert
junk DNA Defective selfish DNA that is of no use to the host cell it inhabits and which can no longer move or express its genes
mini-satellite Another term for a VNTR (variable number tandem repeats)
VNTR See variable number tandem repeats
variable number tandem repeats (VNTR) Cluster of tandemly repeated sequences in the DNA, whose number of repeats differs from one individual to another

DNA that consists of very large numbers of tandem repeats may well have a base composition different from that of the genome as a whole. If so, the satellite DNA will have a different buoyant density from the rest of the DNA, as this property depends on the base composition. DNA may be fractionated according to density by ultracentrifugation in a gradient of the heavy metal salt, cesium chloride (CsCl). Each fraction of DNA forms a band at the position corresponding to its own density. If the %GC varies by 5 percent or more, separate bands are obtained. When mouse DNA is run on a CsCl density gradient, two DNA bands are seen (Fig. 4.09). One contains 92 percent of the DNA with a density of 1.701 gm/cm^3 and the smaller, satellite band contains 8 percent of the DNA with a density of 1.690 gm/cm^3. Satellite DNA was originally defined by this density separation. However, in cases where the average satellite DNA base composition is close to that of the genome as a whole, the satellite DNA cannot be physically separated using a density gradient.

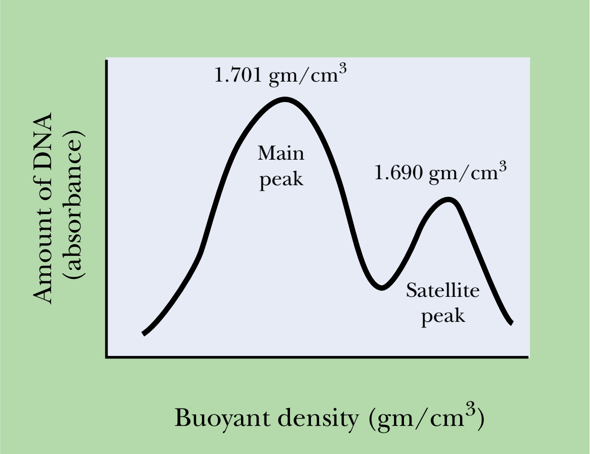


FIGURE 4.09 ***Density Gradient Centrifugation and Satellite Bands***

A cesium chloride gradient will reveal two (or more) bands of fragmented DNA if these differ in density. In this case, the lighter DNA contains sequences that are primarily satellite DNA.

repetitive sequences and other non-coding DNA, including even some introns, may have originated from viral DNA that was inserted into the chromosome of the eukaryotic host cell. Retroviruses, in particular, have played a large role in generating such insertions. In addition, transposable elements (mobile DNA; see Ch. 15) are probably responsible for generating a significant fraction of the repetitive DNA.

Much of our genome consists of the defunct remains of viruses and transposable elements.

The development of repetitive DNA would involve two processes. The original sequences of either viral DNA or transposable elements must have been intact and functional to insert in the first place. Both types of element might replicate and duplicated copies would then be inserted at more locations in the chromosomes of the infected cell. Thus, the numbers of these parasitic sequences would increase. In addition, many of these sequences would mutate, both by base changes and deletions. The result would be a family of related sequences, most of which are no longer functional (Fig. 4.10). Transposable elements that are solely concerned with their own survival and replication, rather than benefiting the host cell in any way are referred to as **selfish DNA**.

Selfish DNA, of no benefit to the host cell, accumulates in the genomes of slow-growing, multi-cellular organisms.

selfish DNA A sequence of DNA that manages to replicate but which is of no use to the host cell it inhabits

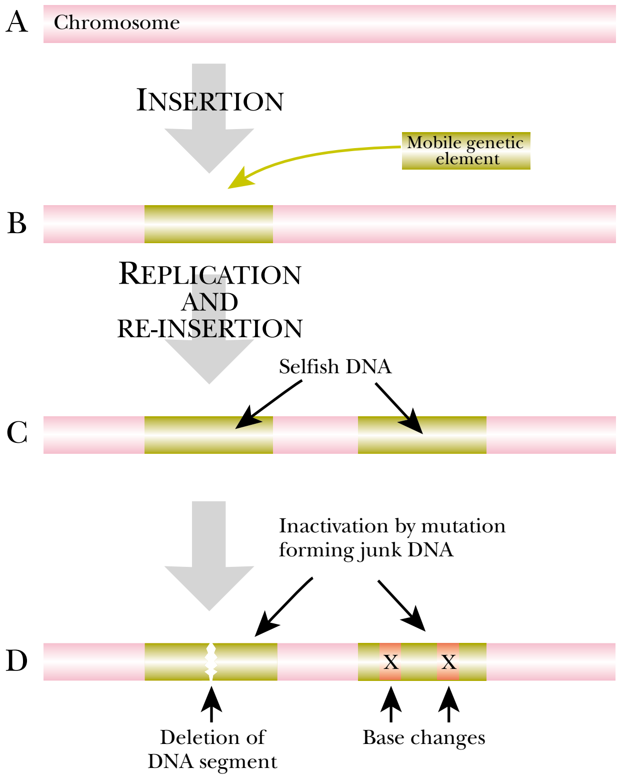


FIGURE 4.10 ***Origins of Selfish DNA and Junk DNA***

1. Insertion of originally mobile DNA; 2. Replication of the inserted DNA; and 3. Deletion and mutation of the inserted DNA appear to be the likely steps in affecting how much junk DNA resides in the genome.

Selfish DNA proliferates through the genome and may be regarded as a subcellular parasite infecting the host chromosomes. The accumulation of selfish DNA depends on two opposing processes, the replication and re-insertion of selfish DNA and its spontaneous deletion. In rapidly dividing, single-celled organisms, such as bacteria, selfish DNA tends to be eliminated, whereas in slowly dividing, multi-cellular organisms it has more opportunity to accumulate. Although originally useless, some defunct selfish DNA sequences may have been put to use by the host cell in non-coding roles such as helping to maintain chromosome structure.

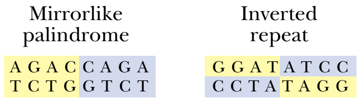


FIGURE 4.11 ***Palindromes and Inverted Repeats***

A mirror-like palindrome and an inverted repeat are shown. Similar colors indicate palindromic or inverted sequences.

Palindromes, Inverted Repeats and Stem and Loop Structures

Palindromes are words or phrases that read the same backwards as forwards. In the case of DNA, which is double stranded, two types of palindromes are theoretically possible. **Mirror-like palindromes** are like those of ordinary text, but involve two strands for DNA. However, in practice, the **inverted repeat** type of palindrome is much more common and of major biological significance. In an inverted repeat, the sequence reads the same forwards on one strand as it reads backwards on the complementary strand (Fig. 4.11).

Inverted repeats are extremely important as recognition sites on the DNA for the binding of a variety of proteins. Many regulatory proteins recognize inverted repeats, as do most restriction and modification enzymes (see Ch 22). In such cases, the inverted repeat usually remains as normal double helical DNA and does not need to be distorted by supercoiling. The term "direct repeat" refers to the situation where the repeated sequences point in the same direction and are on the same strand.

Regulatory proteins often bind to DNA at inverted repeat sequences.

inverted repeat Sequence of DNA that is the same when read forwards as when read backwards, but on the other complementary strand. One type of palindrome
mirrorlike palindrome Sequence of DNA that is the same when read forwards and backwards on the same strand. One type of palindrome
palindrome A sequence that reads the same backwards as forwards

TABLE 4.03 Components of the Eukaryotic Genome

(Numbers of copies given is for the human genome.)

Unique sequences	
Protein encoding genes—comprising upstream regulatory region, exons and introns	
Genes encoding non-translated RNA (snRNA, snoRNA, 7SL RNA, telomerase RNA, Xist RNA, a variety of small regulatory RNAs)	
Non-repetitive intragenic non-coding DNA	
Interspersed Repetitive DNA	
Pseudogenes	
Short Interspersed Elements (SINEs)	
Alu element (300 bp)	~1,000,000 copies
MIR families (average ~130 bp) (mammalian-wide interspersed repeat)	~400,000 copies
Long Interspersed Elements (LINEs)	
LINE-1 family (average ~800 bp)	~200,000–500,000 copies
LINE-2 family (average ~250 bp)	~270,000 copies
Retrovirus like elements (500–1300 bp)	~250,000 copies
DNA transposons (variable; average ~250 bp)	~200,000 copies
Tandem Repetitive DNA	
Ribosomal RNA genes	5 clusters of about 50 tandem repeats on 5 different chromosomes
Transfer RNA genes	multiple copies plus several pseudogenes
Telomere sequences	several kb of a 6 bp tandem repeat
Mini-satellites (= VNTRs)	blocks of 0.1 to 20 kbp of short tandem repeats (5–50 bp), most located close to telomeres
Centromere sequence (α-satellite DNA)	171 bp repeat, binds centromere proteins
Satellite DNA	blocks of 100 kbp or longer of tandem repeats of 20 to 200 bp, most located close to centromeres
Mega-satellite DNA	blocks of 100 kbp or longer of tandem repeats of 1 to 5 kbp, various locations

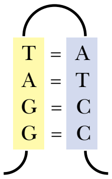


FIGURE 4.12 ***A Hairpin***

If a single strand of DNA containing inverted repeats is folded back upon itself, base pairing occurs forming a hairpin structure.

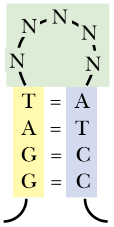


FIGURE 4.13 ***Stem and Loop Motif***

If inverted repeats are separated by a few bases, a stem and loop structure results. The loop contains unpaired bases (NNN).

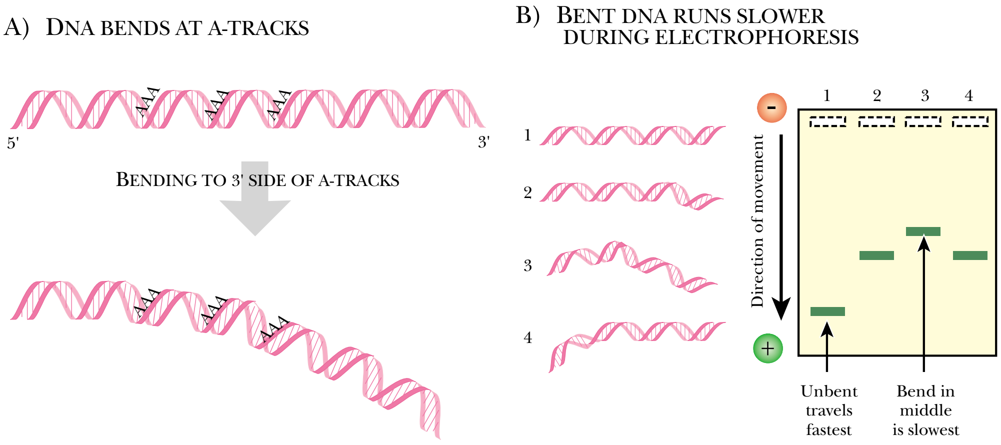
Consider just a single strand of the inverted repeat sequence. Note that the right and left halves of the sequence on each single strand must be complementary to each other. Thus, such a sequence (e.g., GGATATCC) can be folded into a **hairpin** whose two halves are held together by base pairing (Fig. 4.12).

The U-turn at the top of the hairpin is possible, but energetically unfavorable. In practice, normally a few unpaired bases (shown as N = any base, in the diagram below) are found forming a loop at the top of the base paired stem—the so-called **stem and loop** motif. Such a stem and loop can form from one strand of any inverted repeat that has a few extra bases in the middle (Fig. 4.13).

Multiple A-Tracts Cause DNA to Bend

A DNA sequence that contains several runs of A residues (three to five nucleotides long) separated by 10 bp forms bends. Note that the spacing of the A-tracts corresponds to one turn of the double helix. Bending occurs at the 3′-end of the runs of As

hairpin A double stranded base-paired structure formed by folding a single strand of DNA or RNA back upon itself
stem and loop Structure made by folding an inverted repeat sequence



A) DNA BENDS AT A-TRACKS

B) BENT DNA RUNS SLOWER DURING ELECTROPHORESIS

FIGURE 4.14 ***DNA Bending Due to Multiple A Tracts***

(A) Bending of DNA occurs to the 3′-side of A-tracts. (B) Such bending decreases the speed at which DNA travels during electrophoresis. Indeed, the mobility of a DNA molecule of a given length varies depending on the location of bent regions within the molecule. Bends in the middle have greater effect than those close to the ends.

(Fig. 4.14). **Bent DNA** moves more slowly during gel electrophoresis than unbent DNA of the same length (see Ch 21 for electrophoresis).

Bent DNA is found at the origins of replication of some viruses and of yeast chromosomes. It is thought to help the binding of the proteins that initiate DNA replication (see Ch. 5). In addition to "naturally" bent DNA, certain regulatory proteins also bend DNA into U-turns when activating transcription (see Chapters 9 and 10).

Supercoiling is Necessary for Packaging of Bacterial DNA

Bacterial DNA is 1000 times longer than the cell that contains it. The DNA must be supercoiled in order to fit into the cell.

An average bacterial cell is about one millionth of a meter long. The length of the single DNA molecule needed to carry the 4,000 or so genes of a bacterial cell is about one millimeter! Thus, a stretched out bacterial chromosome is a thousand times longer than a bacterial cell. The double helical DNA inside a cell must be **supercoiled** to make it more compact. The DNA, which is already a double helix, is twisted again, as shown in Figure 4.15. The original double helix has a right-handed twist but the supercoils twist in the opposite sense; that is, they are left-handed or **"negative" supercoils**. There is roughly one supercoil every 200 nucleotides in typical bacterial DNA. Negative (rather than positive) supercoiling helps promote the unwinding and strand separation necessary during replication and transcription. [Eukaryotic DNA is also negatively supercoiled, however the mechanism is rather different and involves coiling it around histone proteins as discussed below.]

DNA gyrase puts negative supercoils into the bacterial chromosome.

Negative supercoils are introduced into the bacterial chromosome by DNA gyrase. In the absence of topoisomerase I and topoisomerase IV, the DNA becomes hypernegatively supercoiled. The steady-state level of supercoiling in *Escherichia coli* is maintained by a balance between topoisomerase IV, acting in concert with topoisomerase I, to remove excess negative supercoils and thus acting in opposition to DNA gyrase. A typical bacterial chromosome contains approximately 50 giant loops of supercoiled DNA arranged around a protein scaffold. In Figure 4.16, the single line represents a double helix of DNA and the helixes are the supercoils.

bent DNA Double helical DNA that is bent due to several runs of As
negative supercoiling Supercoiling with a left handed or counterclockwise twist
supercoiling Higher level coiling of DNA that is already a double helix

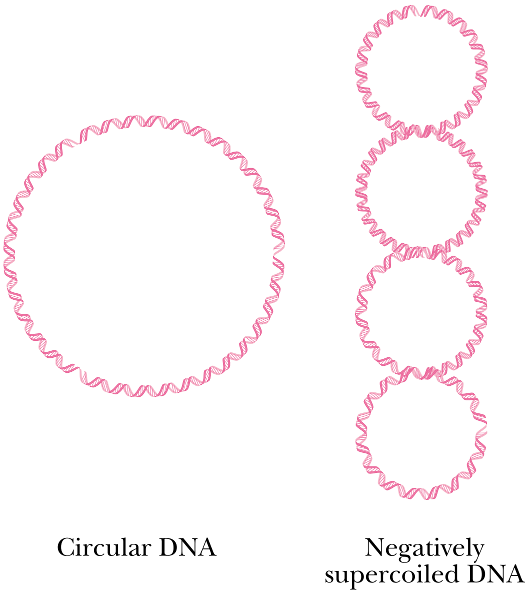


FIGURE 4.15 ***Supercoiling of DNA***

Bacterial DNA is negatively supercoiled in addition to the twisting imposed by the double helix.

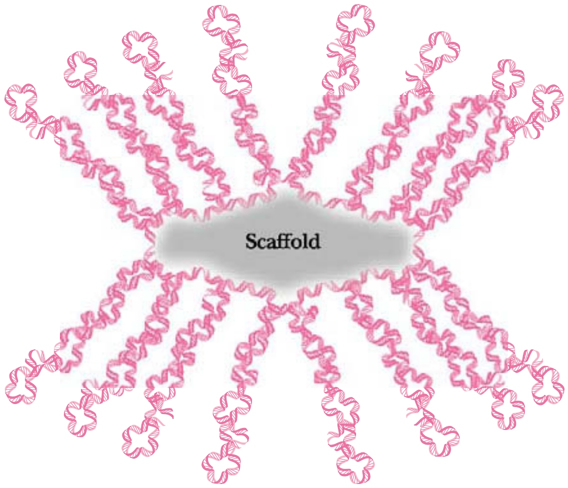


FIGURE 4.16 ***Supercoiling of the Bacterial Chromosome***

Supercoiling of bacterial DNA results in giant loops of supercoiled DNA extending from a central scaffold.

The bacterial chromosome consists of about 50 giant supercoiled loops of DNA.

Bacterial chromosomes and plasmids are double stranded circular DNA molecules and are often referred to as **covalently closed circular DNA**, or **cccDNA**. If one strand of a double stranded circle is nicked, the supercoiling can unravel. Such a molecule is known as an **open circle**.

Topoisomerases and DNA Gyrase

The total amount of twisting present in a DNA molecule is referred to as the **linking number (L)**. This is the sum of the contributions due to the double helix plus the supercoiling. [The number of double helical turns is sometimes known as the **twist**, **T**, and the number of superhelical turns as the **writhe** or **writhing number**, **W**. In this terminology, the linking number, L, is the sum of the twist plus the writhe (L = T + W).]

covalently closed circular DNA (cccDNA) Circular DNA with no nicks in either strand
linking number (L) The sum of the superhelical turns (the writhe, W) plus the double helical turns (the twist, T)
open circle Circular DNA with one strand nicked and hence with no supercoiling
twist, T The number of double helical turns in a molecule of DNA (or double-stranded RNA)
writhe Same as writhing number, W
writhing number, W The number of supercoils in a molecule of DNA (or double-stranded RNA)

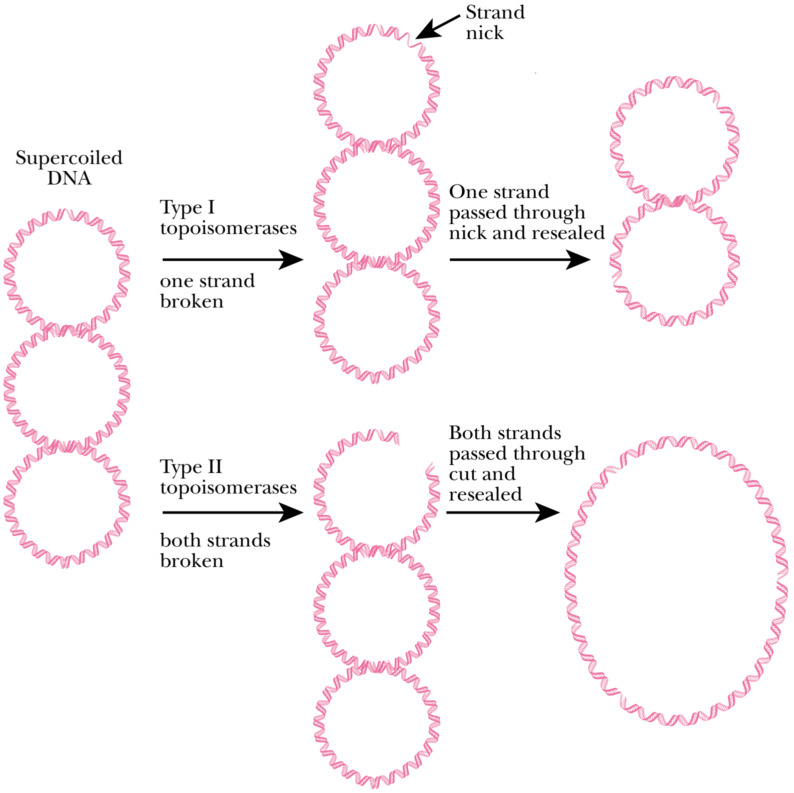


FIGURE 4.17 ***Mechanism of Type I and II Topoisomerases***

The difference in action between topoisomerases of Type I and Type II is in the breakage of strands. Type I breaks only one strand, while Type II breaks both strands. When one strand is broken, the other strand is passed through the break to undo one supercoil. When two strands are broken, double stranded DNA is passed through the break and the supercoiling is reduced by two. After uncoiling, the breaks are rejoined.

Enzymes known as topoisomerases change the level of supercoiling.

The same circular DNA molecule can have different numbers of supercoils. These forms are known as topological isomers, or **topoisomers**. The enzymes that insert or remove supercoils are therefore named **topoisomerases**. **Type I topoisomerases** break only one strand of DNA, which changes the linking number in steps of one. In contrast, **type II topoisomerases** (including **DNA gyrases**) break both strands of the DNA and pass another part of the double helix through the gap. This changes the linking number in steps of two (Fig. 4.17).

DNA gyrase, a type II topoisomerase, introduces negative supercoils into closed circular molecules of DNA, such as plasmids or the bacterial chromosome. Gyrase works by cutting both strands of the DNA, introducing a supertwist and rejoining the DNA strands. Gyrase can generate 1,000 supercoils per minute. As each supertwist is introduced, gyrase changes conformation to an inactive form. Reactivation requires energy, provided by breakdown of ATP. DNA gyrase can also remove negative supercoils (but not positive ones) without using ATP, but this occurs ten times more slowly.

Ciprofloxacin kills bacteria by inhibiting DNA gyrase. It is harmless to animals as they do not use DNA gyrase for compacting their DNA.

DNA gyrase is a tetramer of two different subunits. The GyrA subunit cuts and rejoins the DNA and the GyrB subunit is responsible for providing energy by ATP hydrolysis. DNA gyrase is inhibited by **quinolone antibiotics**, such as **nalidixic acid** and their fluorinated derivatives such as **norfloxacin** and **ciprofloxacin**, which bind to the GyrA protein. An inactive complex is formed in which GyrA protein is inserted into the DNA double helix and covalently attached to the 5′-ends of both broken DNA strands. **Novobiocin** also inhibits gyrase by binding to the GyrB protein and preventing it from binding ATP.

ciprofloxacin A fluoroquinolone antibiotic that inhibits DNA gyrase
DNA gyrase An enzyme that introduces negative supercoils into DNA, a member of the type II topoisomerase family
nalidixic acid A quinolone antibiotic that inhibits DNA gyrase
norfloxacin A fluoroquinolone antibiotic that inhibits DNA gyrase
novobiocin An antibiotic that inhibits type II topoisomerases, especially DNA gyrase, by binding to the B-subunit
quinolone antibiotics A family of antibiotics, including nalidixic acid, norfloxacin and ciprofloxacin, that inhibit DNA gyrase and other type II topoisomerases by binding to the A-subunit
topoisomerase Enzyme that alters the level of supercoiling or catenation of DNA (i.e. changes the topological conformation)
topoisomers Isomeric forms that differ in topology—i.e. their level of supercoiling or catenation
type I topoisomerase Topoisomerase that cuts a single strand of DNA and therefore changes the linking number by one
type II topoisomerase Topoisomerase that cuts both strands of DNA and therefore changes the linking number by two

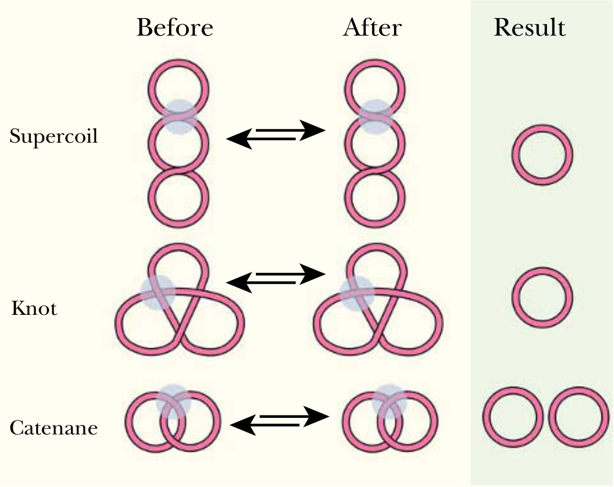


FIGURE 4.18 ***Unlinking of Catenanes by Topoisomerase IV***

Topoisomerases may uncoil, unknot or unlink DNA as well as carrying out the coiling, knotting or interlinking of DNA. Topoisomerases act (at the locations shaded blue) by cutting both strands of the DNA at one location and passing another region of the DNA through the gap.

Catenated and Knotted DNA Must Be Corrected

Circular molecules of DNA may become interlocked during replication or recombination. Such structures are called **catenanes**. The circles may be liberated by certain type II topoisomerases, such as topoisomerase IV (Fig. 4.18) of *E. coli* and related enzymes. Circular DNA molecules may also form knots. Type II topoisomerases can both create and untie knots. Like DNA gyrases, these enzymes are tetramers of two different subunits, one for cutting the DNA and the other for energy coupling. Like gyrase, topoisomerase IV is inhibited by quinolone antibiotics.

Local Supercoiling

When DNA is replicated or when genes are expressed, the double helix must first be unwound. This is aided by the negative supercoiling of the chromosome. However, as the replication apparatus proceeds along a double helix of DNA, it creates positive supercoiling ahead of itself. Similarly, during transcription, when RNA polymerase proceeds along a DNA molecule, it also creates positive supercoiling ahead of itself. For replication and transcription to proceed more than a short distance, DNA gyrase must insert negative supercoils to cancel out the positive ones. Behind the moving replication and transcription apparatus, a corresponding wave of negative supercoiling is generated. Excess negative supercoils are removed by topoisomerase I.

As a result, at any given instant, the extent of supercoiling varies greatly in any particular region of the chromosome. It has been suggested that supercoiling might regulate gene expression. However, only rare examples are known; thus transcription of the gene for DNA gyrase in *E. coli* is regulated by supercoiling. More often, the opposite is the case. Local supercoiling depends largely on the balance between transcription and the restoration of normal supercoiling by gyrase and topoisomerase.

Supercoiling Affects DNA Structure

Supercoiling places DNA under physical strain. This may lead to the appearance of alterations in DNA structure that serve to relieve the strain. Three of these alternate forms are **cruciform structures**; the left-handed double helix, or **Z-DNA**; and the triple helix, or **H-DNA**. All of these structures depend on certain special characteristics within the DNA sequence, as well as supercoiling stress.

catenane Structure in which two or more circles of DNA are interlocked
cruciform structure Cross shaped structure in double stranded DNA (or RNA) formed from an inverted repeat
H-DNA A form of DNA consisting of a triple helix. Its formation is promoted by acid conditions and by runs of purine bases
Z-DNA An alternative form of DNA double helix with left-handed turns and 12 base pairs per turn

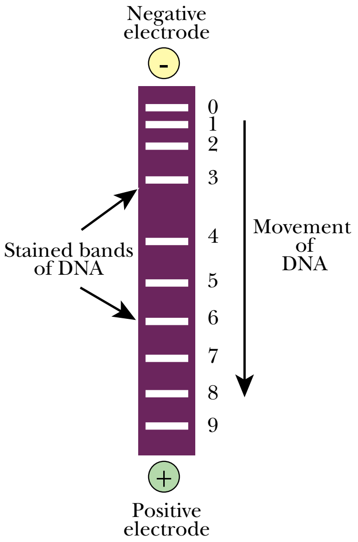


FIGURE 4.19 ***Separation of Supercoiled DNA by Electrophoresis***

Supercoiled DNA molecules, all of identical sequence, were electophoresed to reveal multiple bands, with each band differing in the number of supercoils. The number of supercoils is shown beside the band. Zero refers to open circular DNA, which is not supercoiled at all.

Separation of Topoisomers by Electrophoresis

DNA molecules of different sizes are routinely separated by electrophoresis on agarose gels (see Ch. 21). The mobility of a DNA molecule in such a gel depends on both its molecular weight and its conformation. Heavier molecules travel more slowly but among molecules of the same molecular weight, those that are more compact move faster. Consequently, for a given DNA molecule, supercoiled cccDNA moves faster than open circular DNA, which in turn moves faster than linearized DNA.

For small circular molecules of DNA, it is even possible to separate topoisomers with different numbers of superhelical twists. The more superhelical twists, the more compact the molecule is and the faster it moves in an electrophoretic field (Fig 4.19).

When the DNA gyrase of an *E. coli* strain containing small plasmids is inhibited, it is possible to isolate the plasmids and demonstrate that they are now positively supercoiled. Topoisomerase I continues to remove the excess negative supercoils but the DNA gyrase no longer cancels out the surplus positive supercoils, which therefore accumulate.

The cruciform (cross-like) structures are formed when the strands in a double stranded DNA palindrome are separated and formed into two stem and loop structures opposite each other (Fig. 4.20). The probability that cruciform structures will form increases with the level of negative supercoiling and the length of the inverted repeat. In practice, the four to eight base sequences recognized by most regulatory proteins and restriction enzymes are too short to yield stable cruciform structures. Palindromes of 15 to 20 base pairs will produce cruciform structures. Their existence can be demonstrated because they allow single strand specific nucleases to cut the double helix. (A nuclease is an enzyme that cuts nucleic acid strands; see Ch. 22) Cutting occurs within the small single stranded loop at the top of each hairpin. Cruciform structures partially straighten supercoiled DNA and so the molecule is not folded up as compactly and thus travels slower during gel electrophoresis.

Alternative helical structures are sometimes found for DNA, in addition to the common form, made famous by Watson and Crick and known as B-DNA.

Alternative Helical Structures of DNA Occur

Several double helical structures are actually possible for DNA. Watson and Crick described the most stable and most common of these. It is right handed with 10 base pairs per turn of the helix. The grooves running down the helix are different in depth and referred to as the major and minor grooves. The standard Watson and Crick double helix is referred to as the **B-form** or as **B-DNA** to distinguish it from the other helical forms, **A-DNA**, and Z-DNA (Fig. 4.21). Most of these structures apply not only to double stranded DNA (dsDNA), but also to RNA when it is double stranded.

The A-form helix is found in dsRNA or DNA/RNA hybrids. It has 11 bp per turn—one more than B-DNA.

The **A-form** of the double helix is shorter and fatter than the B-form and has 11 base pairs per helical turn. In the A-form, the bases tilt away from the axis, the minor groove becomes broader and shallower, and the major groove becomes narrower and deeper. Double-stranded RNA or hybrids with one RNA and one DNA strand usually form an A-helix. The extra hydroxyl group, at the 2′ position of ribose, prevents double-stranded RNA from forming a B-helix. Double-stranded DNA tends to form an A-helix only at a high salt concentration or when it is dehydrated. The tendency to form an A-helix also depends on the sequence. The physiological relevance, if any, of A-DNA is obscure. However, double-stranded regions of RNA exist in this form *in vivo*.

A-DNA A rare alternative form of double stranded helical DNA
A-form An alternative form of the double helix, with 11 base pairs per turn, often found for double stranded RNA, but rarely for DNA
B-form or B-DNA The normal form of the DNA double helix, as originally described by Watson and Crick

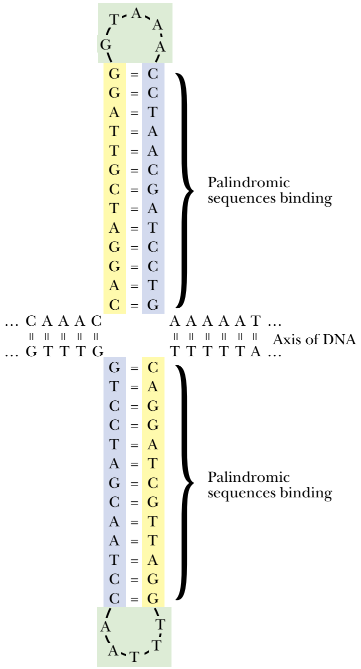


FIGURE 4.20 *Cruciform Structure Formed from an Inverted Repeat*

Because the DNA is palindromic, the strands can separate and base pair with themselves to form lateral cruciform extensions.

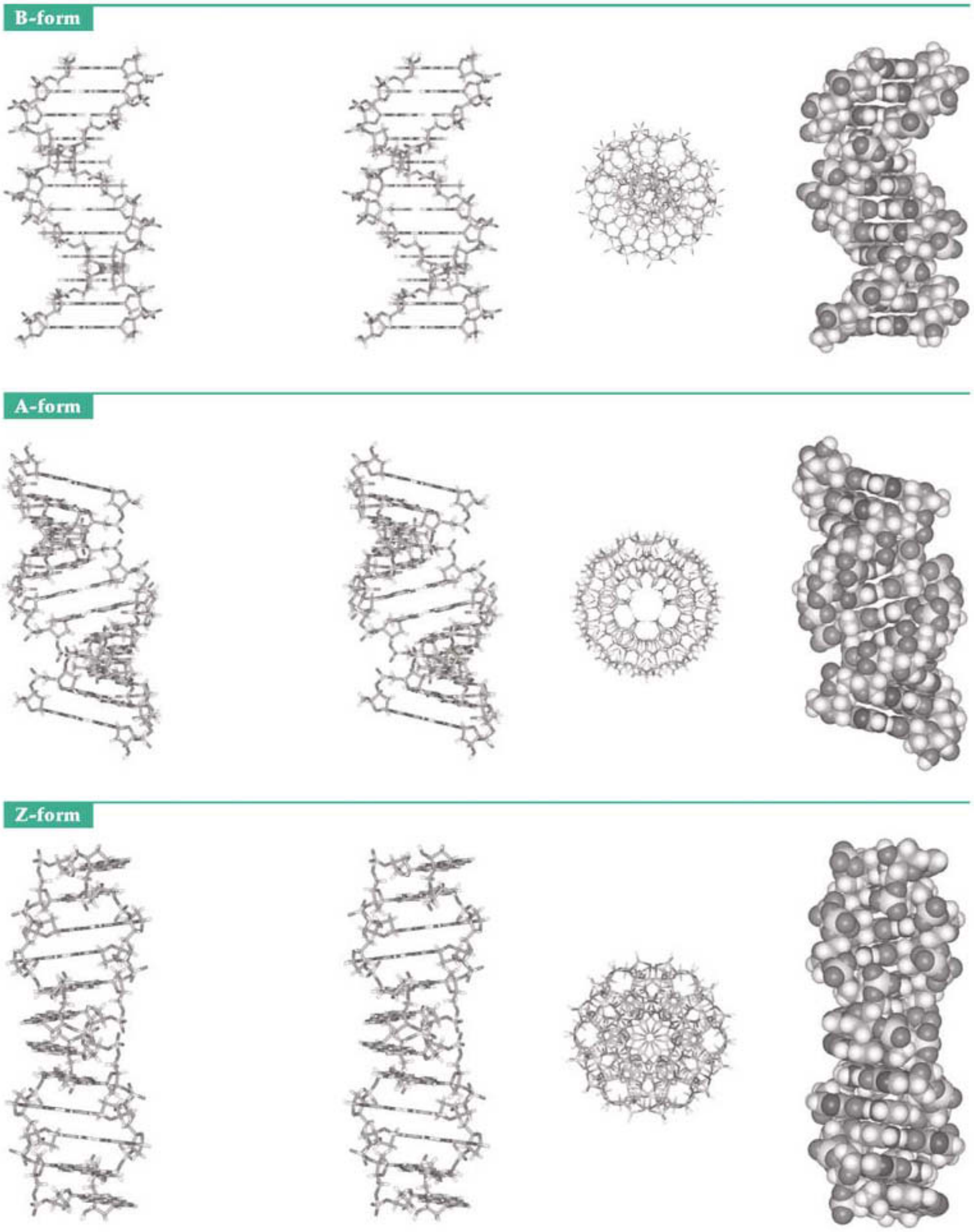


FIGURE 4.21 *Comparison of B-DNA, A-DNA and Z-DNA*

Several structurally different versions of the double helix exist. Shown here are the normal Watson-Crick double helix, the B-form, together with the rarer A-form and Z-DNA form. From left to right: stereoscopic skeleton pairs, end view down the helix axis, and space filling models. Courtesy of Tamar Schlick.

Z-DNA is a left-handed double helix with 12 base pairs per turn.

Z-DNA is a left-handed double helix with 12 base pairs per turn. It is thus longer and thinner than B-DNA and its sugar phosphate backbone forms a zigzag line rather than a smooth helical curve (Fig. 4.21). High salt favors Z-DNA as it decreases repulsion between the negatively charged phosphates of the DNA backbone. Z-DNA is formed in regions of DNA that contain large numbers of alternating GC or GT pairs, such as:

```
GCGCGCGCGCGCGC  or  GTGTGTGTGTGTGT
CGCGCGCGCGCGCG  or  CACACACACACACA
```

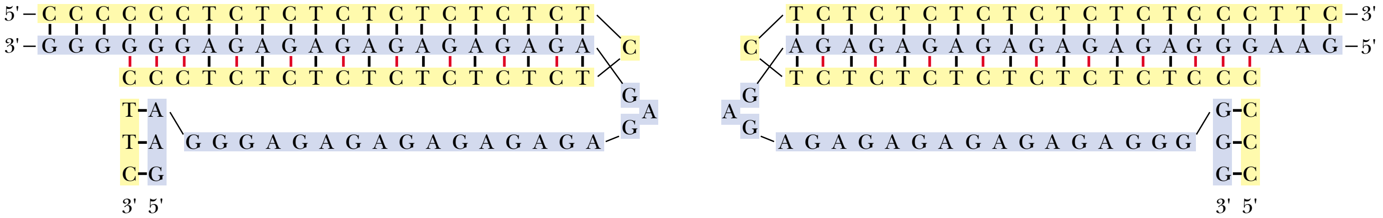

5'– C C C C C C T C T C T C T C T C T C T C T
3'– G G G G G G A G A G A G A G A G A G A G A
C C C T C T C T C T C T C T C T C T C T
C
G
A
G
T-A
T-A
C-G
3' 5'
G G G A G A G A G A G A G A G A G A
T C T C T C T C T C T C T C T C C C T T C –3'
C
A G A G A G A G A G A G A G A G G G A A G –5'
T C T C T C T C T C T C T C T C C C
G
A
G
A G A G A G A G A G A G A G G G
G-C
G-C
G-C
3' 5'

FIGURE 4.22 ***Structure of H-DNA***

This triple helix is formed by GA- and TC-rich regions of a plasmid and is composed of triads of bases.

Such tracts may be abbreviated as $(GC)_n{\cdot}(GC)_n$ and $(GT)_n{\cdot}(AC)_n$. Note that the sequence of each individual strand is written in the 5′- to 3′-direction. That Z-DNA depends specifically on G (not A) alternating with C or T is shown by the fact that DNA with many alternating AT pairs forms cruciform structures, not Z-DNA.

Since Z-DNA is a left-handed helix, its appearance in part of a DNA molecule helps to remove supercoiling stress. As negative supercoiling increases, the tendency for GC- or GT-rich tracts of DNA to take the **Z-form** increases. Its existence can be demonstrated in small plasmids by changes in electrophoretic mobility. Single-strand specific nucleases can cut DNA at the junction between segments of Z-DNA and normal B-DNA.

Short artificial segments of double stranded DNA made solely of repeating GC units take the Z-form even in the absence of supercoiling, provided that the salt concentration is high. (This was how Z-DNA was originally discovered.) Antibodies to Z-DNA can be made by immunizing animals with such linear (GC)n fragments. These antibodies can then be used to show the presence of Z-form (helix) regions in natural DNA, provided it is highly supercoiled.

It has been suggested that regions of Z-helix may be specifically recognized by certain enzymes. An example is the RNA editing enzyme ADAR1, which modifies bases in dsRNA (see Ch. 12 for the details of RNA processing). ADAR1 stands for adenosine deaminase (RNA) type I and it removes the amino group of adenosine, so converting it to **inosine**. It requires a double stranded RNA substrate that, in practice, is formed by folding an intron back onto the neighboring exon. ADAR1 contains separate binding motifs for both DNA and for dsRNA. It has been postulated that the DNA binding domain recognizes Z-DNA because base modification must occur before cutting and splicing of the introns and exons. When RNA polymerase moves along, transcribing DNA into RNA, it generates positive supercoils ahead and negative supercoils behind. Negative supercoiling induces the formation of Z-DNA, especially in GC- or GT-regions. Consequently, a zone of Z-DNA will be found just behind the RNA polymerase. Binding to Z-DNA would ensure that ADAR1 works on newly synthesized RNA.

H-DNA is even more peculiar—it is not a double but a *triple* helix. It depends on long tracts of purines in one strand and, consequently, only pyrimidines in the other strand; e.g.:

```
GGGGGGGGGGGGGG  or  GAGAGAGAGAGAGAG
CCCCCCCCCCCCCC  or  CTCTCTCTCTCTCTC
```

Two such segments are required and may interact forming H-DNA when the DNA is highly supercoiled (Fig. 4.22). In addition, the overall region must be a mirror-like palindrome. H-DNA contains a triple helix, consisting of one purine-rich strand and two pyrimidine-rich strands. The other purine-rich strand is displaced and left unpaired.

inosine A purine nucleoside, found most often in transfer RNA, that contains the unusual base hypoxanthine
Z-form An alternative form of double helix with left-handed turns and 12 base pairs per turn. Both DNA and dsRNA may be found in the Z-form

H-DNA involves strange "sideways" base pairing to form a triple helix.

In H-DNA, adenine pairs with two thymines and guanine pairs with two cytosines. In each case, one pairing is normal, the other sideways (so-called **Hoogsteen base pairs**—hence the name H-DNA). Furthermore, to form the C=G=C triangle, an extra proton (H^+) is needed for one of the hydrogen bonds. Consequently, formation of H-DNA is promoted by acidic conditions. High acidity also tends to protonate the phosphate groups of the DNA backbone, so decreasing their negative charges. This reduces the repulsion between the three strands and helps form a triple helix.

Despite these complex sequence requirements, computer searches of natural DNA have shown that potential sequences that might form triplex H-DNA are much more frequent than expected on a random basis. These are called **PIT (potential intra-strand triplex)** elements. For example, the *E. coli* genome contains 25 copies of a 37 base PIT element. Isolated PIT element DNA does form a stable triplex even at neutral pH. Not surprisingly, the presence of artificial triplexes has been shown to block transcription. This suggests that H-DNA does have some real biological function, although what this is remains obscure.

Histones Package DNA in Eukaryotes

Eukaryotes package their DNA by winding it around specialized proteins, known as histones.

Plants and animals have vastly more DNA than bacteria and must fold this DNA to fit into the cell nucleus. Typical bacteria carry approximately 4,000 genes on a single chromosome, which is about one millimeter long. The chromosome is thus 1,000 times longer than the bacterial cell in which it fits. Eukaryotic chromosomes may be as much as a centimeter long and must be folded up to fit into the cell nucleus, which is five microns across, a necessity for a 2,000-fold shortening. However, eukaryotic chromosomes are not circular, and instead of supercoiling using DNA gyrase, the mechanism of packaging involves winding the DNA around special proteins, the **histones**.

Eukaryotic DNA starts folding by coiling around the histones, positively charged proteins that neutralize the negative charge of the DNA itself. DNA with histones bound to it was named **chromatin** when it was first discovered in chromosomes. Chromatin consists of roughly spherical subunits, the **nucleosomes**, each containing approximately 200 bp of DNA and nine histones, two each of H2A, H2B, H3 and H4 and one of H1 (Fig. 4.23).

200 bp of DNA is wound around nine histones, making up a particle known as a nucleosome. Eukaryotic chromosomes consist of long strings of nucleosomes that are folded up further.

The eight paired histones cluster together and two coils of DNA are wrapped around them. Each coil of DNA is approximately 80 bp long so that this core particle accommodates 160 bp of DNA overall. The remaining 40 bp or so of DNA forms a linker region between neighboring core particles that may vary somewhat in length (from 10 to 100 bp) depending on the DNA sequence. The ninth histone, H1, joins each core particle to the next.

The DNA in the linker region is relatively exposed and can be cut with nucleases specific for dsDNA; these nucleases make double stranded cuts. In practice, micrococcal nuclease is often used to perform this job. Cutting occurs in three stages. First, single nucleosomes with 200 bp of DNA are released, then the linker region is cut off, leaving about 165 bp. Finally the ends of the DNA wound around the core are nibbled away, leaving about 146 bp that are fully protected by the core particle from further digestion.

The core histones, H2A, H2B, H3 and H4, are small, roughly spherical proteins with 102 to 135 amino acids. However, the linker histone, H1, is longer, having about 220 amino acids. H1 has two arms extending from its central spherical domain. The central part of H1 binds to its own nucleosome and the two arms bind to the nucleosomes on either side (Fig. 4.24).

chromatin Complex of DNA plus protein which constitutes eukaryotic chromosomes
histone Special positively charged protein that binds to DNA and helps to maintain the structure of chromosomes in eukaryotes
Hoogsteen base pair A type of nonstandard base pair found in triplex DNA, in which a pyrimidine is bound sideways on to a purine
nucleosome Subunit of a eukaryotic chromosome consisting of DNA coiled around histone proteins
potential intrastrand triplex (PIT) Stretch of DNA that might be expected from its sequence to form H-type triplex DNA

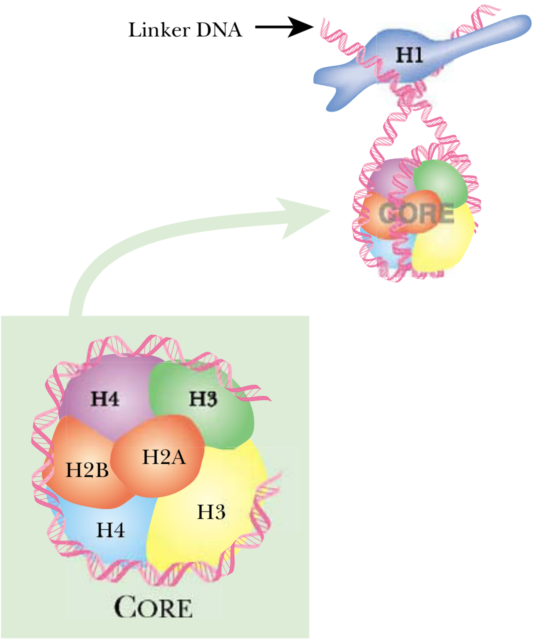


FIGURE 4.23 ***Nucleosomes and Histones***

The basic unit in the folding of eukaryotic DNA is the nucleosome as shown here. A nucleosome is composed of eight histones comprising a core and one separate histone (H1) at the site where the wrapped DNA diverges. The enlarged region shows the packing of histones in the core. The H3-H4 tetramer dictates the shape of the core. Only one of the H2A and H2B dimers is shown; the other is on the other side, hidden from view.

The core histones have a body of about 80 amino acids and a tail of 20 amino acids at the N-terminal end. This tail contains several lysine residues that may have acetyl groups added or removed (Fig 4.25). This is thought to partly control the state of DNA packaging and hence of gene expression. Thus, in active chromatin, the core histones are highly acetylated (see Ch. 10 for further discussion).

Further Levels of DNA Packaging in Eukaryotes

Eukaryotic DNA is so long that it needs several successive levels of folding to fit into the nucleus.

DNA covered with histones and twisted into a series of nucleosomes resembles a string of beads and is sometimes called the "beads on a string" form. However, the folding process continues. The chain of nucleosomes is wound into a giant helical structure with six nucleosomes per turn. It is now known as the **30 nanometer fiber** (Fig. 4.26). In turn, these fibers are looped back and forth. The loops vary in size, averaging about 50 of the helical turns (i.e., about 300 nucleosomes) per loop. The ends of the loops are attached to a protein scaffold, or chromosome axis.

Further folding of chromosomes occurs in preparation for cell division. The precise nature of this is uncertain, but condensed mitotic chromosomes are 50,000 times shorter than fully extended DNA. Highly condensed chromatin is known as **heterochromatin**, appearing dense in the light and electron microscope. In this form it cannot be transcribed (see Ch. 10 for discussion). [Note that some regions of DNA (e.g., satellite DNA near the centromeres) are always found as heterochromatin whereas active regions of the genome condense into heterochromatin during cell division.] An overall summary of DNA folding is presented in Figure 4.27.

Between cell divisions, regions of heterochromatin persist around the centromere and at the ends of the chromosome. These regions include the satellite DNA discussed above and make up about 10 percent of the chromosome. The rest of the chromatin, the **euchromatin**, is in the more extended form shown as a string of beads in panels B and C of Figure 4.27. About 10 percent of this euchromatin is even less condensed and is either being transcribed or is accessible for transcription in the near future (see Ch 10 for details). This is the "active chromatin." During both replication and transcription, the histones are temporarily displaced from short regions of the DNA. After the synthetic enzymes have passed by, the histone cores reassemble on the DNA.

30 nanometer fiber Chain of nucleosomes that is arranged helically, approximately 30 nm in diameter
euchromatin Normal chromatin, as opposed to heterochromatin
heterochromatin A highly condensed form of chromatin that cannot be transcribed because it cannot be accessed by RNA polymerase

TABLE 4.04 Summary of Chromosome Folding

Level of Folding	Consists of	Base Pairs per Turn
DNA double helix	nucleotides	10
Nucleosomes	200 base pairs each	100
30 nanometer fiber	6 nucleosomes per turn	1,200
Loops	50 helical turns per loop	60,000
Chromatid	2,000 loops	

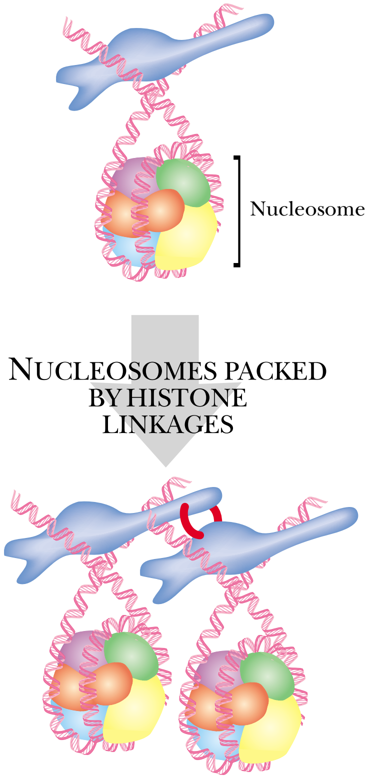


FIGURE 4.24 ***Histone H1 Links Nucleosomes***

The positioning of H1 (blue) above the DNA wrapped around the core particles allows one H1 to bind to another along a linear chain of nucleosomes. This helps in the tighter packing of the nucleosomes.

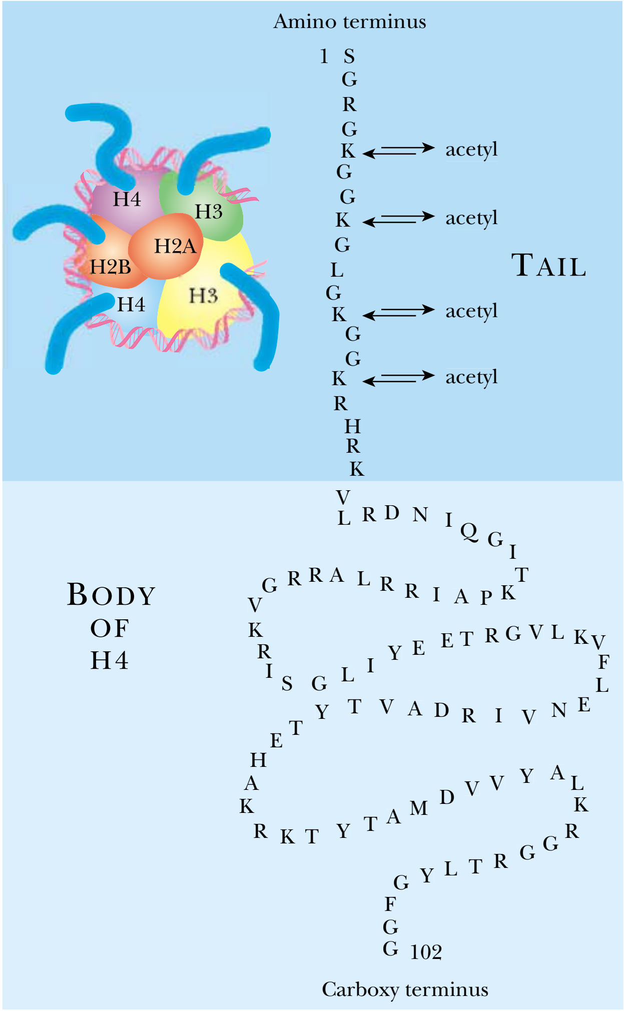

Amino terminus
1 S G R G K
acetyl
G G K
acetyl
G L G K
acetyl
G G K
acetyl
R H R K
TAIL
H4
H3
H2A
H2B
H4
H3
BODY OF H4
V L R D N I Q G I T K P A I R R L A R R G G V K R I S G L I Y E E T R G V L K V F L E N V I R D A V T Y T E H A K R K T V T A M D V V Y A L K R Q G R T L Y G F G G 102
Carboxy terminus

FIGURE 4.25 ***Histone Tails May Be Acetylated***

The N-terminal domains of some of the histone proteins are free for acetylation as indicated by "acetyl." The single letter system for naming amino acids is used.

Histones Are Highly Conserved and Originated Among the Archaebacteria

Of all known proteins, the eukaryotic core histones, especially H3 and H4, are the most highly conserved during evolution. For example, only two amino acids are different, out of 102, between the H4 of cows and peas. The linker histone, H1, is more variable in composition.

Typical bacteria (i.e., the eubacteria) do not possess histones. [Large numbers of "histone-like proteins" are found bound to bacterial chromosomes. Despite the name, these are not homologous in sequence to true histones nor do they form nucleosomes for packaging DNA.] However, some members of the genetically distinct lineage of archaebacteria (e.g., the methane bacteria), do possess histones. Archaeal histones vary significantly from each other. They are 65–70 amino acids long and are missing the tails characteristic of eukaryotic histones. Archaeal nucleosomes accommodate a little under 80 bp of DNA and contain a tetramer of the archaeal histone. They are probably homologous to the $(H3 + H4)_2$ tetramers found in the core of the eukaryotic nucleosome.

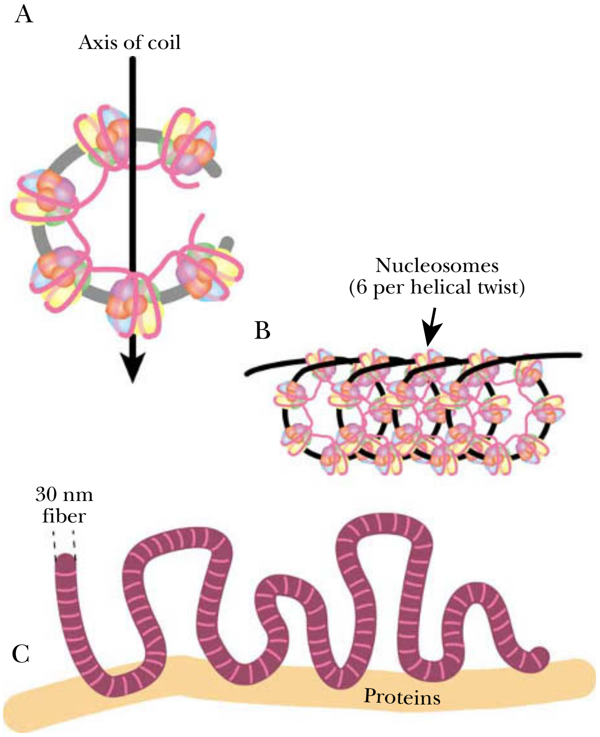


FIGURE 4.26 ***Looping of 30 Nanometer Fiber on Chromosome Axis***

A) A chain of nucleosomes is coiled further with six nucleosomes forming each turn. B) The coiled nucleosomes form a helix, known as a 30 nm fiber. C) The 30 nm fibers form loops that are periodically anchored to a protein scaffolding.

When chromosomes are visible under the light microscope, it is because they are present in a cell that has been caught in the act of dividing. Between cell divisions, most of the DNA is less condensed. It consists of a single, extended molecule of double helical DNA and does not look at all like typical chromosome pictures. Just before cell division, the DNA condenses and folds up, as described above. The typical metaphase chromosome, seen in most pictures, has replicated its DNA some time previously, and is about to divide into two daughter chromosomes, as shown in Figure 4.28. It therefore consists of two identical double helical DNA molecules that are still held together at the centromere. These are known as **chromatids**. Note that between cell

chromatid Single double-helical DNA molecule making up whole or half of a chromosome. A chromatid also contains histones and other DNA-associated proteins.

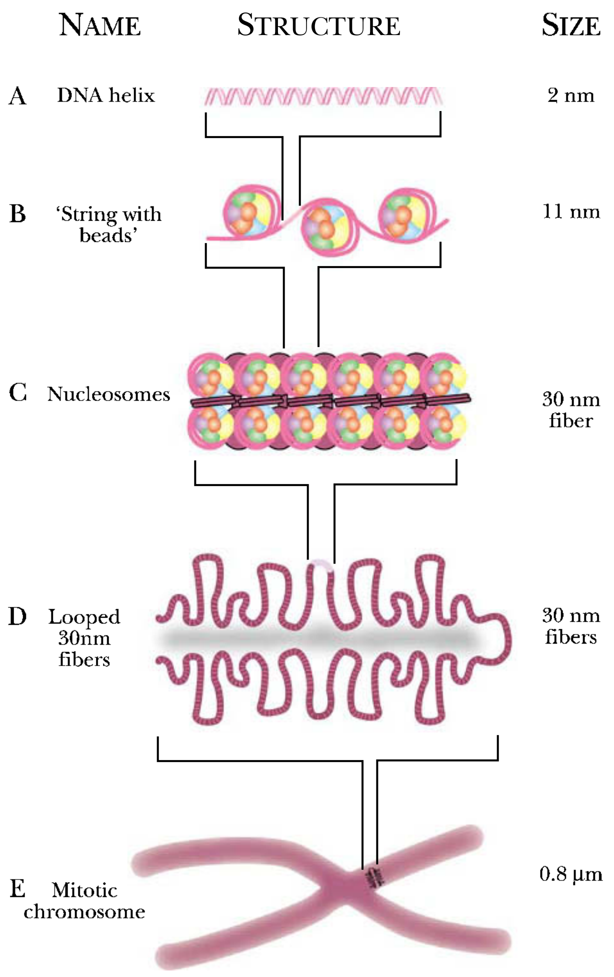


FIGURE 4.27 ***Summary of the Folding of DNA in Eukaryotic Chromosomes***

The DNA helix (A) is wrapped around (B) eight histones (the core). The linker DNA regions unite the nucleosomes to give a "string with beads." This in turn is coiled helically (C) (not clearly indicated) to form a 30 nm fiber. The 30 nm fibers are further folded by looping and attachment to a protein scaffold. Finally, during mitosis the DNA is folded yet again to yield very thick chromosomes.

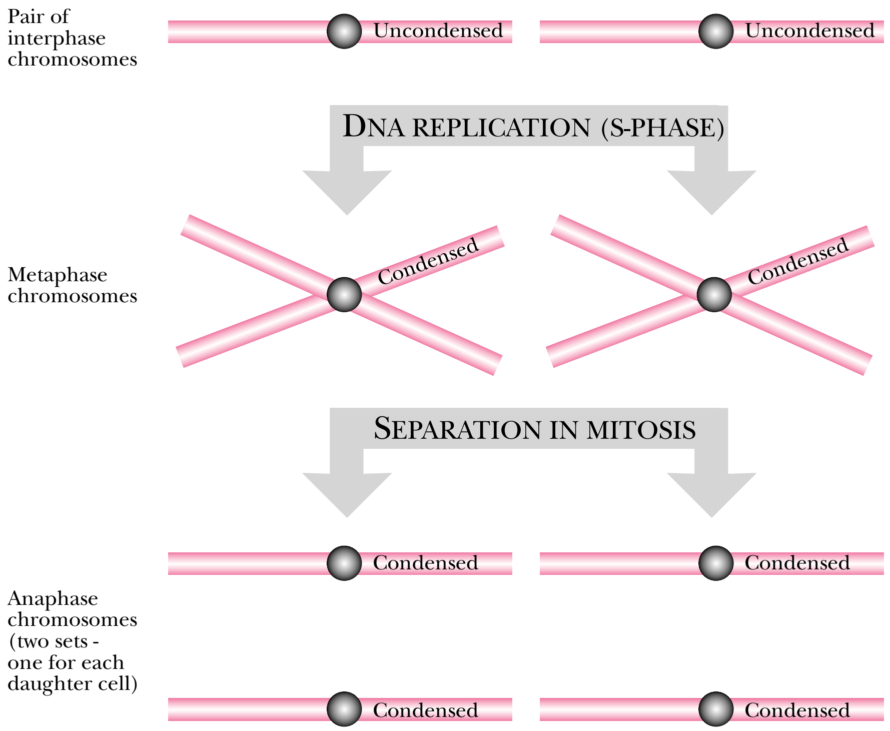


FIGURE 4.28 ***Interphase and Metaphase Chromosomes***

Between rounds of cell division, chromosomes consist of single chromatids, and are referred to as interphase chromosomes. Before the next cell division, the DNA is replicated and each chromosome consists of two DNA molecules or chromatids linked at the centromere. Just prior to mitosis, condensation occurs, making the chromosomes (and chromatids) visible. The chromosomes are best viewed while spread out during the middle part (metaphase) of mitosis. Each daughter cell will acquire one of the chromatids and the process begins anew.

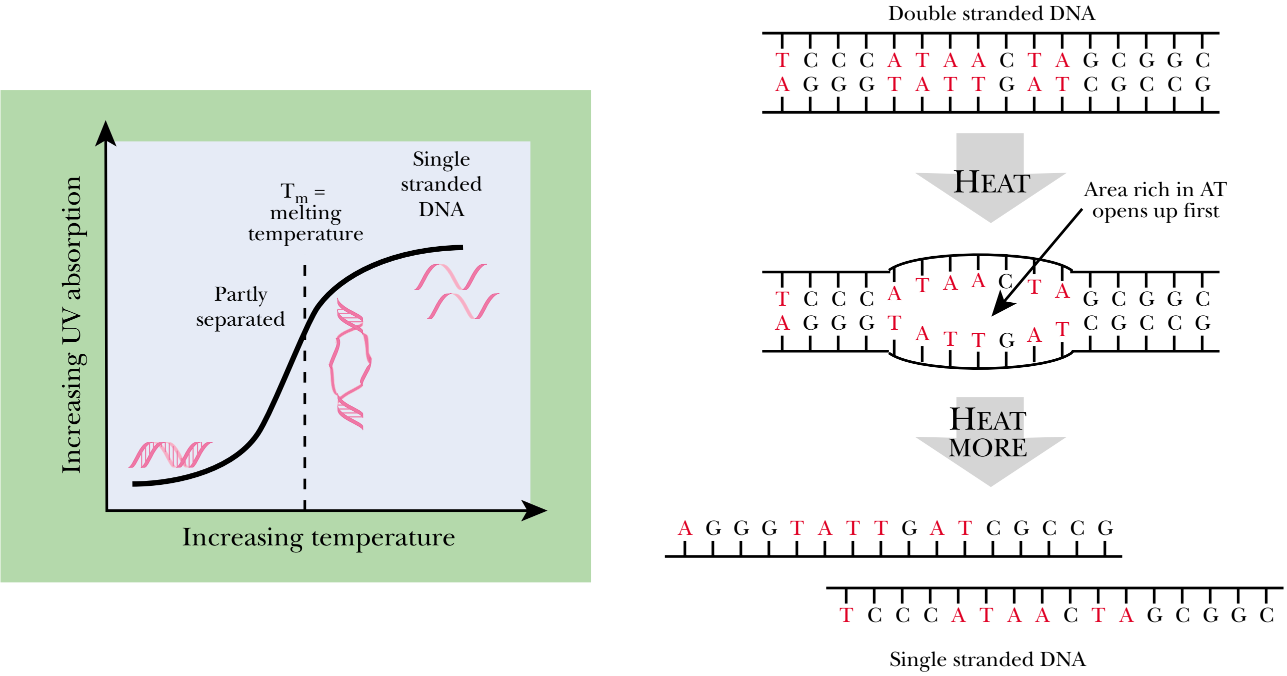

Increasing UV absorption
Tm = melting temperature
Single stranded DNA
Partly separated
Increasing temperature
Double stranded DNA
T C C C A T A A C T A G C G G C
A G G G T A T T G A T C G C C G
HEAT
Area rich in AT opens up first
T C C C A T A A C T A G C G G C
A G G G T A T T G A T C G C C G
HEAT MORE
A G G G T A T T G A T C G C C G
T C C C A T A A C T A G C G G C
Single stranded DNA

FIGURE 4.29 ***Melting of DNA***

DNA strands separate, or "melt," with increasing temperature. This curve shows the measurement of DNA separation by ultraviolet absorption. As the temperature increases more UV is absorbed by the individual strands. The Tm or melting temperature is the point at which half of the double stranded DNA has separated. During the melting process regions rich in A=T melt first since these basepairs have only two hydrogen bonds.

divisions and in non-dividing cells each chromosome consists of only a single chromatid. The term chromatid is mostly needed only to avoid ambiguity when describing chromosomes in process of division. In addition, there are a few unusual cases where giant chromosomes with multiple chromatids are found in certain organisms (e.g. in the salivary glands of flies).

Melting Separates DNA Strands; Cooling Anneals Them

Heating breaks hydrogen bonds and eventually causes the two strands of a DNA double helix to separate—the DNA "melts".

Hydrogen bonds are rather weak, but since a molecule of DNA usually contains millions of base pairs, the added effect of millions of weak bonds is strong enough to keep the two strands together (Fig. 4.29). When DNA is heated, the hydrogen bonds begin to break and the two strands will eventually separate if the temperature rises high enough. This is referred to as "**melting**" or **denaturation** and each DNA molecule has a **melting temperature** (Tm) that depends on its base composition. Therefore, the melting temperature of a DNA molecule is defined strictly as the temperature at the halfway point on the melting curve, as this is more accurate than trying to guess where exactly melting is complete.

The melting temperature is affected by the pH and salt concentration of the solution, so these must be standardized if comparisons are to be made. Extremes of pH disrupt hydrogen bonds. A highly alkaline pH deprotonates the bases which abolishes their ability to form hydrogen bonds and at pH > 11.3 DNA is fully denatured. Con-

denaturation When used of proteins or other biological polymers, refers to the loss of correct 3-D structure
melting When used of DNA, refers to its separation into two strands as a result of heating
melting temperature (Tm) The temperature at which the two strands of a DNA molecule are half unpaired

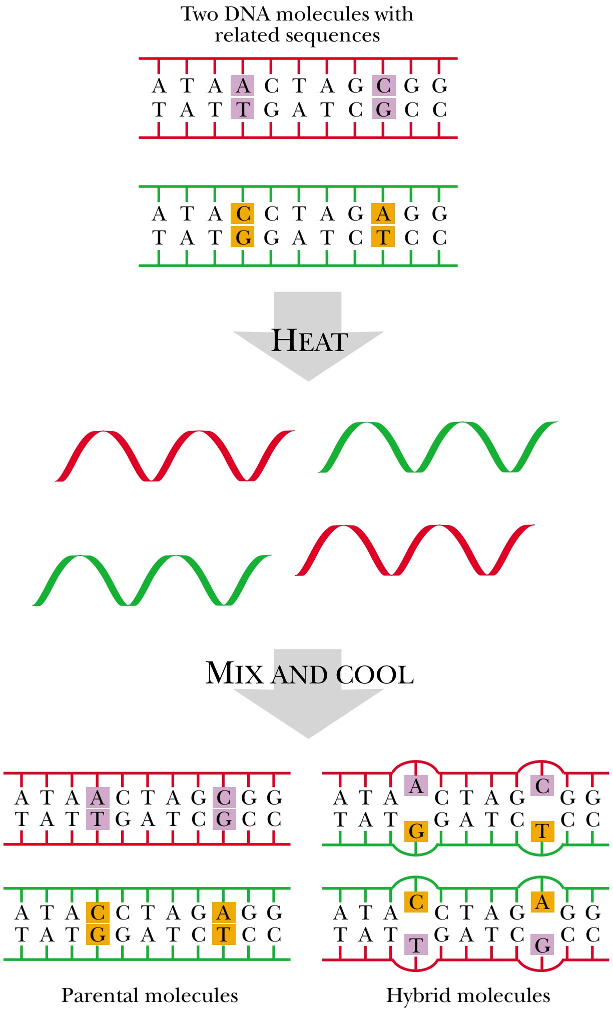


FIGURE 4.30 ***Annealing of DNA***

When pure DNA is heated and cooled the two strands pair up again, i.e., they re-anneal. When DNA from two related sources is heated and re-annealed, hybrid strands may form. The likelihood of hybridization depends on how closely related the two sequences are.

versely, a very low pH causes excessive protonation, which also prevents hydrogen bonding. When DNA is deliberately denatured by pH, alkaline treatment is used because unlike acid, this does not affect the glycosidic bonds between bases and deoxyribose. DNA is relatively more stable at higher ionic concentrations. This is because ions suppress the electrostatic repulsion between the negatively charged phosphate groups on the backbone and hence exert a stabilizing effect. In pure water, DNA will melt even at room temperature.

Melted DNA absorbs more UV light than double-helical DNA.

A spectrophotometer detects the amount absorbed when light is passed through a solution containing DNA. This is compared with the light absorbed by a solution containing no DNA to determine the amount absorbed by the DNA itself. Melting is followed by measuring the absorption of ultraviolet (UV) light at a wavelength of 260 nm (the wavelength of maximum absorption), since disordered DNA absorbs more UV light than a double helix (see Ch. 21).

The more GC base pairs (with three hydrogen bonds) the higher the melting temperature for DNA.

Overall, the higher the proportion of GC base pairs, the higher the melting temperature of a DNA molecule. This is because AT base pairs are weaker, as they have only two hydrogen bonds, as opposed to GC pairs, which have three. In addition, the stacking of GC base pairs with their neighbors is also more favorable than for AT base pairs. In the early days of molecular biology, melting temperatures were used to estimate the percentage of GC versus AT in samples of DNA. DNA base compositions are often cited as the **GC ratio.** The GC content (% G + C) is calculated from the fractional composition of bases as follows:

GC ratio The amount of G plus C relative to all four bases in a sample of DNA. The GC ratio is usually expressed as a percentage

$$\frac{G+C}{A+T+G+C} \times 100\%$$

GC contents for the DNA from different bacterial species vary from 20 percent to 80 percent, with *E. coli* having a ratio of 50 percent. Despite this, there is no correlation between GC content and optimum growth temperature. Apparently this is because the genomes of bacteria are circular DNA molecules with no free ends and this greatly hinders unraveling at elevated temperatures. In fact, small circular DNA molecules, like plasmids, may remain base paired up to 110–120°C. In contrast, the range of GC contents for animals (which have linear chromosomes) is much narrower, from approximately 35–45 percent, with humans having 40.3 percent GC.

As a DNA molecule melts, regions with a high local concentration of AT pairs will melt earlier and GC-rich regions will stay double stranded longer. When DNA is replicated, the two strands must first be pulled apart at a region known as the origin of replication (see Ch. 5). The DNA double helix must also be opened up when genes are transcribed to make mRNA molecules. In both cases, AT-rich tracts are found where the DNA double helix will be opened up more readily.

Upon cooling, the bases in the separated strands of DNA can pair up again and the double helix can re-form.

If the single strands of a melted DNA molecule are cooled, the single DNA strands will recognize their partners by base pairing and the double stranded DNA will re-form. This is referred to as **annealing** or **renaturation**. For proper annealing, the DNA must be cooled slowly to allow the single strands time to find the correct partners. Furthermore, the temperature should remain moderately high to disrupt random H-bond formation over regions of just one or a few bases. A temperature 20–25°C below the Tm is suitable. If DNA from two different, but related, sources is melted and reannealed, **hybrid DNA** molecules may be obtained (Fig. 4.30).

Hybridization of DNA and/or RNA was originally used to estimate the relatedness of different organisms, especially bacteria where the amount of DNA is relatively small, in the days before direct sequencing of DNA became routine. Other uses for hybridization include detection of specific gene sequences and gene cloning. Several extremely useful techniques that are still current and are based on the hybridization of DNA and/or RNA are described in detail in Chapter 21.

Hybrid DNA molecules may be formed by heating and cooling a mixture of two different, but related, DNA molecules.

annealing The re-pairing of separated single strands of DNA to form a double helix
hybrid DNA Artificial double stranded DNA molecule made by pairing two single strands from two different sources
hybridization Pairing of single strands of DNA or RNA from two different (but related) sources to give a hybrid double helix
renaturation Re-annealing of single-stranded DNA or refolding of a denatured protein to give the original natural 3-D structure

CHAPTER FIVE

Cell Division and DNA Replication

Cell Division and Reproduction Are Not Always Identical

When cells divide, the genome must be replicated so each new cell gets a complete set of genes.

Since each cell needs a complete set of genes, an ancestral cell must duplicate its genome before dividing. Each of the two new cells then receives one copy of the genome. Because the genes are made of DNA and are located on the chromosomes, this means that each chromosome must be accurately copied. When a bacterial cell, with a single chromosome divides, each daughter cell receives a copy of the parental chromosome. Division of eukaryotic cells is more complex, as each cell has multiple chromosomes. Not only must all of the chromosomes be duplicated, but a mechanism is needed to ensure that both daughter cells receive identical sets of chromosomes at cell division. This complex process, known as mitosis, is described further below.

Reproduction creates new organisms. Cell division creates new cells. These two processes are only the same for organisms that are single-celled.

When a single-celled organism divides, the result is two new organisms, each consisting of one cell. However, in multi-cellular organisms cell division does not automatically result in the creation of new organisms. When the cells composing a multi-cellular organism divide, they increase the size and/or complexity of the original organism. A distinct process is needed to form new organisms. The term reproduction is used to signify the production of a new individual organism. Thus, in unicellular organisms, cell division and reproduction occur simultaneously, whereas in multicellular organisms cell division and reproduction are two different processes.

In many plants and fungi, clumps of cells may break off or single celled spores may be released from the parental organism and give rise to new individual multicellular organisms. This is known as **asexual** or **vegetative reproduction** as these new individuals will be genetically identical to their parents. This contrasts with sexual reproduction, where each new individual receives roughly equal amounts of genetic information from two separate parents and is therefore a novel genetic assortment. Sexual reproduction is especially characteristic of animals and also occurs in most higher plants and many fungi. Some organisms, particularly plants and fungi possess the ability to reproduce both sexually or asexually. Although they are inextricably entwined in humans and many other animals, it is important to realize that sex and reproduction are two distinct processes from a biological viewpoint.

Strictly speaking, bacteria do not reproduce sexually since new bacteria always result from the division of a single parental cell. Nonetheless, mixing of genes from two individuals may occur in bacteria. However, this occurs in the absence of cell division and involves transfer of a relatively small segment of DNA from one cell (the donor) to another cell (the recipient) (see Ch. 18 for details). Such sideways transfer of DNA, between members of the same generation is sometimes referred to as horizontal gene transmission (see Ch. 15). In contrast, vertical gene transmission is when genes are transmitted from the previous generation to the new generation. Vertical transmission thus includes all forms of cell division and reproduction that create a new copy of the genome, whether sexual or not.

DNA Replication Is a Two-Stage Process Occurring at the Replication Fork

The DNA must be replicated before a cell divides.

The two strands of the double helix must be separated before replication can occur.

Replication is the process by which the DNA of the ancestral cell is duplicated, prior to cell division. Upon cell division, each of the descendants will get one complete copy of the DNA that is identical to its predecessor. The first stage in replication is to separate the two DNA strands of the parental DNA molecule. The second stage is to build two new strands, using each of the two original strands as **templates**. The most fundamental aspect of replication is the base pairing of A with T and of G with C. Each of

asexual or vegetative reproduction Form of reproduction in which there is no reshuffling of the genes between two individuals
replication Duplication of DNA prior to cell division
template strand Strand of DNA used as a guide for synthesizing a new strand by complementary base pairing

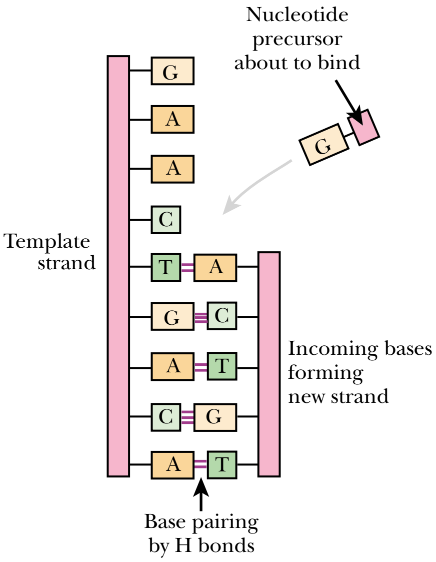

Nucleotide precursor about to bind
G
A
A
C
T
A
G
C
A
T
C
G
A
T
G
Template strand
Incoming bases forming new strand
Base pairing by H bonds

FIGURE 5.01 ***Template Strand and Base Pairing in DNA Replication***

Incoming nucleotides line up on the template strand and are then linked together to form the new strand of DNA. The arriving nucleotides are positioned by base pairing, in which A pairs with T and G binds to C. These base pairs are held together with hydrogen bonds.

New DNA is made using the existing strands as templates for base pairing.

the separated parental strands of DNA serves as a template strand for the synthesis of a new complementary strand. The incoming nucleotides for the new strand recognize their partners by base pairing and so are lined up on the **template strand** (Fig. 5.01). Since A pairs only with T, and since G pairs only with C, the sequence of each original strand dictates the sequence of the new complementary strand.

Synthesis of both new strands of DNA occurs at the **replication fork** that moves along the parental molecule. Amazingly, in *E. coli*, DNA is made at nearly 1,000 nucleotides per second. The replication fork consists of the zone of DNA where the strands are separated, plus an assemblage of proteins that are responsible for synthesis, sometimes referred to as the **replisome**. The result of replication is two double stranded DNA molecules, both with sequences identical to the original one. One of these daughter molecules has the original left strand and the other daughter has the original right strand. The pattern of replication is **semi-conservative**, since each of the progeny conserves half of the original DNA molecule (Fig. 5.02).

Each daughter cell gets one old strand and one new strand of DNA.

Replication is similar, but not exactly the same, in prokaryotes and eukaryotes. DNA replication in bacteria will be covered initially, as this process is better understood and is less complicated than the process in eukaryotes.

Supercoiling Causes Problems for Replication

The strands of the parent DNA molecule cannot be separated until the supercoils and helical twisting have been removed.

Several major problems must be solved to accomplish bacterial DNA replication. First, there are the topological problems due to DNA being not only a double helix but also supercoiled. Because the two strands forming a DNA molecule are held together by hydrogen bonding and are twisted around each other to form a double helix, they cannot simply be pulled apart. The higher level supercoiling further complicates the problem of separating the strands. Consequently, before new DNA can be made, first the supercoiling must be unwound. Next the two strands of the parental DNA double helix must be untwisted (see below). In addition, since the vast majority of bacterial chromosomes are circular, it is important to untangle the two new circles of DNA.

The supercoiled bacterial chromosome is circular and two replication forks proceed in opposite directions around the circle (Fig. 5.03). This process is known as

semi-conservative replication Mode of DNA replication in which each daughter molecule gets one of the two original strands and one new complementary strand
replication fork Region where the enzymes replicating a DNA molecule are bound to untwisted, single stranded DNA
replisome Assemblage of proteins (including primase, DNA polymerase, helicase, SSB protein) that replicates DNA
template strand Strand of DNA used as a guide for synthesizing a new strand by complementary base pairing

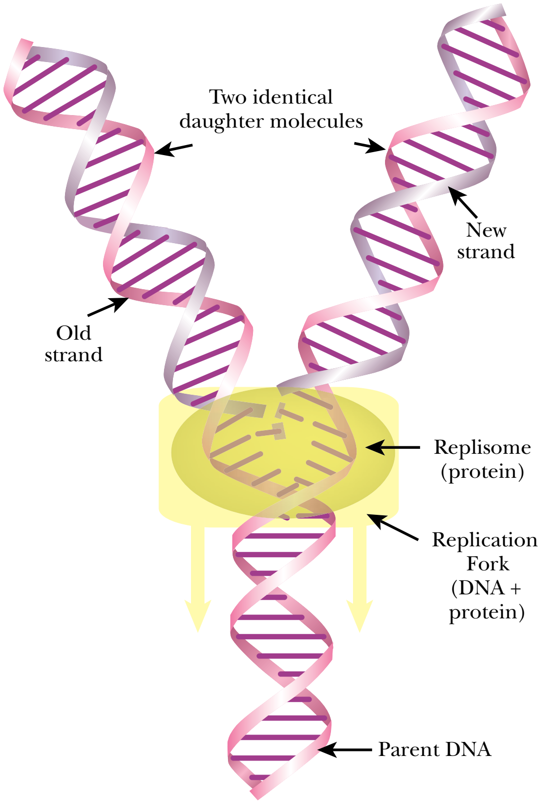

Two identical daughter molecules
New strand
Old strand
Replisome (protein)
Replication Fork (DNA + protein)
Parent DNA

FIGURE 5.02 ***Semi-conservative Replication and the Replisome***

The replication fork is the site of DNA replication and, by definition, includes both the DNA and associated proteins. The assembled proteins, known as the replisome, facilitate the unwinding of the helix and the addition of new nucleotides. The arrows indicate the direction of movement of the replication fork. The synthesis of two DNA helices results from adding a new complementary strand to each one of the separated old strands.

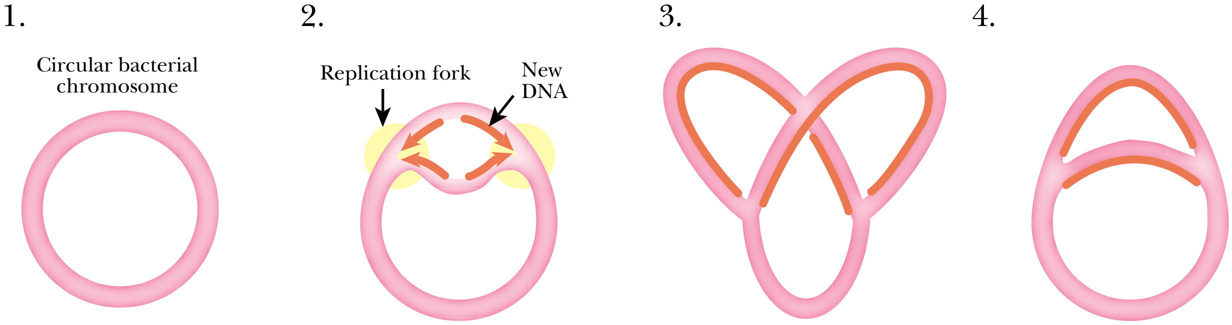


FIGURE 5.03 ***Theta-Replication***

Successive steps in DNA replication are shown for a circular bacterial chromosome. The chromosome (1) begins to replicate using two replication forks (2). Continued replication results in division of the chromosome (3) and its apparent resemblance to θ, the Greek letter, theta (4).

bi-directional replication. A circle observed half way through division looks like the Greek letter theta (θ) and so this mode of replication is also called **theta-replication**.

Supercoiling of DNA has been discussed in detail in Ch. 4. In *E. coli*, the two type II topoisomerases, **DNA gyrase** and **topoisomerase IV**, are involved in DNA replication. As the replication fork proceeds along the DNA, it over-winds the DNA and generates positive supercoils ahead of the replisome. Since the bacterial chromosome is negatively supercoiled, the overwinding introduced by replication is at first cancelled out. However, after replication of about 5 percent of the chromosome, the pre-existing negative supercoils would all be used up. For DNA replication to proceed, the over-winding must be removed. DNA gyrase binds to the DNA ahead of the replica-

bi-directional replication Replication that proceeds in two directions from a common origin
DNA gyrase An enzyme that introduces negative supercoils into DNA, a member of the type II topoisomerase family
theta-replication Mode of replication in which two replication forks go in opposite directions around a circular molecule of DNA
topoisomerase IV A particular topoisomerase involved in DNA replication in bacteria

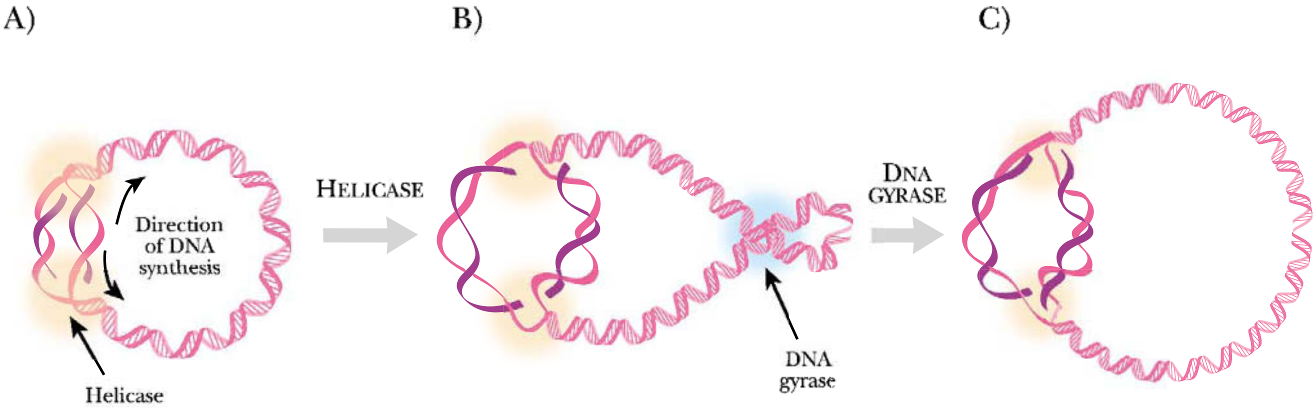


FIGURE 5.04 ***Unwinding of Double Helix and of Supercoils***

For the replication fork to proceed, both the double helix and the supercoils must be unwound. Helicase unwinds the double helix and DNA gyrase removes the supercoiling.

tion fork and introduces negative supercoils that cancel out the positive supercoiling. The net result is that DNA gyrase "removes" supercoiling ahead of the replication fork (Fig. 5.04). Topoisomerase IV may help in this process to some extent, but its main function is disentangling daughter molecules after replication has finished, as described below.

Quinolone antibiotics kill bacteria by inhibiting DNA gyrase. This blocks replication of the DNA.

The quinolone antibiotics (e.g., nalidixic acid and ciprofloxacin) inhibit bacterial DNA replication by acting on the type II topoisomerases, in particular DNA gyrase. Inhibited DNA gyrase remains bound to the DNA at one location and blocks movement of the replication fork. The net effect is cell death.

Strand Separation Precedes DNA Synthesis

The double helix itself is unwound by the enzyme **DNA helicase** (Fig. 5.05). The major helicase of *E. coli* is DnaB protein, which forms hexamers. Helicase does not break the DNA chains; it simply disrupts the hydrogen bonds holding the base pairs together. Energy is needed for this process and helicase cleaves ATP to supply this energy.

Helicase unwinds the DNA helix and SSB protein keeps the strands apart.

The two separated strands of the parental DNA molecule are complementary and so capable of base pairing to each other. In order to manufacture the new strands, the two original strands must be kept apart for use as templates. This is done by **single strand binding protein**, or **SSB protein**, which binds to the unpaired single stranded DNA and prevents the two parental strands from re-annealing (Fig. 5.06). In reality, the single stranded region between helicase and the lagging strand is longer than that between helicase and the leading strand, due to the three dimensional arrangement of the replication fork.

Properties of DNA Polymerase

One new DNA strand is made continuously. The other is made as a series of fragments that must be linked together later.

Polymerases are enzymes that join nucleotides together. Bacterial cells contain several DNA polymerases that have different roles both in DNA replication and in DNA repair (see Ch. 14). Further problems in DNA replication stem from the peculiarities of **DNA polymerase**, the enzyme that is responsible for making new chains of DNA. Firstly, DNA polymerase will only synthesize DNA in a 5′- to 3′- direction. Since the

DNA helicase Enzyme that unwinds double helical DNA
DNA polymerase Enzyme that synthesizes DNA
polymerase Enzyme that synthesizes nucleic acids
single strand binding protein (SSB protein) A protein that keeps separated strands of DNA apart

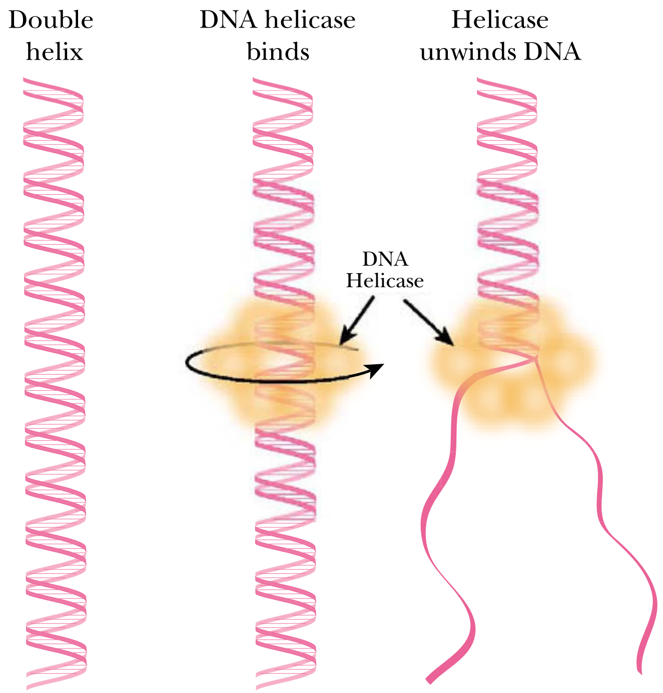


FIGURE 5.05 ***Helicase Unwinds the Double Helix***

To unwind DNA, helicase first binds to DNA and then cleaves the hydrogen bonds connecting base pairs to separate the strands of the helix.

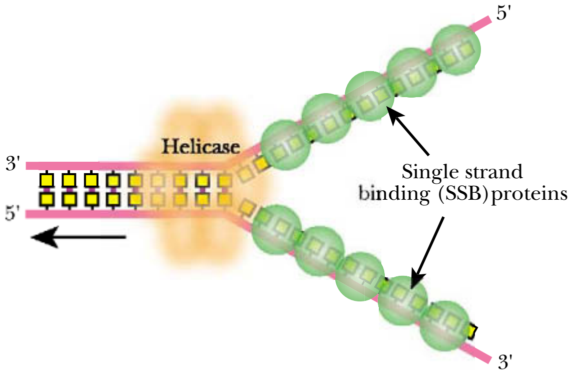


FIGURE 5.06 ***Single Strand Binding Protein Keeps DNA Strands Apart***

Soon after DNA helicase breaks the hydrogen bonds holding the DNA strands together, SSB protein binds to the freed strands to keep them from re-annealing.

DNA polymerase can only make new DNA in one direction. Even stranger, it cannot start new strands of DNA.

New strands of DNA must be started with short segments of RNA, known as primers.

strands in a double helix are anti-parallel, and since a single replication fork is responsible for duplicating the double helix, this means that one of the new strands can be made continuously, but that the other cannot (Fig. 5.07). The strand that is made in one piece is called the **leading strand** and the strand that is made discontinuously is the **lagging strand**.

Secondly, all DNA polymerases lack the ability to initiate a new strand of nucleic acid and can only elongate a pre-existing strand. Consequently, a special mechanism for strand initiation is needed. This involves synthesis of a short **RNA primer** whenever a new DNA strand is begun. To begin a new strand DNA polymerase uses a short RNA primer made by another enzyme. Unlike DNA polymerases, **RNA polymerases** can start new strands. A special RNA polymerase, known as **primase** (DnaG protein) makes the RNA primers that are responsible for strand initiation during DNA

lagging strand The new strand of DNA which is synthesized in short pieces during replication and then joined later
leading strand The new strand of DNA that is synthesized continuously during replication
primase Enzyme that starts a new strand of DNA by making an RNA primer
RNA polymerase Enzyme that synthesizes RNA
RNA primer Short segment of RNA used to initiate synthesis of a new strand of DNA during replication

synthesis in bacteria. Although the leading strand only needs to be started once, the lagging strand is made in short sections and a new RNA primer must be inserted each time a new portion is made. DNA polymerase will build new strands of DNA starting from each RNA primer (Fig. 5.08).

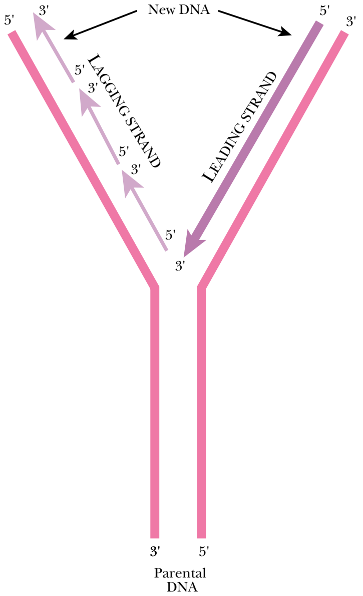


FIGURE 5.07 ***Continuous and Discontinuous Synthesis of DNA at the Replication Fork***

The protein at the replication fork responsible for DNA synthesis, DNA polymerase, always synthesizes DNA in the 5′- to 3′-direction. Therefore one new strand (leading strand) can be made continuously, while the other (lagging strand) must be made discontinuously (i.e., in short segments).

Polymerization of Nucleotides

All nucleic acids, whether DNA or RNA, are synthesized in the 5′- to 3′- direction. Incoming nucleotides are added to the hydroxyl group at the 3′-end of the growing chain. The precursors for DNA synthesis are the **deoxyribonucleoside 5′-triphosphates** (**deoxy-NTPs**), dATP, dGTP, dCTP, and dTTP. Proceeding from the deoxyribose outwards, the three phosphate groups are designated the α, β, and γ phosphates. Upon polymerization, the high energy bond between the α and β phosphates is cleaved, releasing energy to drive the polymerization. The outermost two phosphates (the β- and γ- phosphates) are released as a molecule of pyrophosphate. A new bond is made between the innermost phosphate (the α-phosphate) of the incoming nucleotide and the 3′-OH of the previous nucleotide at the end of the growing chain (Fig. 5.09).

Supplying the Precursors for DNA Synthesis

Precursors for DNA are made from those for RNA by oxidizing the ribose to deoxyribose.

The fact that all new strands of nucleic acid start with a piece of RNA plus the fact that ribonucleotides are made first supports the idea that RNA came first in evolution. This theory, the RNA world, is discussed further in Chapter 20.

The DNA precursors, which contain deoxyribose, are made from the corresponding ribose-containing nucleotides (Fig. 5.10). Reduction of ribose to deoxyribose is catalyzed by the enzyme **ribonucleotide reductase**. This acts on the diphosphate derivatives (ADP, GDP, CDP, and UDP). A **kinase** then adds the third phosphate in the case of dADP, dGDP and dCDP, so giving dATP, dGTP and dCTP. The dUDP follows a different route, as DNA does not contain uracil but instead has thymine, the methyl derivative of uracil. Before methylation, dUDP is converted to dUMP by removal of a phosphate. Then **thymidylate synthetase** adds the methyl group, so converting dUMP to dTMP. Finally, two phosphates are added to give dTTP (Fig. 5.10).

The methyl group of thymine is carried by the **tetrahydrofolate** (**THF**) cofactor, which is oxidized to **dihydrofolate** (**DHF**) during the reaction. The DHF must be reduced back to THF for DNA synthesis to proceed. The enzyme that does this, **dihydrofolate reductase**, is inhibited by the antibiotic **trimethoprim** in bacteria. Moreover, the synthesis of the precursor to the purines, adenine and guanine, also needs a one-carbon fragment carried by THF. So DNA precursor synthesis is actually blocked at two points by trimethoprim.

Eukaryotic dihydrofolate reductase is inhibited by **methotrexate** (**amethopterin**). Since growth of tumors involves rapid cell division and DNA replication by cancer cells, methotrexate is used as an anti-tumor agent. The **sulfonamide** class of antibiotics inhibits synthesis of the **folate** cofactor itself. Animals do not make folate, but require it in their diet, so sulfonamides are harmless to human patients in reasonable doses. Massive doses may cause liver and kidney problems.

deoxyribonucleoside 5′-triphosphate (deoxyNTP) Precursor for DNA synthesis consisting of a base, deoxyribose and three phosphate groups
dihydrofolate (DHF) Cofactor with a variety of roles including making precursors for DNA and RNA synthesis
dihydrofolate reductase Enzyme that converts dihydrofolate back to tetrahydrofolate
folate Cofactor involved in carrying one carbon groups in DNA synthesis
kinase Enzyme that attaches a phosphate group to another molecule
methotrexate (or amethopterin) Anti-cancer drug that inhibits dihydrofolate reductase of animals
ribonucleotide reductase Enzyme that reduces ribonucleotides to deoxyribonucleotides
sulfonamide Antibiotic that inhibits the synthesis of the folate cofactor
tetrahydrofolate (THF) Reduced form of dihydrofolate cofactor that is needed for making precursors for DNA and RNA synthesis
thymidylate synthetase Enzyme that adds a methyl group, so converting the uracil of dUMP to thymine
trimethoprim Antibiotic that inhibits dihydrofolate reductase of bacteria

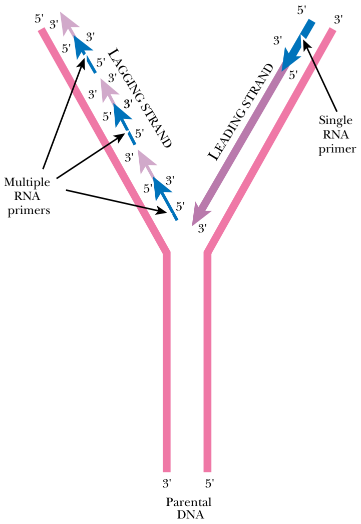


FIGURE 5.08 ***Strand Initiation Requires an RNA Primer***

DNA polymerase cannot begin a new strand but can only elongate. Therefore DNA replication requires an RNA primer to initiate strand elongation. One RNA primer is needed to start the 5′ to 3′ leading strand. In contrast, multiple RNA primers are needed for the 3′ to 5′ lagging strand because this is made in short stretches each running 5′ to 3′. DNA polymerase will then add nucleotides to the end of each RNA primer. Later, the short RNA primers will be removed and replaced by DNA.

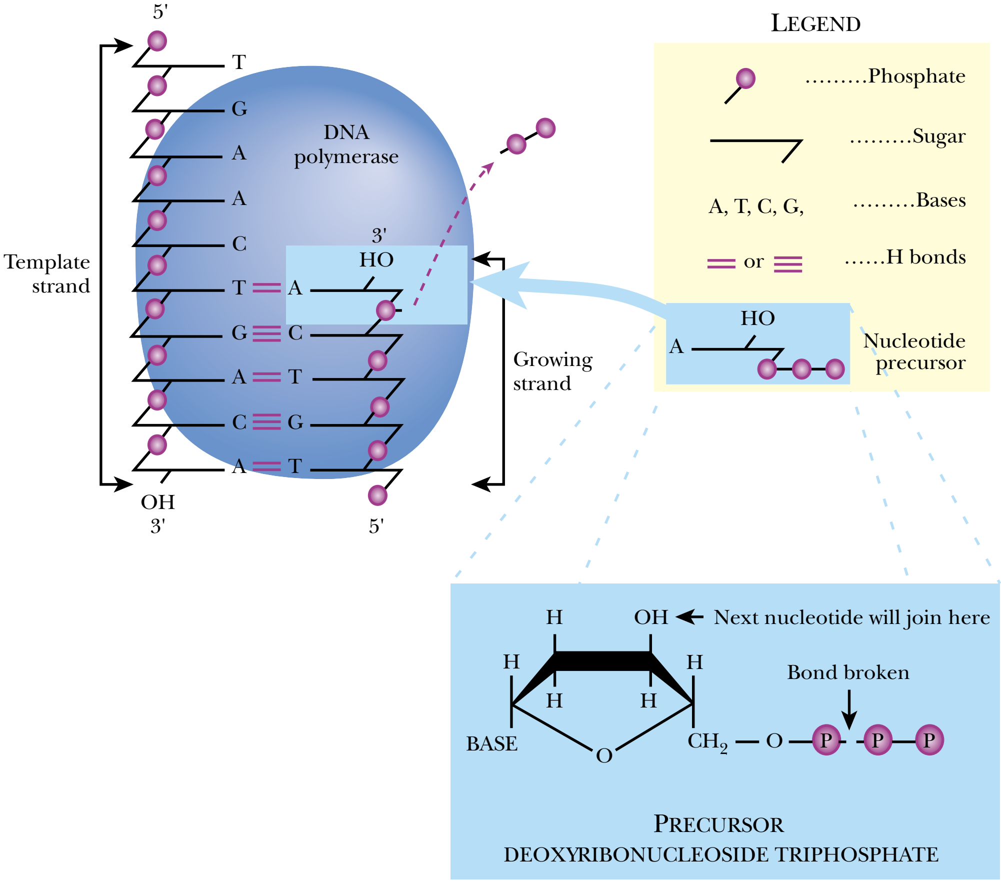


FIGURE 5.09 ***Polymerization of Nucleotides***

Details of the growth of the new DNA strand are shown. At each step, a nucleoside triphosphate is added. During chain elongation, the two outermost phosphate groups of the precursor are cleaved off releasing pyrophosphate. This provides energy for the reaction. The 3′ hydroxyl group of the precursor triphosphate remains available for the next addition to the growing strand.

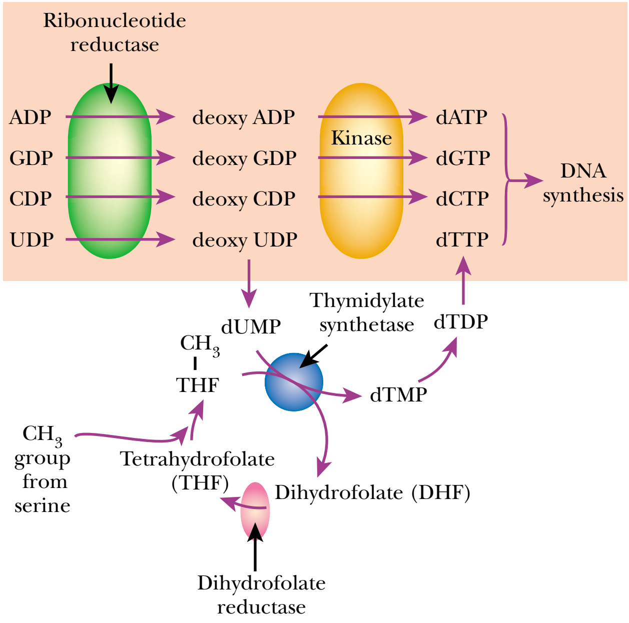


FIGURE 5.10 ***Synthesis of Precursors***

First, ribonucleotide reductase converts the ribonucleoside diphosphates to the deoxy form. Second, a kinase adds a phosphate to form the deoxyribonucleoside triphosphates. The precursor for the thymidine nucleotides is made from a uridine derivative by adding a methyl group that is transferred by the carrier tetrahydrofolate (THF).

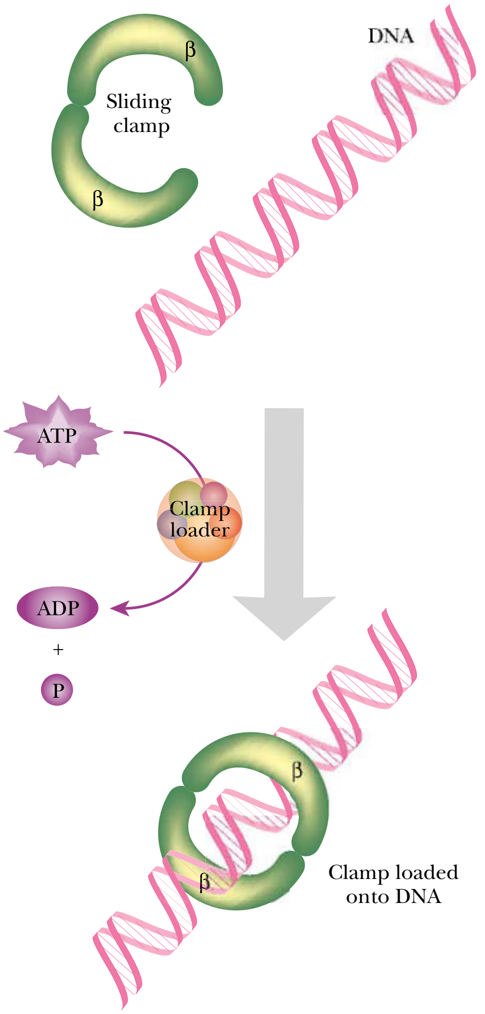


FIGURE 5.11 ***DNA Polymerase III—Sliding Clamp***

The sliding clamp, made of two similar β subunits, encircles the DNA helix. It is loaded onto the DNA with help from the clamp-loading complex.

DNA Polymerase Elongates DNA Strands

DNA polymerase is kept attached to the DNA by a doughnut shaped protein ring that slides up and down the DNA strand.

DNA polymerase not only synthesizes new DNA, but also checks for mismatched base pairs and corrects any errors it has made.

DNA **polymerase III** (**Pol III**) is the major enzyme involved in elongating DNA during chromosome replication. DNA polymerase III has several components. First, the "**sliding clamp**" is shaped like a doughnut and is made from two semicircular subunits of DnaN protein. It slides up and down like a curtain ring on the template strand of DNA (Fig. 5.11) and requires several accessory proteins (HolA, B, C and D plus the γ subunit) known as the **clamp-loading complex** to load the sliding clamp onto the DNA—a process requiring energy.

The "sliding clamp" binds the very large and complex **core enzyme** to the DNA. The core enzyme synthesizes DNA and consists of three subunits, DnaE (α-subunit; polymerization), DnaQ (ε-subunit; proofreading) and HolE (θ-subunit; function uncertain). Two core assemblies, one to make each new strand of DNA, are held together by a pair of tau subunits. A single clamp loader complex is shared by the two core assemblies (Fig. 5.12).

Although hydrogen bonding alone would match bases correctly, the great majority of the time this is not accurate enough for replication of the genome. Consequently, many DNA polymerases possess **kinetic proofreading** ability (Fig. 5.13). This refers to

clamp-loading complex Group of proteins that loads the sliding clamp of DNA polymerase onto the DNA
core enzyme The part of DNA or RNA polymerase that synthesizes new DNA or RNA (i.e. lacking the recognition and/or attachment subunits)
DNA polymerase III (Pol III) Enzyme that makes most of the DNA when bacterial chromosomes are replicated
kinetic proofreading Proofreading of DNA that occurs during the process of DNA synthesis
sliding clamp Subunit of DNA polymerase that encircles the DNA, thereby holding the core enzyme onto the DNA

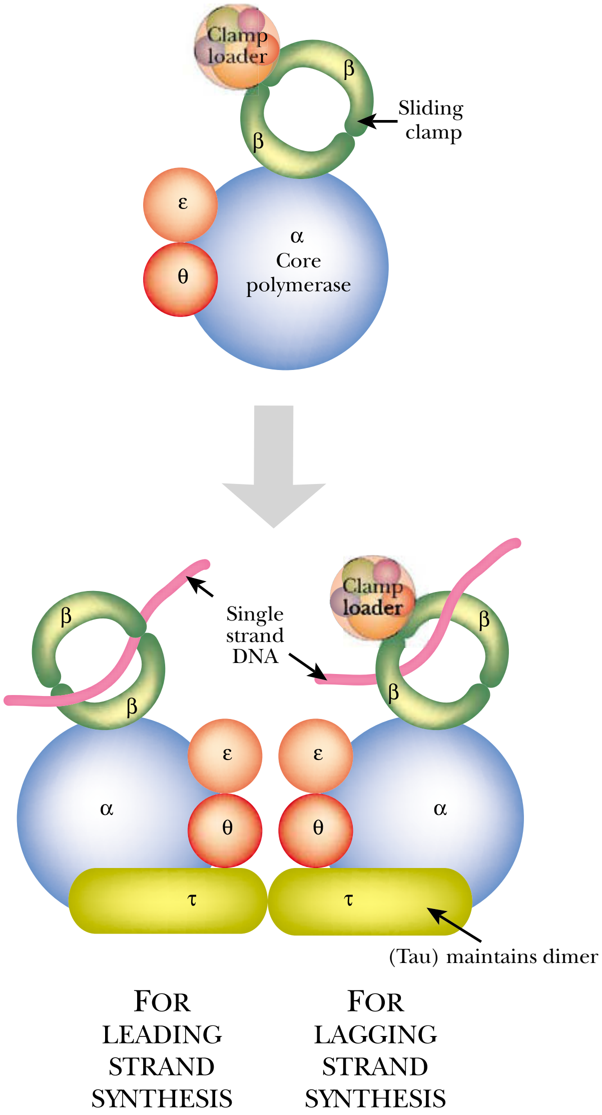


FIGURE 5.12 DNA Polymerase III—Assembly of Subunits

A single core subunit is shown above and its assembly into a dimeric unit is shown below. The dimeric subunit contains only one clamp loader, which is associated with the lagging strand synthetic unit. The two core proteins are bound together by the tau subunit. The components of DNA polymerase III are the β subunits forming the sliding clamp, the epsilon and theta subunits and the clamp loader. When the DNA polymerase assembles, tau subunits join two nearly identical complexes together.

the ability to make corrections on the fly. **Mismatches** are sensed because they cause minor distortions in the shape of the double helix. The polymerase halts upon sensing a mismatch and removes the last nucleotide added. As this was added to the 3′-end of the growing DNA strand, the enzyme activity that removes the offending nucleotide is a **3′-exonuclease**. In the case of DNA polymerase III, proofreading is due to a separate subunit, the DnaQ protein (ε-subunit). (In some other DNA polymerases, proofreading ability resides on the same protein as polymerase activity.) In addition, immediately after replication, the new DNA is checked and if necessary repaired by the **mismatch repair** system (see Ch. 14).

One of the new strands of DNA, the "lagging strand", is made in short fragments that are joined up later.

The Complete Replication Fork Is Complex

The replication fork is defined as all the structural components in the region where the DNA molecule is being duplicated. It includes the zone where the DNA is being untwisted by gyrase and helicase, together with the stretches of single stranded DNA

3′-exonuclease An enzyme that degrades nucleic acids from the 3′-end
mismatch Wrong pairing of two bases in a double helix of DNA
mismatch repair DNA repair system which recognizes and corrects wrongly paired bases

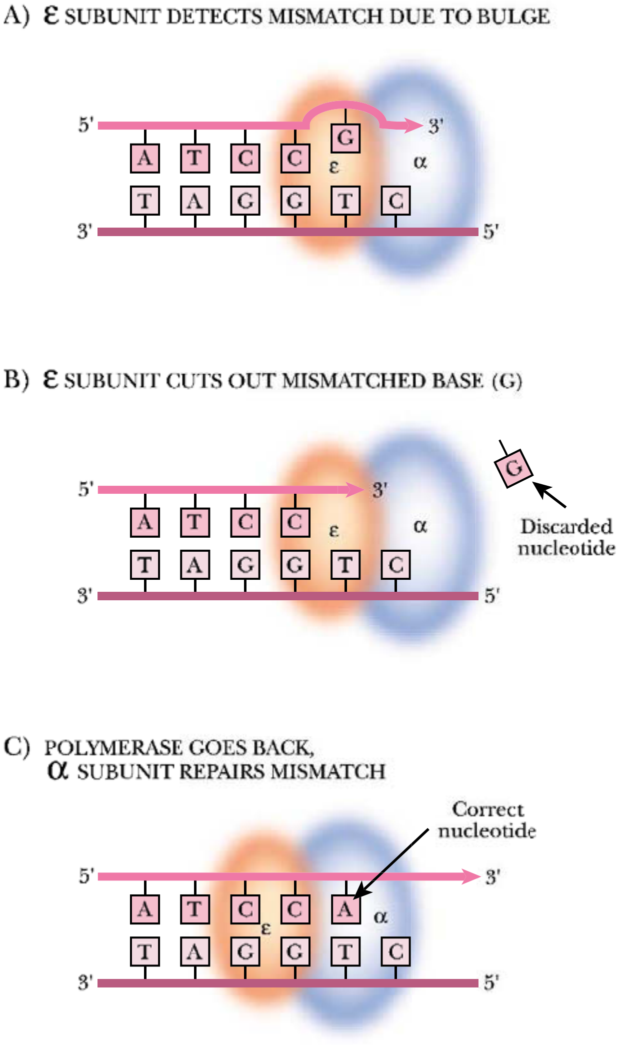

A) ε SUBUNIT DETECTS MISMATCH DUE TO BULGE
5'
3'
G
A
T
C
C
ε
α
T
A
G
G
T
C
3'
5'
B) ε SUBUNIT CUTS OUT MISMATCHED BASE (G)
G
Discarded nucleotide
5'
3'
A
T
C
C
ε
α
T
A
G
G
T
C
3'
5'
C) POLYMERASE GOES BACK, α SUBUNIT REPAIRS MISMATCH
Correct nucleotide
5'
3'
A
T
C
C
A
ε
α
T
A
G
G
T
C
3'
5'

FIGURE 5.13 ***DNA Polymerase III—Proofreading***

The portion of DNA polymerase responsible for checking the accuracy of new nucleotides is the DnaQ protein (ε-subunits). A mismatch is detected by a bulge in the new chain. The incorrect nucleotide is discarded and the appropriate nucleotide added.

Molecular Biology Rarity

Both the tau (τ) and gamma (γ) subunits of DNA polymerase III are encoded by the same gene—*dnaX*. The tau subunit is a normal, full-length product of translation. However, the gamma subunit is made by frameshifting (see Ch. 13) followed by premature termination during translation of the same mRNA. Such differential synthesis of two proteins from a single gene is extremely rare in living cells, but this particular case appears to be widespread among bacteria, including *E. coli*. However, frameshifting to generate two alternative proteins is not so rare among RNA viruses with small genomes. For example, both retroviruses (e.g., HIV) and filoviruses (e.g., Ebolavirus) use this approach—see Ch. 17.

held apart by single strand binding protein (SSB). It also includes two molecules of DNA polymerase III, which are making two new strands of DNA (Fig. 5.14). Although the leading strand is made continuously, the lagging strand is made in short segments of 1,000 to 2,000 bases in length, known as **Okazaki fragments** after their discoverer.

Okazaki fragments The short pieces of DNA that make up the lagging strand

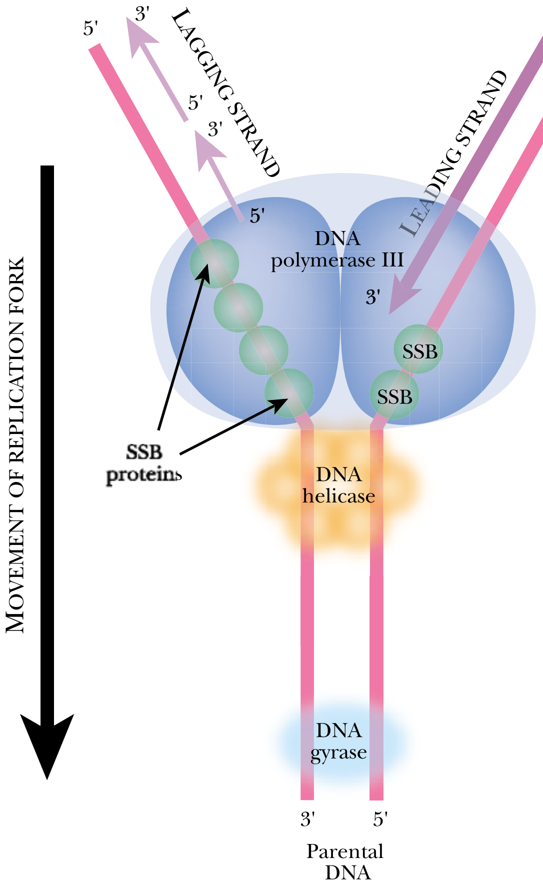


FIGURE 5.14 ***Components of the Replication Fork***

The basic components of the replication fork are the DNA gyrase, DNA helicase, DNA polymerase III, and single strand binding proteins (SSB).

As noted above, during DNA replication, the two new strands must be synthesized in opposite directions. A linear representation of this would imply that the two polymerase assemblies might move apart (Fig. 5.15). In fact, the two molecules of Pol III that are making these two strands are held together by the tau subunits. In order for them both to make new DNA simultaneously, the strands of DNA must be bent or looped, perhaps as shown in panel C of Figure 5.15.

Discontinuous Synthesis of DNA Requires a Primosome

Although the leading strand is synthesized continuously, the lagging strand is composed of multiple pieces, the Okazaki fragments. When synthesis of each new Okazaki fragment is begun, it needs a fresh RNA primer. For priming to occur, **PriA** protein displaces SSB proteins that are bound to the unpaired DNA from a short stretch of DNA and enables the **primase** (DnaG) to bind. The primase then makes a short RNA primer of 11 to 12 bases. This priming complex is sometimes known as the **primosome** (Fig. 5.16).

Each time a new Okazaki fragment is begun, the Pol III assembly that is making the lagging strand releases its grip on the DNA and relocates to start making a new strand of DNA starting from the 3′ end of the RNA primer. This involves disassembly and relocation of the sliding clamp, which is performed by the clamp loading complex. Note that the replisome contains two Polymerase III assemblies, each with its own

PriA Protein of the primosome that helps primase bind
primase Enzyme that starts a new strand of DNA by making an RNA primer
primosome Cluster of proteins (including PriA and primase) that synthesizes a new RNA primer during DNA replication

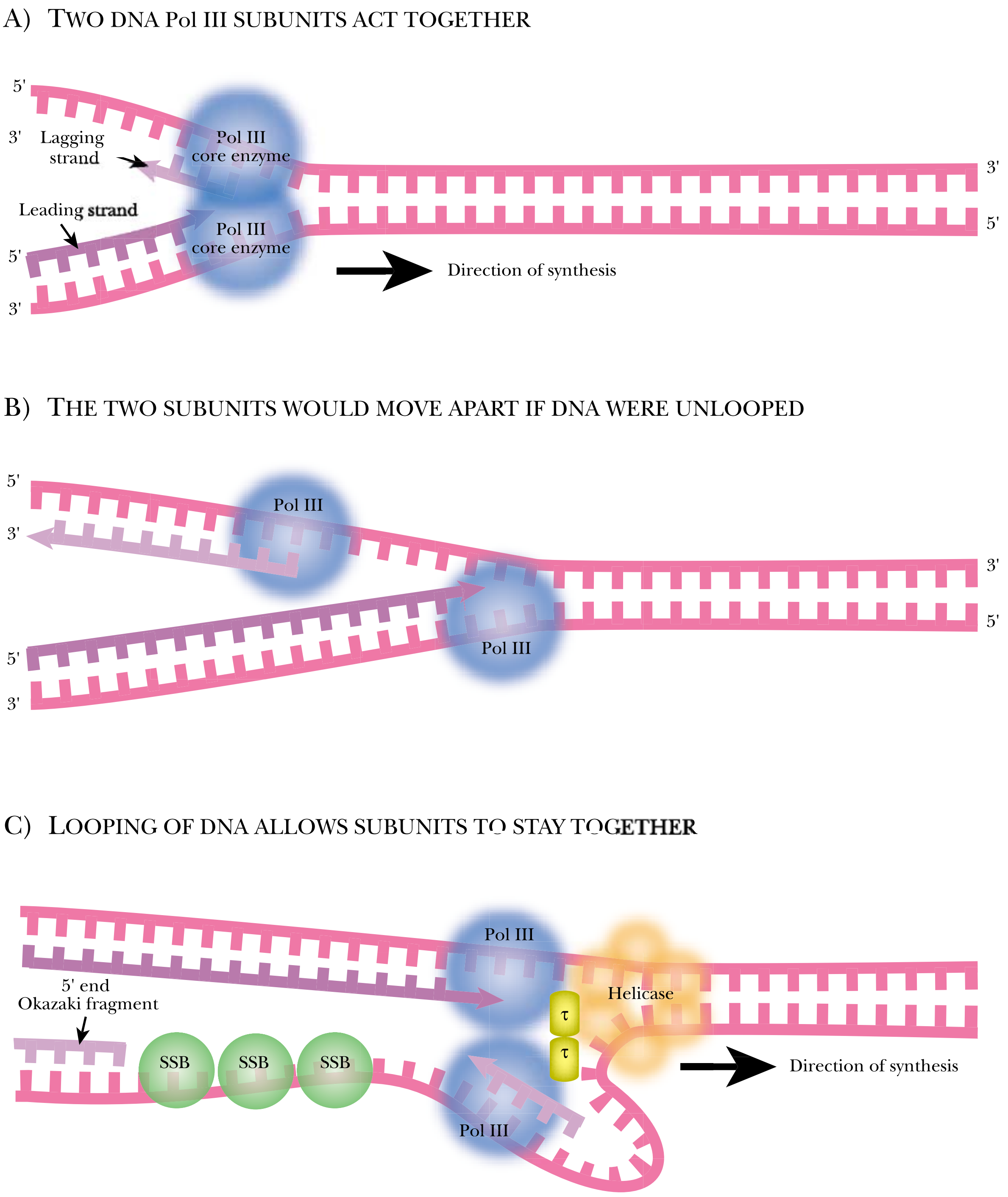


FIGURE 5.15 ***Relative Location of the Subunits of DNA Polymerase III at the Replication Fork***

Antiparallel synthesis dictates that the DNA polymerase III assemblies should change their relative positions as DNA synthesis occurs (compare A with B). Since the two assemblies are held by tau subunits, they cannot in fact move apart. Panel C suggests that the DNA is looped around to allow the DNA polymerase components to remain in contact.

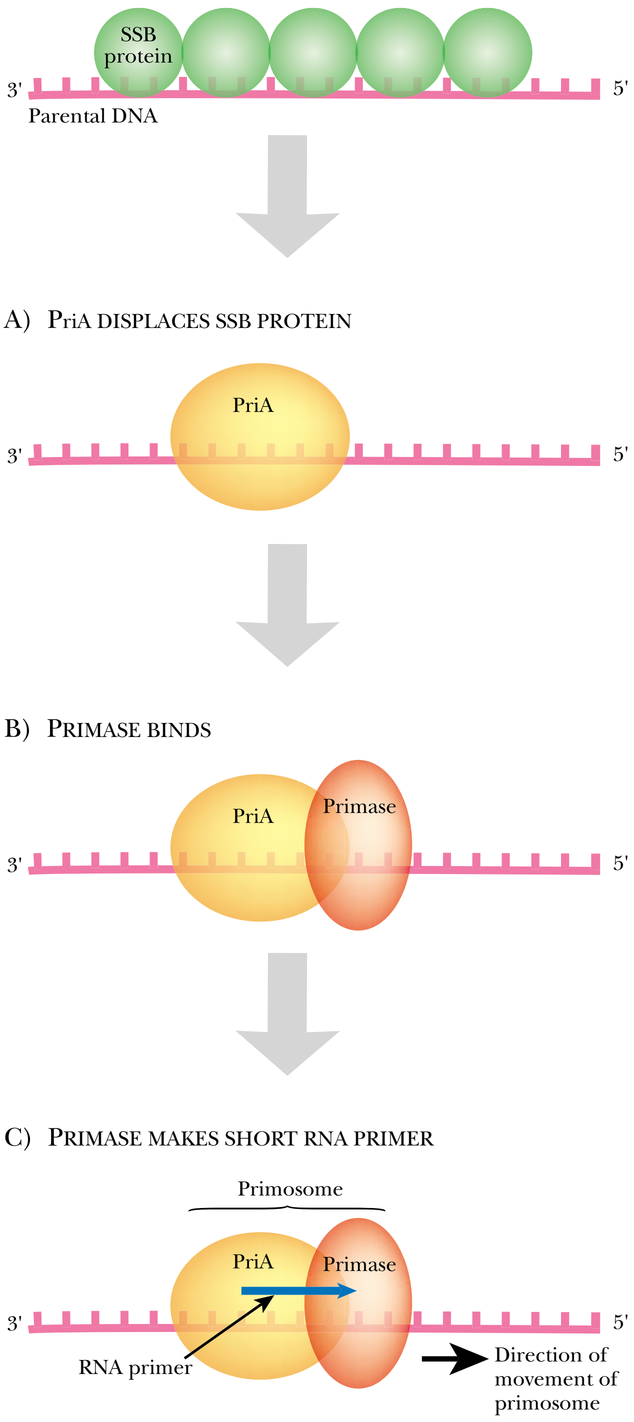


FIGURE 5.16 ***Three Steps in Starting a Primer for a New Okazaki Fragment***

Prior to primer formation, the bases of the parental DNA strand are covered with SSB proteins. A) First, the PriA protein displaces the SSB proteins. B) Second, a primase associates with the PriA protein. C) Lastly, the primase makes the short RNA primer needed to initiate the Okazaki fragment.

sliding clamp, but only a single clamp loader. This is because only the lagging strand needs constant clamp removal and reloading. It is worth repeating that in *E coli*, all of this happens about 1,000 times per second.

The fragments of the discontinuous lagging strand must be linked by removing the RNA primers, filling the gaps with DNA and, finally, joining the ends.

Completing the Lagging Strand

After the replication fork has passed by, the lagging strand is left as a series of Okazaki fragments with **gaps** (that is, spaces from which one or more nucleotides are missing)

gap A break in a strand of DNA or RNA where bases are missing

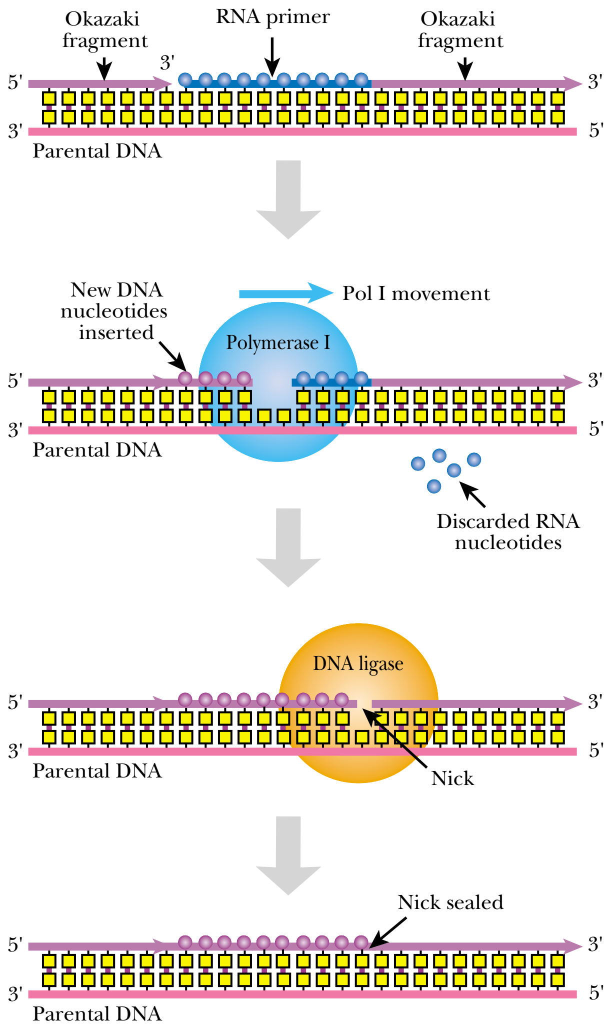


FIGURE 5.17 ***Three Steps in Joining the Okazaki Fragments***

When first made, the lagging strand is composed of alternating Okazaki fragments and RNA primers. The first step in replacing the RNA primer with DNA is the binding of DNA polymerase I to the primer region. As Pol I moves forward it degrades the RNA and replaces it with DNA. Lastly, DNA ligase seals the nick that remains.

between them. The gaps are filled with the RNA primer. Joining the Okazaki fragments to give a complete strand of DNA is accomplished by two—or perhaps three—enzymes working in succession: **Ribonuclease H** (**RNase H**), **DNA polymerase I** (**Pol I**), and **DNA ligase**. Only the last two enzymes are involved in the classical model. DNA polymerase I both degrades the RNA primers and fills the gaps left by the degraded RNA. Finally, DNA ligase joins the ends (Fig. 5.17). It has been suggested that in fact RNase H, which degrades the RNA strand of DNA:RNA double helixes, removes most of each RNA primer and that DNA polymerase I only removes the last few bases of the RNA primers.

Despite being a single polypeptide chain, DNA polymerase I possesses kinetic proofreading ability like Pol III. Pol I also has the unique ability to start replication at a nick in the DNA. The term "**nick**" refers to a break in the nucleic acid backbone with no missing nucleotides. When Pol I finds a nick, it cuts out a small stretch of DNA—or RNA—approximately 10 bases long. It then fills in the gap with new DNA. Pol I is important in completing the lagging strand as well as in DNA repair (see Ch.

DNA ligase Enzyme that joins up DNA fragments end to end
DNA polymerase I (Pol I) Bacterial enzyme that makes small stretches of DNA to fill in gaps between Okazaki fragments or during repair of damaged DNA
nick A break in the backbone of a DNA or RNA molecule (but where no bases are missing)
ribonuclease H (RNase H) Enzyme that degrades the RNA strand of DNA:RNA hybrid double helixes. In bacteria it removes the major portion of RNA primers used to initiate DNA synthesis.

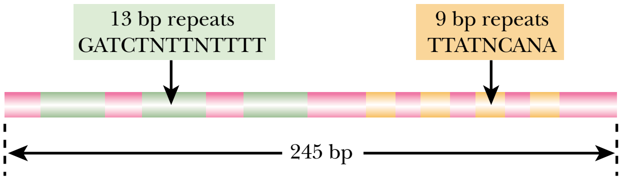


FIGURE 5.18 ***Sequences at the Origin of DNA Replication***

Sequence repeats at the origin of replication are of two varieties, both being AT-rich.

14). This property of Pol I is used in the laboratory for labeling small fragments of DNA with radioactive nucleotides by a procedure known as "**nick translation**." Nicks are introduced into one strand of the DNA by DNaseI (deoxyribonuclease I). DNA polymerase I then starts at the nick and moves along the DNA removing the nucleotides ahead of it and replacing them with the radioactive nucleotides provided. Both DNA polymerase I and DNA ligase have other important uses in genetic engineering (see Ch. 22).

DNA polymerases I and II were discovered before DNA polymerase III, hence the numbering. In retrospect, it is easy to understand why Pol III, with its complex requirements and multiple subunits, took longer to discover than the relatively simple enzymes Pol I and Pol II which are involved in DNA repair.

Chromosome Replication Initiates at *oriC*

Replication of a bacterial chromosome starts at a specific point, the origin of replication.

So far we have discussed the processes involved in the duplication of the DNA double helix. In addition, DNA replication must be synchronized with cell division. This involves starting replication at a specific location on the chromosome and stopping when the chromosome has been successfully copied. Prokaryotic DNA replication starts at a unique site on the chromosome, known as the **origin of replication**, and proceeds in both directions around the circle. The **initiation complex** contains five proteins: DnaA, DnaB, DnaC, gyrase and SSB. Of these, only DnaA is unique to chromosome initiation; the others are also involved in starting new Okazaki fragments. The origin, *oriC*, has three 13-base repeats, each consisting of GATCTNTTNTTTT, followed by four nine-base repeats, each consisting of TTATNCANA. Note that both sequences are AT-rich, a feature which aids strand separation. These are scattered over a 245 base pair region that is required for chromosome initiation (Fig. 5.18).

The first event is binding of **DnaA protein** to the four nine-base sequences. A cluster of 30 or so DnaA proteins is bound and the whole *oriC* region is wrapped around them. Next, the DnaA proteins open the DNA at all three of the 13-base repeats. Next to join is the DnaB helicase. Each hexamer of DnaB is at first accompanied by six DnaC proteins, needed to load DnaB onto the DNA. DnaB displaces DnaA from the single stranded 13-base repeats and begins to unwind the DNA and so create a replication fork (Fig. 5.19). A second DnaB hexamer creates a second replication fork moving in the opposite direction. DNA gyrase is also needed to allow unwinding and SSB proteins are needed to keep the DNA single stranded. DnaB also activates the primase, which then makes an RNA primer at each of the two points where the two leading strands are initiated.

Plasmids have been constructed in the laboratory that carry the chromosomal origin sequences of *E. coli*. This has allowed analysis of initiation in a cell-free system in which individual proteins are added to a DNA molecule of manageable size. The binding of the DnaA/DnaB/DnaC complex to *oriC* has been seen under the electron microscope (Fig. 5.20).

DnaA protein Protein that binds to the origin of bacterial chromosomes and helps initiate replication
initiation complex (for replication) Assemblage of proteins that binds to the origin and initiates replication of DNA
nick translation The removal of a short stretch of DNA or RNA, starting from a nick, and its replacement by newly made DNA
origin of replication Site on a DNA molecule where replication begins

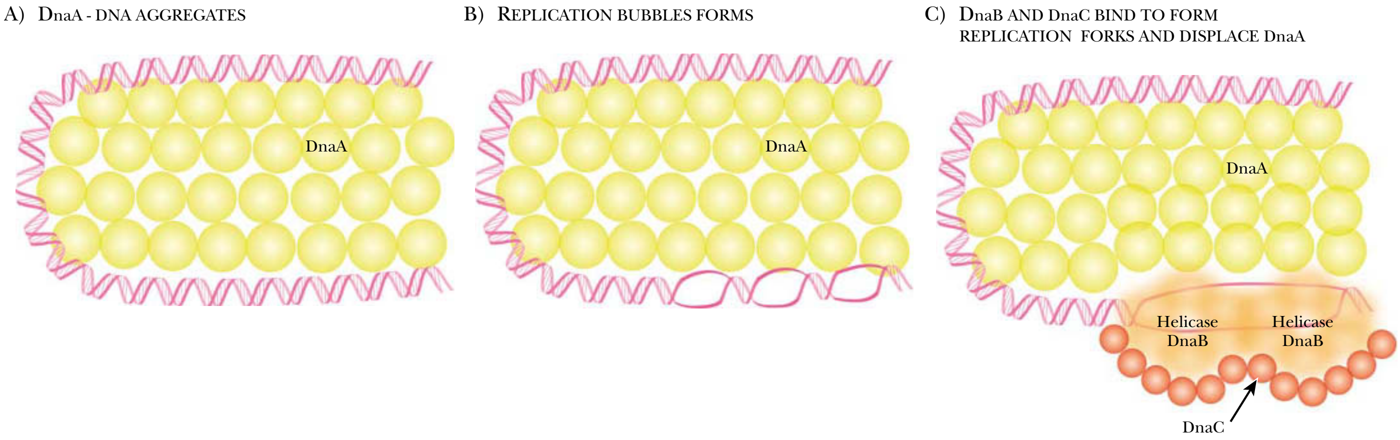

A) DnaA - DNA AGGREGATES
DnaA
B) REPLICATION BUBBLES FORMS
DnaA
C) DnaB AND DnaC BIND TO FORM REPLICATION FORKS AND DISPLACE DnaA
DnaA
Helicase DnaB
Helicase DnaB
DnaC

FIGURE 5.19 ***Three Steps in the Initiation of Replication by DnaA***

A) DnaA protein binds first to the four nine-base repeats, and then to the three 13-base repeats. B) As more DnaA binds, the DNA folds and the three 13-base repeats are unwound. C) Two complexes of DnaB and DnaC bind to the three 13-base repeats. This pushes DnaA away and causes the DNA strand to open all along the AT-rich region. The two DnaB complexes now start two replication forks, each headed in opposite directions around the circular DNA.

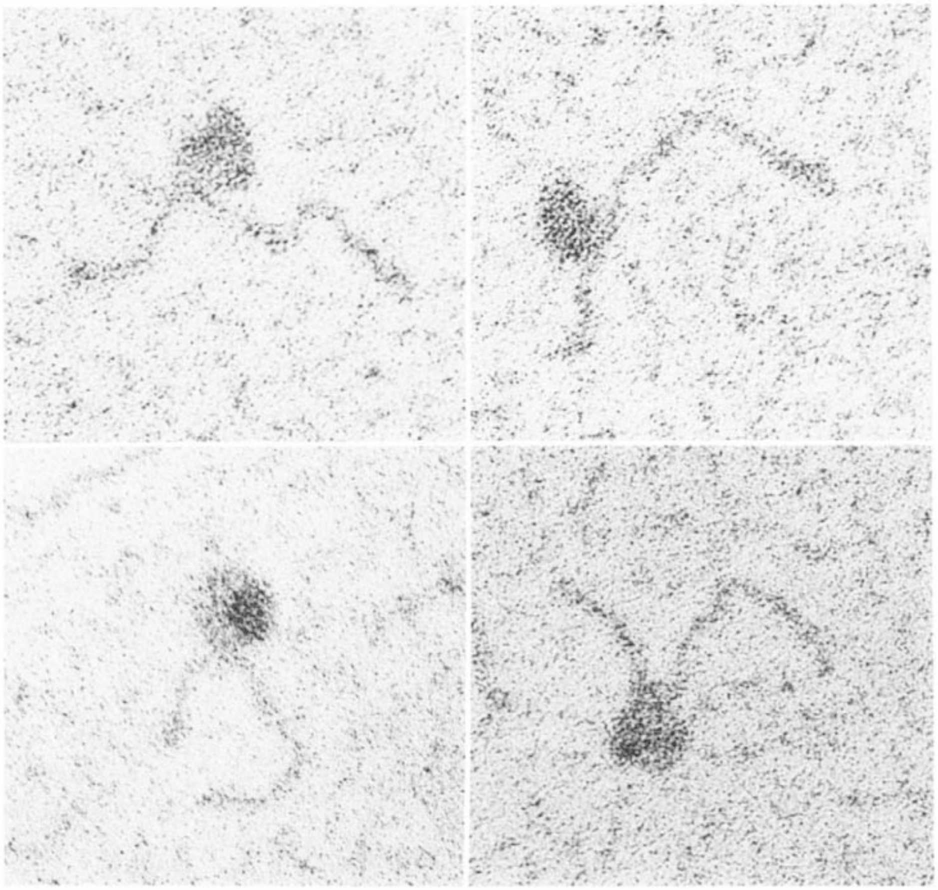


FIGURE 5.20 ***Initiation Complex Bound to* oriC**

The binding of the DnaA/DnaB/DnaC complex to *oriC* has been seen under the electron microscope. A plasmid carrying the *oriC* region was used instead of chromosomal DNA. Complexes were formed on supercoiled DNA of plasmid pCM959 with the proteins DnaA, DnaB, DnaC, plus HU. The complexes were cross-linked and the plasmid DNA was cut with *Ban1* to produce six fragments. The origin, *oriC,* was asymmetrically situated on the 703-bp fragment shown here. From: Funnell, Baker and Kornberg, In vitro assembly of a pre-priming complex at the origin of the *Escherichia coli* chromosome. Journal of Biological Chemistry, 262 (1987) 10327–10334.

TABLE 5.01 Proteins Involved in DNA Replication in *E. coli*

Protein	Gene	Function
DnaA	*dnaA*	Initiation of chromosome division; binds to the origin of replication
Helicase	*dnaB*	Unwinds the double helix
DnaC	*dnaC*	Loading of DNA helicase
SSB	*ssb*	Single strand binding protein
Primase	*dnaG*	Synthesis of RNA primers
RNase H	*rnhA*	Partial removal of RNA primers
Pol I	*polA*	Polymerase I; fills gaps between Okazaki fragments
Polymerase III		DNA polymerase III holoenzyme
α	*dnaE*	strand elongation
ε	*dnaQ*	kinetic proof-reading
θ	*holE*	unknown; part of core enzyme
β	*dnaN*	sliding clamp
τ	*dnaX*	dimerization of core enzyme
γ	*dnaX*	loading of sliding clamp
δ	*holA*	loading of sliding clamp
δ′	*holB*	loading of sliding clamp
χ	*holC*	loading of sliding clamp
ψ	*holD*	loading of sliding clamp
DNA Ligase	*lig*	Seals nicks in lagging strand
DNA Gyrase		Introduces negative supercoils
α	*gyrA*	Makes and seals double strand breaks in DNA
β	*gyrB*	ATP-using subunit
Topoisomerase IV		Decatenation
A	*parC*	Makes and seals double strand breaks in DNA
B	*parE*	ATP-using subunit

DNA Methylation and Attachment to the Membrane Control Initiation of Replication

The control of initiation, especially the correct timing of new rounds of replication, is still unclear. Two factors appear to be involved: methylation of DNA at the origin and attachment to the cell membrane. The *oriC* region contains a total of 11 GATC sequences, which are recognized by the Dam methylase (see Ch. 14). These **palindromic** sites are methylated on the adenine of both strands by the Dam methylase. Before replication, each GATC on the *E. coli* chromosome, including those in the origin of replication, will be fully methylated. Immediately after replication, the old strand is methylated but the new strand has not yet been methylated. The DNA is thus hemi-methylated.

The methylation state of the DNA is involved in controlling new rounds of DNA synthesis. But this is not the whole story.

Over most of the chromosome, full methylation is restored within a minute or two of replication (Fig. 5.21). This brief period allows for use of hemi-methylation as a guide for the mismatch repair system (see Ch. 14). However, the origin of replication is slow to regain full methylation; it takes 10 to 15 minutes. There is a corresponding lag in remethylating the promoter region of the *dnaA* gene. Transcription of this gene is repressed while hemi-methylated; consequently there is also a drop in the amount of DnaA protein available for initiation. Hemi-methylated sites cannot be

palindromic Reading the same backwards as forwards

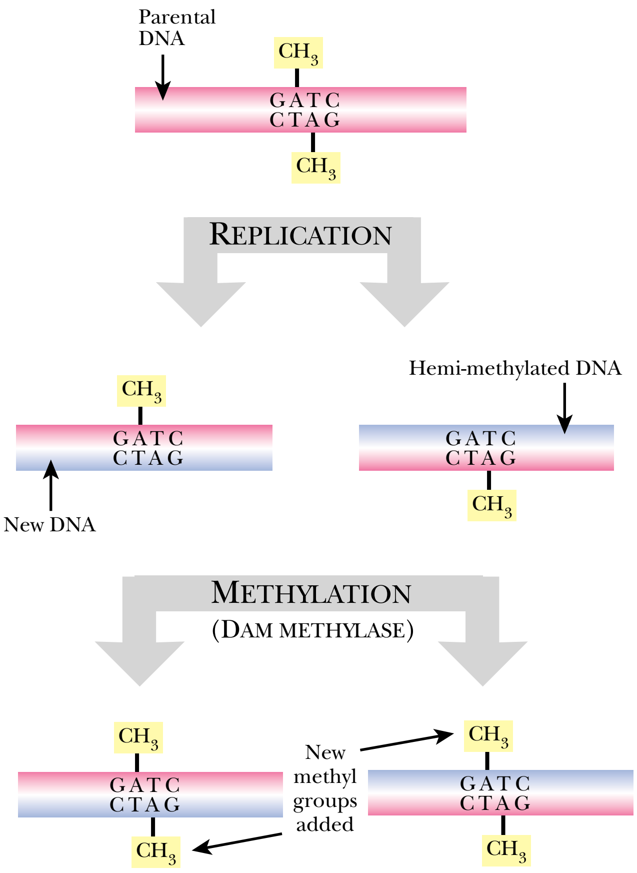


FIGURE 5.21 ***Methylation after DNA Duplication***

Dam methylase recognizes GATC palindromic sites and methylates them prior to DNA strand separation. The complementary sequences are synthesized during DNA replication but are not immediately methylated. Later, Dam methylase adds methyl groups to the newly synthesized sequences.

used to initiate chromosome replication. Hemi-methylated DNA binds to the cell membrane, helped by **SeqA** (**sequestration protein**), whereas fully methylated DNA does not.

The preceding implies that methylation and membrane binding are necessary controlling factors for initiation of replication. This is by no means the whole story, as *dam* mutants of *E. coli*, lacking the ability to add methyl groups, are viable and grow quite well. Thus, an origin with no methylation at all must be functional. Mutants lacking SeqA protein initiate replication more frequently than normal but are also viable. The factor(s) that control membrane binding are still obscure, as is the timing mechanism that oversees how long the origin is bound to the membrane and hidden from Dam methylase.

DNA replication finishes at special sites in the terminus region of the chromosome.

Chromosome Replication Terminates at *terC*

Replication finishes when the two replication forks meet at the **terminus** of the chromosome. This region is surrounded by several ***Ter* sites** that prevent further movement of replication forks (Fig. 5.22). The *Ter* sites act asymmetrically. *TerC, TerB* and *TerF* prevent clockwise movement of forks and *TerA, TerD* and *TerE* prevent counterclockwise movement. Since replication proceeds in two directions in prokaryotes, the meeting of the two replication forks is prevented by sets of *Ter* proteins at the termination region. The two innermost sites (*TerA* and *TerC*) are most frequently used and the outer sites presumably serve as back-ups in case a fork manages to slip

sequestration protein (SeqA) Protein that binds the origin of replication, thereby delaying its methylation
***Ter* site** Site in the terminus region that blocks movement of a replication fork
terminus Region on a chromosome where replication finishes

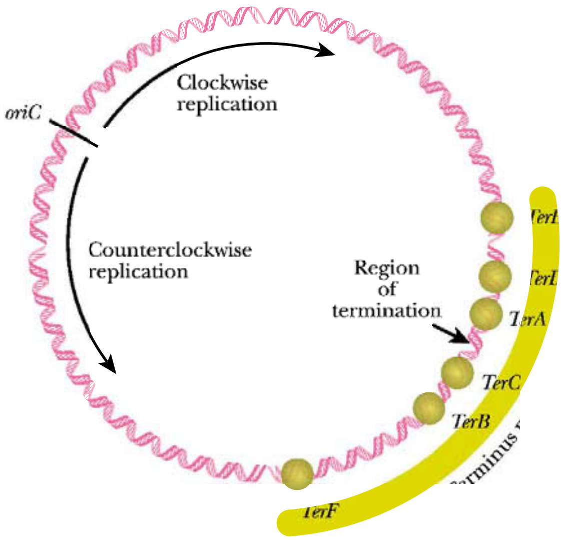


FIGURE 5.22 ***Termination of replication by Tus and Ter Sites***

The circular bacterial chromosome has a termination region, or terminus, with several sites that stop replication forks moving clockwise (*TerF, TerB* and *TerC*) and counterclockwise (*TerE, TerD* and *TerA*).

past *TerA* or *TerC*. For example, the clockwise replication fork might pass through *TerE*, *TerD*, and *TerA* sites with no problem, until it is halted by *TerC*.

The *Ter* sites have a 23 bp consensus sequence that binds **Tus protein**. This blocks the movement of the DnaB helicase and brings movement of the replication fork to a halt. Tus binds asymmetrically and stops movement from one direction only. It can be displaced from the DNA by a fork coming from the other direction. However, the whole terminus region of *E. coli* (including the gene for the Tus protein, which is located next to *TerB*) can be deleted without apparent ill effects. This implies that the replication forks do not need to meet at a *Ter* site and can terminate successfully wherever they collide.

Disentangling the Daughter Chromosomes

When a circular chromosome finishes replicating, the two new circles may be physically interlocked or *catenated* (see Ch. 4). Such catenanes must be separated so that each daughter cell receives a single chromosome upon cell division (Fig. 5.23). Decatenation of interlocked circles is carried out by topoisomerase IV. Although the terminology is confusing, Topo IV is actually a type II topoisomerase whose mode of action is similar to DNA gyrase. Topo IV is found behind the replication fork, where it untangles the newly formed DNA as replication proceeds. It can also decatenate finished DNA circles, both of chromosomes and plasmids.

If circular molecules of DNA become interlocked, specific enzymes are needed to untangle them.

A related problem sometimes results from recombination. The two growing circular chromosomes may recombine even during the process of replication. Each pair of crossovers or exchanges of genetic material may cause the growing circles to become interlocked. If there is an even number of crossovers, Topo IV can disentangle the circles and no harm is done. However, an odd number of crossovers will covalently join the two circles of DNA (Fig. 5.24). The covalent dimer must be separated by the **crossover resolvase**, XerCD, which uses the ***dif* sites** on the two chromosomes to introduce a final crossover. This gives, in effect, an even number of crossovers.

crossover resolvase Bacterial enzyme that separates covalently fused chromosomes
***dif* site** Site on bacterial chromosome used by crossover resolvase to separate covalently fused chromosomes
Tus protein Bacterial protein that binds to *Ter* sites and blocks movement of replication forks

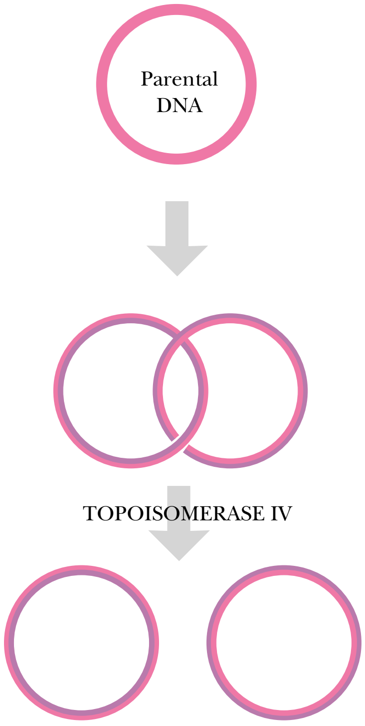


FIGURE 5.23 ***Decatenation by Topoisomerase IV***

Topoisomerase IV aids in unlinking two newly replicated circles of DNA. Replication of the circular bacterial chromosome may give two catenated circles of DNA. Topo IV unlinks these to give decatenated circles.

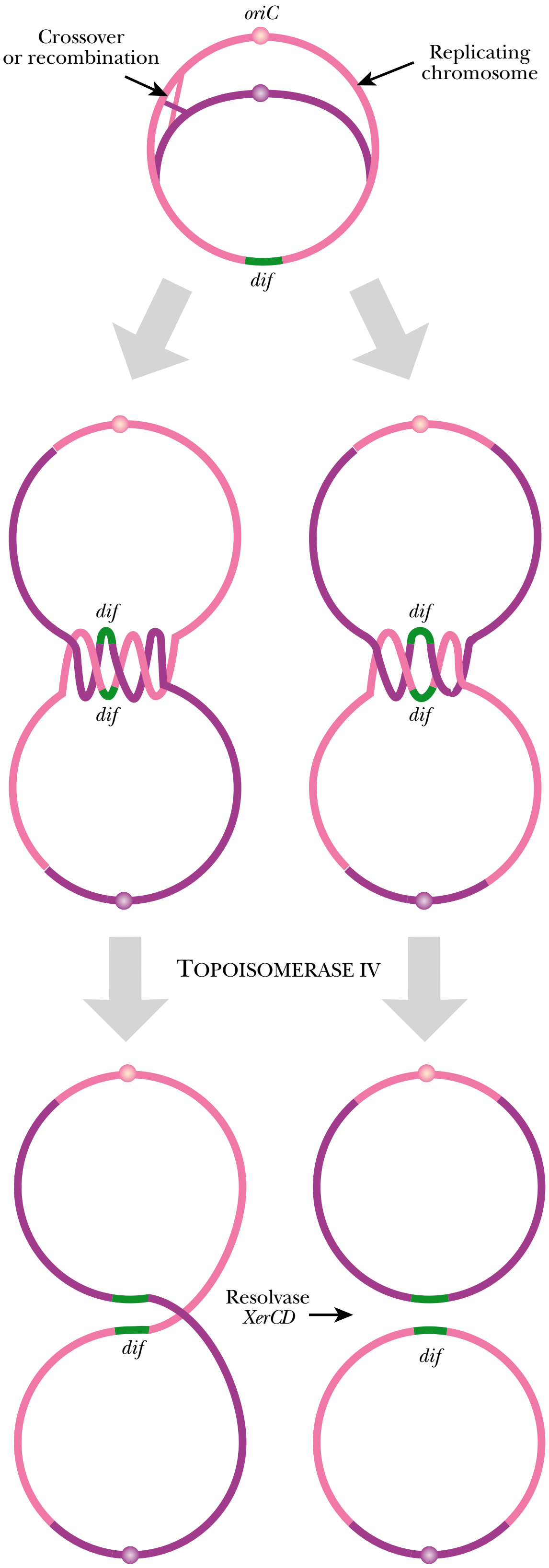


FIGURE 5.24 ***Recombination Causes Problems***

The number of recombination sites determines how chromosomes are unwound. TopoIV can readily untangle circles with even numbers of crossovers, whereas Resolvase XerCD is additionally needed to separate circles with odd numbers of crossovers.

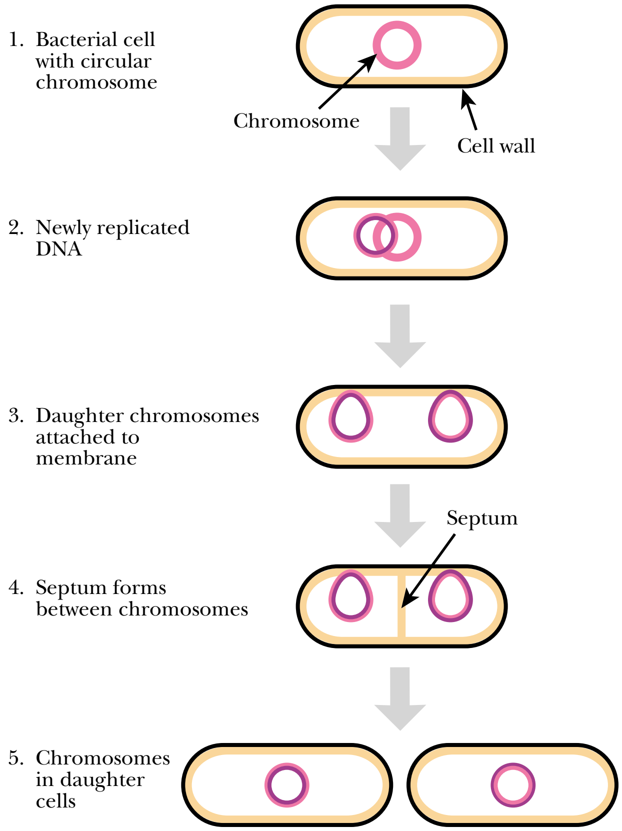

1. Bacterial cell with circular chromosome
Chromosome
Cell wall
2. Newly replicated DNA
3. Daughter chromosomes attached to membrane
Septum
4. Septum forms between chromosomes
5. Chromosomes in daughter cells

FIGURE 5.25 ***Elongation of the Cell Separates Chromosomes***

Segregation of chromosomes is caused by elongation of the cell. Subsequently, a partition, or septum, is formed that completes cell division.

Cell Division in Bacteria Occurs after Replication of Chromosomes

Bacterial cells grow longer and replicate their DNA simultaneously. Then they divide.

Bacteria divide by **binary fission**, or splitting. Bacteria have only one chromosome, which lies within the cytoplasm. Bacterial cell division is thus relatively simple and may be divided for convenience into four stages, although some of the processes overlap slightly:

1. Replication of the chromosome
2. Partition of the daughter chromosomes
3. Cell elongation
4. Separation of the two daughter cells by formation of a cross-wall.

Replication proceeds in both directions at once around the circular bacterial chromosome. Eventually, the two replication forks meet and merge, yielding two new circular chromosomes. These are attached to the cell membrane, probably at their origins, as suggested above. As the cell elongates, the chromosomes are pulled apart (Fig. 5.25). The final step of cell division is the formation of a cross-wall, or **septum**.

How Long Does It Take for Bacteria to Replicate?

The time required for an *E. coli* cell to divide, the **generation time**, ranges from 20 minutes to several hours, depending on the conditions. Despite this, duplication of the

binary fission Simple form of cell division, by splitting down the middle, found among bacteria
generation time The time from the start of one cell division to the start of the next
septum Cross-wall that separates two new bacterial cells after division

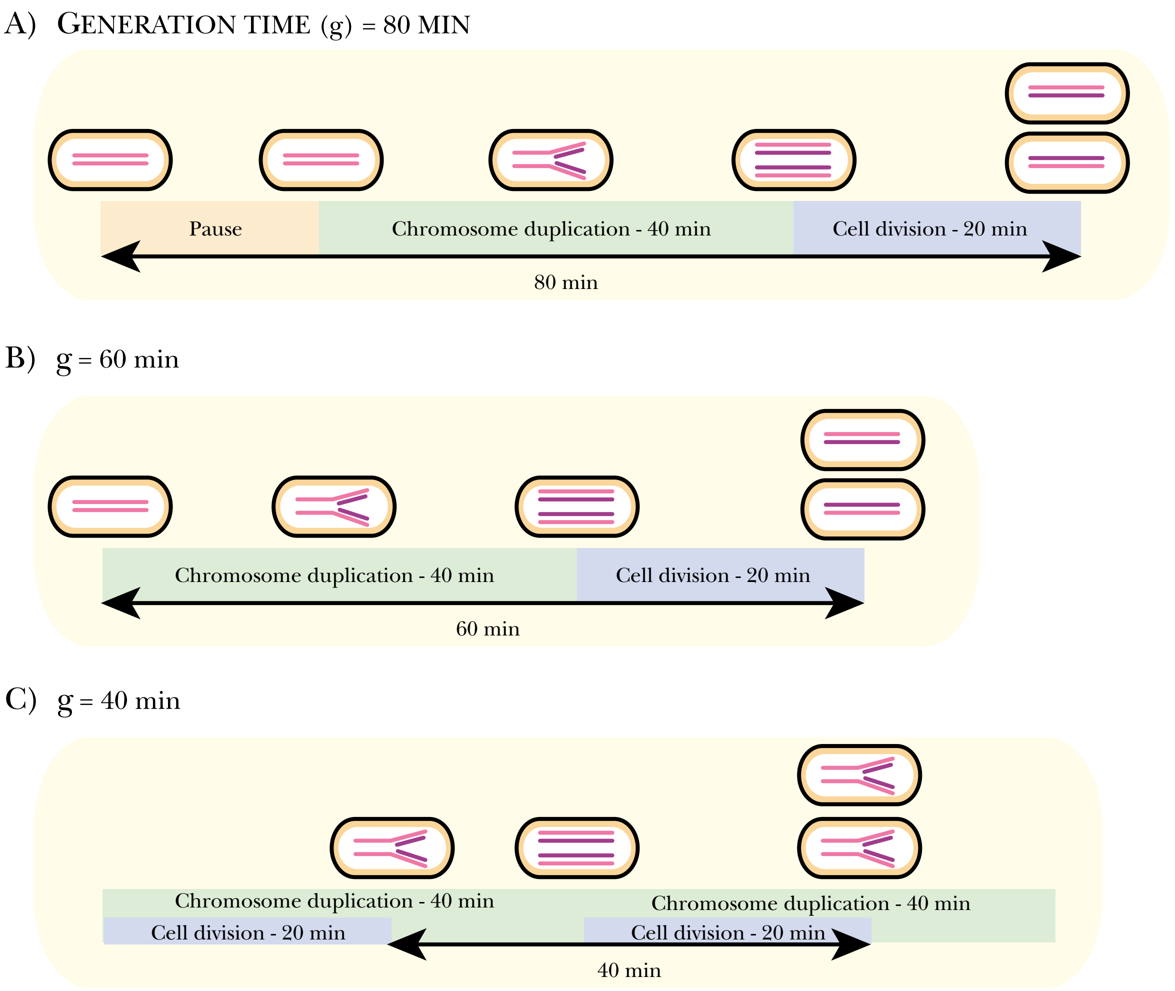

A) GENERATION TIME (g) = 80 MIN
Pause
Chromosome duplication - 40 min
Cell division - 20 min
80 min
B) g = 60 min
Chromosome duplication - 40 min
Cell division - 20 min
60 min
C) g = 40 min
Chromosome duplication - 40 min
Cell division - 20 min
Chromosome duplication - 40 min
Cell division - 20 min
40 min

FIGURE 5.26 ***Cell Division and Chromosome Replication***

A) Cell division in more than 60 minutes allows for a time gap after cell division and the start of the next DNA replication. B) Cell division in 60 minutes requires 20 minutes for replication and 40 minutes for completion of cell division. C) Cell division in less than 60 minutes requires that a new cycle of replication be initiated before the last replication is completed.

chromosome always takes 40 minutes and the time from termination of replication to completion of cell division takes 20 minutes. If the generation time is less than 60 minutes, one or more rounds of chromosome replication must overlap. This means that a new cycle of replication may start before the previous one has finished. Cells in rapidly dividing cultures of bacteria therefore contain multiple but incomplete copies of the chromosome. If the generation time is longer than 60 minutes, there is a gap between division of the cell and initiation of the next round of chromosome replication (Fig. 5.26).

The Concept of the Replicon

Nucleic acid molecules that survive and divide must have both an origin of replication and be circular (or have ends that are protected).

A **replicon** is any DNA (or RNA) molecule that is capable of surviving and replicating itself inside a cell. A replicon must possess an origin of replication where replication is initiated. A replicon must also be an intact, "complete" molecule of DNA (or RNA) with ends that are protected from being attacked by a cell's defense system.

replicon Molecule of DNA or RNA that contains an origin of replication and can self-replicate

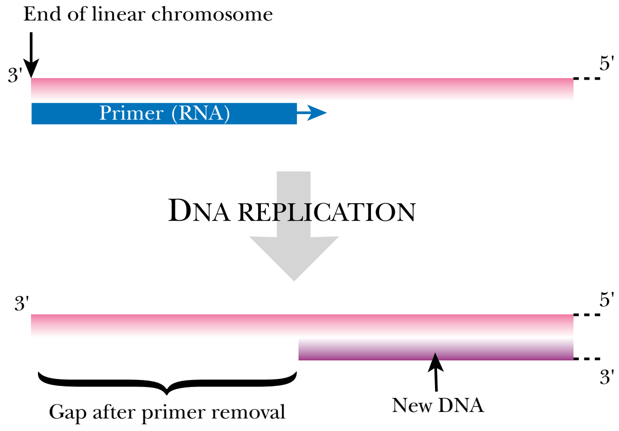


FIGURE 5.27 *The 5′-End is Potentially Lost in Replication of Linear DNA*

When an RNA primer is removed after initiating a strand of linear DNA the gap cannot be filled by DNA as there is no upstream 3′ hydroxyl to accept nucleotides. Thus, linear DNA would be shortened during each replication cycle.

Although chromosomes are clearly replicons, they are by no means the only ones. Virus genomes replicate when inside their host cell. Consequently, their nucleic acid qualifies as a replicon. Since some virus genomes consist of RNA, this means that the definition of replicon must include both DNA and RNA.

In prokaryotes, replicons are usually closed circles of DNA that have no ends. In most bacteria, linear molecules of DNA are degraded by **exonucleases**. These are enzymes that degrade nucleic acids one nucleotide at a time, starting from one end or the other. Consequently, linear segments of DNA that enter a bacterial cell during conjugation or transformation (see Ch. 18) will eventually be degraded. If some of the genetic information carried on such DNA is to survive, it must recombine onto a circular replicon, similar in form to bacterial chromosome.

Despite this, a few bacteria contain linear chromosomes. These have a variety of individual adaptations to protect the ends from endonucleases. *Borrelia burgdorferi*, which causes **Lyme disease**, has hairpin sequences at the ends of its linear chromosome. *Streptomyces lividans*, a soil organism, has proteins covalently attached to the ends of its DNA.

Extra DNA molecules, known as plasmids, are found in many bacteria. They are usually circular and much smaller than chromosomes.

Plasmids are another group of replicons. They are extra self-replicating molecules of DNA that are not necessary for survival of the host cell (see Ch. 16). Plasmids are usually circular, although linear plasmids occur in *Borrelia* and *Streptomyces*, the same bacteria that contain linear chromosomes. Circular replicons are occasionally found in eukaryotic cells, including plasmids such as the 2μ circle of yeast. Mitochondria and chloroplasts also contain their own genomes, or replicons, which are circular molecules of self-replicating DNA.

Replicating Linear DNA in Eukaryotes

Replicating a linear molecule of DNA requires certain adaptations. Since DNA polymerases can only elongate, not initiate, new strands of DNA must be initiated with an RNA primer. Since synthesis always proceeds from 5′ to 3′, one of these RNA primers must be located right at the 5′-end of each new strand when replicating linear DNA (Fig. 5.27). When this terminal RNA primer is removed, it cannot be replaced with DNA, as there is no strand for DNA polymerase to elongate. If nothing was done to overcome this problem, the molecule of DNA would grow shorter, by the length of an average RNA primer, with each round of replication. Circular prokaryotic DNA molecules do not have ends and so do not have this problem.

exonuclease Enzyme that cleaves nucleic acid molecules at the end and usually removes just a single nucleotide
Lyme disease Infection caused by *Borrelia burgdorferii* and transmitted by ticks
plasmid Accessory molecule of nucleic acid capable of self-replication. Does not normally carry genes needed for existence of host cell. Usually consists of double stranded circular DNA but occasional plasmids that are linear or made of RNA exist.

One solution to the problem of initiation of replication in linear DNA is to use a **protein primer** at the ends. We normally think of DNA polymerase as only being able to extend nucleic acid chains. However, strictly speaking, DNA polymerase can add nucleotides only to a free —OH group. Although this free —OH group is normally furnished by DNA itself, or by an RNA primer, some DNA polymerases *can* add nucleotides to a free —OH group on specific proteins. This solution is used by several viruses and for the linear plasmids and chromosomes of *Streptomyces*.

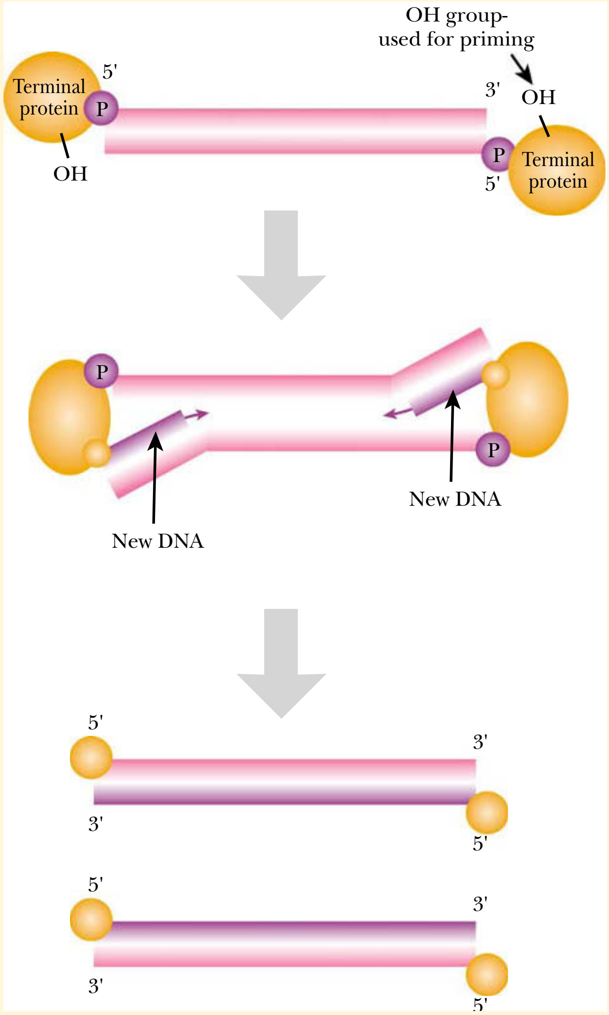


FIGURE 5.28 ***Protein Primers for the Ends of Linear DNA***

The terminal proteins of some viruses and plasmids bind to the 5′-end of linear DNA. These proteins have special OH groups that allow priming of DNA synthesis. The result is complete replication of the linear DNA, without "end-shortening."

Eukaryotic DNA is linear and needs special structures, the telomeres, to protect its ends.

Eukaryotes have adapted to the problem of replicating linear DNA by developing structures known as **telomeres** that are located at the ends of their chromosomes. Telomeres consist of multiple tandem repeats (from 20 to several hundred) of a short sequence, usually of six bases (TTAGGG, in vertebrates including humans). During

protein primer Protein used instead of RNA as a primer for DNA synthesis in some bacteria and viruses
telomere Specific repetitive sequence of DNA found at the end of linear eukaryotic chromosomes

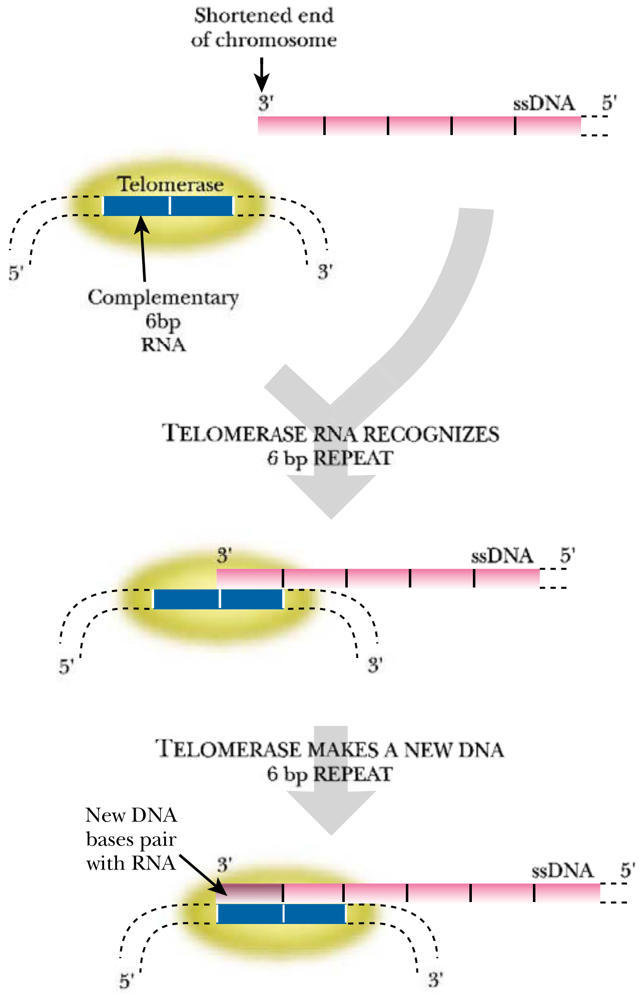


FIGURE 5.29 ***Telomerase Replaces Repeats at the Ends of Chromosomes***

Telomerase RNA recognizes the tandem repeats at the end of linear DNA. The RNA of telomerase sticks out beyond the chromosome ends and serves as a template for addition of new DNA repeats that will repair the segment lost during the last DNA replication.

each replication cycle, the chromosomes are, in fact, shortened and several of the telomere repeats are lost. However, no unique coding information is lost. Furthermore, in cells where the enzyme **telomerase** is present, the lost DNA is later replaced by adding several of the six-base-pair units to the 3′-end after each replication cycle (Fig. 5.29). Telomerase carries a small segment of RNA, complementary to the six-base-pair telomere repeat. This allows it to recognize the telomeres and reminds it what sequence to add.

After telomerase has elongated the 3′-ends, the complementary strand can be filled in by normal RNA priming followed by elongation by DNA polymerase and joining by **ligase**. The telomere repeats also protect the ends of chromosomes against degradation by exonucleases.

Telomere repeat sequences have been remarkably conserved throughout evolution although some variation is seen. The characteristic TTAGGG repeat of vertebrates is also found in the protozoan *Trypanosoma*, while the sequence in the protozoans *Paramecium* and *Tetrahymena*, TTGGGG, differs by only one base. Many insects have the five base repeat, TTAGG, whereas the flowering plant *Arabidopsis* has a seven base repeat, TTTAGGG. However, recent data indicates that this is not typical of all plants, indeed, several monocotyledonous plants have the same TTAGGG repeat as vertebrates. Among the fungi *Aspergillus nidulans* has TTAGGG whereas its close relative

ligase Enzyme that joins up DNA fragments end to end
telomerase Enzyme that adds DNA to the telomere of a eukaryotic chromosome

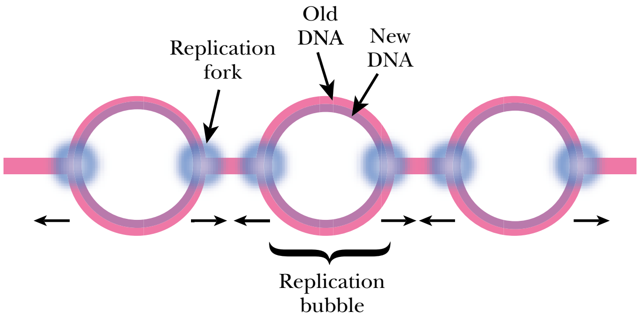


FIGURE 5.30 *Eukaryotic Chromosome Replication Bubbles*

Numerous openings in the DNA, or replication bubbles occur at the sites of replication in eukaryotic chromosomes. The longer replication continues, the larger the bubbles. The bubbles eventually merge together which separates the newly replicated DNA molecules (not shown).

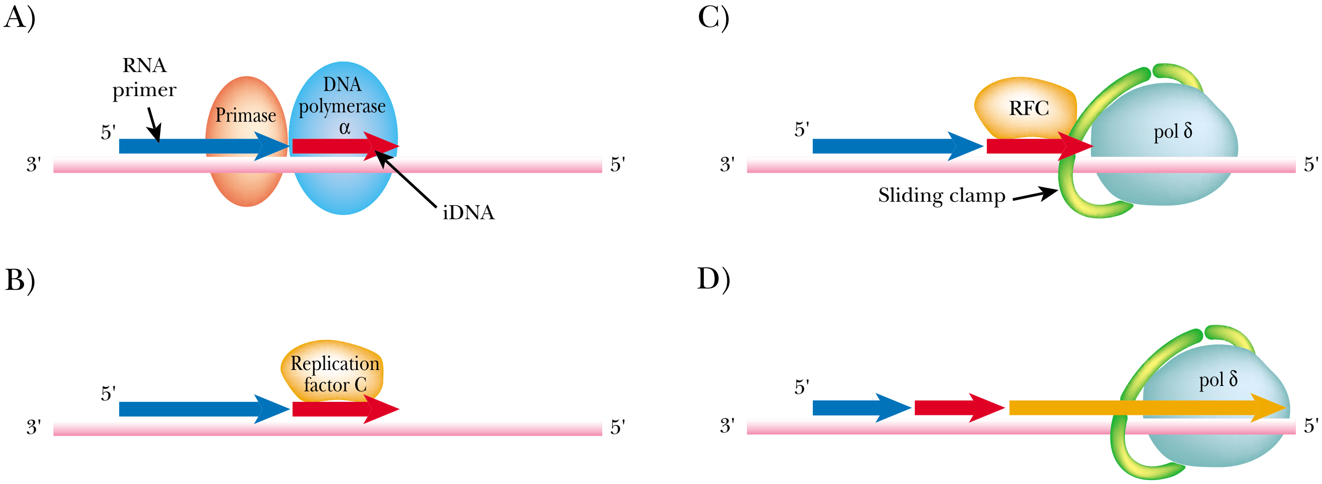


FIGURE 5.31 *Eukaryotic DNA Replication*

A) The process begins with the binding of a primase which produces an RNA primer. Subsequently, DNA polymerase α binds and initiates a short segment of DNA called initiator DNA (iDNA).
B) Replication factor C associates with the iDNA and is C) helpful in positioning DNA polymerase δ.
D) DNA polymerase δ elongates the new DNA strand.

Aspergillus oryzae has a double-length repeat—TTAGGGTCAACA. One strange exception to this general pattern is the fruit fly, *Drosophila*, which has telomeres consisting of tandem sequences generated by successive transposition of two retrotransposons (HeT-A and TART) instead of being synthesized by telomerase.

Eukaryotic Chromosomes Have Multiple Origins

Eukaryotic chromosomes are much longer than bacterial ones and have multiple replication origins.

Eukaryotic chromosomes are often very long and have numerous replication origins scattered along each chromosome. Replication is bi-directional, as in bacteria. A pair of replication forks starts at each origin of replication and the two forks then move in opposite directions (Fig. 5.30). The bulges where the DNA is in the process of division are often called **replication bubbles**.

A vast number of replication origins function simultaneously during eukaryotic DNA replication. For example, there are estimated to be between 10,000 and 100,000 replication origins in a dividing human somatic cell. This creates major problems in synchronization. Synthesis at each origin must be coordinated to make sure that each chromosome is completely replicated. Conversely, each origin must initiate once and once only during each replication cycle in order to avoid duplication of DNA segments

replication bubble (replication eye) Bulge where DNA is in the process of replication that have already been replicated. This is achieved by a protein complex, known as replication licensing factor (RLF), which binds to the DNA next to each origin before each replication cycle and is displaced during replication. Only when RLF is present is DNA replication permitted.

Synthesis of Eukaryotic DNA

Synthesis of eukaryotic DNA resembles that in bacteria in most general respects:
a) it is semi-conservative
b) one strand is made in short fragments
c) RNA primers are needed to start new strands.

The synthesis of DNA in eukaryotes is less well investigated than in bacteria. Nonetheless, the same general principles apply, although there are differences in detail from the bacterial scheme. In eukaryotes, **semi-conservative replication** occurs. One new strand is made continuously and the other in fragments. Both strands are made simultaneously by a replisome consisting of a helicase plus two DNA polymerase assemblies. A sliding clamp holds the polymerase on the DNA. An RNA primer is required.

In animal cells, two DNA polymerases (α and δ) are involved in chromosome replication (Fig. 5.31). DNA **polymerase α** is responsible for initiation of new strands. It is accompanied by two smaller proteins that make the RNA primer. After the RNA primer has been made, polymerase α elongates it by a short piece of DNA only three or four bases long (the **initiator DNA**, or "iDNA").

Another protein, **Replication factor C** (**RFC**), then binds to the iDNA and loads DNA **polymerase δ** plus its sliding clamp (**PCNA protein**) onto the DNA. Two assemblies of DNA polymerase δ elongate the two new strands. The sliding clamp of animal cells is a trimer (not a dimer as in bacteria) that forms a ring surrounding the DNA. It was named PCNA, for proliferating cell nuclear antigen, before its role was fully known.

Linking of the Okazaki fragments differs significantly between animal and bacterial cells. In animals, there is no equivalent of the dual function polymerase I of bacteria. The RNA primers are removed by an exonuclease (MF1) and the gaps are filled by the DNA polymerase δ that is working on the lagging strand. As in bacteria, the nicks are sealed by DNA ligase.

Cell Division in Higher Organisms

The presence of a nucleus complicates cell division in eukaryotes. The result is a complex cell cycle that includes dissolution and reassembly of the nucleus as well as duplication of the chromosomes.

The eukaryotic cells of higher organisms face further problems during cell division. Not only do they have multiple chromosomes, but these are inside the nucleus, separated from the rest of the cell by the nuclear membrane. Consequently, an elaborate process is needed to disassemble the nucleus, replicate the chromosomes and partition them among the daughter cells. This process is mitosis and involves several operations:

1. Disassembly of the nuclear membrane of the mother cell
2. Division of the chromosomes
3. Partition of the chromosomes
4. Reassembly of nuclear membranes around each of the two sets of chromosomes
5. Final division of the mother cell, or **cytokinesis**.

cytokinesis Cell division
DNA polymerase α Enzyme that makes short segment of initiator DNA during replication of animal chromosomes
DNA polymerase δ Enzyme that makes most of the DNA when animal chromosomes are replicated
initiator DNA (iDNA) Short segment of DNA made just after the RNA primer during replication of animal chromosomes
PCNA protein The sliding clamp for the DNA polymerase of eukaryotic cells (PCNA = proliferating cell nuclear antigen)
replication factor C (RFC) Eukaryotic protein that binds to initiator DNA and loads DNA polymerase δ plus its sliding clamp onto the DNA
semi-conservative replication Mode of DNA replication in which each daughter molecule gets one of the two original strands and one new complementary strand

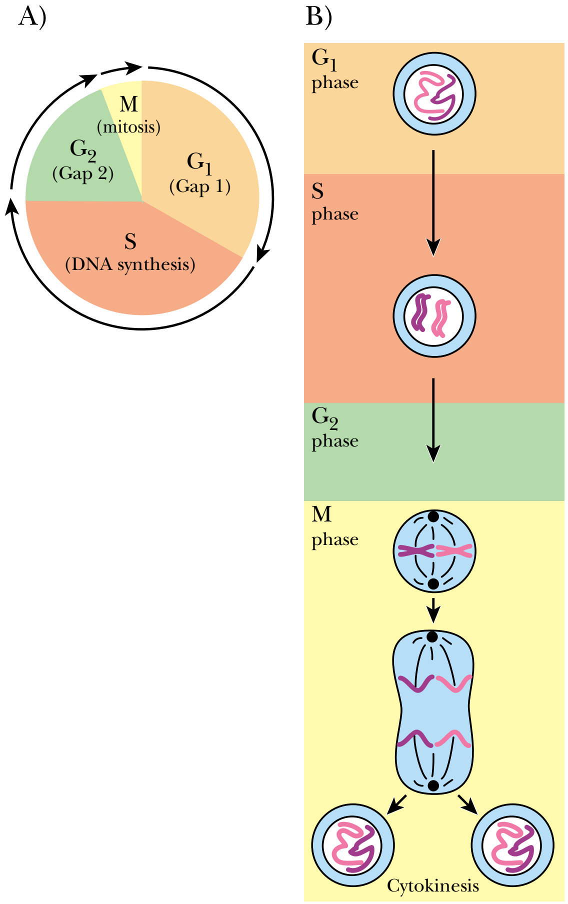


FIGURE 5.32 ***The Eukaryotic Cell Cycle***

DNA replication occurs during the S phase of the cell cycle but the chromosomes are actually separated later, during mitosis or M phase. The S and M phases are separated by G1 and G2.

Mitosis itself is only one of several phases of the eukaryotic **cell cycle** (Fig. 5.32). The process of DNA replication described above takes place in the synthetic, or **S-phase**, of the cell cycle. The S-phase is separated from the actual physical process of cell division (**mitosis**) by two gap phases, or **G-phases**, in which nothing much appears to happen (except the normal processes of cellular activity and metabolism). Together, G1, S and G2 constitute **interphase**.

cell cycle Series of stages that a cell goes through from one cell division to the next
G1 phase Stage of the eukaryotic cell cycle following cell division; cell growth occurs here
G2 phase Stage of the eukaryotic cell cycle between DNA synthesis and mitosis: preparation for division
interphase Part of the eukaryotic cell cycle between two cell divisions and consisting of G1-, S- and G2- phases
mitosis Division of eukaryotic cell into two daughter cells with identical sets of chromosomes
S-phase Stage in the eukaryotic cell cycle in which chromosomes are duplicated

CHAPTER SIX

Transcription of Genes

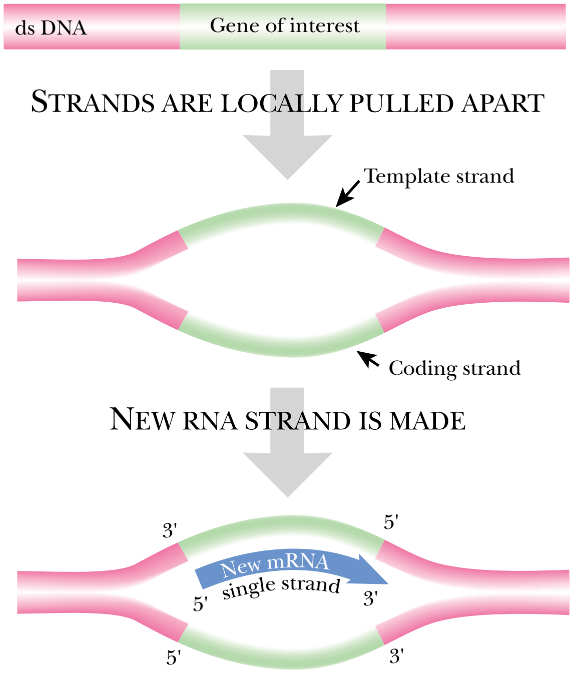


FIGURE 6.01 ***Transcription in Its Simplest Form***

The two strands of the DNA to be transcribed are separated locally. The top strand serves as a template for building a new RNA molecule.

Genes are Expressed by Making RNA

DNA "merely" stores genetic information. Putting the information to use requires RNA and (usually) protein.

Messenger RNA carries the information for making proteins from the genes to the cytoplasm.

The DNA double helix must be opened up for RNA polymerase to read the template strand and make RNA.

For a cell to operate, its genes must be expressed. The word "expressed" means that the gene products, whether proteins or RNA molecules, must be made. The DNA molecule that carries the original copy of the genetic information is used to store genetic information but is not used as a direct source of instructions to run the cell. Instead, working copies of the genes, made of RNA, are used. The transfer of information from DNA to RNA is known as **transcription** and RNA molecules are therefore sometimes referred to as transcripts. Genes may be subdivided into two major groups: those whose final product is an RNA molecule (e.g., tRNA, rRNA, assorted regulatory RNAs—see below) and those whose final product is protein. In the latter case, the RNA transcript acts as an intermediary and a further step is needed to convert the information carried by the RNA to protein. This process is discussed in Ch. 8. The type of RNA molecule that carries genetic information encoding a protein from the genes into the rest of the cell is known as **messenger RNA**, or mRNA. Since the great majority of genes encode proteins, we will deal with these genes first.

For a gene to be transcribed, the DNA, which is double stranded, must first be pulled apart temporarily, as shown in Figure 6.01. Then, RNA is made by **RNA polymerase**. This enzyme binds to the DNA at the start of a gene and opens the double helix. Finally, it manufactures an RNA molecule.

The sequence of the RNA message is complementary to the **template strand** of the DNA from which it is synthesized. Apart from the replacement of thymine in DNA with uracil in RNA, this means that the sequence of the new RNA molecule is identical to the sequence of the **coding strand** of DNA; that is, the strand not actually used as a template during transcription. Note that RNA, like DNA, is synthesized in the 5′- to 3′- direction (Fig. 6.02). Other names for the template strand are the *non-coding* or *anti-sense* strand; other names for the coding strand are *non-template* or *sense* strand. Only one of the strands of DNA is copied in any given transcribed region. [But note

coding strand The strand of DNA equivalent in sequence to the messenger RNA (same as plus strand)
messenger RNA (mRNA) The molecule that carries genetic information from the genes to the rest of the cell
RNA polymerase Enzyme that synthesizes RNA using a DNA template
transcription Process by which information from DNA is converted into its RNA equivalent
template strand Strand of DNA used as a guide for synthesizing a new strand by complementary base pairing

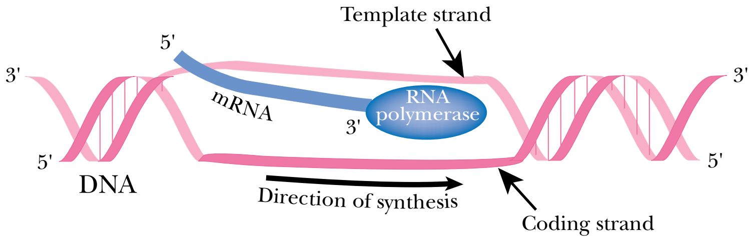


FIGURE 6.02 ***Naming the Basic Components Involved in Transcription***

The DNA is shown in its double helical form. After local separation of the strands, the new RNA is synthesized so that it base pairs with one of the DNA strands—the template strand. The other DNA strand is inactive and is called the coding strand. The enzyme RNA polymerase synthesizes single-stranded RNA in the 5′- to 3′-direction. The sequence of bases in the RNA is the same as in the coding strand of DNA, except that uracil substitutes in RNA for thymine in DNA. The bases of the mRNA are complementary to those of the template DNA strand; note that uracil base pairs with adenine.

that the two different strands of the DNA may each be used as templates in different regions of the chromosome.]

Short Segments of the Chromosome Are Turned into Messages

Each mRNA carries information from only a short stretch of the DNA.

Although a chromosome carries hundreds or thousands of genes, only a fraction of these are in use at any given time. In a typical bacterial cell, about 1,000 genes, or about 25% of the total, are expressed under any particular set of growth conditions. Some genes are required for the fundamental operations of the cell and are therefore expressed under most conditions. These are known as **housekeeping genes**. Other genes vary in expression in response to changes in the environment. During cell growth and metabolism, each gene or small group of related genes is used to generate a separate RNA copy when, and if, it is needed. Consequently, each cell contains many different RNA molecules, each carrying the information from a short stretch of DNA.

In the cells of higher organisms, which have many more genes than do bacteria, the proportion of genes in use in a particular cell at a particular time is much smaller. Different cells of multi-cellular organisms express different selections of genes depending on their specialized roles. For example, in the human female, genes related to the menstrual cycle are largely unique and expressed in a timed sequence to provide functionality to the organism. In addition, gene expression varies with the stage of development. Embryonic genes are often expressed only at certain times. Thus, the control of gene expression is much more complex in higher organisms, although the basic principles are the same.

Terminology: Cistrons, Coding Sequences and Open Reading Frames

A **cistron** is a **structural gene**, which is a coding sequence or segment of DNA encoding a polypeptide. It was defined originally as a genetic unit by complementation using the *cis/trans* test. Nowadays, the terms cistron and structural gene also include DNA sequences that code for RNA molecules that function as RNA without being translated into proteins (e.g., rRNA, tRNA, snRNA, etc.). An **open reading frame** (**ORF**)

cistron Segment of DNA (or RNA) that encodes a single polypeptide chain
housekeeping genes Genes that are switched on all the time because they are needed for essential life functions
open reading frame (ORF) Sequence of bases (either in DNA or RNA) that can be translated (at least in theory) to give a protein
structural gene Sequence of DNA (or RNA) that codes for a protein or for an untranslated RNA molecule

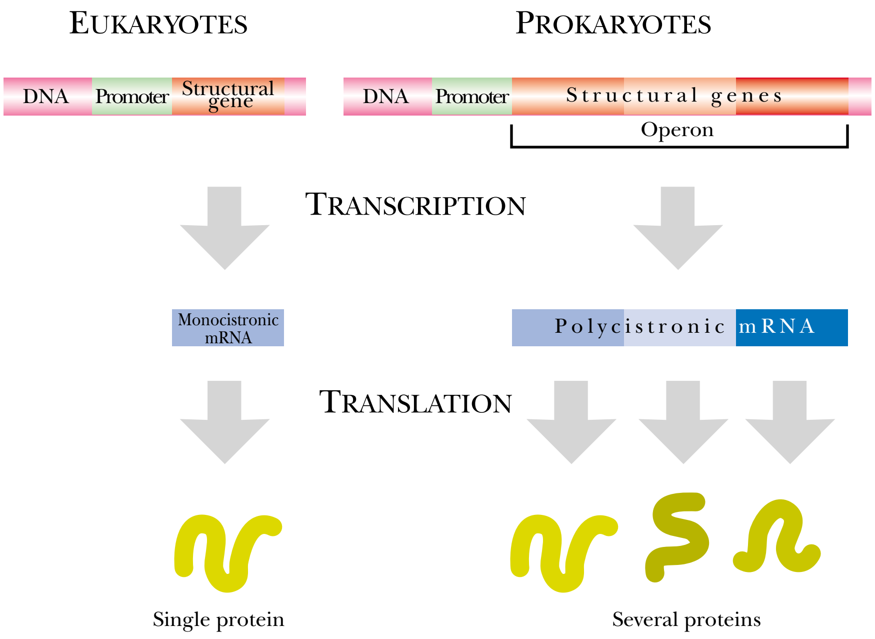


FIGURE 6.03
Monocistronic Versus Polycistronic mRNA

The typical situation in eukaryotes is to have one structural gene produce monocistronic RNA and this, in turn, be translated into a single protein. In bacteria, it is common to see several structural genes transcribed under the control of a single promoter. The RNA produced is polycistronic and yields several separate proteins.

is any sequence of bases (in DNA or RNA) that could, in theory, encode a protein. The ORF is "open" in the sense that it does not contain any stop codons that would interrupt its translation into a polypeptide chain (although, of course, every ORF ends in a stop codon). (Any cistron that encodes a protein must also be an ORF, whereas a cistron that encodes an untranslated RNA is not an ORF.)

In eukaryotes, each mRNA carries only a single gene. In prokaryotes, several genes may be carried on the same mRNA.

In eukaryotes, each gene is transcribed to give a separate mRNA and each mRNA molecule therefore encodes the information for only a single protein and is known as **monocistronic mRNA** (Fig. 6.03). In bacteria, clusters of related genes, known as **operons**, are often found next to each other on the chromosome and are transcribed together to give a single mRNA, which is therefore called **polycistronic mRNA**. Thus, a single bacterial mRNA molecule may encode several proteins, usually with related functions, such as the enzymes that oversee the successive steps in a metabolic pathway.

How Is the Beginning of a Gene Recognized?

Before starting transcription, RNA polymerase binds to the promoter, a recognition sequence in front of the gene.

Transcription will first be described in bacteria because it is simpler. The principles of transcription are similar in higher organisms, but the details are more complicated, as will be shown below. The major differences between prokaryotes and eukaryotes occur in the initiation and regulation of transcription, rather than in the actual synthesis of RNA. In front of each gene is a region of regulatory DNA that is not itself transcribed. This contains the **promoter**, the sequence to which RNA polymerase binds (Fig. 6.04), together with other sequences involved in the control of gene expression. This stretch of DNA in front of a gene (i.e., at the 5′-end) is often referred to as the **upstream region**. Note also that the first base of the mRNA of a protein-encoding gene is not the first base of the protein coding sequence. Between these two points there is a short stretch known as the **5′-untranslated region**, or 5′-UTR, meaning it will not be translated to form protein (Fig. 6.04). At the far end of the mRNA there is another short

5′-untranslated region (5′-UTR) Region of an mRNA between the 5′-end and the translation start site
monocistronic mRNA mRNA carrying the information of a single cistron, that is a coding sequence for only a single protein
operon A cluster of prokaryotic genes that are transcribed together to give a single mRNA (i.e. polycistronic mRNA)
polycistronic mRNA mRNA carrying the information of multiple cistrons, that is coding sequences for several proteins
promoter Region of DNA in front of a gene that binds RNA polymerase and so promotes gene expression
upstream region Region of DNA in front (i.e. beyond the 5′-end) of a structural gene; its bases are numbered negatively counting backwards from the start of transcription

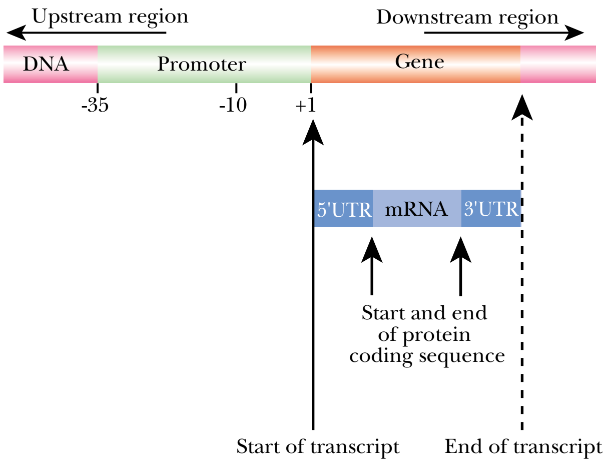


FIGURE 6.04 ***Upstream and Downstream Regions***

Genes and their regulatory regions are divided into upstream and downstream portions. The upstream portion contains the promoter. The downstream region begins with the information for the 5′-untranslated component, then the structural gene. The messenger RNA begins with the 5′-untranslated region (5′-UTR), then the coding sequence for the protein. Transcription begins by definition at the first base after the promoter. The upstream region, including the promoter, is given negative numbers counting backward from the beginning of transcription.

region, beyond the end of the protein coding sequence, that is not translated. This is the **3′-untranslated region**, or 3′-UTR.

The sigma subunit of bacterial RNA polymerase recognizes the promoter. The core enzyme makes RNA.

Bacterial RNA polymerase consists of two major components, the **core enzyme** (itself made of four subunits) and the **sigma subunit**. The core enzyme is responsible for RNA synthesis whereas the sigma subunit is largely responsible for recognizing the promoter. The sigma subunit, recognizes two special sequences of bases in the promoter region of the coding (non-template) strand of the DNA (Fig. 6.05). These are known as the **–10 sequence** and the **–35 sequence** because they are found by counting backward approximately 10 and 35 bases, respectively, from the first base that is transcribed into mRNA. [Previously, the –10 sequence was known as the **Pribnow box**, after its discoverer. This name is rarely used nowadays.]

Strong promoters usually have sequences close to consensus.

The consensus sequence for the –10 sequence is TATAA and the consensus sequence at –35 is TTGACA. (Consensus sequences are found by comparing many sequences and taking the average.) Although a few highly expressed genes do have the exact consensus sequences in their promoters, the **–10** and **–35** region sequences are rarely perfect. However, as long as they are wrong by only up to three or four bases, the sigma subunit will still recognize them. *The strength of a promoter depends partly on how closely it matches the ideal consensus sequence*. Strong promoters are highly expressed and are often close to consensus. Promoters further away from the consensus sequence will be expressed only weakly (in the absence of other factors—but see below).

Promoter sequences vary in different organisms.

In practice, consensus sequences for regulatory sites on DNA such as promoters will vary from one group of organisms to another. Thus, the –10 and –35 consensus sequences given above are for *Escherichia coli* and related bacteria. Both the consensus sequences and the proteins that recognize them will diverge in more distantly related organisms. This is of practical importance when genes from one organism are expressed in another as a result of biotechnological manipulations. Consequently, it is

–10 region Region of bacterial promoter 10 bases back from the start of transcription that is recognized by RNA polymerase
3′-untranslated region (3′-UTR) Sequence at the 3′-end of mRNA, downstream of the final stop codon, that is not translated into protein
–35 region Region of bacterial promoter 35 bases back from the start of transcription that is recognized by RNA polymerase
core enzyme Bacterial RNA polymerase without the sigma (recognition) subunit
Pribnow box Another name for the –10 region of the bacterial promoter
sigma subunit Subunit of bacterial RNA polymerase that recognizes and binds to the promoter sequence

FIGURE 6.05 ***Sigma Recognizes the –10 and –35 Sequences***

The sigma protein binds to both the –10 and –35 sequences of the promoter, thereby establishing a constant position with respect to the start of transcription.

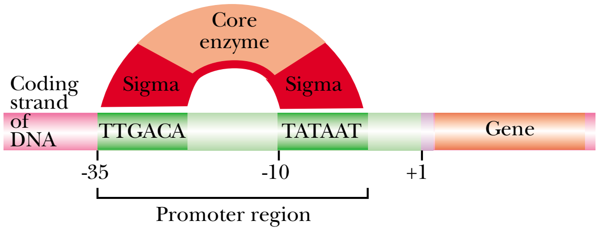
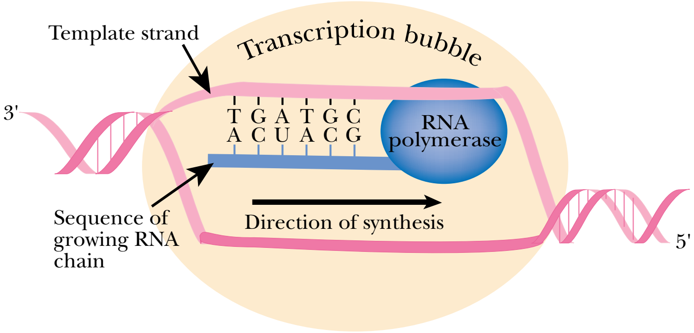


FIGURE 6.06 ***Elongation of the mRNA***

The beginning of RNA synthesis is shown. The DNA strands have separated at the transcription bubble. Synthesis of six bases of RNA complementary to those of the DNA template strand occurs while RNA polymerase remains at the promoter site.



often helpful to supply consensus promoters (or other regulatory sequences) that work well in the host organism to achieve high expression of a cloned gene from a foreign source. The use of "expression vectors" to optimize gene expression during cloning is discussed in more detail in Ch. 22.

Manufacturing the Message

RNA polymerase opens up the DNA to form the transcription bubble.

Once the sigma subunit has bound to a promoter, the RNA polymerase core enzyme opens up the DNA double helix locally to form the **transcription bubble**. Note that the –10 sequence, TATAAT, consists of AT base pairs and this assists in the melting of the DNA into single strands. After the DNA helix has been opened, a single strand of RNA is generated using one of the DNA strands as a template for matching up the bases. Once the RNA polymerase has bound to the DNA and initiated a new strand of RNA, the sigma subunit is no longer needed and often (though not always) detaches from the DNA, leaving behind the core enzyme. The RNA polymerase actually remains at the promoter until the new strand is eight or nine bases long. At this point, sigma leaves and the core enzyme is free to move forward and elongate the mRNA (Fig. 6.06).

The core enzyme moves ahead, manufacturing RNA and leaving sigma behind at the promoter.

The first transcribed base of the mRNA is normally an A (as in Fig. 6.06). This special A is usually flanked by two pyrimidines, most often giving the sequence CAT. Sometimes the first transcribed base is a G, but almost never a pyrimidine. Synthesis of mRNA is from 5′ to 3′ and proceeds at about 40 nucleotides per second. This is much slower than DNA replication (~1,000 bp/sec), but roughly equivalent to the rate of polypeptide synthesis (15 amino acids per second).

The core enzyme of RNA polymerase consists of four subunits, two α plus β and β′ (Fig. 6.07). The β and β′ subunits comprise the catalytic site of the enzyme. The α subunit is required partly for assembly and partly for recognizing promoters. RNA polymerase has a deep groove through the middle that can accommodate about 16 bp of DNA in the case of bacteria and about 25 bp in the case of eukaryotes such as yeast, whose RNA polymerase is larger. A thinner groove, roughly at right angles to the first, may hold the newly constructed strand of RNA.

transcription bubble Region where DNA double helix is temporarily opened up so allowing transcription to occur

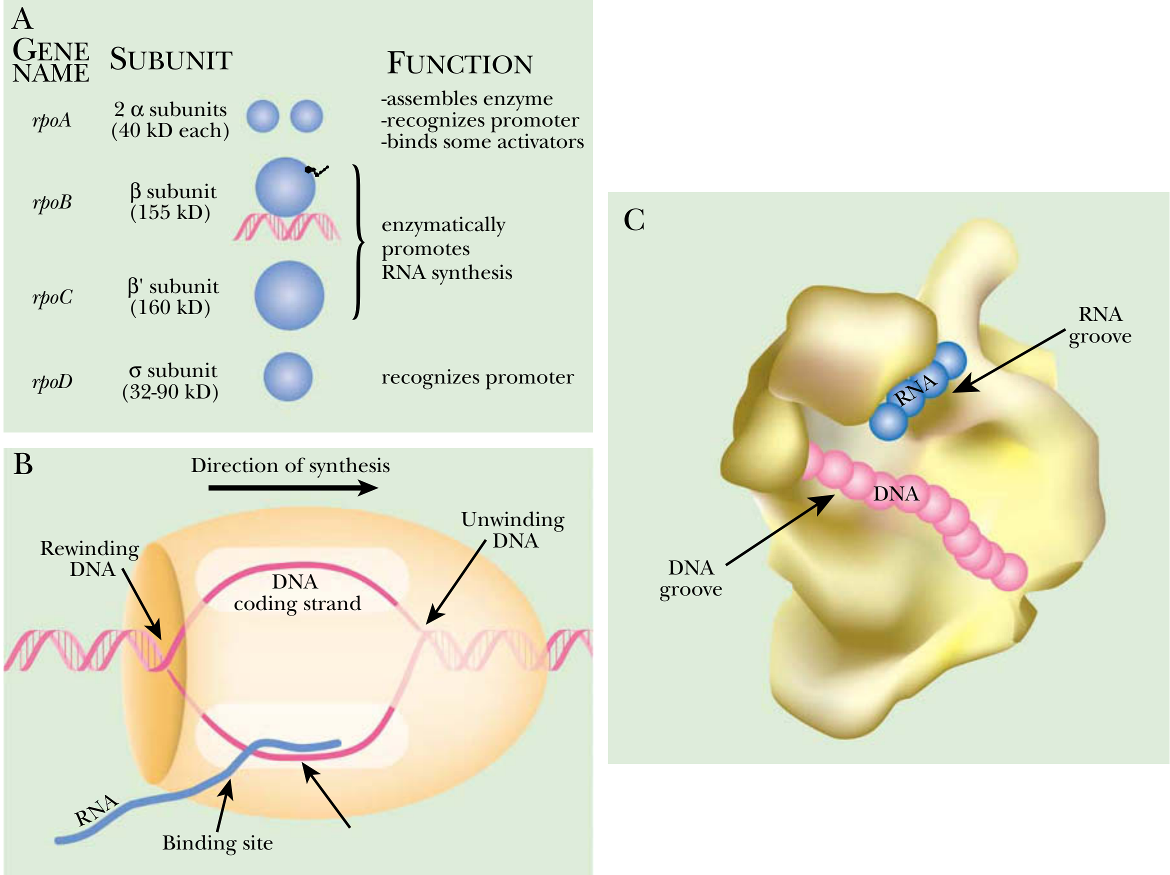


FIGURE 6.07 ***Structure of RNA Polymerase***

A) Bacterial RNA polymerase has four types of subunits and three functional specificities.
B) Topography of functions performed by RNA polymerase. C) The structure of yeast RNA polymerase reveals a groove that may be used by DNA as it moves relative to the polymerase and a potential groove for the newly formed RNA.

The negative supercoiling of the chromosome promotes opening of DNA during transcription. As RNA polymerase moves along the DNA, it winds the DNA more tightly ahead of itself, creating positive supercoils. It also leaves partly unwound DNA behind, which generates negative supercoils. To restore normal levels of supercoiling, DNA gyrase inserts negative supercoils ahead of RNA polymerase and topoisomerase I removes negative supercoils behind RNA polymerase (see Ch. 4).

RNA Polymerase Knows Where to Stop

The end of a gene is marked by a terminator sequence that forms a hairpin structure in the RNA.

Just as there is a recognition site at the front of each gene, so there is a special **terminator** sequence at the end. The terminator is in the template strand of DNA and consists of two inverted repeats separated by half a dozen bases, and followed by a run of A's. The sequence of the mRNA will be the same as the non-template strand of DNA except for the substitution of U for T. Thus the string of A's in the DNA template strand gives rise to a run of U's at the 3′-end of the mRNA (Fig. 6.08). Note that in the DNA, the two inverted repeat sequences are on opposite strands. Although researchers often talk as if the mRNA has inverted "repeats," its second "repeat" is actually the complement of the first. Because of this, such inverted repeats on the

terminator DNA sequence at end of a gene that tells RNA polymerase to stop transcribing

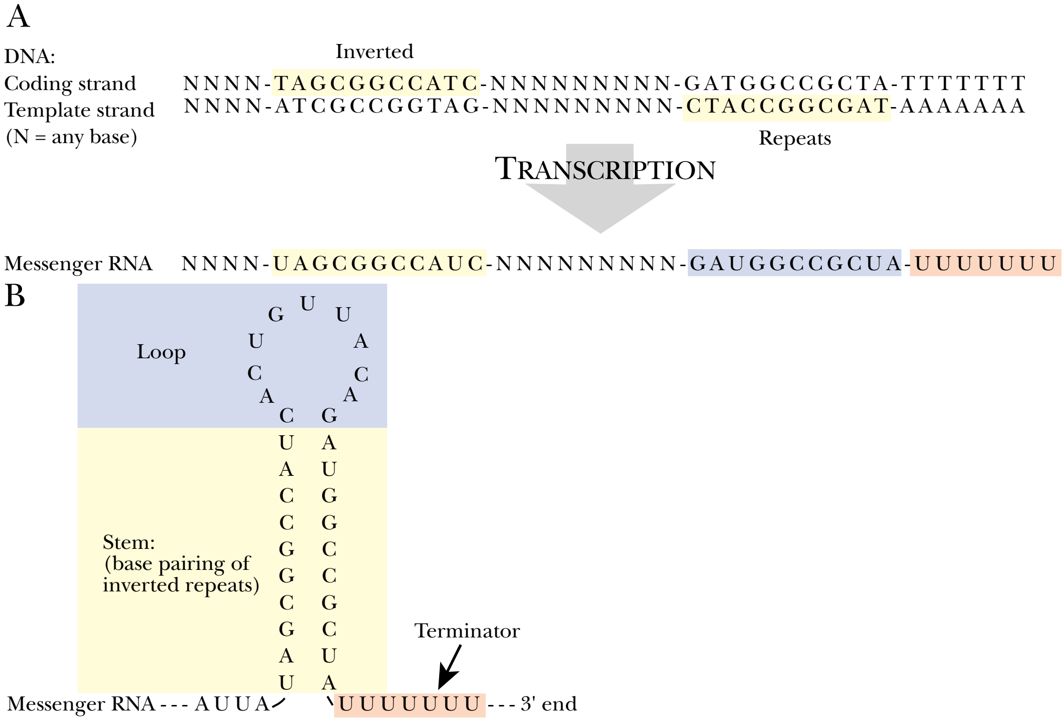


FIGURE 6.08 ***The Terminator Sequence is Transcribed into RNA***

A) The signal for RNA polymerase to stop is shown in both the DNA and the RNA transcribed from it. The terminator consists of an inverted repeat separated by approximately 10 bases from a run of U's. B) The complementary bases form the stem of the hairpin, with the intervening bases forming the loop.

same strand of an RNA molecule can pair up to generate a stem and loop or "hairpin" structure (Fig. 6.08).

Once the RNA polymerase reaches the stem and loop, it pauses. Long RNA molecules contain many possible hairpin structures that cause RNA polymerase to slow down or stop briefly, depending on the size of the hairpin. This provides an opportunity for termination, but if there is no string of U's, the RNA polymerase will start off again. However, a string of U's paired with a string of A's in the template strand of DNA is a very weak structure, and the RNA and DNA fall apart while the RNA polymerase is idling (Fig. 6.09). Pausing varies in length, but is around 60 seconds for a typical terminator.

Termination may actually occur at several possible positions in the middle or end of the run of U's. In other words, the RNA polymerase "stutters" and the precise location of termination may vary slightly between different molecules of the same mRNA. Once the DNA and RNA have separated at the terminator structure, the RNA polymerase falls off and departs to find another gene (Fig. 6.09).

Two classes of terminators exist. **Rho-dependent terminators** need **Rho (ρ) protein** to separate the RNA polymerase from the DNA. **Rho-independent**, or "intrinsic," terminators do not need Rho or any other factor to cause termination. Most terminators in *E. coli* do not need Rho. In contrast, Rho-dependent terminators are relatively frequent in bacteriophages.

Rho protein is a specialized helicase that uses energy from ATP to unwind a DNA/RNA hybrid double helix. It consists of a hexamer of six identical subunits that recognizes and binds to a sequence of 50 to 90 bases located upstream of the termi-

A sub-class of terminators require a recognition protein, known as Rho, to function.

Rho (ρ) protein Protein factor needed for successful termination at certain transcriptional terminators
Rho-dependent terminator Transcriptional terminator that depends on Rho protein
Rho-independent terminator Transcriptional terminator that does not need Rho protein

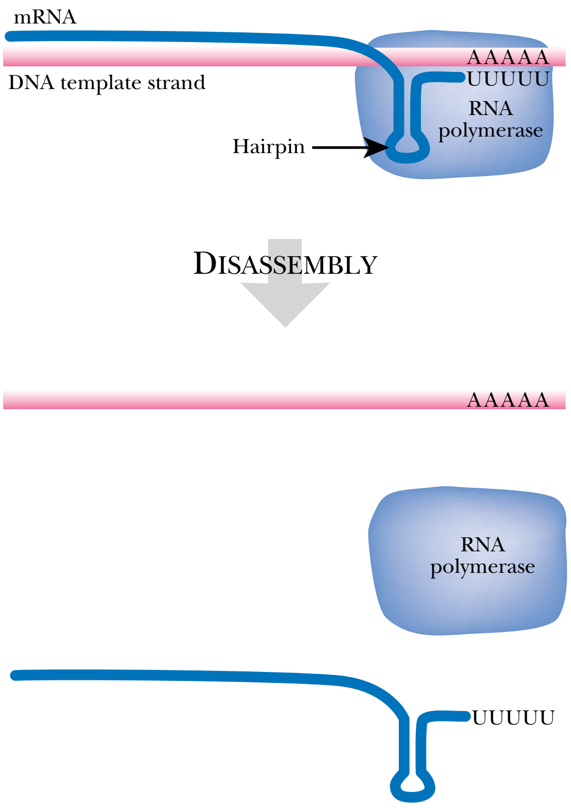


FIGURE 6.09 ***Termination of Messenger RNA***

When the mRNA reaches the hairpin of the terminator it pauses; when it reaches the AAAAA sequence it falls off the template strand along with the newly synthesized RNA.

nator in the mRNA. The Rho hexamer does not form a closed ring, but instead is split open and resembles a lock washer in structure. The RNA sequence for Rho binding is poorly defined but is high in C and low in G. Rho can only bind to the growing mRNA chain once the RNA polymerase has synthesized the C-rich/G-poor recognition region and moved on. Rho moves along the RNA transcript and catches up with the RNA polymerase at the terminator stem and loop structure where the RNA polymerase pauses (Fig. 6.10). Rho then unwinds the DNA/RNA helix in the transcription bubble and separates the two strands.

How Does the Cell Know Which Genes to Turn On?

Housekeeping genes are switched on all the time.

Some genes, known as housekeeping genes, are switched on all the time; i.e., they are expressed "**constitutively**." In bacteria, these often have both their –10 and –35 region promoter sequences very close or identical to consensus. Consequently, they are always recognized by the sigma subunit of RNA polymerase and are expressed under all conditions. Other constitutive promoters are further from consensus and expressed less strongly. Nonetheless, if only relatively low amounts of the gene product are needed, this is acceptable.

Some genes need activator proteins to switch them on.

Genes that are only needed under certain conditions sometimes have poor recognition sequences in the –10 and –35 regions of their promoters. In such cases, the promoter is not recognized by sigma unless another accessory protein is there to help (Fig. 6.11). These accessory proteins are known as gene **activator proteins** and are different for different genes. Each activator protein may stimulate the transcription of one or more genes. A group of genes that are all recognized by the same activator protein will be expressed together under similar conditions, even if the genes are at different places on the DNA. Higher organisms have many genes that are expressed differently

activator protein Protein that switches a gene on
constitutive gene Gene that is expressed all the time

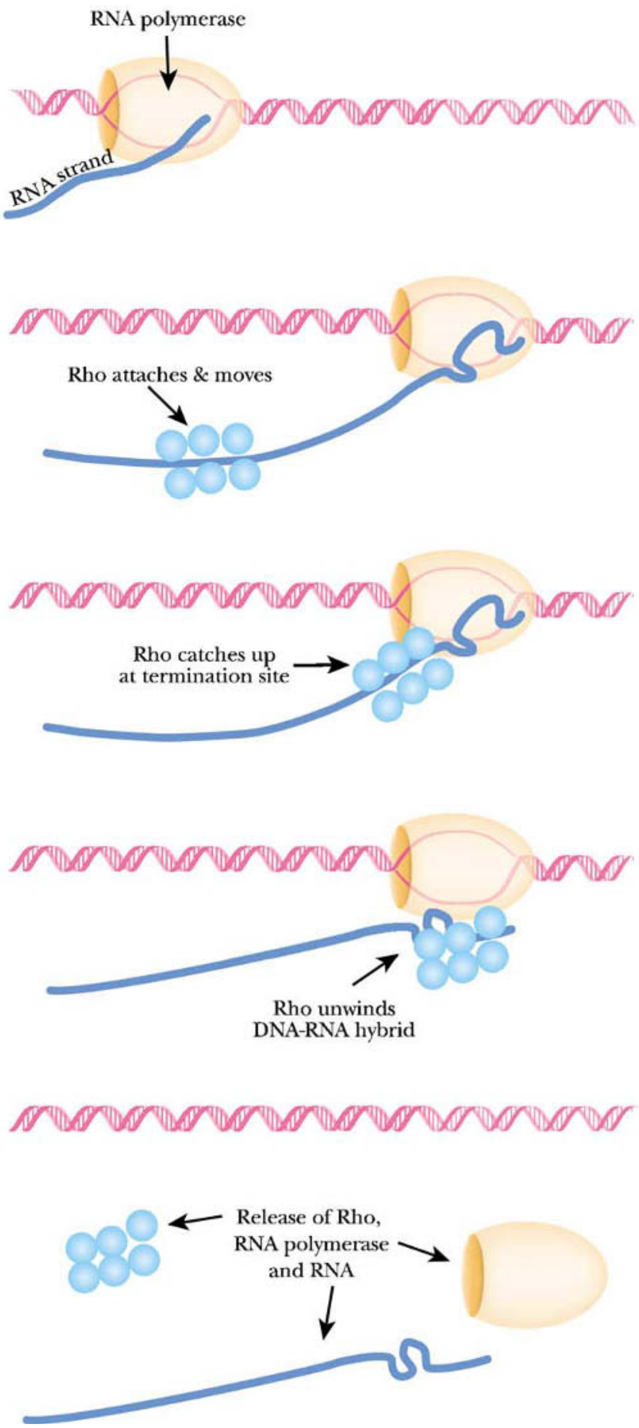

RNA polymerase
RNA strand
Rho attaches & moves
Rho catches up at termination site
Rho unwinds DNA-RNA hybrid
Release of Rho, RNA polymerase and RNA

FIGURE 6.10 ***Termination by Rho***

Rho first binds to the growing messenger RNA. When the RNA polymerase pauses at the termination site, Rho catches up and untwists the newly formed mRNA strand from the DNA. Subsequently, the mRNA and RNA polymerase fall off the DNA and Rho detaches from the mRNA.

in different tissues. As a result, eukaryotic genes are often controlled by multiple activator proteins, more specifically known as **transcription factors** (see below).

What Activates the Activator?

Long ago, the Greek philosopher Plato pondered the political version of this question: "Who will guard the guardians?" In living cells, especially in more complex higher

transcription factor Protein that regulates gene expression by binding to DNA in the control region of the gene

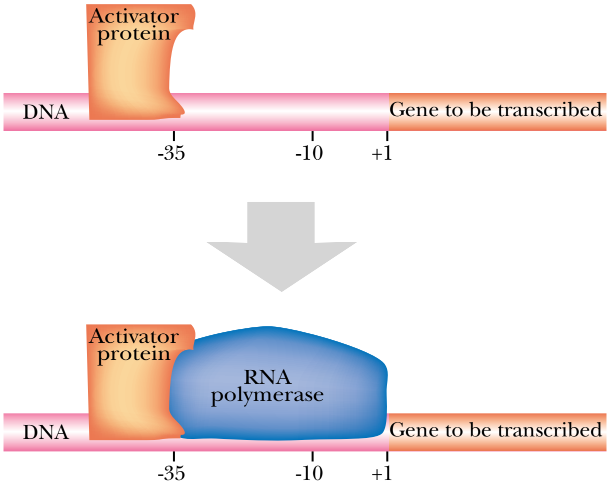


FIGURE 6.11 ***Gene Activator Proteins***

The activator protein first binds to the promoter region of the gene. Once bound, the activator protein facilitates the binding of the RNA polymerase. Gene transcription then commences.

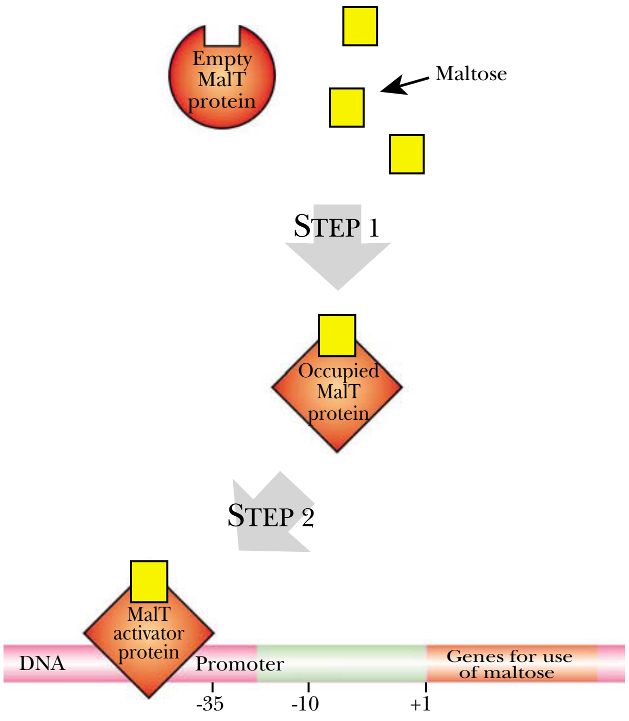


FIGURE 6.12 ***MalT Changes Shape upon Binding Maltose***

The MalT protein has a binding site complementary in shape to the sugar maltose. In step 1, MalT binds to maltose, which causes MalT to change shape. In step 2, the new conformation of MalT protein allows it to bind to DNA at a specific sequence found only in certain promoters. The gene thus activated is involved in the metabolism of maltose.

organisms, there may indeed be a series of regulators, each regulating the next. What is the initial event? The cell must respond to some outside influence or must be influenced by other internal processes. The regulation of gene expression will be considered in more detail in Chapters 9 and 10. This chapter will be limited to a discussion of the basic mechanisms needed for a promoter to be functional.

Activator proteins often change shape in response to small molecules. Only one conformation binds to DNA.

As a simple example of an activator, consider the use of maltose by *Escherichia coli*. Maltose is a sugar made originally from the starch in malt and many other sources. It can be used by *E. coli* to satisfy all of its needs for energy and organic material. An activator protein, MalT, detects maltose and binds to it (Fig. 6.12). This causes the MalT protein to change shape, exposing its DNA-binding site. The original "empty" form of MalT cannot bind to DNA. The active form (MalT + maltose) binds to a specific sequence of DNA found only in the promoter region of genes needed for growth on maltose. The presence of MalT helps RNA polymerase bind to the promoter and transcribe the genes. The small molecule, in this case maltose, which causes gene

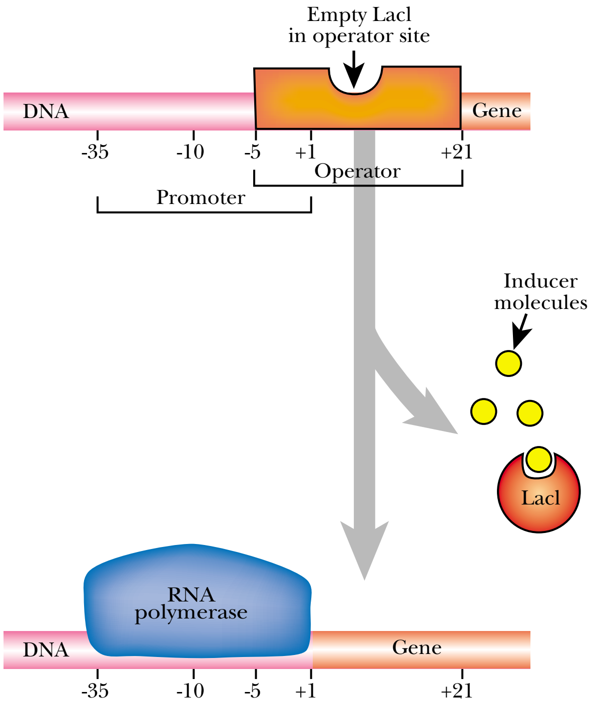


FIGURE 6.13 ***Principle of Negative Regulation by a Repressor***

The LacI protein is bound to the operator site, within the promoter region of a gene that affects lactose metabolism. The inducer binds to LacI, changing its conformation and causing its release from the DNA. The RNA polymerase is then free to transcribe the gene.

expression is known as the **inducer**. The result of this is that the genes intended for using maltose are only induced when this particular sugar is available. The same general principle applies to most nutrients, although the details of the regulation may vary from case to case.

Negative Regulation Results from the Action of Repressors

Repressors are proteins that switch genes off.

Genes may be controlled by positive or **negative regulation**. In **positive regulation**, an activator protein binds to the DNA only when the gene is to be turned on. In **negative regulation**, a **repressor** protein binds to the DNA and insures that the gene is turned off. Only when the repressor is removed from the DNA can the gene be transcribed. The site where a repressor binds is called the **operator** sequence. Like activator proteins, repressor proteins alternate between DNA-binding and nonbinding forms. In this case, binding of the inducer to the repressor causes it to change from its DNA-binding form to the nonbinding form.

Historically, negative regulators were discovered before activators. The best known example is the lactose repressor, the **LacI protein** (Fig. 6.13). Lactose is another sugar, found in milk, on which bacteria such as *E. coli* can grow. When no lactose is available, the LacI protein binds to its operator sequence, which overlaps part of the promoter and the front part of the coding region for the genes for using lactose. When lactose is present, the LacI protein changes shape and is released from the DNA and the lactose genes are induced. Overall, the result is the same as for maltose: when lactose is available, the genes for using it are switched on and when there is no lactose, the genes are turned off.

inducer Small signal molecule that binds to a regulatory protein and thereby causes a gene to be switched on
LacI protein Repressor that controls the *lac* operon
negative regulation Regulatory mode in which a repressor keeps a gene switched off until it is removed
operator Site on DNA to which a repressor protein binds
positive regulation Control by an activator that promotes gene expression when it binds
repressor Regulatory protein that prevents a gene from being transcribed

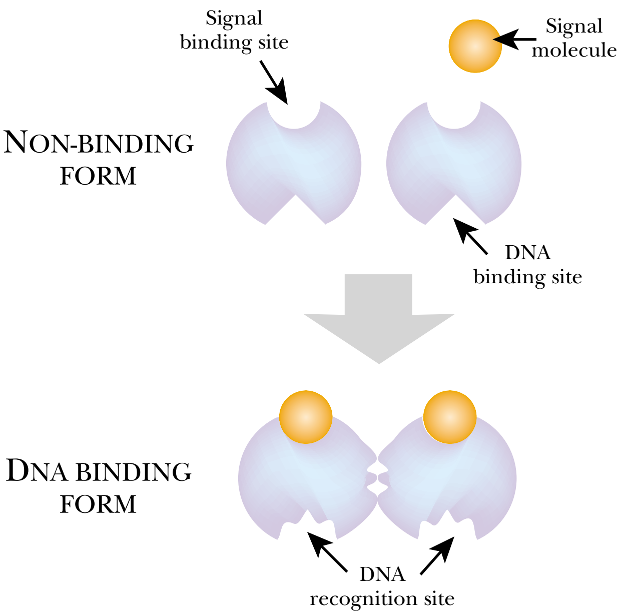


FIGURE 6.14 ***Allosteric Protein Binds a Signal Molecule and Changes Shape***

The two subunits shown have a signal-binding site and a DNA-binding site. When the signal molecule binds to the subunits, they pair and change conformation They are then able to bind to DNA.

The detailed mechanism by which repressors prevent transcription varies considerably and is often unknown. The repressor sometimes blocks the binding of RNA polymerase to the promoter, simply by getting in the way (steric hindrance). An example is the well-studied CI repressor of bacteriophage lambda. Sometimes the repressor may bind further downstream, inside the structural gene. In this case, RNA polymerase can still bind to the promoter but is prevented from moving forward and transcribing the gene. Sometimes, even though their binding sites in the DNA sequence overlap, the RNA polymerase and the repressor both bind the DNA simultaneously. [Remember that the DNA double helix is 3-dimensional and that two proteins may therefore bind to the same linear segment, if they occupy separate locations around its surface.] Indeed, an example of this is the LacI repressor. In this case the RNA polymerase actually binds more tightly in the presence of repressor, but is locked in place and cannot open the DNA to initiate transcription.

Many Regulator Proteins Bind Small Molecules and Change Shape

Small molecules may control gene expression by binding to regulatory proteins.

Whether a regulator protein is an activator or a repressor, it needs a signal of some sort. One of the most common ways to do this is by using some small molecule that fits into a binding site on the regulatory protein (Fig. 6.14). This is called the **signal molecule**. In the case of using a nutrient for growth, an obvious and common choice is the nutrient molecule itself. [In prokaryotes the DNA binding protein often binds the signal molecule directly. In eukaryotes, where the DNA is inside the nucleus, things are often more complex, and multiple proteins are involved. The signal molecule is often bound by proteins in the cell membrane or cytoplasm and the signal is then transmitted to the nucleus. The DNA binding protein itself normally stays in the nucleus and upon receiving the signal, is converted to its DNA-binding form by phosphorylation.]

When a regulator protein binds its signal molecule, it changes shape (Fig. 6.14). Regulator proteins have two alternative forms, the DNA-binding form and the non-binding form. Binding or loss of the signal molecule causes the larger protein to flip-flop between its two alternative shapes. Proteins that change in activity by changing

signal molecule Small molecule that exerts a regulatory effect by binding to a regulatory protein

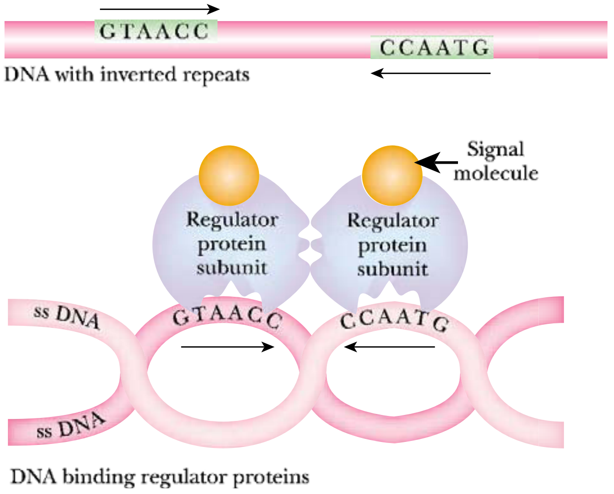

GTAACC
CCAATG
DNA with inverted repeats
Signal molecule
Regulator protein subunit
Regulator protein subunit
ss DNA
GTAACC
CCAATG
ss DNA
DNA binding regulator proteins

FIGURE 6.15 ***Regulator Binds at an Inverted Repeat—Principle***

At sites where regulator proteins bind there is often an inverted repeat with both DNA strands participating as shown. If the subunits of regulator protein are identical, they each recognize one of the inverted repeats and pair so that the same regions of each subunit face each other.

shape in this manner are called **allosteric proteins**. Examples include some enzymes, transport proteins and regulators. Allosteric proteins have multiple subunits that change shape in concert (Fig. 6.14). Usually there is an even number of subunits, most often two or four. All of the subunits bind the signal molecule and then they all change shape together.

Recognition sites on DNA are often inverted repeats. Separate subunits of the regulator protein each bind one of the repeat sequences.

Since there is an even number of protein subunits, the recognition site on the DNA for regulator proteins is often duplicated. In this case, the recognition site is usually an inverted repeat, often referred to as a palindrome. This is because the subunits of the regulator protein bind to each other head to head rather than head to tail (Fig. 6.15). Consequently, the two protein molecules are pointing in opposite directions. Because they have identical binding sites for DNA, they recognize the same sequence of bases but in opposite directions on the two strands of the DNA. The two half-sites are usually separated by a spacer region of several bases, whose identity is free to vary. The two half-sequences of such recognition sites are not always exact matches.

One regulator protein subunit binds to the recognition sequence on the template strand of the DNA double helix, and its partner binds to the same sequence but on the non-template strand of the DNA pointing in the opposite direction. This is simpler in practice than it sounds, precisely because the DNA molecule is helical. Although the two recognition sequences are on different strands of DNA, they end up on the same face of the DNA molecule due to its helical twisting (Fig. 6.15).

An example of a palindromic recognition site is the *lac* operator sequence which is bound by the LacI repressor (a tetramer). This sequence runs from –6 to +28 relative to the start of transcription. It is not exactly symmetrical. The two half sequences are: TGTGTGgAATTGTgA and, running in the opposite sense on the other strand, TGTGTGaAATTGTtA (capital letters indicate matching bases). The two half-sites are separated by five base pairs. The left hand half of this site binds the LacI protein more strongly than the right hand side. A stronger operator sequence could be generated by changing the right hand half-site to exactly match the left.

Transcription in Eukaryotes Is More Complex

Eukaryotes have three RNA polymerases that specialize in which type of genes they transcribe.

Since typical eukaryotic cells have 10 times as many genes as do bacteria, the whole process of transcription and its regulation is more complex. For a start, eukaryotes have three different RNA polymerases, unlike bacteria which have just one. The three RNA polymerases transcribe different categories of nuclear genes. In addition, mitochondria and chloroplasts have their own RNA polymerases, which resemble the bacterial enzyme.

allosteric protein Protein that changes shape when it binds a small molecule

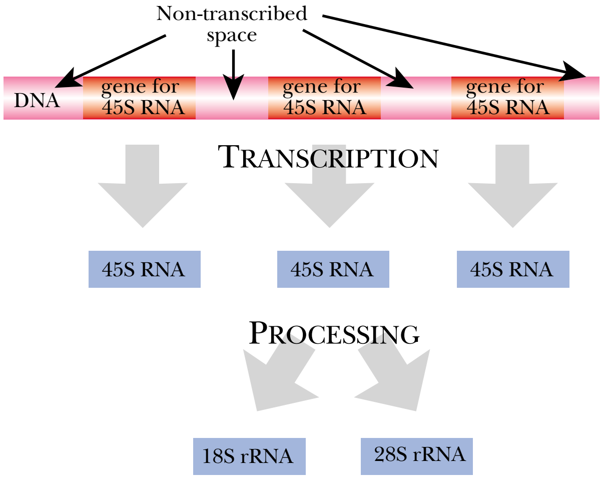


FIGURE 6.16 ***Clusters of Ribosomal RNA Genes***

The genes for rRNA are located at multiple sites along the DNA. A single transcribed unit of DNA yields an initial RNA molecule of 45S. The 45S RNA is processed to yield the final 18S and 28S subunits of rRNA.

RNA polymerase I transcribes the genes for the two large ribosomal RNA molecules and **RNA polymerase III** transcribes the genes for tRNA, 5S rRNA and a few other small RNA molecules. **RNA polymerase II** transcribes most eukaryotic genes that encode proteins and as a result is subject to the most complex regulation. Since ribosomal RNA and transfer RNA are needed all the time by all types of cells, RNA polymerases I and III operate constitutively in most cell types.

Many transcription factors are involved in controlling gene expression in eukaryotes.

A variety of proteins, known as **transcription factors** are also needed for the correct functioning of RNA polymerases. Transcription factors may be divided into general transcription factors and specific transcription factors. General transcription factors are needed for the transcription of all genes transcribed by a particular RNA polymerase, and are typically designated TFI, TFII, TFIII followed by individual letters. The I, II, and III refer to the corresponding RNA polymerase (see below). Specific transcription factors are needed for transcription of particular specific gene(s) under specific circumstances. [Proteins such as the sigma subunit of bacterial RNA polymerase may also be regarded as transcription factors, however, this terminology is usually only used for eukaryotes.]

Transcription of rRNA and tRNA in Eukaryotes

The genes for the two large ribosomal RNAs are present in multiple copies, from seven in *E. coli* to several hundred in higher eukaryotes (Fig. 6.16). In bacteria, the copies are dispersed, but in eukaryotes they form clusters of tandem repeats. In humans, there are clusters of rRNA genes on five separate chromosomes. The 18S and the 28S rRNA are transcribed together as a single large RNA (45S RNA) that is cleaved to release the two separate ribosomal RNA molecules. Between these transcription units are non-transcribed spacer regions. In eukaryotic cells, the rRNA genes have their own RNA polymerase to transcribe them, RNA polymerase I.

Eukaryotes contain many copies of the genes for ribosomal RNA. These are found in clusters and are transcribed by RNA polymerase I.

Synthesis of rRNA is localized to a special zone of the nucleus known as the nucleolus. Here the rRNA precursor is both transcribed and processed into 18S and 28S rRNA. These rRNA molecules then bind proteins, giving ribonucleo-protein particles. This yields a dense granular region when seen under the microscope (Fig. 6.17). The segments of chromosomes associated with the nucleolus were named "**nucleolar organizers**." It is now known that these correspond to the clusters of rRNA genes.

nucleolar organizer Chromosomal region associated with the nucleolus; actually a cluster of rRNA genes
RNA polymerase I Eukaryotic RNA polymerase that transcribes the genes for the large ribosomal RNAs
RNA polymerase II Eukaryotic RNA polymerase that transcribes the genes encoding proteins
RNA polymerase III Eukaryotic RNA polymerase that transcribes the genes for 5S ribosomal RNA and transfer RNA
transcription factor Protein that regulates gene expression by binding to DNA in the control region of the gene

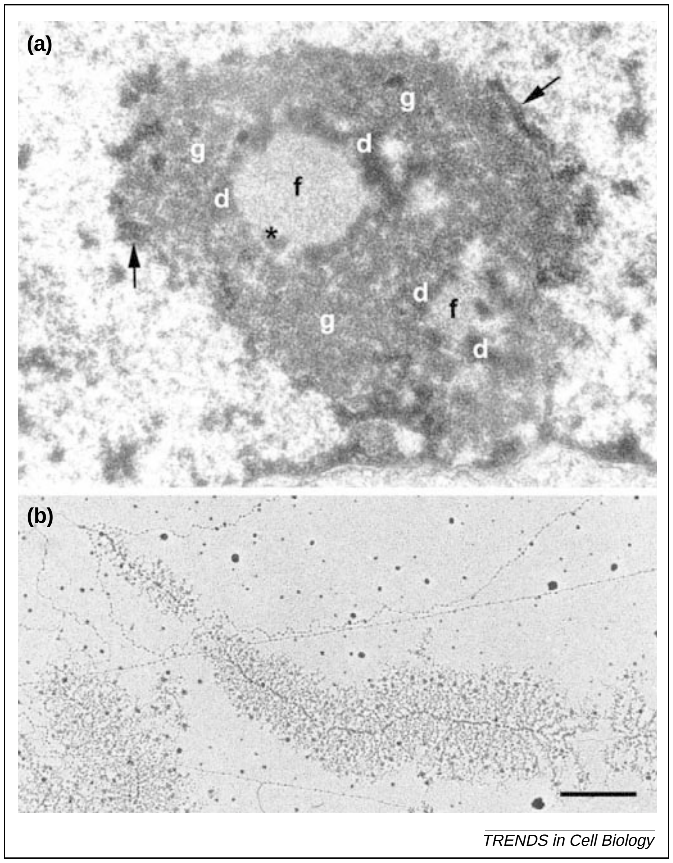


FIGURE 6.17 ***Ribosomal RNA is Made in the Nucleolus***

A) Electron micrograph of a thin-sectioned nucleolus from a mouse cell fixed in situ. Black arrows indicate peri-nucleolar condensed chromatin and the asterisk shows dense fibrillar components (d) clumping around fibrillar centers (f). Granular regions (g) of newly made ribonucleoproteins are also marked. Image provided by Ulrich Scheer, University of Würzburg. B) Spread Christmas tree structure (4 microns long) from a mouse cell is shown at the same magnification as (A). Bar represents 0.5 micron. From: Raska I., Oldies but goldies: searching for Christmas trees within the nucleolar architecture. Trends in Cell Biology 13 (2003) 517–525.

Although most promoters are AT rich, presumably because the weaker base pairs help in opening up the DNA, the promoter for RNA polymerase I is unusual in containing many GC pairs. There are two GC-rich regions, the core promoter and the upstream control element, that are 80 to 90 percent identical in sequence (Fig. 6.18). Both are recognized by protein UBF1 (Upstream Binding Factor 1), a single polypeptide. After UBF1 has bound, another protein, selectivity factor SL1, binds next to it. SL1 consists of four polypeptides, one of which, TBP (TATA Binding Protein), is also required for RNA polymerases II and III (see below). Once UBF1 and SL1 are in place, RNA polymerase I can bind. It is uncertain how the binding of UBF1 and SL1 at the upstream control element helps initiation in the case of RNA polymerase I. However, in similar cases, the DNA is known to bend around, bringing the upstream element into direct contact with the promoter region.

RNA polymerase III transcribes genes for small non-coding RNAs, in particular tRNA and 5S rRNA.

RNA polymerase III is responsible for making 5S rRNA and transfer RNA. It also makes some small nuclear RNAs, while other snRNAs are transcribed by RNA polymerase II (see below). The promoters for 5S rRNA and tRNA are unique and somewhat bizarre in being internal to the genes. Transcription of these genes requires the binding of either of two proteins known as TFIIIA and TFIIIC to a region over 50 bp downstream from the start site (Fig. 6.19). Once these have bound, they enable TFIIIB to bind to the region around the start of transcription. TFIIIB consists of three polypeptides, including TBP, and positions RNA polymerase III correctly at the start site.

As the promoters for RNA polymerase I and RNA polymerase III illustrate, recognition factor sites may be upstream or downstream from the start of transcription. However, in both cases, a positioning factor (SL1 or TFIIIB, respectively) is required to make sure that the polymerase starts transcribing at the correct place. These positioning factors thus play a similar role to that of the sigma factor in bacteria.

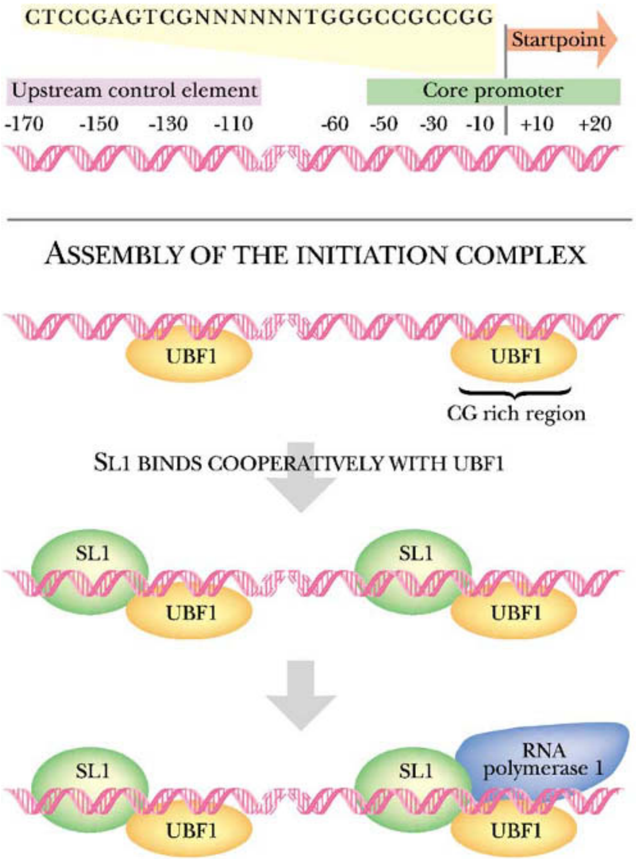


FIGURE 6.18 ***RNA Polymerase I Transcribes rRNA Genes***

The promoter for RNA polymerase I has an upstream control element and a core promoter, the latter rich in GC sequences. The UBF1 protein recognizes and binds to both the upstream control element and the core promoter. Subsequently, SL1 binds to the DNA in association with UBF1. Finally, RNA polymerase I binds and transcription commences. How this binding pattern facilitates transcription of rRNA is not known.

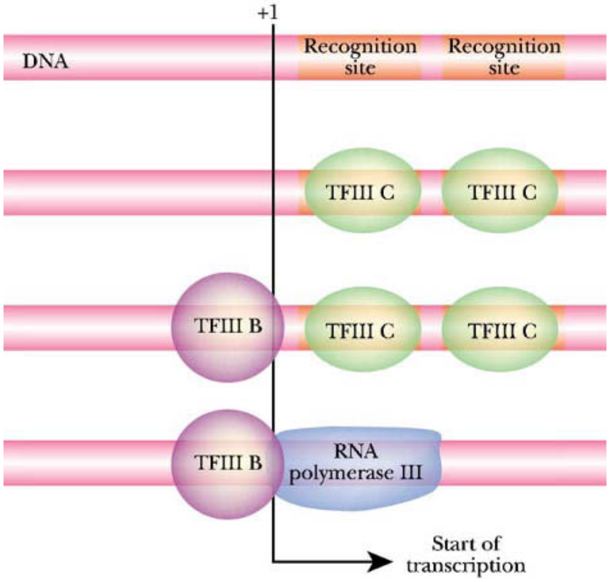


FIGURE 6.19 ***Internal Promoter for RNA Polymerase III***

The gene for 5S rRNA is transcribed using a promoter located within the gene itself. The recognition sites are downstream of the start site. TFIIIC (or TFIIIA) binds to both sites and this induces TFIIIB to bind to the promoter near the start site. Only after TFIIIB binds can RNA polymerase III bind.

Transcription of Protein-Encoding Genes in Eukaryotes

RNA polymerase II transcribes genes that code for proteins.

RNA polymerase II transcribes most eukaryotic genes that encode proteins. Recognition of the promoter and initiation of transcription by RNA polymerase II requires a number of general transcription factors. In addition, since many protein-encoding genes vary markedly in expression, a variety of specific transcription factors are needed for expression of certain genes under particular circumstances. For example, in a multi-

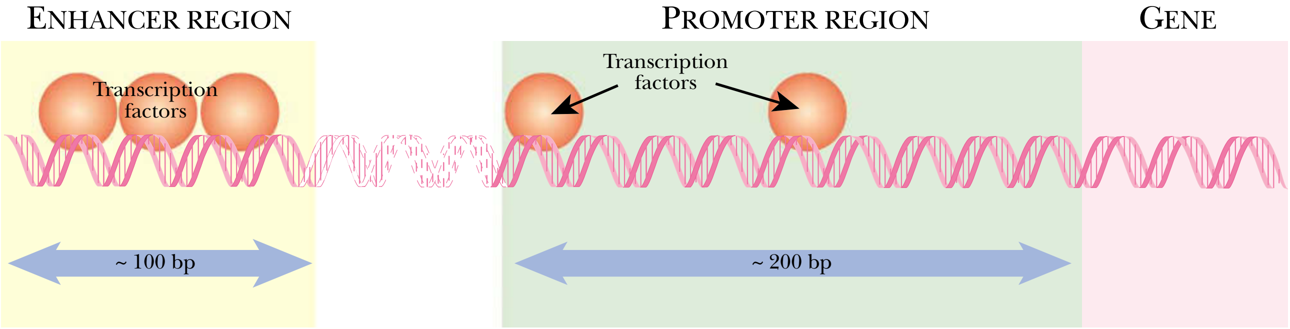


FIGURE 6.20 ***Promoter and Enhancer***

Although one RNA polymerase is used to transcribe most protein encoding genes, specificity is controlled by transcription factors and their recognition sequences. The promoter region is close to the start site and usually binds several transcription factors. In addition, extra transcription factors bind to regions known as enhancers. These may be far upstream of the promoter, as shown, or may be located downstream. Binding of the transcription factors to their recognition sequences influences polymerase activity and gene expression.

cellular organism, different cell types produce different types of proteins. Thus, red blood cells produce hemoglobin, whereas white blood cells make antibodies. Further, protein production often varies during development. Fetal hemoglobin is different from the adult version.

Some transcription factors bind to the promoter region, others to distant enhancer sequences.

The assorted transcription factors bind to and recognize specific sequences on the DNA. These DNA sequences are of two major classes, those comprising the promoter itself and a variety of **enhancer** sequences (Fig. 6.20). The general transcription factors for RNA polymerase II (TFII factors) bind to the promoter region. However, although some of the specific transcription factors also bind to the promoter region, others bind to the enhancer.

In eukaryotes, many protein-encoding genes are interrupted by introns. These are removed at the RNA stage. Consequently, transcription of DNA to give RNA does not yield messenger RNA directly. The RNA that results from transcription is known as the **primary transcript** and must be processed as described in Chapter 12 to give mRNA. The present discussion will therefore be limited to the transcription of genes by RNA polymerase II to give the primary transcript.

The TATA box is the critical sequence that allows RNA polymerase II to recognize the promoter.

Promoters for RNA polymerase II consist of three regions, the **initiator box**, the **TATA box** and a variety of **upstream elements** (see below). The initiator box is a sequence found at the site where transcription starts. The first transcribed base of the mRNA is usually A with a pyrimidine on each side, as in bacteria. The consensus is weak: YYCAYYYYY (where Y is any pyrimidine). About 25 base pairs upstream from this is the TATA box, an AT-rich sequence, which is recognized by the same factor TBP (**TATA binding protein** or **TATA box factor**) that is needed for binding of RNA polymerases I and III. TBP is unusual in binding in the minor groove of DNA. (Almost all DNA-binding proteins bind in the major groove). On both sides of the TATA box are GC-rich regions (Fig. 6.21).

TBP is found in three different protein complexes, depending on whether RNA polymerase I, II or III is involved. In the present case, TBP forms part of a transcription factor complex known as TFIID that is needed to recognize promoters specific for RNA polymerase II. Several other TFII complexes are also needed for RNA polymerase II function. TFIIA and TFIIB bind next. Then at last RNA polymerase II itself

enhancer Regulatory sequence outside, and often far away from, the promoter region that binds transcription factors
initiator box Sequence at the start of transcription of a eukaryotic gene
primary transcript RNA molecule produced by transcription before it has been processed in any way
TATA binding protein (TBP) Transcription factor that recognizes the TATA box
TATA box Binding site for a transcription factor that guides RNA polymerase II to the promoter in eukaryotes
TATA box factor Another name for TATA binding protein
upstream element DNA sequence upstream of the TATA box in eukaryotic promoters that is recognized by specific proteins

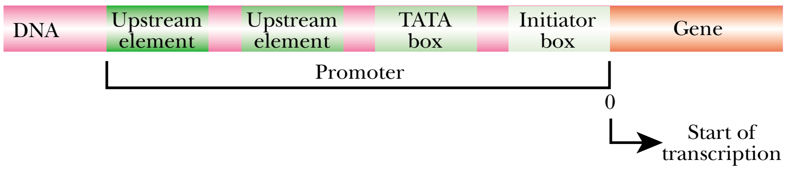


FIGURE 6.21 ***Eukaryotic Promoter Components—Initiator and TATA Boxes***

The promoter for RNA polymerase II has an initiator box at the start site and a TATA box slightly upstream of this. Further upstream there are normally several upstream elements (two are shown here).

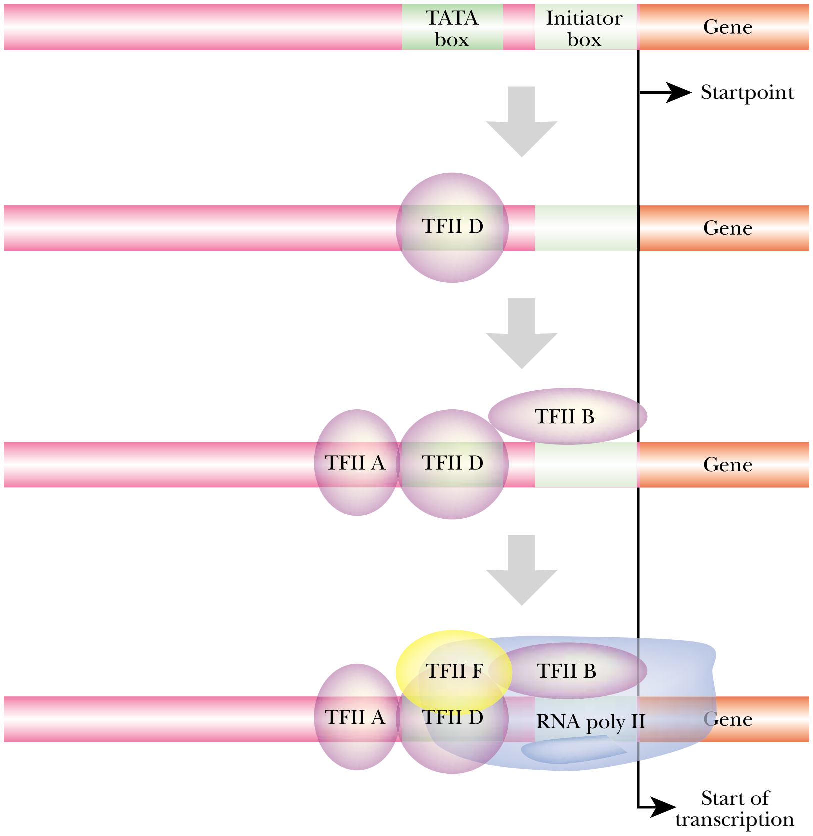


FIGURE 6.22 ***Binding of RNA Polymerase II to Promoter***

Starting with TFIID, which contains TATA binding protein, the components of the TFII complex bind one after another. Finally TFIIF helps RNA polymerase II to bind to the DNA.

arrives, accompanied by TFIIF which probably helps RNA polymerase bind (Fig. 6.22). At this point RNA polymerase II can initiate synthesis of RNA. However, it is not yet free to move away from the promoter.

Release of RNA polymerase II from the promoter and elongation of the RNA requires three more TFII complexes, TFIIE, TFIIH and TFIIJ. In particular, TFIIH must phosphorylate the tail of RNA polymerase before it can move (Fig. 6.23). The tail, or **CTD (carboxy-terminal domain)**, consists of a seven-amino acid sequence (Tyr Ser Pro Thr Ser Pro Ser) repeated approximately 50 times. This may be phosphorylated on the serine or threonine residues. All of the TFII complexes except for TFIIH are left behind as RNA polymerase moves forward.

Like bacterial RNA polymerase, the eukaryotic RNA polymerases all have multiple subunits. RNA polymerase II has more than 10 subunits and shares three of these

CTD (carboxy-terminal domain) Repetitive region at the C-terminus of RNA polymerase II that may be phosphorylated

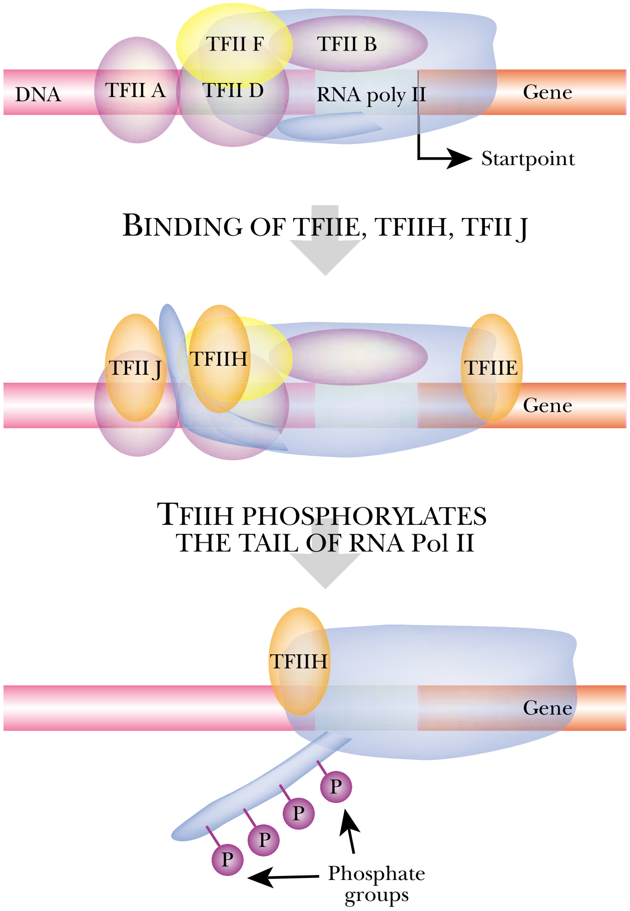


FIGURE 6.23 ***RNA Polymerase II Moves Forward from the Promoter***

Before RNA polymerase II can move forward, the binding of other factors must occur. One of these, TFIIH, phosphorylates the tail of RNA polymerase II. The tail changes position with respect to the body of RNA polymerase II. The other factors leave and RNA polymerase moves along the DNA and begins the process of transcription.

with RNA polymerases I and III. The largest subunit of RNA polymerase II is related to the β′ subunit of bacterial RNA polymerase and possesses the CTD tail. In addition, the assorted TFII complexes each consist of several polypeptide chains. Thus the initiation complex for RNA polymerase II includes over 20 polypeptides.

Upstream Elements Increase the Efficiency of RNA Polymerase II Binding

Upstream elements close to the promoter bind a range of specific transcription factors.

RNA polymerase II can bind and initiate transcription at minimal promoter consisting of just an initiator and TATA boxes. However, this is extremely inefficient unless upstream elements are also present. There are many different upstream elements. They are typically five to ten base pairs long and located from 50 to 200 bases upstream of the start site. There may be more than one upstream element in a given promoter and the same upstream element may be found at different places in different promoters.

The TFII proteins are *general* transcription factors because they are always required. In contrast, *specific* transcription factors affect only certain genes and are involved in regulating gene expression in response to a variety of signals (Fig. 6.24). The upstream elements are the recognition sites for specific transcription factors. These usually make contact with the transcription apparatus via TFIID, TFIIB or TFIIA, not by directly touching RNA polymerase II itself. Most commonly, binding is to TFIID. Binding of the specific transcription factors helps assembly of the transcription apparatus and therefore increases the frequency of initiation.

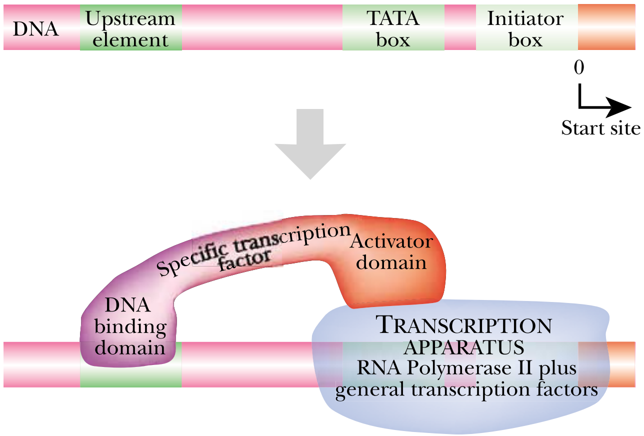


FIGURE 6.24 ***Upstream Elements Facilitate Transcription***

The upstream elements make contact with one domain of an activator protein. The activator also binds to the transcription apparatus near the start site.

TABLE 6.01 General Transcription Factors Associated with RNA Polymerase II

TBP	binds to TATA box, part of TFIID
TFIID	includes TBP, recognizes Pol II specific promoter
TFIIA	binds upstream of TATA box; required for binding of RNA Pol II to promoter
TFIIB	binds downstream of TATA box; required for binding of RNA Pol II to promoter
TFIIF	accompanies RNA Pol II as it binds to promoter
TFIIE	required for promoter clearance and elongation
TFIIH	phosphorylates the tail of RNA Pol II, retained by polymerase during elongation
TFIIJ	required for promoter clearance and elongation

Common upstream elements include the GC box, CAAT box, AP1 element and Octamer element. The GC box (GGGCGG) is often present in multiple copies. Despite being nonsymmetrical, the GC box works in either orientation and is recognized by the SP1 factor. Some upstream elements are recognized by more than one protein. In these cases, different transcription factors are often present in different tissues. For example, the Oct-1 and Oct-2 proteins both recognize the Octamer element. Oct-1 is found in all tissues but Oct-2 only appears in immune cells, where it helps activate genes encoding antibodies.

In eukaryotes most genes are subject to positive control. Repressors are rare and usually behave differently to those in bacteria.

The specific transcription factors that bind to the upstream elements and enhancers are thus gene activator proteins. Repressors are rare in eukaryotes. Furthermore, when found, they do not bind to the DNA directly and block the binding of the RNA polymerase. Instead, they bind to some component of the growing transcription apparatus and block further assembly.

Enhancers Control Transcription at a Distance

Enhancer sequences are located far away from the genes they control.

Enhancers are sequences that are involved in gene regulation, especially during development or in different cell types. Enhancers do exactly what their name indicates—they enhance the initiation of transcription as a result of binding specific transcription factors. Enhancers often consist of a cluster of recognition sites and therefore bind several proteins. Some recognition sites (e.g., Octamer and AP1) are found in both enhancers and as upstream elements in promoters.

Although enhancers are sometimes close to the genes they control, more often they are found a considerable distance, perhaps thousands of base pairs away, not even

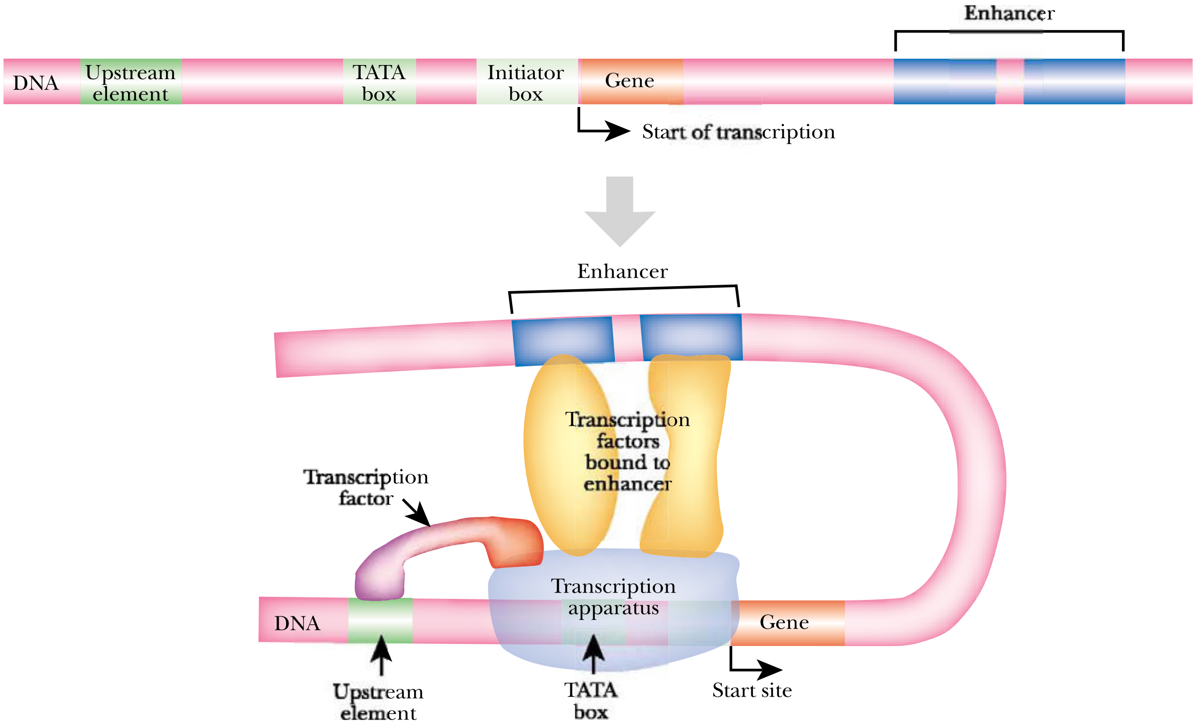


FIGURE 6.25 ***Looping Model for Enhancer***

The enhancer shown here is located downstream of the start site. To enhance transcription, the enhancer first binds several transcription factors. Subsequently, the DNA forms a loop allowing the enhancer to make contact with the transcription apparatus via the bound transcription factors.

associated with the gene. Enhancers may be located either upstream or downstream from the promoter and the position may vary from case to case. In addition, enhancers function equally well in either orientation. Experiments in which enhancers have been moved have shown that an enhancer will increase transcription from any promoter within its neighborhood. These properties imply that the enhancer must make contact with the transcription apparatus. When an enhancer switches a gene on, the DNA between it and the promoter loops out as shown in Figure 6.25.

CHAPTER SEVEN

Protein Structure and Function

Proteins Are Formed from Amino Acids

Formation of Polypeptide Chains

Twenty Amino Acids Form Biological Polypeptides

Amino Acids Show Asymmetry around the Alpha-carbon

The Structure of Proteins Reflects Four Levels of Organization

The Secondary Structure of Proteins Relies on Hydrogen Bonds

The Tertiary Structure of Proteins

A Variety of Forces Maintain the 3-D Structure of Proteins

Cysteine Forms Disulfide Bonds

Multiple Folding Domains in Larger Proteins

Quaternary Structure of Proteins

Higher Level Assemblies and Self-Assembly

Cofactors and Metal Ions Are Often Associated with Proteins

Nucleoproteins, Lipoproteins and Glycoproteins Are Conjugated Proteins

Proteins Serve Numerous Cellular Functions

Protein Machines

Enzymes Catalyze Metabolic Reactions

Enzymes Have Varying Specificities

Lock and Key and Induced Fit Models Describe Substrate Binding

Enzymes Are Named and Classified According to the Substrate

Enzymes Act by Lowering the Energy of Activation

The Rate of Enzyme Reactions

Substrate Analogs and Enzyme Inhibitors Act at the Active Site

Enzymes May Be Directly Regulated

Allosteric Enzymes Are Affected by Signal Molecules

Enzymes May Be Controlled by Chemical Modification

Binding of Proteins to DNA Occurs in Several Different Ways

Denaturation of Proteins

Proteins Are Formed from Amino Acids

The nucleic acids DNA and RNA are largely concerned with storing and distributing genetic information, and so are termed informational macromolecules. In contrast, proteins are biological polymers that carry out most of the cell's routine functions. Some proteins are *structural* or take part in maintaining cell shape and in carrying out cell movements; others *transport* nutrients and wastes; others function *enzymatically* by generating energy or by carrying out biochemical reactions, including the synthesis of nucleotides and their assembly into nucleic acids.

Molecules such as DNA and mRNA, whose primary role is to carry information, are linear with a regular repeating structure. [Note that the converse, possessing a repetitive linear structure, does not imply that a molecule carries information. Plastics such as polyethylene and a few specialized structural proteins such as collagen also have this type of structure.] Molecules that are structural and function in transport and/or as enzymes are normally folded into complex three-dimensional (3-D) structures. These include both proteins and certain specialized RNA molecules such as rRNA and tRNA. Nonetheless, both proteins and folded RNA molecules are first made as linear polymers and are folded later.

Proteins are linear polymers made from amino acids. Most proteins fold into complex 3D structures.

Proteins consist of a linear chain of monomers, known as amino acids, and are folded into a variety of complex 3-D shapes. A chain of amino acids is called a polypeptide chain. What is the difference between a polypeptide chain and a protein? Firstly, some proteins consist of more than one polypeptide chain and secondly, many proteins contain additional components such as metal ions or small organic molecules known as cofactors in addition to their polypeptide portion (see below).

Formation of Polypeptide Chains

Twenty different **amino acids** are used in making proteins. They all have a central carbon atom, the **alpha carbon**, surrounded by an amino group, a carboxyl group, a hydrogen atom and a side chain or **R-group**, as shown in Figure 7.01A (proline is an exception—see below). The simplest amino acid is **glycine** (Fig 7.01B) in which the R-group is just a single hydrogen atom. In solution, under physiological conditions, the amino group and the carboxyl group of amino acids are both ionized, to give a **zwitterion** or **dipolar ion** with one positive and one negative charge (Fig. 7.01C).

Amino acids are joined together by **peptide bonds** (Fig. 7.02) to give a **polypeptide chain**. The first amino acid in the chain retains its free amino (NH_2) group and this end is therefore called the **amino- or N-terminus** of the polypeptide chain. The last amino acid to be added is left with a free carboxy (COOH) group, so this end is the **carboxy-** or **C-terminus**. When synthesized, the polypeptide is elongated from the amino terminus toward the carboxy terminus.

Twenty Amino Acids Form Biological Polypeptides

The twenty amino acids found in proteins possess a variety of different chemical groups (Fig. 7.03). This wide choice of possible monomers makes proteins very versatile, with

alpha- (α-) carbon Central carbon atom of an amino acid that carries both the amino group and the carboxyl group
amino acid Monomer from which polypeptide chains are built
amino- or N-terminus The end of a polypeptide chain that is made first and that has a free amino group
carboxy- or C-terminus The end of a polypeptide chain that is made last and has a free carboxy-group.
dipolar ion Same as zwitterion; a molecule with both a positive and a negative charge
glycine The simplest amino acid
peptide bond Type of chemical linkage holding amino acids together in a polypeptide chain
polypeptide chain A polymer that consists of amino acids
R-group Any unspecified chemical group; in particular the side chain of an amino acid
zwitterion Same as dipolar ion; a molecule with both a positive and a negative charge

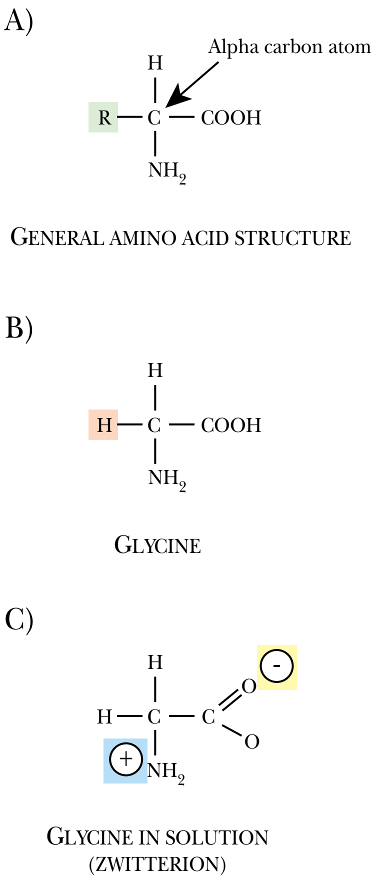


FIGURE 7.01 ***General Structure of Amino Acids***

A) A generalized amino acid contains an alpha carbon atom, an R-group, an NH_2 group and a COOH group. B) Glycine is the simplest amino acid with an H atom as the R-group. C) When glycine is placed in solution at neutral pH it ionizes to form a zwitterion.

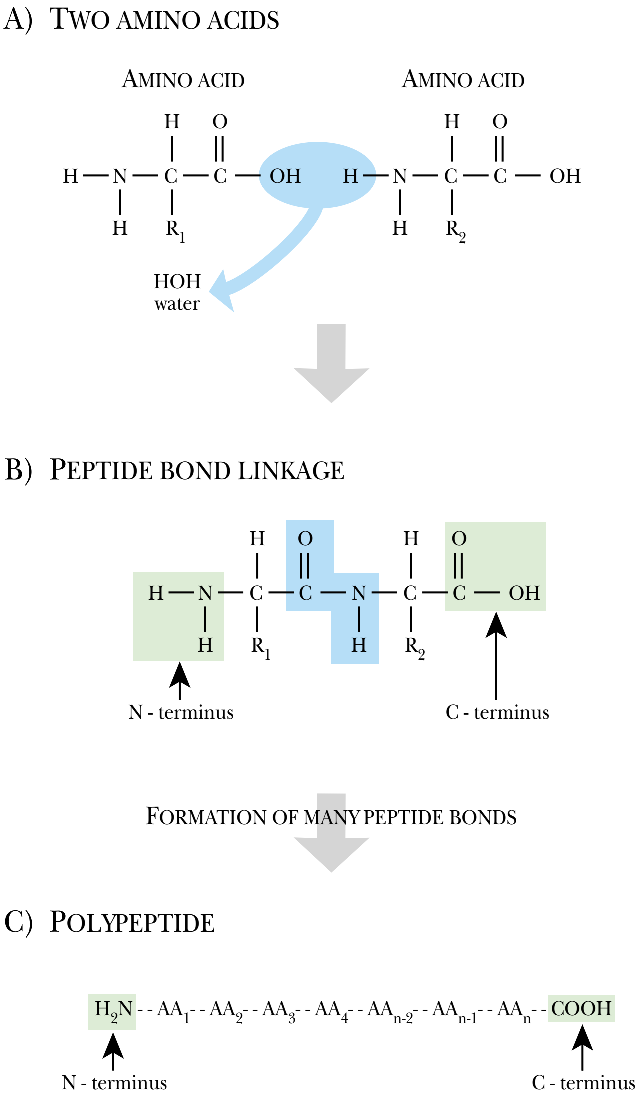


FIGURE 7.02 ***Polypeptide Chain Is Made of Amino Acids***

A) Two generic amino acids are shown. The R-groups, R1 and R2 represent the side chains of any of the 20 different amino acids that make up proteins. Each amino acid has an amino (NH_2) and carboxyl (COOH) group. B) A peptide bond is formed between one NH_2 and one COOH group, with water eliminated in the process. C) Successive amino acids are joined in a similar manner by peptide bonds to form a polypeptide. The polypeptide contains an amino (or N-) terminus and a carboxyl (or C-) terminus. The side chains of the successive amino acids are labeled R1, R2, R3, etc.

> The twenty amino acids that comprise proteins vary greatly in their chemical and physical properties.

a wide range of properties and capabilities, including a great variety of possible 3-D structures. The amino acids may be classified into groups depending on their physical and chemical characteristics. The major division is between those with **hydrophilic** (water-loving) and those with **hydrophobic** (water-hating) R-groups. Glycine has only a single hydrogen atom as its side chain, so it does not really fit into either group.

The hydrophilic amino acids may be subdivided into basic, acidic and neutral. Basic amino acids contribute a positive charge to the protein whereas acidic residues provide a negative charge. Strictly, this refers to the situation in solution within the physiological pH range. Neutral polar residues have side chains that are capable of forming

hydrophilic Water-loving; readily dissolves in water
hydrophobic Water-hating; repelled by water and dissolves in water only with great difficulty

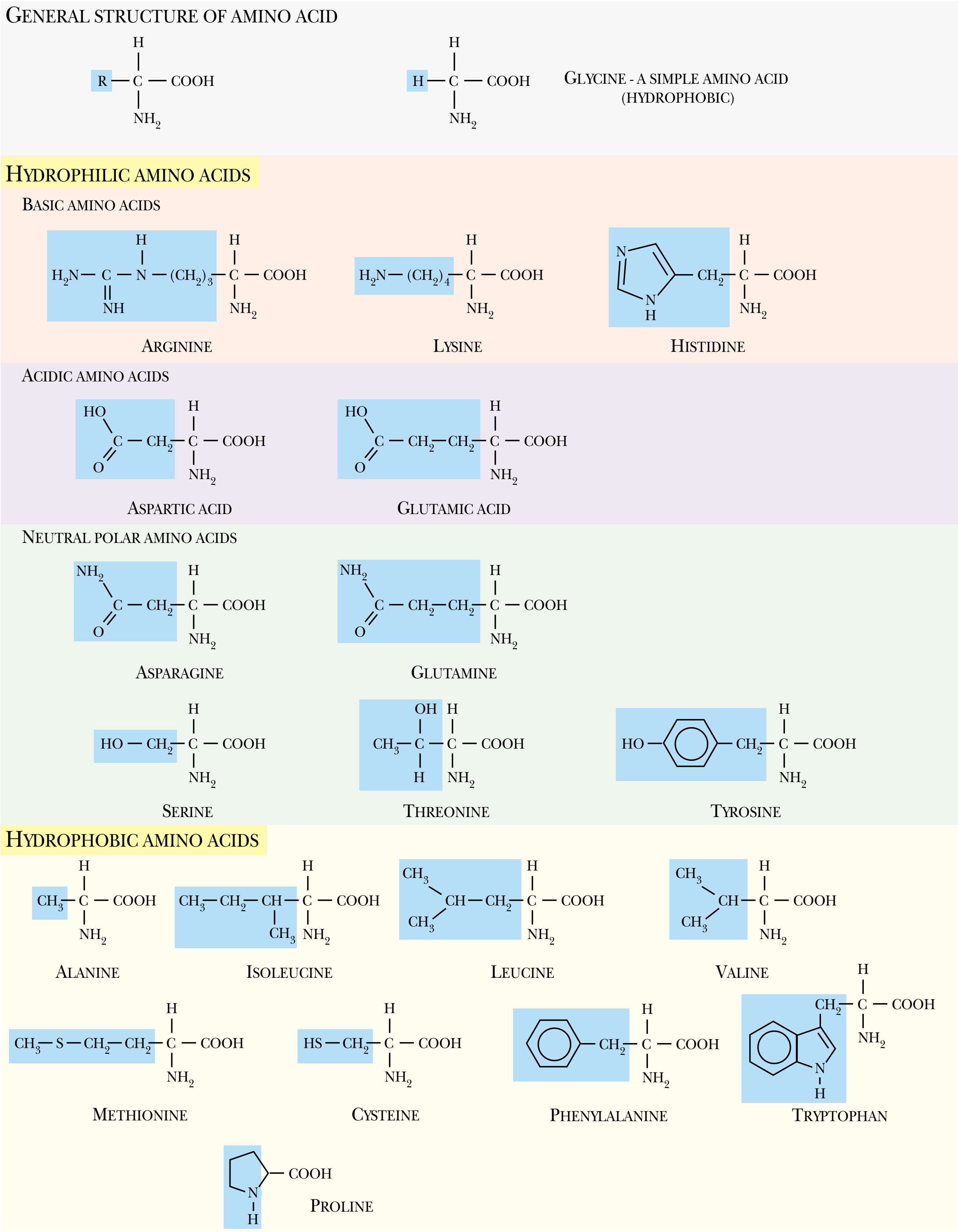


FIGURE 7.03 ***The Twenty Amino Acids Found in Proteins***

Amino acids can be grouped by their physical and chemical properties. The R-group for each amino acid is highlighted.

hydrogen bonds. The side chains of the hydrophilic amino acids carry chemical groups that can take part in reactions. The active sites of enzymes (see below) often contain serine, histidine and basic or acidic amino acids.

The hydrophobic amino acids may be subdivided into those that are aliphatic (Ala, Leu, Ile, Val, Met) and those containing aromatic rings (Phe, Trp and Tyr). These amino acid residues are largely structural in function, except for tyrosine which has a hydroxyl group attached to its aromatic ring and can therefore take part in a variety of reactions. The classification of tyrosine is ambiguous as its hydroxyl group is polar in nature. Proline has a non-aromatic ring and, strictly speaking, is an imino acid (rather than an amino acid) since it has an —NH— (imino) group as part of a ring (rather than a free amino group).

Two of the hydrophobic amino acids, Met and Cys, contain sulfur. When a polypeptide chain is first synthesized, methionine is always the first amino acid, although it may be trimmed off later. Cysteine is important for 3-D structure as it forms disulfide bonds (see below). The free **sulfhydryl group** of cysteine is highly reactive and is often used in enzyme active sites or to attach various chemical groups to proteins (see below).

The twenty different amino acids normally found in proteins may be represented by both three-letter and one-letter abbreviations (Table 7.01). These mostly correspond to the first letter(s) of the name, but since several amino acids sometimes start with the same letter of the alphabet, the others need a little imagination. They are used especially when writing out protein sequences. The amides, asparagine and glutamine, are relatively unstable and break down easily into the corresponding acids, aspartate and glutamate. Consequently, many analyses do not distinguish the acids from their amides. The abbreviations Asx and Glx were invented for these ambiguous pairs.

Amino Acids Show Asymmetry around the Alpha-carbon

Amino acids occur as pairs of optical isomers. Natural proteins are made from the L-isomers.

Apart from glycine, the amino acids have four different chemical groups surrounding the central (alpha) carbon atom. This is called a **chiral** or **asymmetric center** (Fig. 7.04). Consequently such amino acids exist as two alternative mirror-image isomers with different "chirality" or "handedness." A pair of mirror-image isomers is known as **enantiomers** or **optical isomers**. They are referred to as the **L-** and **D- forms**.

The designations L- and D- refer to the fact that solutions of chiral molecules rotate the plane of polarization of light in either a left-handed (L- = levorotatory) or right handed (D- = dextrorotatory) direction. The optical rotation is cancelled out if equal amounts of both L- and D- forms are present. An equal mixture of L- and D- isomers is known as a **racemic mixture** and enzymes that interconvert the L- and D- isomers of a molecule are referred to as **racemases**. A molecule that has multiple asymmetric centers, such as a polypeptide chain, will have many possible stereoisomers, since each center may exist in L- or D- conformations.

The D-isomers of amino acids are found naturally in bacterial cell walls and some antibiotics.

The amino acids found in proteins are all of the L- form. Although L- amino acids are sometimes referred to as the "natural" isomers, D- amino acids do exist in nature. The **peptidoglycan** that is found in bacterial cell walls contains several different

asymmetric center Carbon atom with four different groups attached. This results in optical isomerism
chiral center Same as asymmetric center
enantiomers A pair of mirror-image optical isomers (i.e., D- and L-isomers)
L- and D-forms The two isomeric forms of an optically active substance; also called L- and D-isomers
optical isomers Isomers where the molecules differ only in their 3D arrangement and consequently affect the rotation of polarized light
peptidoglycan Polymer that makes up eubacterial cell walls; consists of long chains of sugar derivatives, cross-linked at intervals with short chains of amino acids
racemase An enzyme that interconverts the D- and L-isomers of an optically active substance
racemic mixture Mixture of equal amounts of both D- and L-isomers of an optically active substance
sulfhydryl group -SH; Chemical group of sulfur and hydrogen

TABLE 7.01 Amino Acids and Their Properties

Amino Acid	3-Letter Code	1-Letter Code	Physical Properties
Alanine	Ala	A	hydrophobic
Arginine	Arg	R	basic
Asparagine	Asn	N	neutral polar
Aspartic acid	Asp	D	acidic
Cysteine	Cys	C	hydrophobic
Glutamic acid	Glu	E	acidic
Glutamine	Gln	Q	neutral polar
Glycine	Gly	G	—
Histidine	His	H	basic
Isoleucine	Ile	I	hydrophobic
Leucine	Leu	L	hydrophobic
Lysine	Lys	K	basic
Methionine	Met	M	hydrophobic
Phenylalanine	Phe	F	hydrophobic
Proline	Pro	P	hydrophobic
Serine	Ser	S	neutral polar
Threonine	Thr	T	neutral polar
Tryptophan	Trp	W	hydrophobic
Tyrosine	Tyr	Y	neutral polar/ hydrophobic
Valine	Val	V	hydrophobic
Aspartic acid or Asparagine	Asx	B	
Glutamic acid or Glutamine	Glx	Z	
Unspecified amino acid		X	

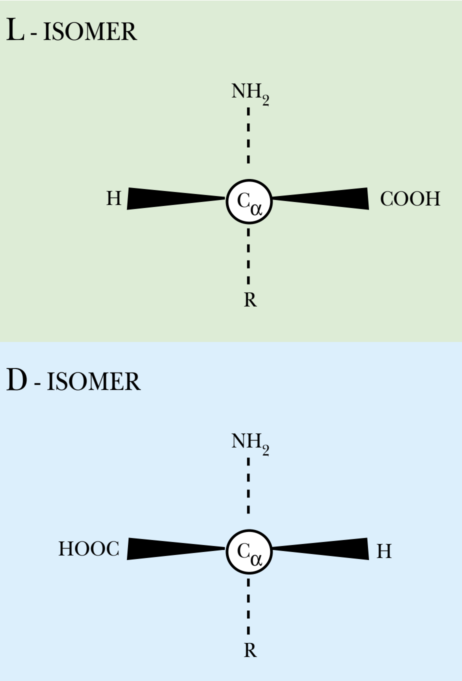


FIGURE 7.04 ***The L- and D-Forms of an Amino Acid***

The four groups are arranged around the alpha carbon differently in the L-form and the D-form of an amino acid. Although they share the same molecular formula, one is the mirror image of the other.

D-amino acids and several peptide antibiotics, also made by prokaryotes (e.g., bacitracin, polymixin B, actinomycin D), contain D- amino acids.

The Structure of Proteins Reflects Four Levels of Organization

The linear polypeptide chains must be folded into the correct 3-D structure to function properly. Furthermore, many proteins are assembled from more than one polypeptide chain and many also have **cofactors** or **prosthetic groups**—associated molecules that are not made of amino acids. The final shape of a protein is determined by its amino acid sequence, so proteins with similar sequences have similar 3-D conformations.

Linear polypeptide chains are folded up to give the final 3D structures.

Typical polypeptides are 300–400 amino acids long. Polypeptides much smaller or much larger are less common. However, many hormones and growth factors, such as insulin, do consist of relatively short polypeptide chains. Individual polypeptides with more than a thousand amino acids are very rare and very large proteins tend to consist of several separate polypeptide chains rather than a single long chain.

The structures of biological polymers, both proteins and nucleic acids, are often divided into four levels of organization:

1. **Primary structure** is the order of the monomers; i.e., the sequence of the amino acids for a protein, or of the nucleotides in the case of DNA or RNA.
2. **Secondary structure** is the folding or coiling of the original polymer chains by means of hydrogen bonding. In the case of proteins, the hydrogen bonds are between the atoms of the polypeptide backbone.
3. **Tertiary structure** is the further folding that gives the final 3-D structure of a single polymer chain. In the case of proteins, this involves interactions between the R groups of the amino acids.
4. **Quaternary structure** is the assembly of several separate polymer chains.

The Secondary Structure of Proteins Relies on Hydrogen Bonds

By definition, the secondary structure is folding that depends solely on hydrogen bonding. In DNA, hydrogen bonding occurs between base pairs and is the basis of the double helix. In proteins, hydrogen bonding occurs between the peptide groups that form the backbone of the polypeptide (Fig. 7.05). The polypeptide chain must be folded around to bring two peptide groups alongside each other. The hydrogen on the nitrogen of one peptide group is then bound to the oxygen of the other. [Note that hydrogen bonds also contribute to tertiary structure, but here they are not the only or even the major forces involved.]

Most of the secondary structure found in proteins is due to one of two common secondary structures, known as the **α- (alpha) helix** and the **β- (beta) sheet**. Both structures allow formation of the maximum possible number of hydrogen bonds and are therefore highly stable.

alpha- (α-) helix A helical secondary structure found in proteins
β- (beta-) sheet A flat sheet-like secondary structure found in proteins
cofactor Extra chemical group bound (often temporarily) to a protein but which is not part of the polypeptide chain
primary structure The linear order in which the subunits of a polymer are arranged
prosthetic group Extra chemical group bound (often covalently) to a protein but which is not part of the polypeptide chain
quaternary structure Aggregation of more than one polymer chain to form a final structure
secondary structure Initial folding up of a polymer into a regular, repeating structure, due to hydrogen bonding
tertiary structure Final 3-D folding of a polymer chain

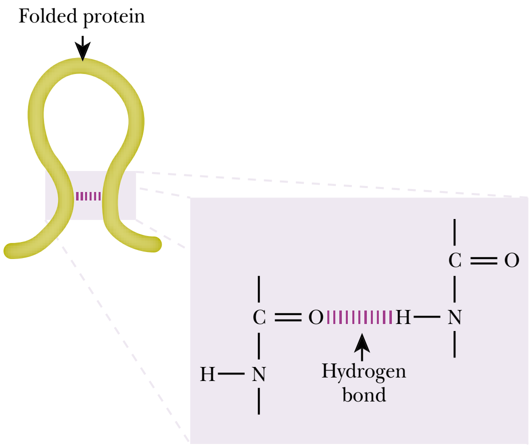


FIGURE 7.05 ***Hydrogen Bonding between Peptide Groups***

Two peptide bonds of a polypeptide chain may be aligned to form a hydrogen bond by looping the polypeptide chain around.

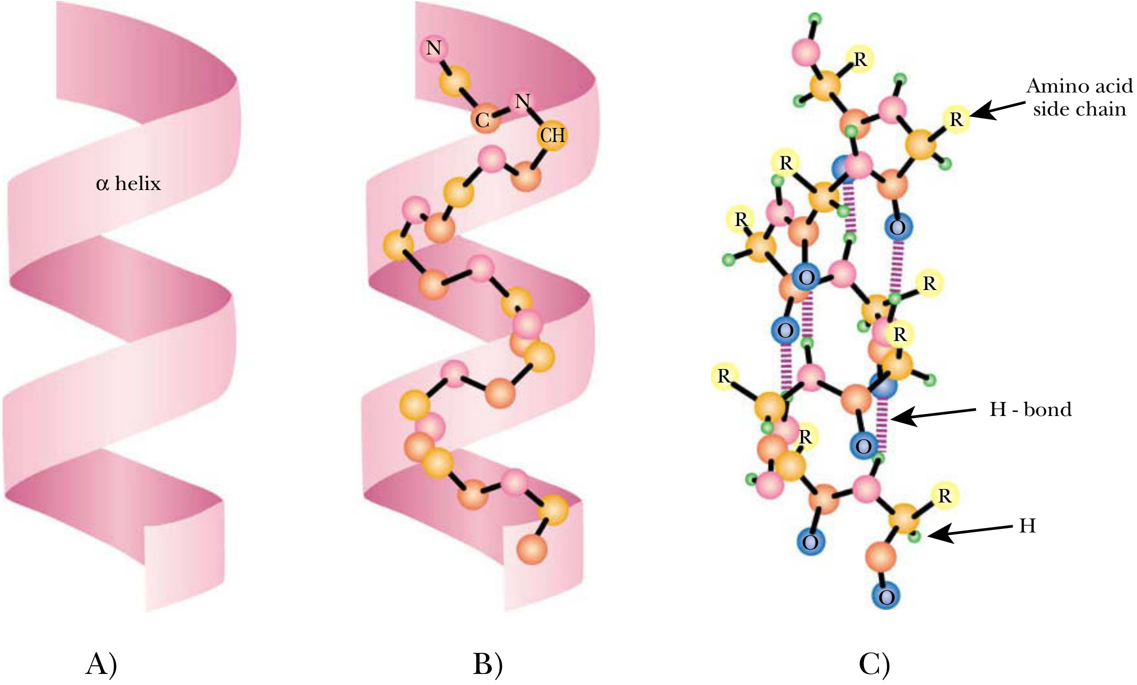


FIGURE 7.06 ***The Alpha Helix***

A) The general shape of an α-helix. B) The carbon backbone of the polypeptide chain. C) The hydrogen bonds between peptide groups.

Hydrogen bonding is responsible for the formation of alpha-helix and beta-sheet structures in proteins.

In the α-helix (Fig. 7.06), a single polypeptide chain is coiled into a right-handed helix and the hydrogen bonds run vertically up and down, parallel to the helix axis. In fact, the hydrogen bonds in an α-helix are not quite parallel to the axis. They are slightly tilted relative to the helix axis because there are 3.6 amino acids per turn rather than a whole number. The pitch (repeat length) is 0.54 nm and the rise per residue is about 0.15 nm.

The hydrogen bonds hold successive twists of the helix together and run from the C═O group of one amino acid to the NH group of the fourth amino acid residue down the chain. The α-helix is very stable because all of the peptide groups (—CO—NH—) take part in two hydrogen bonds, one up and one down the helix axis. A right-handed helix is most stable for L- amino acids. (A stable helix cannot be formed

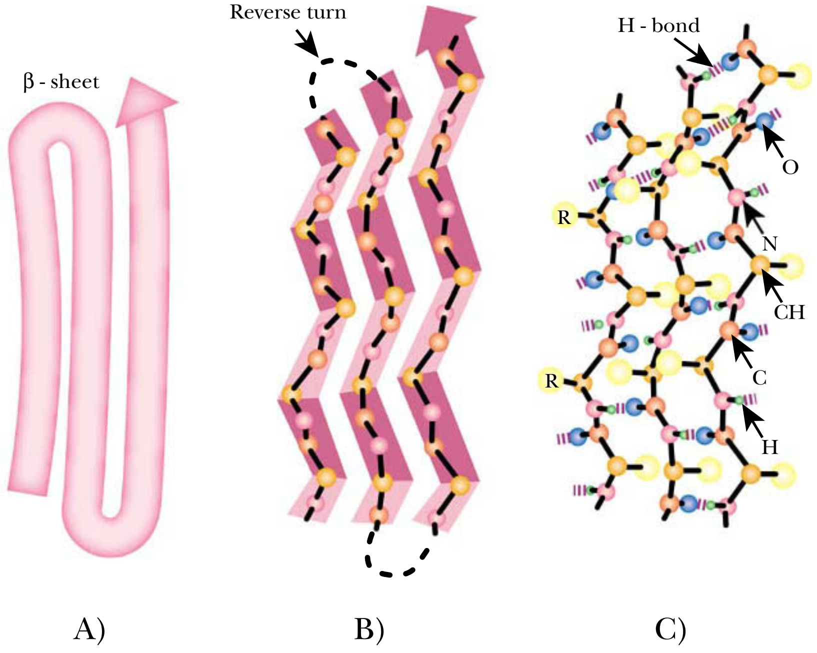
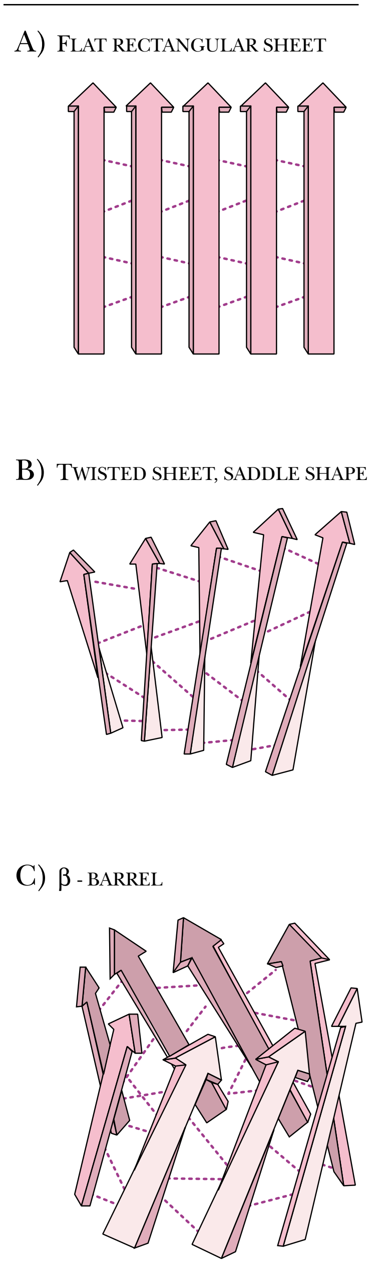


FIGURE 7.07 ***The Beta Sheet—Hydrogen Bonding***

A) A polypeptide chain is shown folded back and forth three times to form a flattened sheet. B) The polypeptide backbone of the sheet is depicted demonstrating how the β-sheet forms a zigzag in three dimensions. C) A ball and stick model shows the hydrogen bonding between the strands of the β-sheet.


A) FLAT RECTANGULAR SHEET

B) TWISTED SHEET, SADDLE SHAPE

C) β - BARREL

FIGURE 7.08 ***The Beta Sheet Conformations***

A) Flat sheet. B) Twisted sheet. C) Barrel.

with a mixture of D- and L- amino acids, although a stable left-handed helix could theoretically be formed from D- amino acids).

The R-groups extend outwards from the tightly packed helical polypeptide backbone. Of the 20 amino acids, Ala, Glu, Leu and Met are good α-helix formers but Tyr, Ser, Gly and Pro are not. Proline is totally incompatible with the α-helix, due to its rigid ring structure. Furthermore, when proline resides are incorporated, no hydrogen atoms remain on the nitrogen atom that takes part in peptide bonding. Consequently, proline residues interrupt hydrogen-bonding patterns. In addition, two bulky residues or two residues with the same charge that lie next to each other in the polypeptide chain will not fit properly into an α-helix. Overall, the α-helix forms a solid cylindrical rod.

The β-sheet is also held together by hydrogen bonding between peptide groups but in this case the polypeptide chain is folded back on itself to give a flattish zigzag structure (Fig. 7.07). Like the α-helix, the β-sheet is very stable because all of the peptide groups (except for those on the edge of the sheet) take part in two hydrogen bonds. In the β-sheet, these go sideways from each peptide group, one to each side. The sections of the polypeptide chain which lie side by side are usually, although not always, antiparallel and the R-groups lie alternately above and below the zigzagging plane of the β-sheet.

A variety of β-sheet conformations is known (Fig. 7.08). Although some β-sheets are flat, most known β-sheets are twisted (in a right-handed manner) so that the sheet structure as a whole is not flat but curved. In some cases, β-sheets may curve around so that the last strand bonds to the first, forming a barrel structure.

A **reverse turn** (also known as a β-turn or β-bend) is where the polypeptide chain turns back upon itself. Beta-sheets have reverse turns at the ends of each segment, but such turns are also found frequently in other places. Pro and Gly are often found in

reverse turn Region of polypeptide chain that turns around and goes back in the same direction

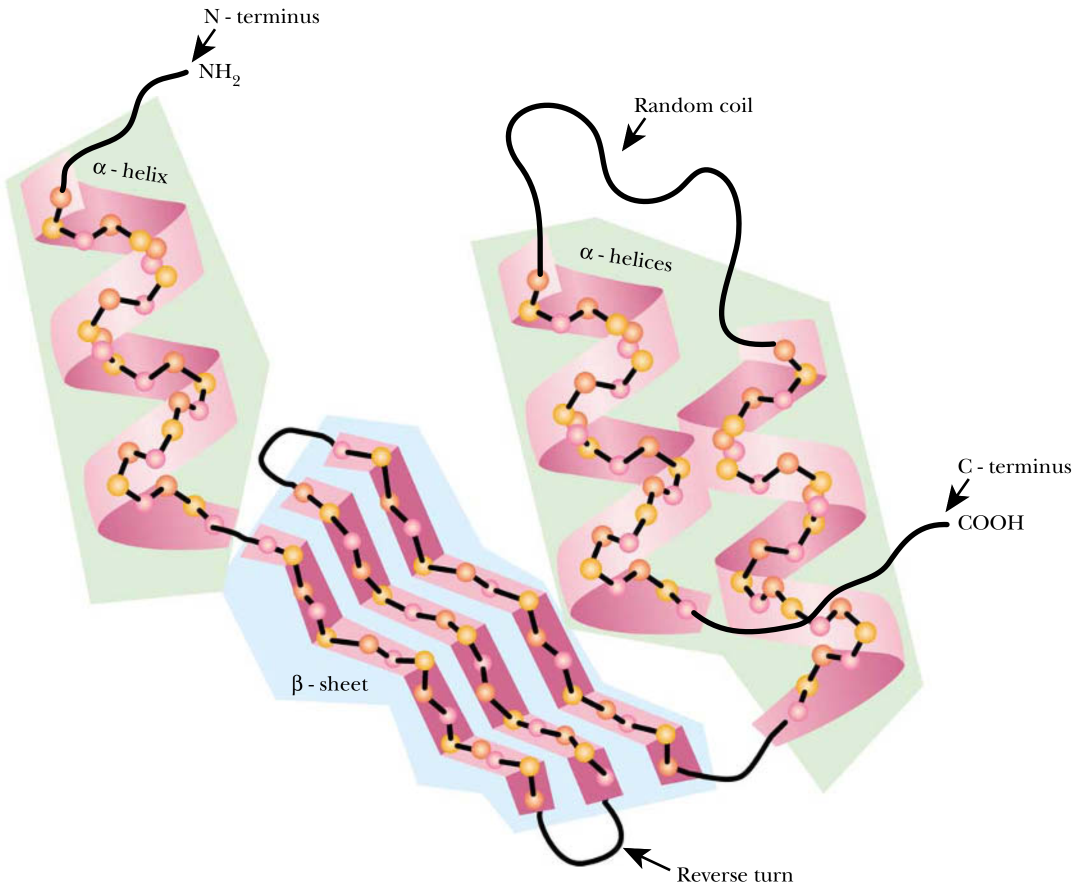


FIGURE 7.09 ***Modular Arrangement of α-Helices and β-Sheets***

A stylized polypeptide chain. Both α-helical regions and a β-sheet segment are shown in this folded polypeptide chain. These modular segments are linked by reverse turns and regions of random coil.

reverse turns. Those regions of protein that do not form secondary structures are referred to as "**random coil**" although they are, of course, not truly random, but merely irregular.

The Tertiary Structure of Proteins

Further folding of the polymer chain constitutes the tertiary structure. In a nucleic acid this would be the supercoiling. In a protein, the polypeptide chain, with its preformed α-helix and β-sheet regions, is folded to give the final 3-D structure. In general, polypeptide chains with similar amino acid sequences fold to give similar 3-D structures. Tertiary folding depends on interactions between the side chains of the individual amino acids. Since there are 20 different amino acids, a large variety of final 3-D conformations is possible, although most polypeptides are roughly spherical.

The α-helix and β-sheet modules form the basic structural units of the protein (Fig. 7.09). They are linked by loops of random coil of various lengths. Many of these loops are at the surface of the protein, exposed to a solvent and contain predominantly charged or polar amino acid residues. The rigid ring structure of proline causes an approximately 90° change in direction of the polypeptide backbone. Consequently,

random coil Region of polypeptide chain lacking secondary structure

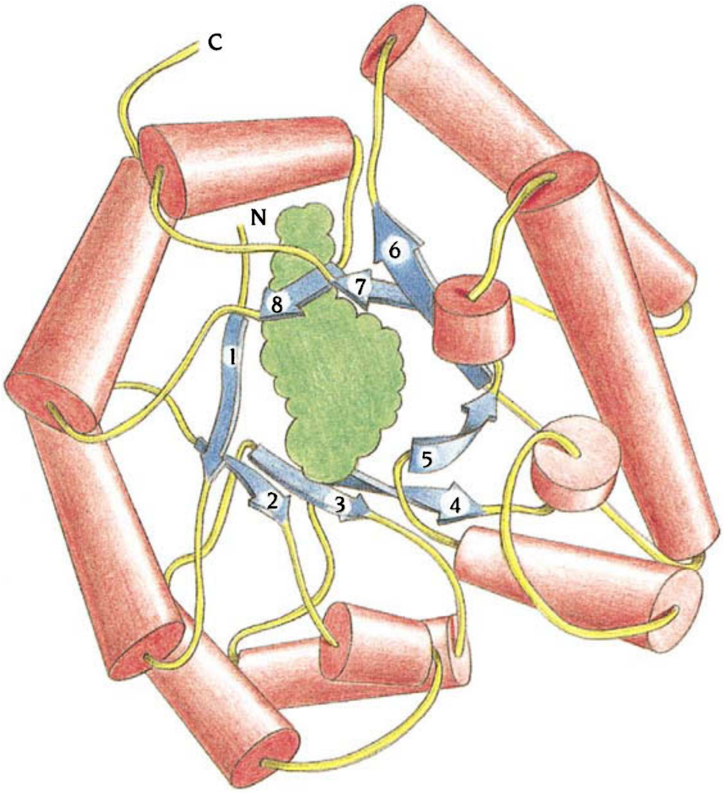


FIGURE 7.10 ***Arrangement of an α/β-Barrel Domain***

Schematic diagram of the α/β barrel domain of the enzyme methylmalonyl CoA mutase. Alpha helices are red and beta strands are blue. The inside of the barrel is lined by small hydrophilic side chains (Ser and Thr) which allows space for the substrate coenzyme A (green) to bind along the axis of the barrel. From: Introduction to Protein Structure by Brandon & Tooze, 2nd ed., 1999. Garland Publishing, Inc., New York and London.

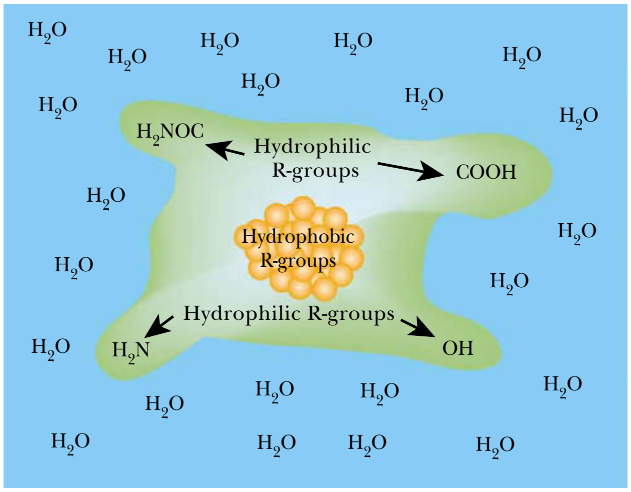


FIGURE 7.11 ***The Oil Drop Model of Protein Structure***

A simplified model to illustrate that hydrophilic groups are exposed to the water environment and that most hydrophobic groups are centrally positioned. Note that in real life many of the hydrophilic groups will form hydrogen bonds to the surrounding water molecules. In addition some hydrophilic groups will ionize, forming charged ions.

proline disrupts secondary structures and contributes to overall folding by forming bends. Examination of their 3-D structures has shown that the thousands of known proteins are in fact built from relatively few structural motifs. Such motifs generally consist of several α-helices and/or β-sheets joined to form a useful and recognizable structure (Fig. 7.10).

This 3-D folding is largely driven by two factors acting in concert. Many of the amino acids have R-groups that are very water-soluble (hydrophilic). These side chains prefer to be on the surface of the protein so they can dissolve in the water surrounding the protein and make hydrogen bonds to water molecules. In contrast, R-groups that are water repellent (hydrophobic) huddle together inside the protein away from the water (Fig. 7.11). Since hydrophobic molecules are greasy and insoluble, this

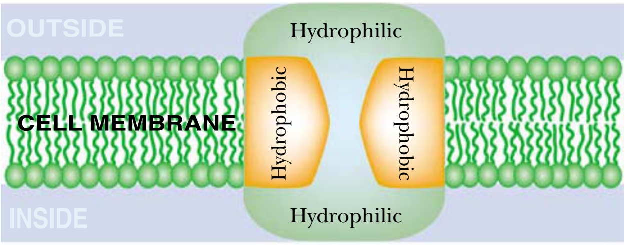


FIGURE 7.12 ***Inverse Conformation of a Membrane Protein***

The hydrophilic regions of membrane proteins interact with the aqueous cell interior and exterior. Their hydrophobic regions face the hydrophobic phospholipds of the membrane.

arrangement is known as the **oil drop model** of protein structure. The terms hydrophobic interaction, hydrophobic bonding or apolar bonding all refer to the tendency of non-polar groups to cluster together and avoid contact with water.

Hydrophobic amino acids tend to cluster together inside proteins so promoting 3D folding.

The formation of hydrophobic bonds is driven mostly by effects on water structure, not by any inherent attraction of hydrophobic groups for each other. Strictly, the term hydrophobic (which means "fearing/disliking water") is misleading as it is the water which dislikes the dissolved non-polar groups. Exposed hydrocarbon residues exert an organizing effect on surrounding water molecules. This decreases the entropy of the water and is thermodynamically unfavorable. Removal of hydrocarbon residues allows the water to return to its less organized H-bonding structure, which results in a large increase in entropy (approximately 0.7 Kcal per methylene group removed). Thus removing a leucine side chain from contact with water releases 3.5 Kcal/mole. Because the hydrophobic interaction depends on entropy, the strength of hydrophobic bonding increases with temperature, unlike most other forms of bonding which become less stable at higher temperatures.

Proteins that are inserted deeply into membranes show a conformation that is the inverse of the standard oil drop. They are hydrophobic on the surfaces where they contact the membrane lipid. Their hydrophilic residues are mostly clustered internally, but some are found at the surface in those regions where the protein emerges from the membrane (Fig. 7.12).

A Variety of Forces Maintain the 3-D Structure of Proteins

In addition to the major influence of hydrophobic interactions in the core of the protein and the hydrogen bonding of hydrophilic side chains to water, a variety of other effects are important (Fig. 7.13). These include hydrogen bonds, ionic bonds, van der Waals forces, and disulfide bonds.

Hydrogen bonds may form between the R-groups of two nearby amino acids. Those amino acids with hydroxyl, amino or amide groups in their side chains can take part in such hydrogen bonding. Similarly, ionic bonds (—NH_3^+ ^-OOC—) may form between the R-groups of basic and acidic amino acid residues (Fig. 7.13). Relatively few of the possible ionic interactions occur in practice. This is because most polar groups are on the surface of the protein and form hydrogen bonds to water.

Hydrophilic amino acids often end up on the surface of folded proteins where they make contact with water.

Van der Waals forces hold molecules or portions of molecules together if they fit well enough to approach very closely. Van der Waals forces are weak and decrease very rapidly with distance. Consequently, they are only significant for fairly large regions that are complementary in shape. Disulfide bonds between cysteines are also sometimes important in maintaining 3-D structure (see below).

oil drop model Model of protein structure in which the hydrophobic groups cluster together on the inside away from the water

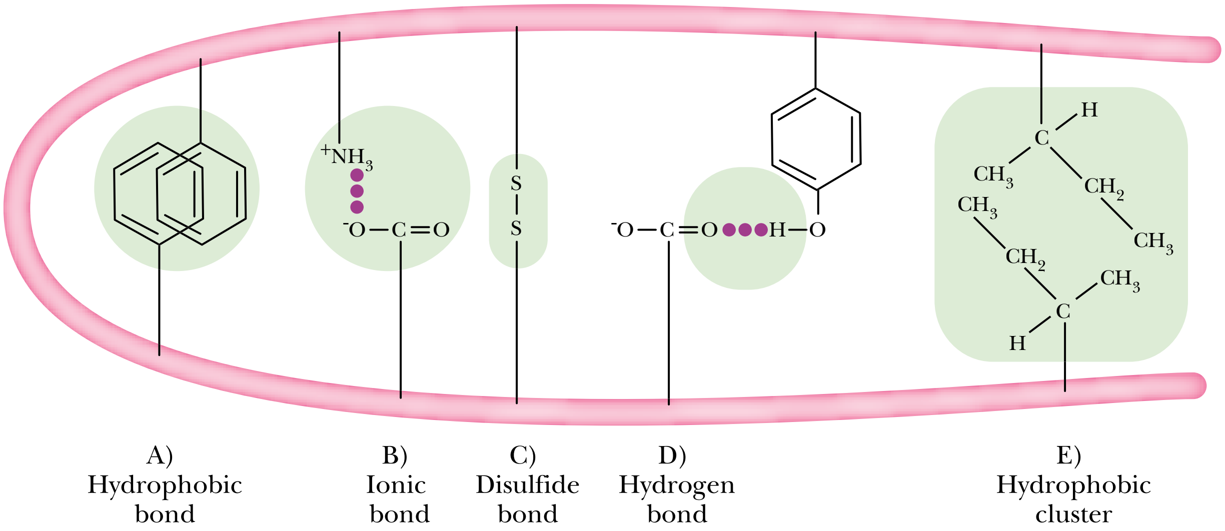


FIGURE 7.13 ***Some Forces that Maintain 3-D Structure of Proteins***

Other forces that maintain the 3-D structure include: A) hydrophobic ring stacking, B) ionic bonds, C) disulfide bonds, D) hydrogen bonds and E) hydrophobic clustering.

Cysteine Forms Disulfide Bonds

Disulfide bonds between two cysteine residues can stabilize protein structures.

Under oxidizing conditions, the sulfhydryl groups of two cysteines can form a **disulfide bond**. The dimer consisting of two cysteines (pronounced "cystEEn") joined by a disulfide bond is known as cystine (pronounced "cystYne") (Fig. 7.14). Disulfide bonds between cysteine residues are important in maintaining 3-D structure in certain cases (see Fig. 7.13, above). Disulfide bonds may hold together two regions of the same polypeptide chain (intrachain disulfide bond for tertiary structure) or may be used to hold together two separate polypeptides (interchain disulfide bond for quaternary structure).

Since disulfides are easily reduced to sulfhydryl groups inside cells, they are of little use in stabilizing intracellular proteins. Disulfide bonds are mostly used to stabilize extracellular proteins that are exposed to more oxidizing conditions. The classic examples are the antibodies that circulate in the blood of vertebrates. Secreted enzymes, such as **lysozyme**, or hormones, such as insulin, also rely on disulfide bonds. Single-celled organisms make relatively few extracellular proteins compared to higher multicellular organisms and consequently employ disulfide bonding much less.

Multiple Folding Domains in Larger Proteins

In long proteins folding may occur separately in different regions of the polypeptide chain.

Long polypeptide chains may contain several regions that fold up more or less independently and are joined by linker regions with little 3-D structure. Such regions are known as **domains** and may be from 50–350 amino acids long (Fig. 7.15). Short proteins may have a single domain and extremely long proteins may occasionally have up to a dozen.

Note that proteins begin to fold before they are completely made. As soon as a sufficient length of polypeptide chain to form a 3-D structure has emerged from the ribosome, it folds. Hence domains fold up independently, one after the other. (Because

disulfide bond A sulfur to sulfur bond formed between two sulfhydryl groups, in particular between those of cysteine, and which binds together two protein chains
domain (of protein) A region of a polypeptide chain that folds up more or less independently to give a local 3D-structure
lysozyme Enzyme that degrades peptidoglycan, the cell wall polymer of bacteria

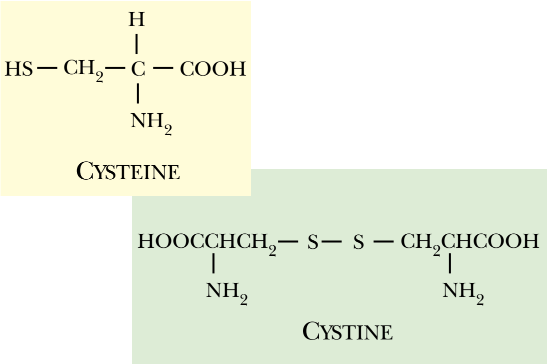


FIGURE 7.14 ***Cysteine and Cystine***

Two cysteines will join by a disufide bond to form cystine.

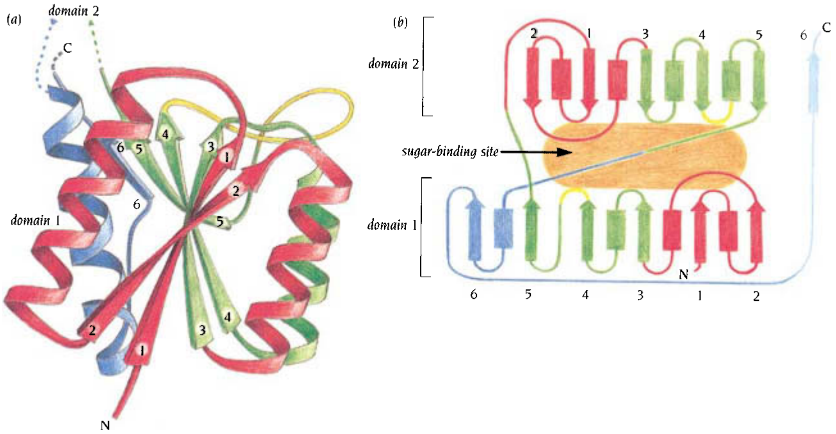


FIGURE 7.15 ***Two Domains in Arabinose Binding Protein***

The arabinose binding protein of *E. coli* contains two open twisted α/β domains of similar structure. A) Schematic diagram of a single domain. B) Topology diagram showing the orientation of the two domains and the crevice between them in which the arabinose molecule binds. From: Introduction to Protein Structure by Brandon & Tooze, 2nd ed., 1999. Garland Publishing, Inc., New York and London.

the sequence of folding is different, this also means that the refolding a denatured polypeptide often differs significantly from the original folding.)

Many transcription factors consist of two domains—one that binds DNA and another that binds the signal molecule. When the signal molecule is bound, it changes the shape of its own domain (Fig. 7.16). The change in conformation is then transmitted to the DNA-binding domain, which also changes shape. Thus, although they fold separately, domains do interact physically.

Quaternary Structure of Proteins

Many proteins consist of several individual polypeptide chains. This is especially true of proteins whose total molecular weight is much greater than 50,000 daltons (i.e. around 400 amino acids). [Although occasional polypeptide chains are found with a 1,000 or more amino acids, they are relatively rare.] The assembly of these multiple subunits yields the quaternary structure (Fig. 7.17). (Proteins with only one polypeptide chain have no quaternary structure.) The subunits, or **protomers**, are usually present as an even number, most often two or four. The terms dimer, trimer, tetramer, oligomer and multimer refer to structures with two, three, four, few/several and multiple subunits, respectively. Less than 10 percent of **multimeric** proteins have an odd number of subunits. The subunits may be all identical or all different or several each of two (or more) different types. For example, the lactose repressor consists of four identical subunits, whereas hemoglobin has two α-subunits and two β-subunits. The

Many proteins consist of subunits, usually an even number.

multimeric Formed of multiple subunits
protomer A single polymer chain that is itself a subunit for a higher level of assembly

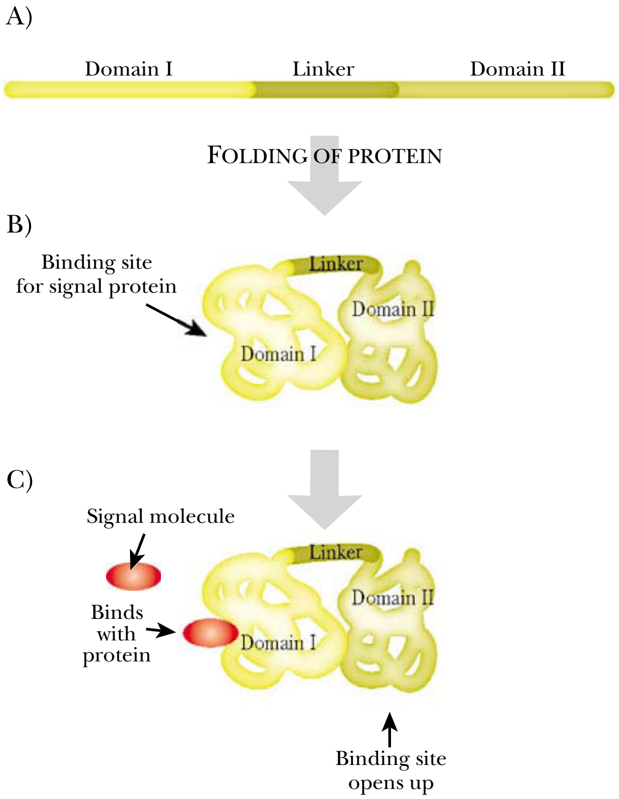


FIGURE 7.16 ***Interactions between Protein Domains to Activate a DNA-Binding Site***

A) The unfolded protein shows two domains (I and II) and a linker region. B) When folded, both domain I and domain II form binding sites. C) A signal molecule binds to domain I and changes its conformation. The interaction between the two domains triggers domain II to change shape so opening up its binding site.

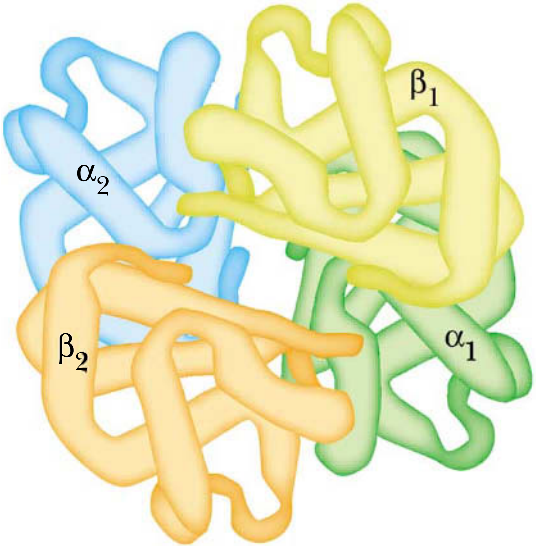


FIGURE 7.17 ***Hemoglobin—An Example of a Heterotetramer***

Two α-subunits and two β-subunits form the hemoglobin tetramer. The two α-subunits are identical as are the two β-subunits. However, the α-subunits differ from the β-subunits, although both types of chain are related in amino acid sequence and have a similar overall shape.

prefixes homo- (same) and hetero- (different) are sometimes used to indicate whether the subunits are the same or different. Thus the lactose repressor is a homo-tetramer, whereas hemoglobin is a hetero-tetramer.

The same hydrophobic forces largely responsible for tertiary structure are involved in the assembly of multiple subunits. Soluble proteins, with only a single polypeptide chain, fold so that almost all of their hydrophobic residues are hidden in the interior. In the case of subunit proteins, the polypeptide chains are folded, leaving a cluster of hydrophobic residues exposed to the water at the protein surface (Fig. 7.18). This is an unfavorable arrangement and when two polypeptide chains with exposed hydrophobic patches come into contact with each other, they tend to stick together, rather like hook-and-loop fasteners. As noted above, the hydrophobic force

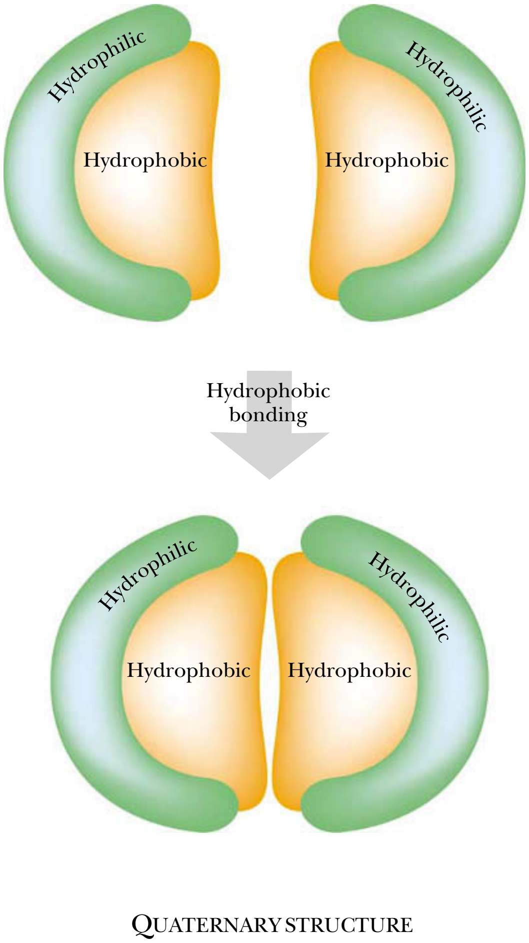


FIGURE 7.18 ***Hydrophobic Force Drives Assembly of Subunits***

Two proteins with hydrophobic regions will often bind together so that their hydrophobic regions become internalized, away from the surrounding water, in the dimer.

is weaker at lower temperatures, hence many subunit proteins tend to come apart at low temperatures.

Higher Level Assemblies and Self-Assembly

Subunits that are designed with more than one bonding region can be used to build rings or chains. Long chains of identical protein subunits are often twisted helically, as in bacterial flagella or in human collagen. If a helix of protein subunits has wide coils that are packed close together, it will form a hollow cylinder (Fig. 7.19). The protein shell of certain viruses (e.g. tobacco mosaic virus) and the microtubules that are found in eukaryotic cells are both constructed in this manner. In some cases, merely mixing the subunits allows the final structure to form. This is known as **self-assembly** and is true of the coat of tobacco mosaic virus, for example. In other cases, these higher level structures require other proteins and cofactors to help assembly.

Cofactors and Metal Ions Are Often Associated with Proteins

To function properly, many proteins need extra components, called cofactors or prosthetic groups, which are not themselves proteins. Many proteins use single metal atoms as cofactors; others need more complex organic molecules. Strictly speaking, prosthetic

self-assembly Automatic assembly of protein subunits without need of any outside assistance

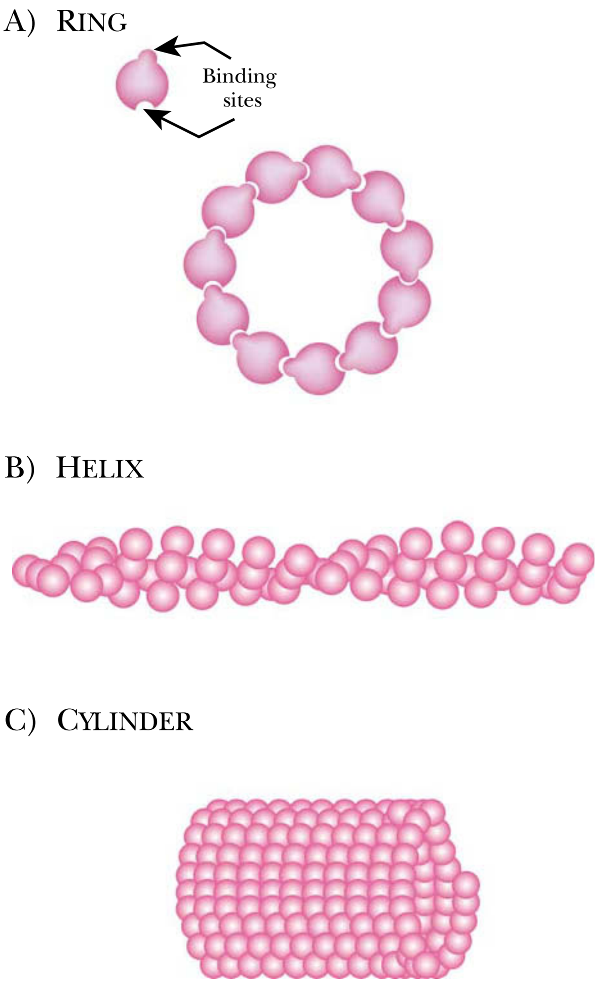


FIGURE 7.19 ***Protein Assemblies: Rings, Chains and Cylinders***

A) Joining protein subunits in a circle forms a ring. B) A helical chain, like that of actin, allows assembly of a very long thin structure from globular subunits. C) Winding protein subunits into a helix forms a cylinder, like that found in microtubules.

groups are fixed to a protein, whereas cofactors are free to wander around from protein to protein. However, this classification breaks down because the same organic cofactor may be covalently attached to one enzyme but non-covalently associated with another. Consequently, the terms are often used loosely. A protein without its prosthetic group is referred to as an **apoprotein**.

For example, oxygen carrier proteins such as hemoglobin have a cross-shaped organic cofactor called heme that contains a central iron atom. The heme is bound in the active site of the apoprotein, in this case globin, and so hemoglobin results. Oxygen binds to the iron atom at the center of the heme and the hemoglobin carries it around the body. Prosthetic groups are often shared by more than one protein; for example, heme is shared by hemoglobin and by myoglobin, which receives oxygen and distributes it inside muscle cells.

Most bacteria and plants are able to synthesize their own cofactors. However, many organic enzyme cofactors cannot be made by animals and consequently they or their immediate precursors must be provided in the diet. Such cofactors and/or their precursors are then referred to as vitamins (Table 7.02). A few cofactors, such as heme, can be synthesized by animals and are therefore not vitamins. Conversely, not all vitamins are cofactors or their precursors. For example, vitamin D gives rise to a hormone. Vitamin A confuses the classification scheme as it is partly converted to retinaldehyde—a protein cofactor—and partly to retinoic acid—a hormone. Vitamin C does not act directly as a cofactor but is needed to keep metal ions (such as Cu and Fe) that do act as cofactors in their reduced states. A further complication is that certain cofactors

apoprotein That portion of a protein consisting only of the polypeptide chains without any extra cofactors or prosthetic groups

X-Ray Crystallography is Used to Solve 3D Structures

X-ray crystallography, also known as X-ray diffraction, is used to solve the 3-D structure of molecules, in particular proteins and nucleic acids. It was X-ray diffraction that first revealed that DNA was twisted into a double helix. Knowing their 3-D shapes allows us to understand better how biological molecules fit together and interact.

When a beam of X-rays is shone through a substance, the X-rays are scattered by the atoms they encounter. If the target substance is a crystal with a regular structure, the scattering of the X-rays will give rise to a regular, though complex, diffraction pattern (Fig. 7.20). In practice, the crystal is rotated into a variety of positions on a computer-controlled stage. The diffraction patterns are recorded and, after computer analysis, are used to generate a 3-D atomic map of the protein molecule.

X-ray crystallography needs large well-formed crystals of highly purified protein. Nowadays, molecular cloning and over-expression of the gene encoding it allow us to obtain plenty of the protein under investigation. However, getting nice crystals is often difficult for such massive complex molecules as proteins, especially those whose 3-D shapes are irregular. X-ray crystallography is sophisticated and time consuming and each protein must be purified and examined individually. Nonetheless, X-ray structures are available for a significant number of proteins. As of early 2004, the Protein Data Bank (www.rcsb.org/pdb/) lists approximately 24,000 structures of which 19,000 were provided by X-ray analysis and 3,000 by NMR. Since many proteins are members of related families, once a 3-D structure is available for one, it provides insight into the conformation of a whole series of related molecules.

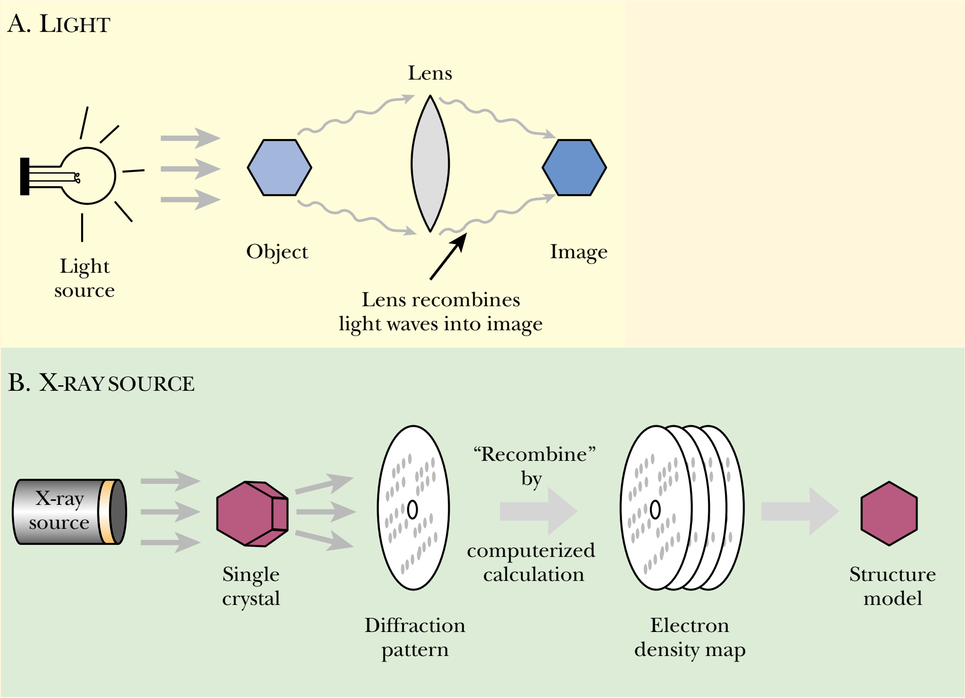


FIGURE 7.20 ***X-ray Crystallography***

A) As light passes through an object, the light waves are distorted from their original pathway. Using a lens allows the distorted light waves to be refocused into an image of the object. B) X-rays are also distorted when they pass through an object. The structure of the proteins within a single crystal dictates the pattern in which the X-rays are diffracted. Moving the crystal around alters the pattern in which the X-rays are diffracted, and all the different patterns can be combined into one to form a model of the protein's actual structure.

TABLE 7.02 Organic Cofactors and Vitamins

Vitamin/Parent compound	Active form/Cofactor	Function
Vitamins that are enzyme cofactors or their precursors		
Vitamin A = Retinol	Retinaldehyde	vision
Vitamin B1 = Thiamine	Thiamine pyrophosphate	decarboxylations
Vitamin B2 = Riboflavin	Flavin adenine dinucleotide Flavin mononucleotide	many redox reactions
Vitamin B3 = Niacin (refers to nicotinamide and/or nicotinic acid)	Nicotinamide adenine dinucleotide	redox reactions (degradative)
	Nicotinamide adenine dinucleotide phosphate	redox reactions (biosynthetic)
Vitamin B5 = Pantothenic acid	Coenzyme A 4′-Phosphopantetheine	acylation reactions fatty acid synthesis
Vitamin B6 = Pyridoxine, pyridoxal, or pyridoxamine	Pyridoxal phosphate	amino acid metabolism
Vitamin B12 = Cobalamin	Methyl Cobalamin Deoxyadenosyl Cobalamin	methyl group carrier rearrangements
Biotin (a B vitamin)	Biotin	carboxylations
Folic acid (a B vitamin)	Tetrahydrofolate	redox reactions and one carbon carrier
Vitamin K Vitamin K1 = Phylloquinone Vitamin K2 = Menaquinone Vitamin K3 = Menadione	Phylloquinone Menaquinone Menaquinone	post-translational carboxylation of glutamate
Vitamins that are not enzyme cofactors or their precursors		
Vitamin A = Retinol	Retinoic acid	hormone/regulator
Vitamin C = Ascorbic acid	Ascorbic acid	antioxidant
Vitamin D = Ergocalciferol (D2) or Cholecalciferol (D3)	Calcitriol	hormone controlling Ca and P metabolism
Vitamin E = Tocopherol	Tocopherol	antioxidant
Cofactors that are not vitamins (i.e. can be made by animals)		
Heme	Heme	oxygen carrier
Lipoic acid	Lipoamide	redox reactions 2-carbon carrier
Biopterin	Tetrahydrobiopterin	Phe metabolism

can be made by some animals but not others. Thus vitamin C is a vitamin for humans but is not a dietary requirement for most other mammals.

Nucleoproteins, Lipoproteins and Glycoproteins Are Conjugated Proteins

Certain proteins are linked to other large molecules, such as lipids, carbohydrates or nucleic acids.

Conjugated proteins are proteins that are linked to molecules of other types. For example, **nucleoproteins** are complexes of protein and nucleic acid, **lipoproteins** are proteins with lipid attached and **glycoproteins** have carbohydrate components.

conjugated protein Complex of protein plus another molecule
glycoprotein Complex of protein plus carbohydrate
lipoprotein Complex of protein plus lipid
nucleoprotein Complex of protein plus nucleic acid

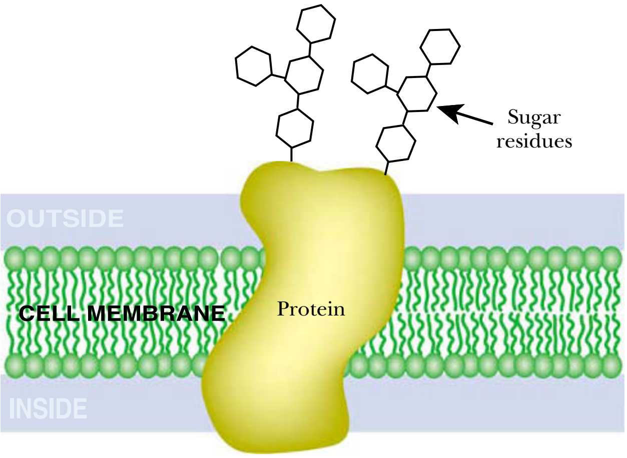


FIGURE 7.21 ***A Membrane Glycoprotein***

A glycoprotein in the cytoplasmic membrane of an animal cell is shown protruding from both sides of the membrane. At the exterior surface, several sugar residues project from the protein into the extra-cellular space.

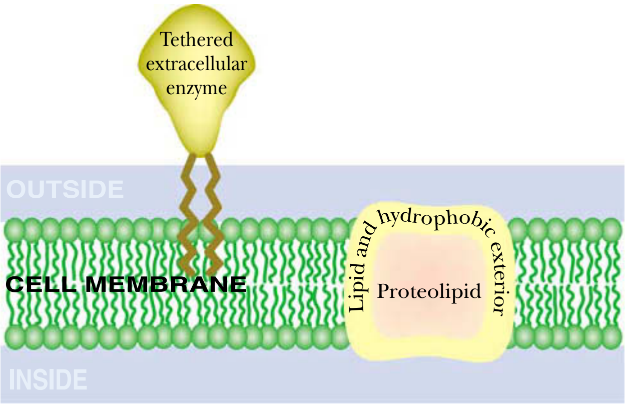


FIGURE 7.22 ***Extracellular Enzyme Tethered by Lipid Tail***

A lipoprotein is held to the membrane surface by one or more lipid tails that penetrate into the lipid bilayer of the membrane. The lipid tails are related to those of the phospholipids that comprise the membrane itself.

Many glycoproteins are found at the surface of cells. They carry short carbohydrate chains consisting of several sugar molecules, which usually project outward from the cell (Fig. 7.21). The sugar chain is usually linked via the hydroxyl group of serine or threonine or the amide group of asparagine. Glycoproteins often function in *cell-to-cell adhesion*, especially in organisms (animals) that lack rigid cell walls. In addition, the carbohydrate portion of glycoproteins is often the key factor in *cellular recognition*. For example, sperm recognize egg cells by binding to the carbohydrate part of a surface glycoprotein. Recognition by the immune system often depends on the precise structure of the carbohydrate chains of glycoproteins. For example, the A- and B-antigens of the well-known ABO blood typing system are actually protein-borne carbohydrate chains that differ in the presence or absence of a single sugar.

Many lipoproteins are attached to membranes by their lipid tails (Fig. 7.22). An example is the **β-lactamase** found in many gram-positive bacteria, such as *Bacillus*. The β-lactamase protects the cell by destroying antibiotics of the β-lactam family, which includes penicillin. Since the target for penicillin action is the cell wall, the protective enzyme needs to be outside the cell. The lipid tail makes sure it does not drift away into the surrounding medium.

Proteolipids are a specialized subclass of lipoproteins that are extremely hydrophobic and are insoluble in water. They are soluble in organic solvents and are found in the hydrophobic interior of membranes. These properties are not solely due to attached lipid groups. In part, their hydrophobicity is due to a high percentage of

β-lactamase Enzyme that destroys antibiotics of the β-lactam class that includes penicillins and cephalosporins
proteolipid A type of lipoprotein that is extremely hydrophobic and found in the interior of membranes

hydrophobic amino acid residues. Instead of being hidden inside the protein, many of these are exposed on the surface.

Proteins Serve Numerous Cellular Functions

Proteins play a wide variety of functional roles in the cell.

Proteins make up about 60% of the organic matter of living organisms. They are responsible for most of the metabolic reactions and many of the structural components of cells. Not surprisingly, there is colossal variety in the functional role of proteins. Nonetheless, proteins may be subdivided into several major categories:

1. Enzymes
2. Structural proteins
3. Binding proteins (transport, carrier, and storage proteins)
4. Mechanical proteins
5. Information processing proteins

Most metabolic reactions are catalyzed by proteins known as enzymes.

Enzymes are proteins that catalyze chemical reactions. These are discussed in more detail below. Many of the characteristics of enzymes, such as the presence of binding pockets for small molecules and the ability to change shape, are shared by other proteins.

Many sub-cellular structures consist largely or partly of **structural proteins**. The filaments of the flagella with which bacteria swim around, the microtubules used to control traffic flow inside cells of higher organisms, the fibers in connective tissue, and the outer coats of viruses (see Ch. 17) are examples of structures built using proteins.

Specialized proteins are known that control the structure of water. Fish that live in polar regions have **antifreeze proteins** to keep their blood from freezing. These proteins bind to ice surfaces and prevent the growth of ice crystals. Conversely, surface proteins of certain bacteria, known as **ice nucleation factors**, promote the formation of ice crystals and are important in causing frost damage to plants. Damaging plant cells releases nutrients on which the bacteria can grow and may also allow colonization of plant tissue by the bacteria.

Some proteins carry nutrients or other molecules across membranes or around the organism.

Binding proteins bind small molecules but unlike enzymes they do not carry out a chemical reaction. Nonetheless, they also need "**active sites**" to accommodate the small molecules. **Transport proteins** or **carrier proteins** are involved in moving their substrates around within the organism. Transport proteins or **permeases** are located in membranes and transport their target molecules across the membrane. Nutrients, such as sugars, must be transported into the cells of all organisms, whereas waste products are deported. Many permeases consist of a bundle of α-helical segments (often 7 or 11) that crosses the membrane and is joined by regions of random coil (Fig. 7.23). Most permeases require energy to operate.

In contrast to permeases, carrier proteins are soluble, not membrane bound. Most transport nutrients within the bodies of multicellular organisms. These carrier proteins are sometimes extracellular and found in the bloodstream or other body fluids. In other

active site Special site or pocket on a protein where other molecules are bound and the chemical reaction occurs
antifreeze protein Protein that prevents freezing of blood, tissue fluids or cells of organisms living at sub-zero temperatures
binding protein Protein whose role is to bind another molecule
carrier protein Protein that carries other molecules around the body or within the cell
enzyme A protein that catalyzes a chemical reaction
ice nucleation factor Protein found on surface of certain bacteria that promotes the formation of ice crystals
permease A protein that transports nutrients or other molecules across a membrane
structural protein A protein that forms part of a cellular structure
transport protein Protein that transports another molecule across membranes or from one cell to another

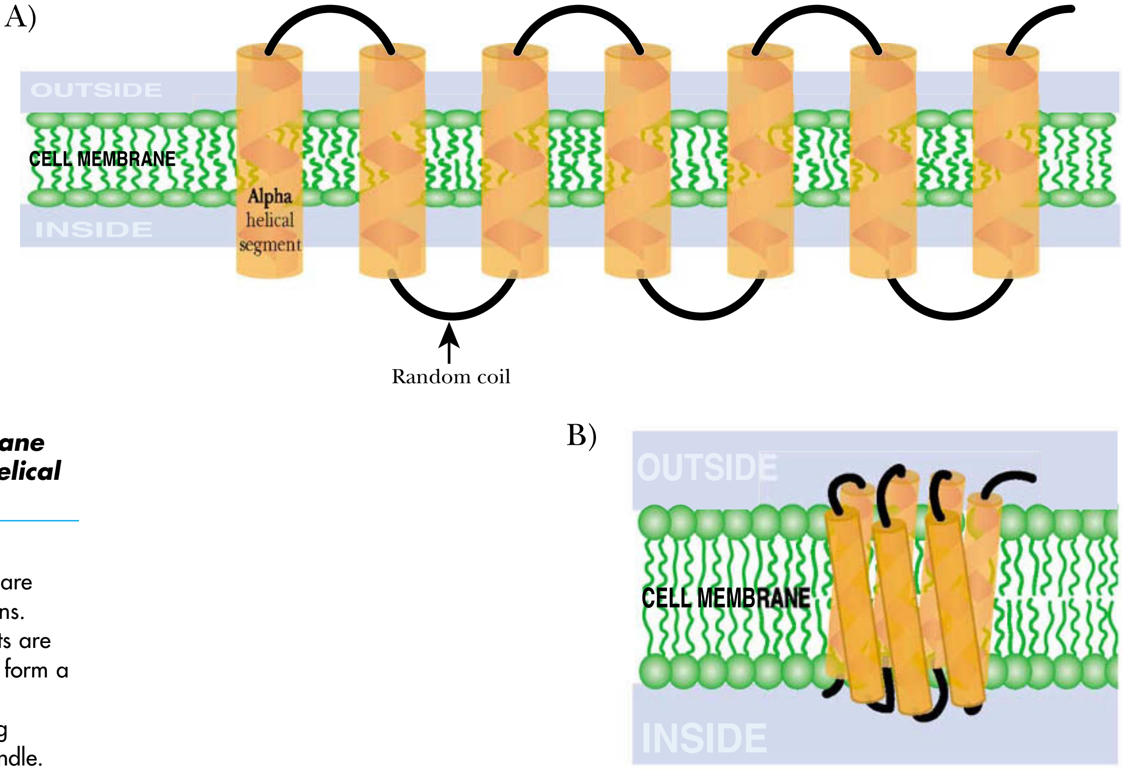


FIGURE 7.23 ***Membrane Permease Made of Helical Segments***

A) Seven α-helical segments traverse the membrane and are joined by random coil regions.
B) The seven helical segments are actually bundled together to form a transport channel, with the transported molecule passing through the center of the bundle.

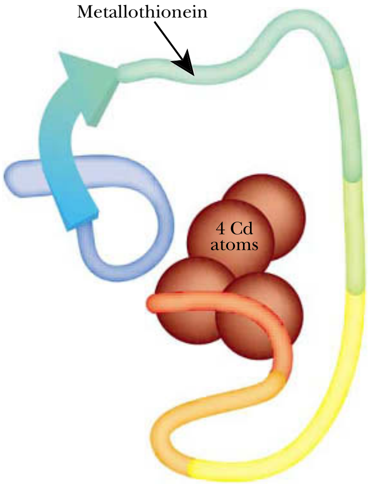


FIGURE 7.24 ***Metallothionein Is a Metal Binding Protein***

This short protein is capable of sequestering four cadmium atoms.

Movement of both muscles and flagella is due to proteins that contract. This consumes energy in the form of ATP.

cases, carrier proteins are found inside specialized cells that themselves travel around the body—such as red blood cells. Classic examples of carrier proteins are hemoglobin and myoglobin, located in the red blood cells and muscle tissue respectively, which are responsible for carrying oxygen in animals.

Storage proteins sequester nutrients or other molecules, but instead of transporting them, they store them. For example, **ferritin** in animal cells, and the corresponding bacterioferritin in bacteria, store iron. **Metallothionein** binds assorted heavy metal ions, and its role is largely protective (Fig. 7.24). The metallothionein gene is induced by traces of heavy metals and has a very strong promoter that is widely used in genetic engineering. Protective metal-binding proteins are also found in certain bacteria. They are sometimes exploited in biotechnological processes for extraction of metals such as gold or uranium from ore or industrial waste streams.

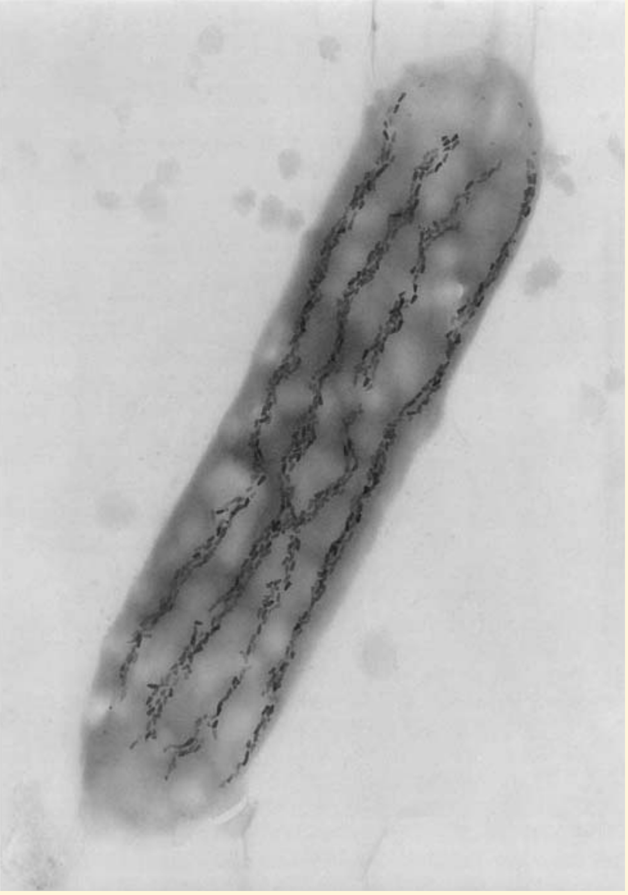
The immune systems of higher animals have many highly specialized binding proteins that function in protection against infection by invading bacteria and viruses. These include antibodies and T-cell receptors.

Mechanical proteins are sometimes classified as specialized structural proteins or as enzymes. They perform physical work at the expense of biological energy (usually by hydrolyzing ATP or GTP). Their energy consumption results in a reversible change of conformation. Proteins of the myosin family that contract upon energization are found in muscle fibers and the filaments of eukaryotic flagella. Actins are found in muscle with myosin where the two proteins participate in a contractile function such as shortening of a muscle (Fig. 7.26), but they are also involved in cellular movements such as endocytosis and amoeboid motion.

ferritin An iron storage protein
mechanical protein Protein that uses chemical energy to perform physical work
metallothionein Protein that protects animal cells by binding toxic metals

FIGURE 7.25
Magnetotactic Bacteria Contain Magnetosomes to Bind Iron

Transmission electron micrograph of *Magnetobacterium bavaricum*, a rod-shaped magnetotactic bacterium from Lake Chiemsee (Upper Bavaria) with four bundles of chains of magnetosomes. Individual cells containing up to 1000 hook-shaped magnetosomes yield magnetic moments as high as (10–60)x10^{12} Gauss ccm, which is one to two orders of magnitude more than the values characteristic of other magnetotactic bacteria. The large electron-opaque bodies inside the cell consist of sulfur. The magnetosomes are made of magnetite and measure, on average, 100 nm. Courtesy of Prof. Michael Winklhofer, Institut für Geophysik, Theresienstrasse 41, D-80333 München, Germany.



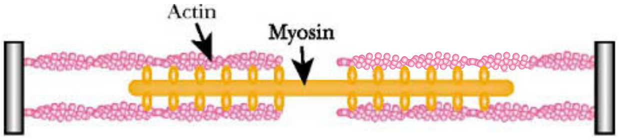


FIGURE 7.26 ***Mechanical Proteins***

Myosin and actin interact to cause contraction. The participating proteins move relative to one another but do not shorten individually. The movements bring the end attachments of the actin filaments closer together. Numerous units such as those shown here are attached end-to-end, causing skeletal muscle to shorten.

Bacterial flagella do not operate by contraction of filamentous proteins. Instead, the base of the flagellum consists of a protein ring that rotates as it consumes energy. The filament is attached to the rotating ring and makes a helical lashing motion that drives the bacterium along.

Chaperone proteins, or **chaperonins**, assist other proteins in folding correctly. Some chaperonins are involved in protein export and prevent premature folding of proteins that are to be secreted through membranes (see Ch. 8). Other chaperonins attempt to re-fold proteins that have become denatured due to high temperature or other environmental stresses that damage proteins (see the heat shock response, Ch. 9).

chaperonin Protein that helps other proteins to fold correctly

Information processing proteins is a rather artificial assemblage of proteins that can all be alternatively classified as enzymes or binding proteins. However, they are sometimes considered together in order to emphasize their role in transmitting biological information. They include cell surface receptors, signal transmission proteins and regulatory proteins that control the expression of genes at various levels. The functions of these proteins are discussed in detail in the appropriate chapters (especially Ch. 9 through Ch. 11). In this chapter the 3-D structural motifs of proteins involved in the binding of proteins to DNA are discussed further below.

Certain proteins emit or absorb light and are sometimes included among the information processing proteins. **Luciferases** send a signal from one organism to another by emitting light. Luminous bacteria glow to attract deep-sea fish to swallow them so they can take up residence in the intestines of these fish. Fireflies flash as part of their mating strategy. Both bacterial and insect luciferases have been used as reporter enzymes in genetic analysis (see Ch. 25). **Green fluorescent protein** (**GFP**) is made by fluorescent jellyfish. GFP is also used as a **reporter** in genetic analysis. By reporter, it is meant that the green fluorescence can be used to monitor gene expression and/or to localize areas where a particular gene is expressed (see Ch. 25). Rhodopsins are proteins that absorb light. They are used in the eyes of both invertebrate and vertebrate animals to detect light.

Protein Machines

Many cellular processes such as DNA synthesis, RNA splicing or protein degradation are performed by groups of a dozen or more proteins that operate in a coordinated manner. When such groups of proteins stay associated together, consume energy and carry out carefully controlled movements as well as catalyzing chemical reactions, they are sometimes referred to as protein machines.

Subcellular machines are assemblies of protein and, often, RNA, that carry out complex tasks in the cell.

There is a trend to name these assemblies to rhyme with ribosome. Thus the replisome moves along the chromosome while synthesizing new DNA (see Ch. 5), the spliceosome is responsible for RNA splicing (see Ch. 12) and the proteasome carries out protein degradation (see below). However, as Chapter 8 will show, ribosomal RNA plays a more critical part in protein synthesis than do the ribosomal proteins. RNA is also involved in spliceosome function. Thus the term “protein machine” is misleading; a better name might be subcellular machine.

Referring to biological components as machines may perhaps be due to influence from the field of nanotechnology. This refers to the ability to manipulate matter by precisely placing atoms and molecules. Nanotechnology relies on molecular-sized nanomachines. These machines would be programmed to reproduce themselves in the millions and then place atoms precisely to build other molecules. These molecules could then be assembled together into whatever components are needed. Perhaps to be truly trendy and take into account the role of RNA as well as proteins, the term bionanomachine should be introduced.

Enzymes Catalyze Metabolic Reactions

Beta-galactosidase splits a range of molecules that consist of galactose linked to another component.

Enzymes are proteins that catalyze chemical reactions but are not consumed in the process. Virtually all metabolic reactions depend on enzymes. An enzyme first binds the reacting molecule, known as its **substrate**, and then performs a chemical operation upon it. Some enzymes bind only a single substrate molecule; others may bind two or more, and react them together to give the final product. Many enzyme-catalyzed reactions are reversible; that is, the enzyme speeds up reaction in either direction.

green fluorescent protein (GFP) A jellyfish protein that emits green fluorescence and is widely used in genetic analysis
luciferase Enzyme that consumes energy and generates light
reporter gene Gene that is used in genetic analysis because its product is convenient to assay or easy to detect
reporter protein A protein that is easy to detect and gives a signal that can be used to reveal its location and/or indicate levels of gene expression
substrate Molecule that binds to an enzyme and is the target of enzyme action

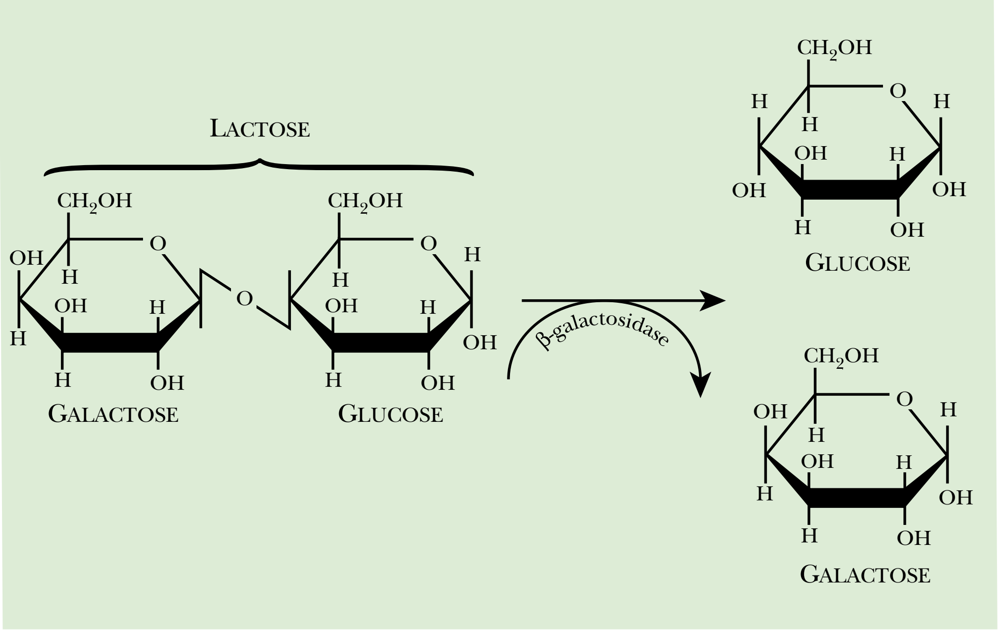


FIGURE 7.27 ***β-Galactosidase Splits Lactose***

Lactose is split by the enzymatic action of β-galactosidase into glucose and galactose.

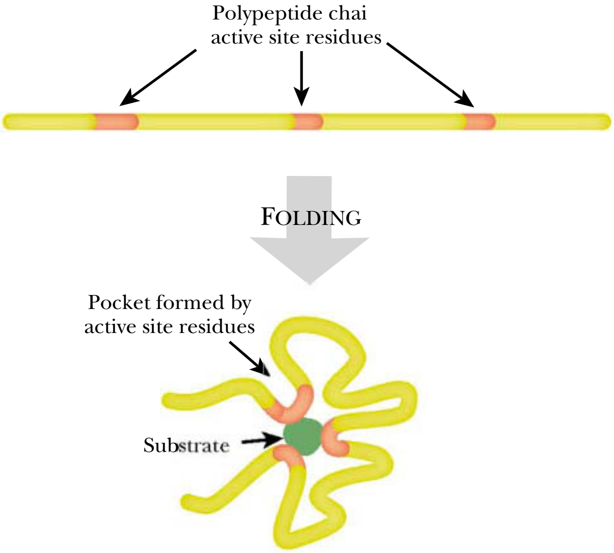


FIGURE 7.28 ***Active Site Consists of Residues Far Apart in the Sequence***

Before a polypeptide chain is folded, the amino acid residues that cooperate to carry out the enzymatic reaction are often far apart in the sequence. Upon correct folding, the critical amino acids are brought together and form the active site.

Both the substrate and cofactors (if needed) bind to the active site of the enzyme.

The most famous enzyme in molecular biology is **β-galactosidase**, encoded by the *lacZ* gene of the bacterium *Escherichia coli*. This enzyme is so easy to assay that it is widely used in genetic analysis (see Ch. 25 for details). One of the natural substrates of β-galactosidase is the sugar lactose, made by linking together the two simple sugars, glucose and galactose. β-Galactosidase hydrolyses lactose into the two simpler sugars (Fig. 7.27).

Substrates bind to the enzyme at the active site, a pocket or cleft in the protein, where the reaction occurs. The active site is the result of precise folding of the polypeptide chain so that amino acid residues that may have been far apart in the linear sequence can come together to cooperate in the enzyme reaction (Fig. 7.28 and Fig. 7.29).

Many enzymes rely on cofactors or prosthetic groups to assist in catalysis. These may be organic molecules, such as NAD (nicotinamide adenine dinucleotide), which

β-galactosidase Enzyme that splits lactose and related compounds

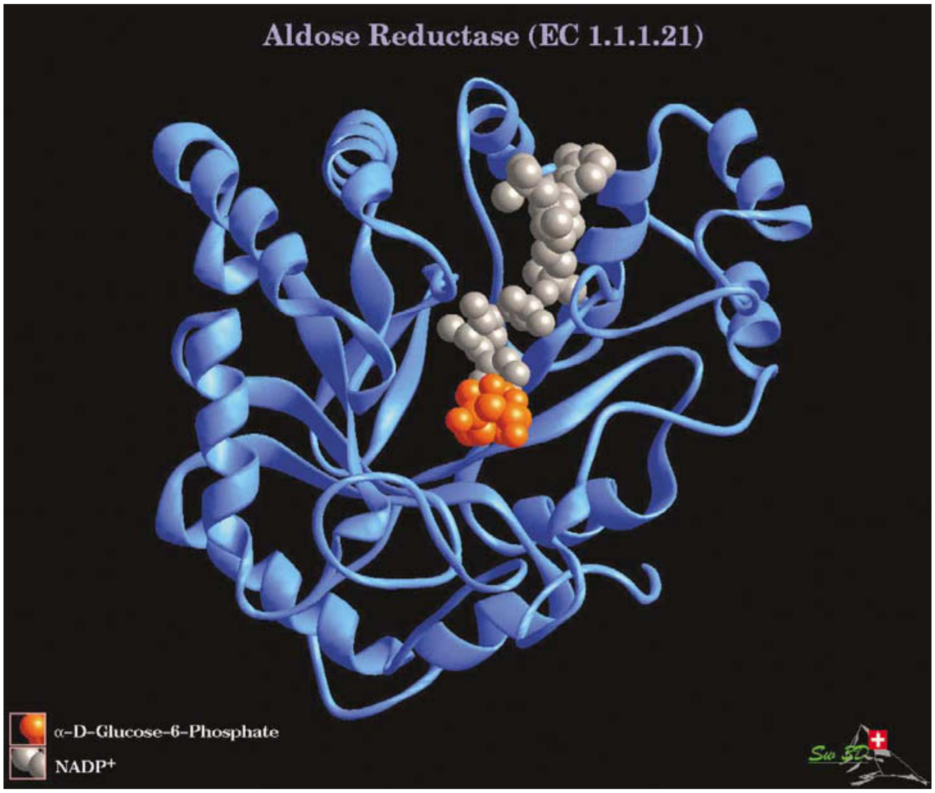


FIGURE 7.29 ***Enzyme Substrate Complex of Aldose Reductase***

A three-dimensional computer model of aldose reductase (EC 1.1.1.21) shown binding its substrates, glucose-6-phosphate (orange) and NADP (gray). The image was generated by Dr. Manuel C. Peitsch at the Glaxo Institute for Molecular Biology in Geneva, Switzerland, using coordinates from the Brookhaven Protein Data Bank.

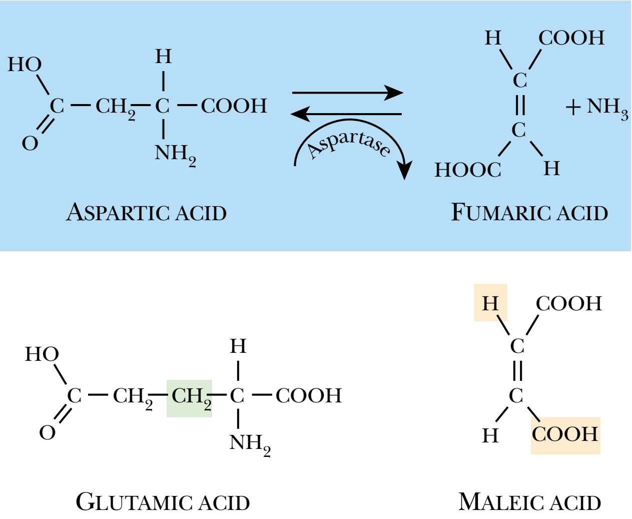


FIGURE 7.30 ***Some Enzymes, Such as Aspartase, Can Distinguish Isomers***

L-Aspartate is transformed to fumarate plus NH_3 by the enzyme aspartase. This enzyme will not transform "look-alike" substrates such as glutamic acid nor structural isomers like maleic acid. Furthermore, aspartase can distinguish optical isomers and only uses the L-form of aspartate.

carries the hydrogen atoms added to (or removed from) the substrate by many enzymes. Metal ions are also common. Zinc ions are found at the active site of most enzymes that synthesize or degrade nucleic acids. The Zn^{2+} binds to the negatively charged phosphate groups and weakens the critical bonds.

Enzymes Have Varying Specificities

Some enzymes are extremely specific and will use only a single substrate. Aspartase catalyzes the inter-conversion of aspartic acid and fumaric acid (Fig 7.30). Aspartase will use only aspartate (not similar amino acids such as glutamate) and it uses only the

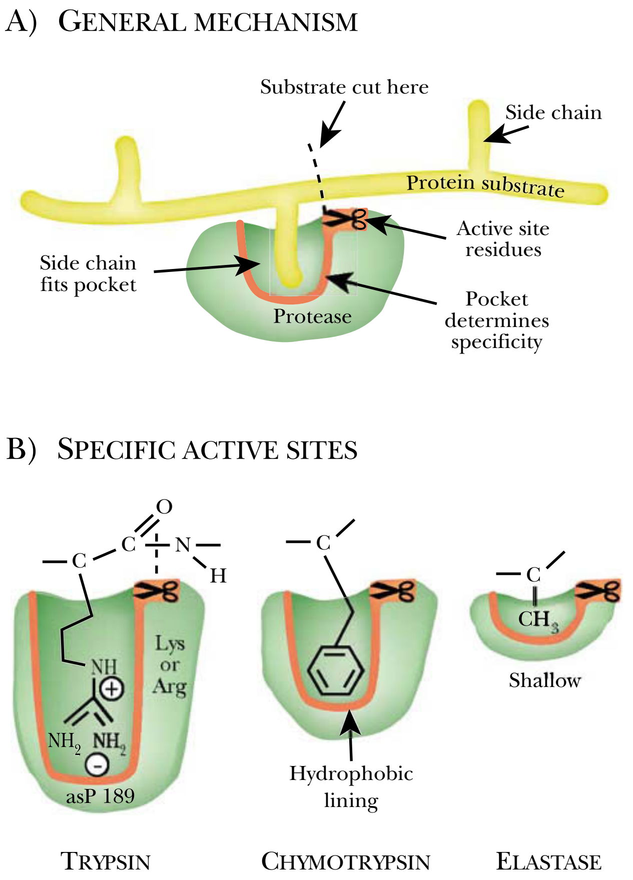


FIGURE 7.31 ***Active Site Specificity of Proteases***

A) In general, a side chain of an amino acid fits into the pocket of a protease and a clipping mechanism nearby breaks the amino acid chain. B) The pockets have different properties in different protease enzymes. A negative charged pocket as in trypsin binds positively charged residues such as arginine. The hydrophobic pocket of chymotrypsin attracts hydrophobic residues and the shallow pocket of elastase only fits small residues such as glycine or alanine.

Biosynthetic enzymes are often highly specific, in contrast to degradative enzymes, which often show wide specificity.

L-isomer of aspartate. Aspartase can also catalyze the reverse reaction, but it will add only NH_3 to fumarate and will not use maleate, the *cis*-isomer of fumarate.

Other enzymes have much broader substrate specificity. Broad specificity is shown by many degradative enzymes; for example, alkaline phosphatase, which removes phosphates from a wide range of molecules, or carboxypeptidase, which snips the C-terminal amino acid off many polypeptide chains. Most enzymes have intermediate specificity. For example, alcohol dehydrogenase from *E. coli* will act on C3 and C4 alcohols as well as its natural substrate, ethanol. Again, β-galactosidase uses several compounds in which galactose is linked to another chemical group, such as lactose, ONPG and X-gal (see below, Fig. 7.36).

Differences in specificity between similar enzymes are largely determined by the nature of the active site. The active site pocket can vary in size and shape and in the chemical nature of the amino acids comprising it. For example, the three digestive enzymes trypsin, chymotrypsin and elastase all split polypeptide chains by the same catalytic mechanism (Fig 7.31). However, the active site of trypsin has a negative charge at the bottom, so trypsin cuts after positively charged residues (e.g., Lys or Arg). In chymotrypsin, the active site pocket is lined by hydrophobic groups and so this enzyme cuts after hydrophobic residues (e.g., Phe or Val). In elastase, the active site pocket is very small and so elastase cuts after residues with small side chains (e.g., Ala).

L-isomer That one of a pair of optical isomers that rotates light in an anticlockwise direction

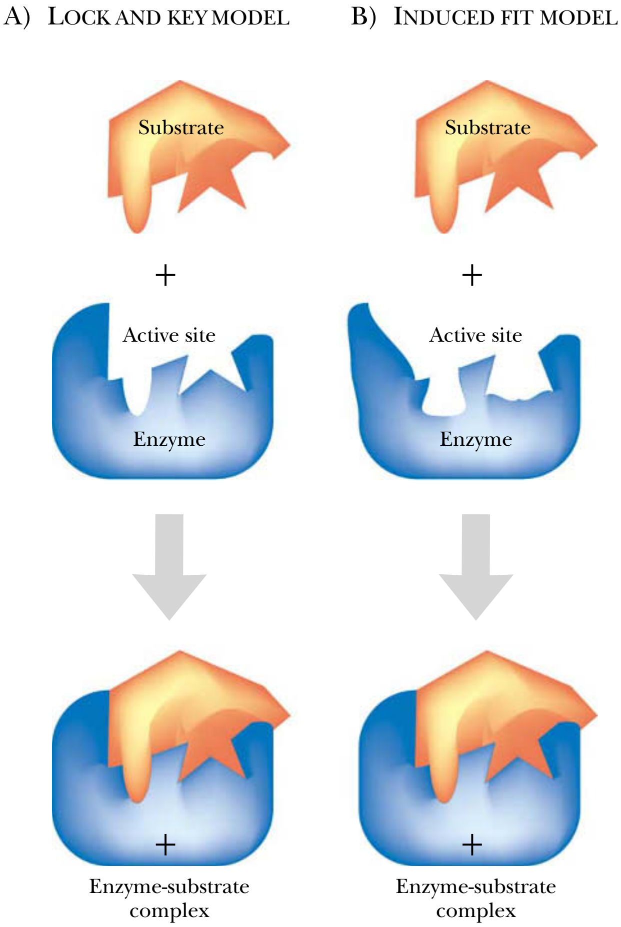


FIGURE 7.32 *Lock and Key Versus Induced Fit*

A) The lock and key model says that the active site and the substrate must fit perfectly. B) The induced fit model proposes that the conformation of the active site will change upon binding to the substrate.

Lock and Key and Induced Fit Models Describe Substrate Binding

Two different models have been proposed to explain the binding of substrate in the active site of enzymes. In the **lock and key model**, the active site of the enzyme fits the substrate precisely. In contrast, the **induced fit** model proposes that the binding of the substrate induces a change in enzyme conformation so that the two fit together better (Fig. 7.32).

Enzymes may be found that operate by both mechanisms. For example, the enzyme chymotrypsin degrades proteins in the digestive tract. Almost no detectable structural change occurs when chymotrypsin binds its substrate. Carboxypeptidase also degrades proteins, but by snipping amino acids off from the carboxyl end. When carboxypeptidase binds substrate, this causes a tyrosine (at position 248) to move 12 Angstroms to a position where it is in physical contact with the substrate and can now play a direct catalytic role.

Enzymes Are Named and Classified According to the Substrate

Nowadays enzyme names end in "ase." Some enzymes, such as trypsin or lysozyme, were named before this convention was introduced and so have irregular names. Some major categories of enzymes of interest in molecular biology are listed in Table 7.03.

induced fit When the binding of the substrate induces a change in enzyme conformation so that the two fit together better
lock and key model Model of enzyme action in which the active site of an enzyme fits the substrate precisely

TABLE 7.03 Types of Enzymes and Their Roles

Enzyme Type	Type of Reaction Catalyzed
Hydrolase	Splits substrate by hydrolysis. Includes many sub-categories, including nucleases, proteases, glycosidases and phosphatases.
Nuclease	Cleaves nucleic acids by hydrolyzing phosphodiester bonds between nucleotides. Includes restriction enzymes that cut DNA at specific sequences.
Protease	Cleaves polypeptide chains by hydrolyzing peptide bonds between amino acids.
Glycosidase	Cleaves polysaccharides by hydrolyzing glycoside bonds between sugars. Includes β-galactosidase, lysozyme and cellulase.
Phosphatase	Removes phosphate groups by hydrolysis.
β-Lactamase	Hydrolase that inactivates antibiotics of the β-lactam group (penicillins and cephalosporins) by opening the lactam ring.
ATPase	General name for enzymes that hydrolyse ATP so releasing energy that is used to drive a reaction
Synthase	General name for a biosynthetic enzyme. Includes ligases, polymerases, transferases and other classes.
Ligase	Forms new bonds by joining fragments together. The term "ligase" alone is usually understood to refer to DNA ligase.
Polymerase	Type of ligase that synthesizes polymers by linking the subunits. Includes DNA polymerase and RNA polymerase.
Transferase	Transfers chemical groups from one molecule to another. Includes methylases, acetylases, kinases and others.
Methylase	Transferase that adds methyl groups to a molecule. Includes modification enzymes that methylate DNA.
Acetylase	Transferase that adds acetyl groups to a molecule. Includes histone acetyl transferase (HAT), which acetylates histones.
Kinase	Transferase that adds phosphate groups to a molecule. Includes protein kinases, which attach phosphate groups to proteins.
Isomerase	Interconverts the isomers of a molecule.
Racemase	Specialized isomerase that interconverts the D- and L- isomers of an optically active molecule such as an amino acid.
Oxidoreductase	Catalyzes oxidation and reduction reactions. Includes dehydrogenases, which remove hydrogen atoms from molecules.

Enzymes Act by Lowering the Energy of Activation

Enzymes provide alternative reaction routes of lower energy for organic reactions.

In any chemical reaction, the reactants are converted to the products via the reaction intermediate or **transition state**. In an enzyme catalyzed reaction, the substrate(s) is bound by the enzyme and the reaction occurs within the active site. In either case, energy may be needed to prime the reaction. This **transition state energy**, $\Delta G^{\ddagger}$, is the difference in free energy between the starting materials and the transition state (Fig. 7.33). When $\Delta G^{\ddagger}$ is positive, as shown, energy must be supplied to reach the transition state. Even if this energy is released once the overall reaction is over, the higher the transition state energy, the slower the reaction. Enzymes cannot change the overall free energy of a reaction, i.e., the energy difference between reactants and products. The role of an enzyme is to *lower the transition state energy*, either by stabilizing the

transition state Another term for the activated intermediate in a chemical reaction
transition state energy Energy difference between the reactants and the activated reaction intermediate or transition state

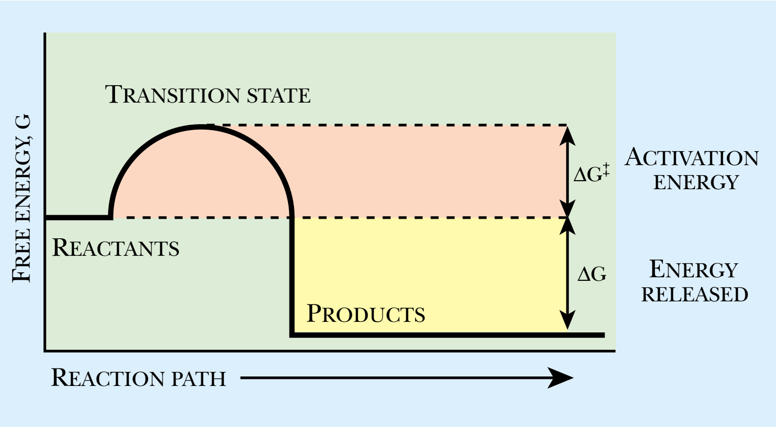


FIGURE 7.33 ***Energy is Required to Reach the Transition State***

Initiation of a chemical reaction requires an input of energy to reach the transition state. The required energy is called the energy of activation, or transition state energy, $\Delta G^{\ddagger}$, and is needed even if the reaction is exothermic (as shown) and will eventually release more energy overall than originally put in. The net energy released is ΔG.

intermediates in the active site, or by providing an alternative reaction mechanism that proceeds by a pathway of lower energy.

The rate of a reaction may be slow without an enzyme. Rate increases by enzymes range from 10^8 to 10^{20} relative to the uncatalyzed, spontaneous reaction (e.g., 10^9 for alcohol dehydrogenase; 10^{16} for alkaline phosphatase). A high enzymatic rate occurs when enzymes can position the substrate correctly relative to the catalytic groups in the active site. Several factors are involved in enzyme rate increases:

1. *Proximity*—(up to 10^6 fold increase). The enzyme binds the substrate so that the susceptible bond is very close to the catalytic group in the active site. The local concentration of substrate in the active site may be as much as 50 Molar whereas its concentration in the cytoplasm may be less than 1 mM. Chemical reaction rates are proportional to the concentrations of the reactants.
2. *Orientation*—(up to 10^2 fold). When the substrate is bound, the reacting groups must be properly oriented. The orbital steering hypothesis suggests that the binding of the substrate(s) to the enzyme aligns the reactive groups so that the relevant molecular orbitals involved in bond formation overlap. This increases the probability of forming the transition state.
3. *Covalent intermediates*—(around 10^{10} fold). The strategy is to lower the transition state energy "hump" by taking an alternative reaction pathway. Consider the transfer of a chemical group "X" from molecule A to molecule B. Here "X" is a molecular fragment such as a phosphate group, acyl group or glycosyl group.

$$AX + B \rightarrow A + BX$$

The enzyme may react in two steps, picking up group "X" from molecule A and then transferring it to molecule B in a second reaction:

$$\begin{gathered} AX + Enz \rightarrow Enz\text{-}X + A \\ Enz\text{-}X + B \rightarrow Enz + BX \end{gathered}$$

4. *Acid-base catalysis*—(around 10^{10} fold). Acid-base catalysis is due to proton donors (acids) or proton acceptors (bases), that donate protons to or remove protons from the reaction intermediate. Enzymes use the side chains of acidic or basic amino acids to attack the substrate. An acidic group in one part of the active site may donate a proton and a basic group in another part of the active site may remove another proton from the reaction intermediate. This referred to as concerted acid-base catalysis. Most hydrolytic enzymes use acid/base catalysis.
5. *Metal ion catalysis*—(around 10^{10} fold). Many enzymes have metal ions at the active site. Sometimes these are used as redox centers, as in cytochromes. Often,

however, the metal ion is not reduced or oxidized, but acts to stabilize negative charges on the reaction intermediate. Thus Zn^{2+} in the active site of carboxypeptidase polarizes the C=O of the peptide bond which is about to be broken.

6. *Distortion of the substrate*—(up to 10^8 fold). Binding to the enzyme may distort the substrate. The active site of the enzyme may fit the transition state intermediate better than the substrate. Once the substrate is bound it will be forced into the shape of the reaction intermediate. Distortion is difficult to demonstrate since the strain is imposed only after the substrate has bound to the enzyme.

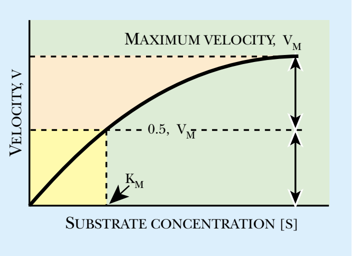


FIGURE 7.34 ***Saturation Kinetics***

As the substrate concentration increases, the velocity of the reaction reaches a maximum (V_m). The K_m is the substrate concentration that yields half the maximum velocity.

The Rate of Enzyme Reactions

The maximum velocity and the Michaelis constant summarize an enzymes kinetic properties.

The principles of chemical kinetics apply to enzyme reactions, with certain modifications. Typically, reaction rates are proportional to the concentrations of the reactants. For enzymes, the rate is proportional to substrate concentration, [S], only at low substrate concentrations. At higher substrate concentrations all the active sites of the enzyme will be filled by substrate and the enzyme can work no faster. The enzyme is said to be **saturated** and has reached its **maximum velocity**, or $\mathbf{V_{max}}$ or $\mathbf{V_m}$. The resulting relation between [S] and rate is a hyperbolic curve (Fig. 7.34).

The V_{max} depends on the nature of the chemical reaction and hence depends on the chemical properties of the substrate and the enzyme active site. The V_{max} may vary with pH and temperature. The more enzyme, the more substrate it can handle, so V_{max} is also proportional to the amount of enzyme. The $\mathbf{K_m}$, or **Michaelis constant**, is the substrate concentration that gives half maximal velocity. It largely depends on the affinity of the substrate for the active site and is independent of the enzyme concentration. The lower the K_m, the higher the affinity of the substrate for the enzyme and the faster the reaction (— at low substrate concentrations; at high substrate concentrations the K_m becomes irrelevant). These factors are summarized in the **Michaelis-Menten equation**:

$$\text{Rate (V)} = V_{max} \times (S)/(K_m + [S])$$

Substrate Analogs and Enzyme Inhibitors Act at the Active Site

Substrate analogs bind to the enzyme active site instead of the substrate. Some inhibit the enzyme, others react in a way similar to the natural substrate.

Analogs are molecules resembling natural substances sufficiently well to bind to the active site of the enzyme. Some analogs act as alternative substrates for the enzyme. Other analogs bind to the active site, but instead of reacting, they block the active site and inhibit the enzyme. Such analogs are known as **competitive inhibitors**, as they compete with the true substrate for binding to the enzyme. The extent of inhibition depends both on the relative concentrations and the relative affinities of the substrate and the inhibitor.

The active site of many, perhaps most, enzymes is designed to better fit the reaction intermediate or transition state, rather than the substrate itself. This tends to

analog A chemical substance that mimics another well enough to be mistaken for it by biological macromolecules, in particular enzymes, receptor proteins, or regulatory proteins
competitive inhibitor Chemical substance that inhibits an enzyme by mimicking the true substrate well enough to be mistaken for it
$\mathbf{K_m}$ See Michaelis constant
maximum velocity ($\mathbf{V_m}$ or $\mathbf{V_{max}}$) Velocity reached when all the active sites of an enzyme are filled with substrate
Michaelis constant ($\mathbf{K_m}$) The substrate concentration that gives half maximal velocity in an enzyme reaction. It is an inverse measure of the affinity of the substrate for the active site
Michaelis-Menten equation Equation describing relationship between substrate concentration and the rate of an enzyme reaction
saturated (Referring to enzymes) When all the active sites are filled with substrate and the enzyme cannot work any faster

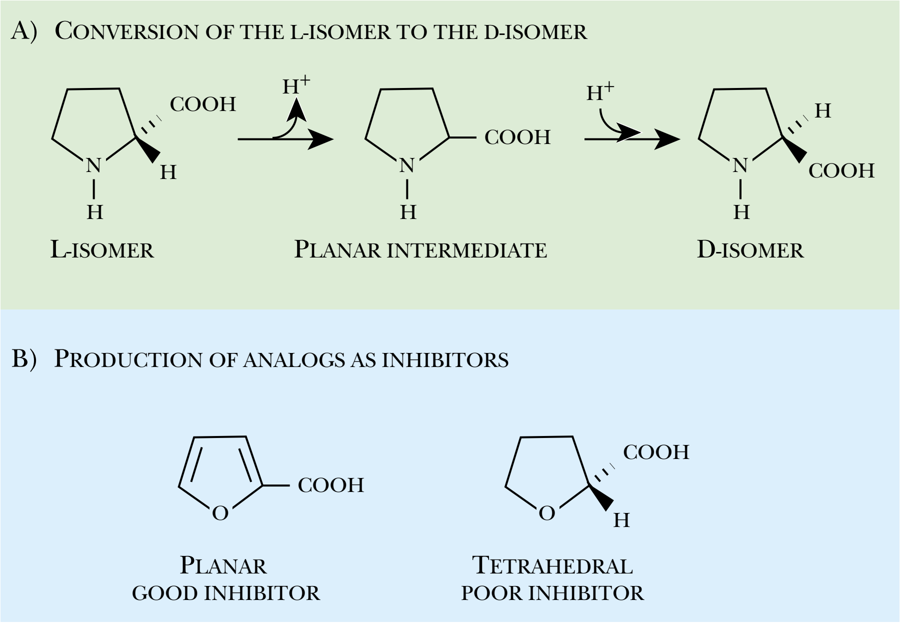


FIGURE 7.35 ***Proline Racemase Produces a Planar Intermediate***

A) In the conversion of the L-isomer to the D-isomer of proline, the intermediate product is flat. This reaction is reversible. B) The best analogs for inhibiting the reaction are planar since these most closely resemble the flat transition state intermediate.

distort the substrate in the required direction and help the reaction along. Consequently, some of the best competitive inhibitors are molecules that mimic the reaction intermediate. These are sometimes known as "**transition state analogs**." Proline racemase inter-converts the L- and D- isomers of proline by removing an H-atom and replacing it in a different configuration (Fig. 7.35). The substrate and product are both tetrahedral about the α-carbon but the transition state is planar. The best competitive inhibitors are flat ring compounds rather than ones that look most like the substrate (Fig. 7.35).

Beta-galactosidase splits ONPG giving a yellow color and X-gal to release a blue dye.

The enzyme β-galactosidase splits many molecules in which galactose is linked to another molecule (refer to Fig. 7.27, above). To take advantage of this, researchers can use a substance called **ONPG (*ortho*-nitrophenyl galactoside)**, which consists of *ortho*-nitrophenol linked to galactose. When ONPG is split, the result is galactose, which is colorless, and *ortho*-nitrophenol, which is bright yellow (Fig. 7.36). Using ONPG allows researchers to monitor the level of β-galactosidase by measuring the appearance of the yellow color. Similarly, **X-gal** is split by β-galactosidase into a blue dye plus galactose (see Ch. 25 for applications). Compounds that are themselves colorless (or only pale) but react to release strongly colored products are known as **chromogenic substrates**.

Some inhibitors react with the enzyme and inactivate it permanently.

Irreversible inhibition occurs when an inhibitor covalently modifies an enzyme, usually at the active site. Covalent inhibitors are not always analogs of the substrate. Some irreversible inhibitors react with components of the active site, other inhibitors react with other regions of the protein and may affect protein conformation, solubility or other properties. For example, the nerve agent, sarin, used in chemical warfare covalently inhibits enzymes that have serine in the active site (Fig. 7.37). The nerve agent reacts with the serine and blocks the active site permanently, inactivating the enzyme. Many hydrolytic enzymes, including proteases, rely on active serine residues. So does acetylcholine esterase, the enzyme that splits the neurotransmitter acetylcholine. Inhibition of this causes paralysis of neuromuscular junctions and ultimately death from respiratory failure.

chromogenic substrate Colorless or pale substrate that is converted to a strongly colored product by an enzyme
irreversible inhibition Type of inhibition in which an enzyme is permanently inactivated by a chemical change
***ortho*-nitrophenyl galactoside (ONPG)** Artificial substrate for β-galactosidase that yields a yellow color upon cleavage
transition state analog Enzyme inhibitor that mimics the reaction intermediate or transition state, rather than the substrate
X-gal Artificial substrate for β-galactosidase that yields a blue color upon cleavage

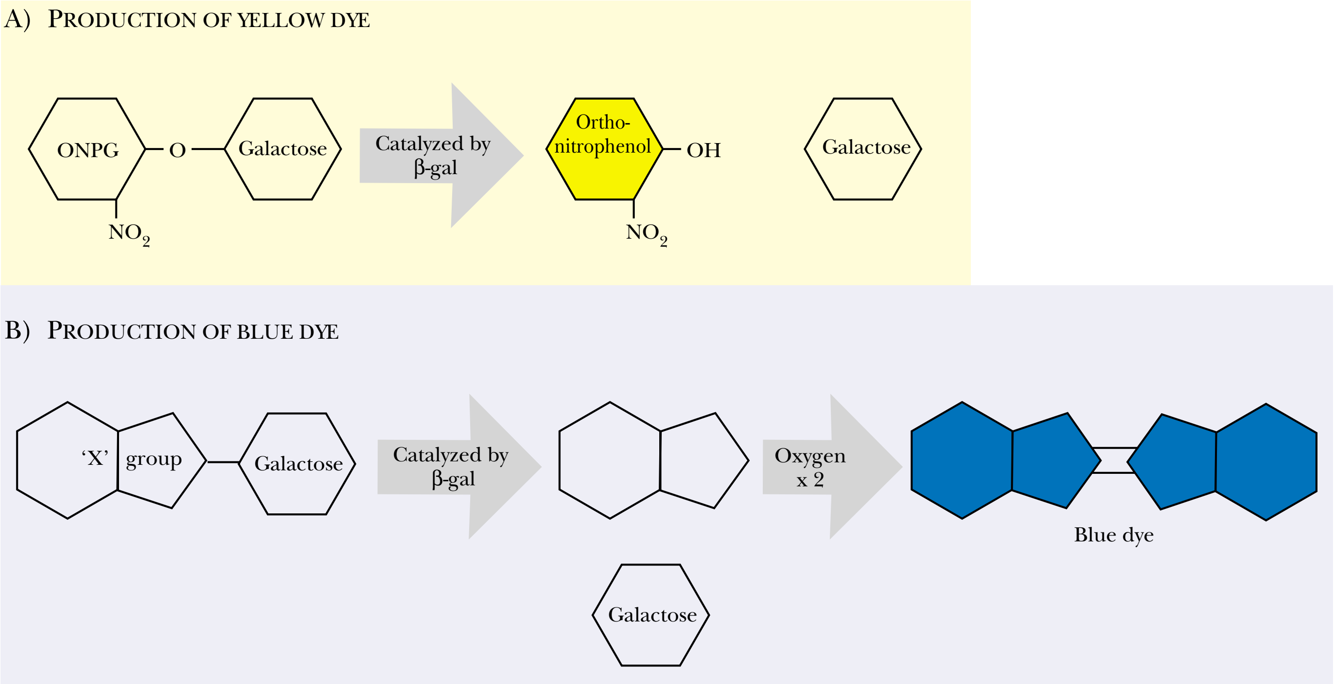


FIGURE 7.36 β-*Galactosidase Splits ONPG and X-Gal*

β-Galactosidase is an enzyme that splits off galactose from other molecules to which it is attached. A) The substrate, ONPG, is used in molecular biology to measure the level of β-galactosidase activity since the reaction product, ortho-nitrophenol, is yellow and can be easily measured. B) X-Gal is another chromogenic substrate for β-galactosidase that is split releasing galactose and a fragment that reacts with oxygen in air yielding a blue dye.

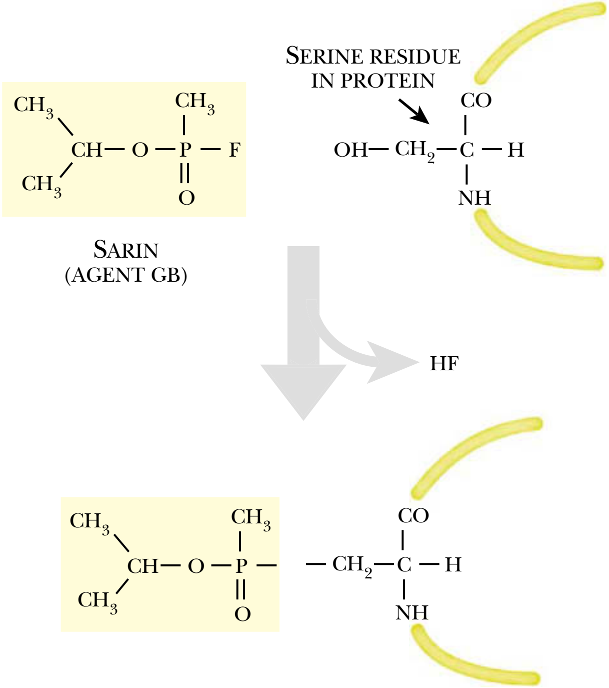


FIGURE 7.37 *Covalent Inhibition of Serine Enzyme by Sarin*

The "nerve gas" sarin interacts with serine residues in the active site of proteins. It blocks the active site and prevents entry of the true substrate into the active site.

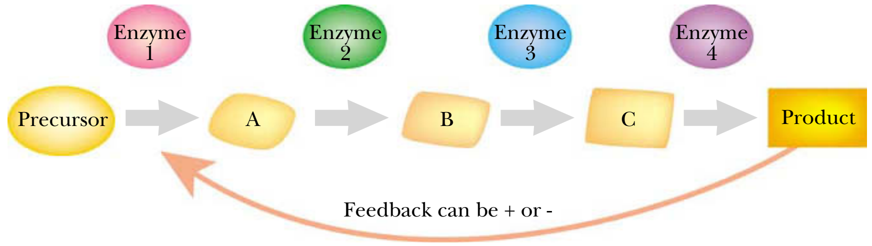


FIGURE 7.38 ***Feedback Inhibition of Metabolic Pathways***

Feedback occurs when the product of an enzyme-catalyzed pathway interacts with one of the enzymes in the pathway, usually the first, to inhibit (negative feedback) or promote (positive feedback) the activity of that enzyme. The activities of the other enzymes in the pathway (#2, #3 and #4) are usually unaffected.

Enzymes May Be Directly Regulated

The activity of an enzyme or other protein may be controlled at the genetic level by deciding whether or not to synthesize the protein. Such genetic regulation is discussed in later chapters. More rapid cellular responses are possible if the activity of an existing protein is controlled. Although most of the following examples apply to enzymes, it should be realized that all types of proteins, including regulatory proteins, transport proteins and mechanical proteins, can be similarly modified and their activity regulated.

The rate at which the majority of enzymes work depends simply on the level of substrate that is available and how much enzyme protein is present. The latter, of course, depends on the level of expression of the gene encoding the enzyme. However, a significant proportion of enzymes are also *directly* regulated in a variety of ways. This has the advantage that it is more or less instantaneous. Such regulation is usually reversible, so that an enzyme that is temporarily inactive is not destroyed, but can be used again later if needed.

Enzymes located at critical points in metabolic pathways are often regulated by altering the activity of the protein.

Regulated enzymes are usually found at the beginnings, ends or branch points of metabolic pathways. Control is exerted at these critical points and the enzymes in between merely operate automatically. Many biosynthetic pathways are regulated by **negative feedback** (Fig. 7.38). The final product of the pathway inhibits the first enzyme in the pathway. This way, when sufficient product has accumulated, the cell does not waste materials making any more.

End products of metabolic pathways often inhibit the enzyme that performs the first step of the pathway.

Many biosynthetic pathways are branched and lead to several final products. In this case, the enzymes at each branch point may be controlled by the products of the separate branches. An example is the synthesis of the aspartate family of amino acids. Aspartate is the precursor to four other amino acids (Lys, Met, Thr, and Ile), all of which inhibit, by feedback, both their own branch and the pathway as a whole (Fig. 7.39).

Allosteric Enzymes Are Affected by Signal Molecules

Enzymes that are regulated by binding small molecules, at a site away from the active site, alternate between two forms with different 3-D conformations and are known as **allosteric enzymes**. Interconversion between the active and inactive form involves a change in shape and usually also assembly or disassembly of the protein subunits that make up the enzyme. In the active form, the active site is available for substrate, whereas in the inactive form, the active site is often blocked or altered in shape so that

allosteric protein Protein that changes shape when it binds a small molecule
negative feedback Form of negative regulation where the final product of a pathway inhibits the first enzyme in the pathway

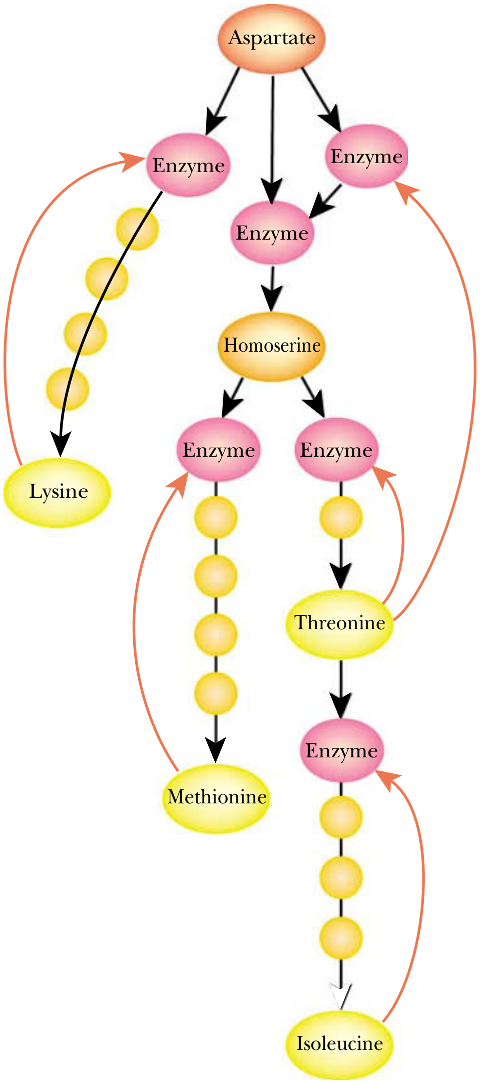


FIGURE 7.39 ***Feedback Inhibition of Aspartate Family Pathway***

Aspartate can be converted into four other amino acids. Negative feedback, as shown by arrows, is exerted on the production of the four amino acids shown in yellow.

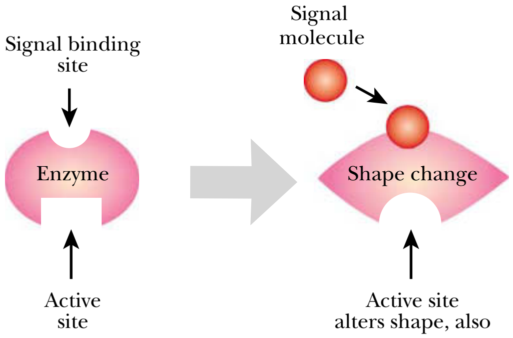


FIGURE 7.40 ***Allosteric Enzymes Change Shape***

An enzyme that has both an active site and an allosteric site is altered in shape when a signal molecule binds. The altered shape also affects the configuration of the active site.

substrate cannot bind. The small molecules that regulate allosteric enzymes bind at special regulatory sites called allosteric sites because when they are occupied, this triggers the enzyme to change shape (Fig. 7.40).

Allosteric enzymes generally do not follow standard Michaelis-Menten enzyme kinetics. Plots of rate versus substrate concentration are S-shaped or sigmoid, not hyperbolic. Allosteric changes may alter the K_m or the V_m or both. Allosteric effectors are usually unrelated chemically to the substrate of the enzyme they control. As noted above, they are often the end products of metabolic pathways in which the enzyme is involved.

Allosteric enzymes may be inhibited or activated upon binding small signal molecules.

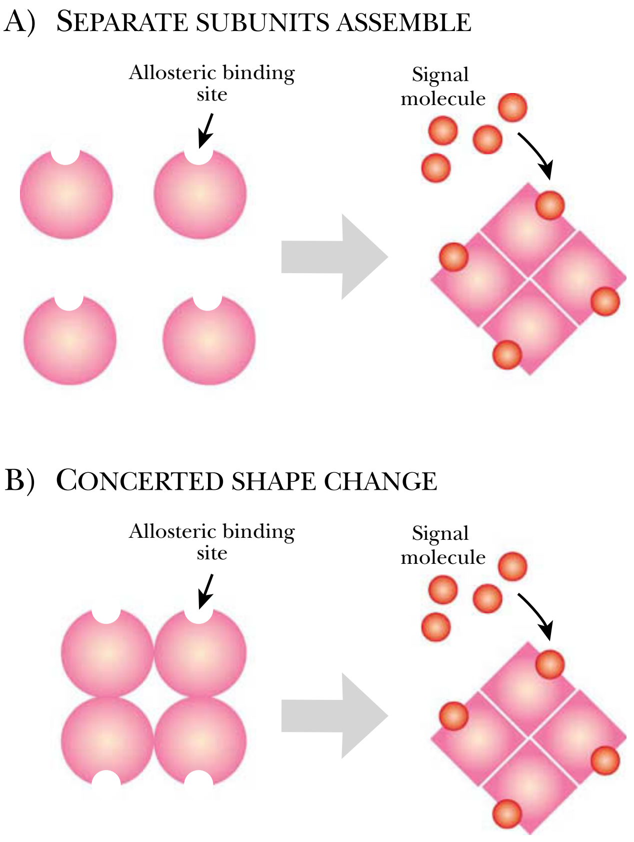


FIGURE 7.41 ***Allosteric Enzymes Have Multiple Subunits***

A) Four subunits of an allosteric enzyme are induced to bind together forming the active enzyme when the signal molecule binds to the allosteric site. B) An allosteric enzyme with four subunits already associated binds the signal molecule and undergoes a concerted shape change.

Allosteric enzymes may be regulated negatively or positively or both. For example, the enzyme phosphofructokinase (PFK) is a major control point for glycolysis. When there is plenty of ATP, this inhibits PFK, whereas the build-up of AMP and/or ADP signals that the cell is running low on energy and activates PFK. Thus, some allosteric enzymes have two different allosteric sites, one for an activator and the other for an inhibitor. In the case of PFK, ATP is the negative allosteric effector and AMP is the positive allosteric effector.

All allosteric enzymes consist of multiple subunits. When the allosteric effector binds, it changes the shape of the subunit to which it binds. In some cases, this also affects the assembly of the subunits (Fig. 7.41). One form of the **allosteric protein** exists as monomers and the other as multimers. In other cases, the subunits stay together. In this case, the shape change may be transmitted from one subunit to the next (which, as a bonus, can now bind the allosteric effector more easily). The subunits are said to undergo a *concerted shape change*.

Many DNA-binding proteins are also allosteric. They change their shape and their ability to bind to DNA when they bind small regulatory molecules. This allows the regulation of gene expression in response to a variety of chemical stimuli, as discussed in Chapters 9 and 10.

Enzymes May Be Controlled by Chemical Modification

Some proteins change conformation and activity after binding a small molecule. In other cases, a shape change is caused by modifying the protein chemically. Normally this occurs when a chemical group is added—one that can be removed again later. Phosphate groups are the most common examples of small molecules that affect shape change, but other groups including acetyl, methyl, and adenyl (AMP) may be effective.

allosteric protein Protein that changes shape when it binds a small molecule

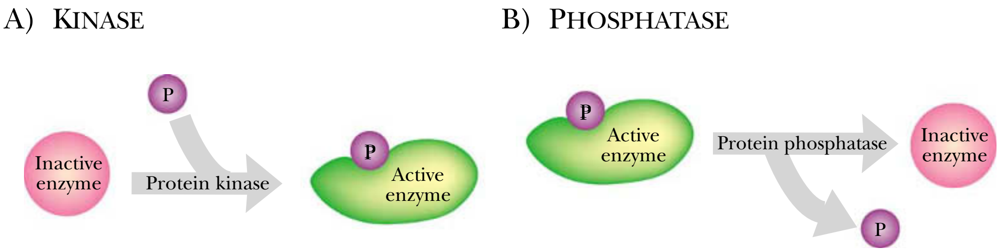


FIGURE 7.42 ***Addition and Removal of Phosphate to and from Proteins***

A) An inactive enzyme may be made active by the addition of a phosphate group by a protein kinase. A shape change occurs that makes the phosphorylated enzyme active. B) The active enzyme is altered back to the inactive conformation when the phosphate group is removed by a phosphatase.

The activity of many proteins is altered by adding or removing phosphate groups.

Phosphate groups are attached to proteins by enzymes known as **protein kinases** and are removed by protein **phosphatases** (Fig. 7.42). Control of enzyme activity by covalent modification is relatively rare in bacteria but extremely common in animal cells. About a third of all the 10,000 or so different proteins in an animal cell are phosphorylated at any given instant. When animal cells receive signals from outside, they often respond by phosphorylating a particular set of proteins that includes both enzymes and transcription factors.

The classic example of control by phosphorylation is the synthesis and breakdown of **glycogen** by animal cells. Glycogen is a storage carbohydrate that is split to give glucose when cells need energy. It is made by glycogen synthase and broken down by glycogen phosphorylase (Fig. 7.43). Both enzymes are controlled by phosphorylation. Glycogen phosphorylase is active when phosphorylated, whereas glycogen synthase is inactive. As is often the case, there is a cascade of enzymatic reactions involving two protein kinases. When the cells need energy, protein kinase A phosphorylates both glycogen synthase and phosphorylase kinase, which in turn phosphorylates glycogen phosphorylase.

Potentially dangerous enzymes are often activated by cleavage of inactive precursors.

Some enzymes are activated by cleavage of a precursor protein to yield active enzyme. The digestive enzymes trypsin, chymotrypsin and pepsin are synthesized as longer precursors known as trypsinogen, chymotrypsinogen and pepsinogen. Only in the intestine, and safely outside the cells that made them, are they activated by cleavage of the polypeptide chain. Unlike control by binding small molecules or phosphorylation, this sort of activation is non-reversible.

Binding of Proteins to DNA Occurs in Several Different Ways

A wide range of proteins binds to DNA. These proteins are involved in DNA replication, in gene expression and its control, in protection and repair of DNA, and a variety of other processes. Understanding the properties of DNA-binding proteins is of major importance in biotechnology, where they are used to control the expression of cloned genes. DNA-binding proteins are also of relevance in molecular medicine, especially in such areas as cancer and aging. Despite the great variety of DNA-binding proteins, there are some common themes in how these proteins interact with DNA.

Many DNA binding proteins recognize specific base sequences. Such recognition sequences are often (but not always) inverted repeats.

Although some DNA-binding proteins are relatively non-specific, many recognize and bind to specific base sequences in the DNA. Almost all DNA-binding proteins fit into the major groove of DNA as this allows them to recognize and make contact with the bases. When several sites recognized by a particular DNA-binding protein are compared, they are found to have very similar, though rarely identical, sequences. Many

glycogen Storage carbohydrate found both in bacteria and in the livers of animals
phosphatase An enzyme that removes phosphate groups
protein kinase An enzyme that adds phosphate groups to another protein

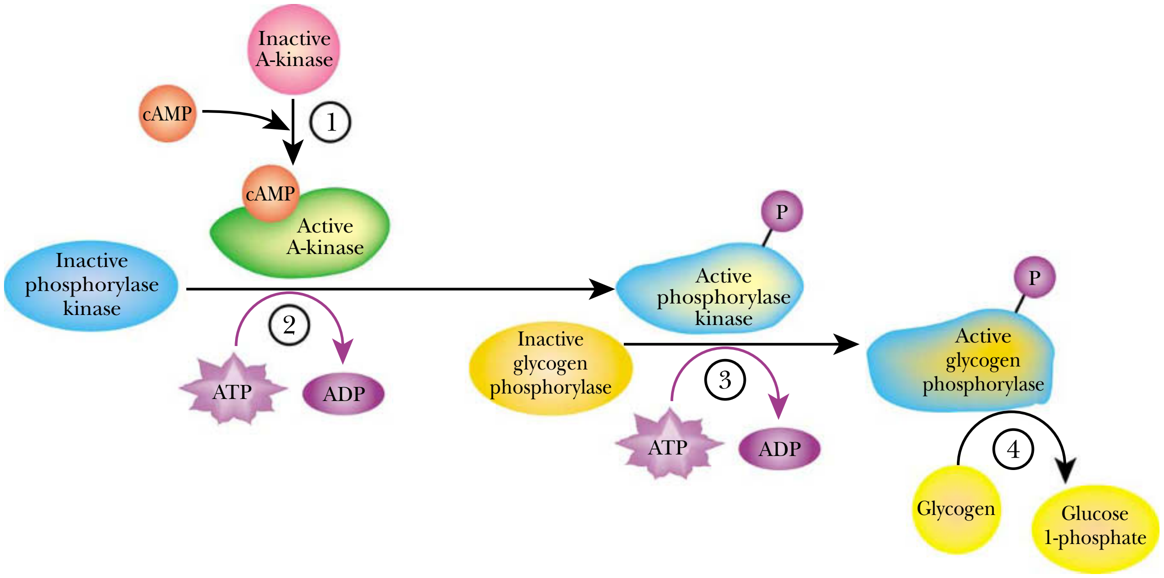

Inactive A-kinase
cAMP
1
cAMP
Active A-kinase
Inactive phosphorylase kinase
2
ATP
ADP
P
Active phosphorylase kinase
Inactive glycogen phosphorylase
3
ATP
ADP
P
Active glycogen phosphorylase
4
Glycogen
Glucose 1-phosphate

FIGURE 7.43 ***Control of Glycogen Synthesis and Breakdown***

A four-step process is necessary to breakdown glycogen to release glucose as glucose 1-phosphate. 1) Inactive protein kinase A is activated upon binding to cyclic AMP (cAMP). 2) Activated protein kinase A uses ATP to change the inactive phosphorylase kinase to the active phosphate-bound form. 3) Activated phosphorylase kinase converts inactive glycogen phosphorylase to the active phosphorylated form. 4) Ultimately, active glycogen phosphorylase converts glycogen to glucose 1-phosphate, which is the first substrate for the process of glycolysis.

DNA-binding proteins, including both regulatory proteins and enzymes that cut or modify DNA, recognize palindromes or inverted repeats in the DNA. In this case, proteins that consist of single subunits often bind to inverted repeats that are 4 to 8 bp long overall. Proteins that consist of paired subunits usually bind to inverted repeats that have two 5- or 6-base repeats separated by half a dozen bases whose sequence is relatively unimportant (Fig. 7.44). These relatively short palindromes do not form hairpins or stem and loop structures.

Just a few structural motifs are responsible for binding DNA in a large number of different DNA binding proteins.

A vast number of different transcription factors and other regulatory proteins bind DNA by means of a relatively small number of DNA-binding domains. The best known of these motifs are the helix-turn-helix, helix-loop-helix, leucine zipper and zinc finger.

Both the **helix-turn-helix (HTH)** and the similar but distinct **helix-loop-helix (HLH)** motifs consist of two α-helices joined by a loop (Fig. 7.45). The turn or loop is shorter for the HTH motif and longer for the HLH domain. In each case, one of the α-helices fits into the major groove of the DNA double helix and makes contact with the bases. In the HTH motif, it is the second of the two helices (counting from the N-terminal end) that is responsible for DNA binding whereas in the HLH motif it is the first α-helix. Proteins with these motifs usually bind as dimers to inverted repeats in the DNA (Fig. 7.46).

The HTH domain is widely used by both prokaryotes and eukaryotes whereas the HLH motif is found mostly in eukaryotes. For example, the HTH motif is found in the Crp global activator of *E. coli* and in both the CI and Cro (Fig. 7.47) regulatory proteins of bacteriophage lambda. Eukaryotic transcription factors that recognize homeobox sequences and control development in multi-cellular animals use an HTH motif. [Homeobox sequences are found in the regulatory regions of genes involved in overseeing spatial and temporal development in animals (see Ch. 19).] Although the rest

helix-loop-helix (HLH) One type of DNA-binding motif common in proteins
helix-turn-helix (HTH) One type of DNA-binding motif common in proteins

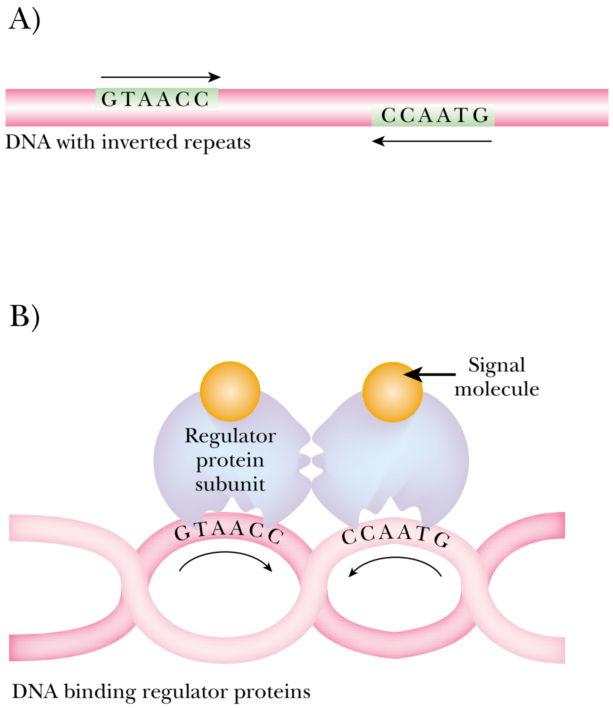

A)
GTAACC
CCAATG
DNA with inverted repeats
B)
Signal molecule
Regulator protein subunit
GTAACC
CCAATG
DNA binding regulator proteins

FIGURE 7.44 *Binding of Proteins to Inverted Repeats on DNA*

A) Double-stranded DNA with a 5-base inverted repeat. B) A protein dimer has bound to the inverted repeat sequences on the two different strands of DNA. Note how the helical twisting of the DNA brings the two recognition sequences together and so allows the two protein subunits to bind side by side.

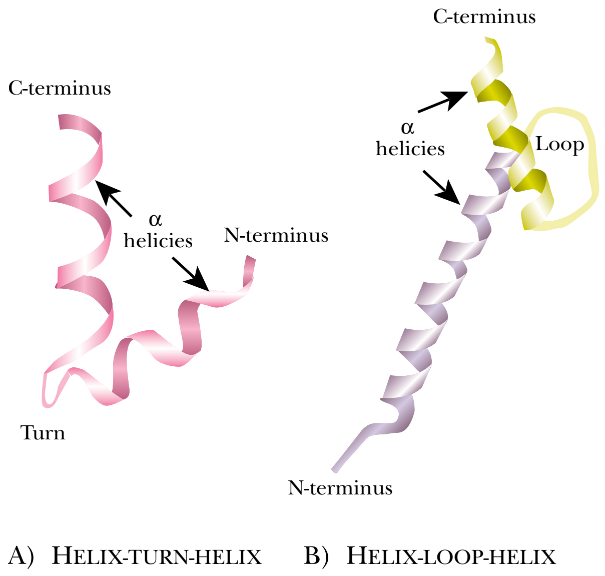


FIGURE 7.45 *Helix-Turn-Helix (HTH) and Helix-Loop-Helix (HLH) Motifs*

A simple bend versus a loop in the protein is the structural feature distinguishing between these DNA-binding proteins.

of the DNA-binding domain is different, the HTH motif that actually binds to the DNA is almost identical to that found in the lambda CI repressor (Fig. 7.46).

The **leucine zipper** is found in many eukaryotic transcription factors, including the Fos, Jun and Myc proteins that are involved in control of cell division and carcinogenesis. A leucine zipper motif consists of an α-helix with leucine residues every seventh amino acid. In addition, the amino acids halfway between the leucines are usually hydrophobic. Because there are 3.6 amino acids per turn, these hydrophobic residues form a strip down the side of the α-helix (Fig. 7.48). Two such α-helices can bind together by their hydrophobic strips forming a zipper structure. The actual binding of DNA is due to basic residues in front of the zipper region.

leucine zipper One type of DNA-binding motif common in proteins

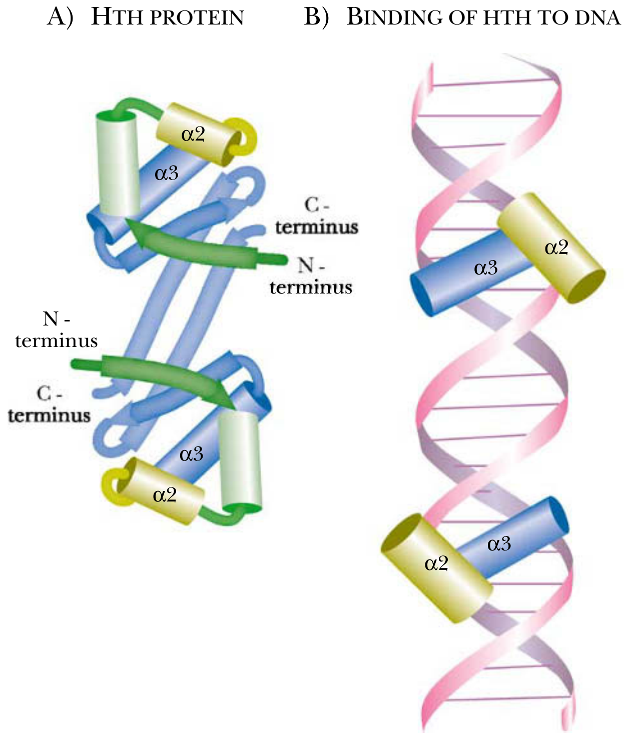


FIGURE 7.46 *Binding of Helix-Turn-Helix (HTH) Motif to DNA*

A typical HTH protein is a dimer with two sets of α-helices, labeled α2 and α3, that actually bind to the DNA. B) The pairs of α-helices fit into two adjacent major groves in the DNA. In panel B only the α2 and α3 helices are indicated to show how they interact with the DNA. The HTH shown is the phage lambda CI repressor.

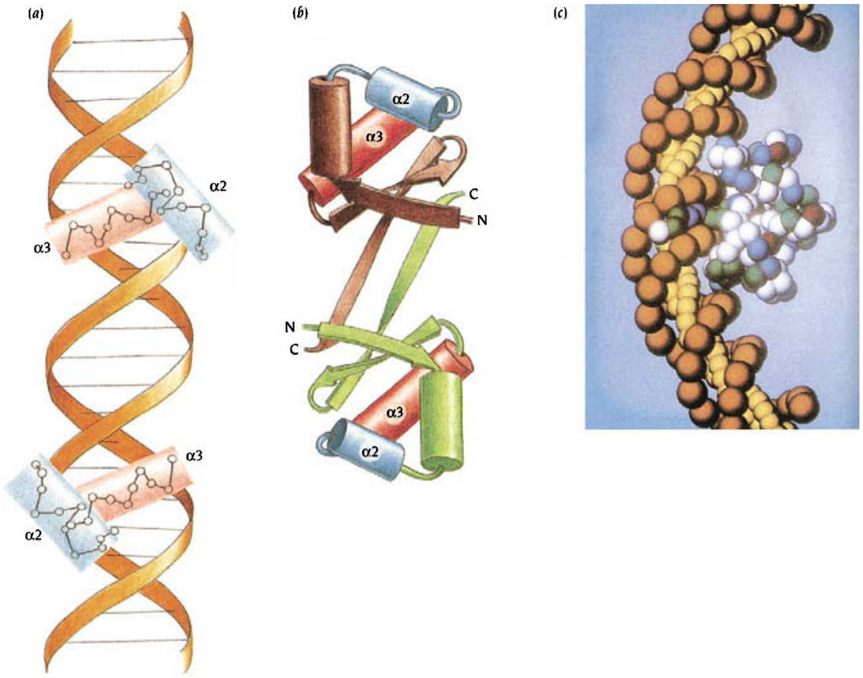


FIGURE 7.47 *Helix-Turn-Helix Motif of Cro Protein from Lambda*

Lambda Cro protein is shown bound to DNA (orange). A) The two HTH recognition helices (red) of Cro sit in the major groove of the DNA according to the model of Brian Matthews. B) Schematic diagram of the Cro dimer. C) Space-filling model of Cro dimer bound to bent B-DNA. The sugar-phosphate backbone of DNA is orange and the bases are yellow. From: Introduction to Protein Structure by Brandon & Tooze, 2nd ed., 1999. Garland Publishing, Inc., New York and London.

Zinc fingers are widely distributed in DNA binding proteins. Each zinc finger recognizes three bases in the DNA.

A **zinc finger** consists of a central zinc atom with a segment of 25–30 amino acid residues arranged around it (Fig. 7.49). In the classic version of the Zn finger, the Zn is bound to two cysteines, which lie in a very short piece of β-sheet—a β-hairpin—and two histidines, which lie in a short α-helix. The far end of the α-helix protrudes into the major groove of the DNA. Over a thousand zinc finger proteins are known, and

zinc finger One type of DNA-binding motif common in proteins

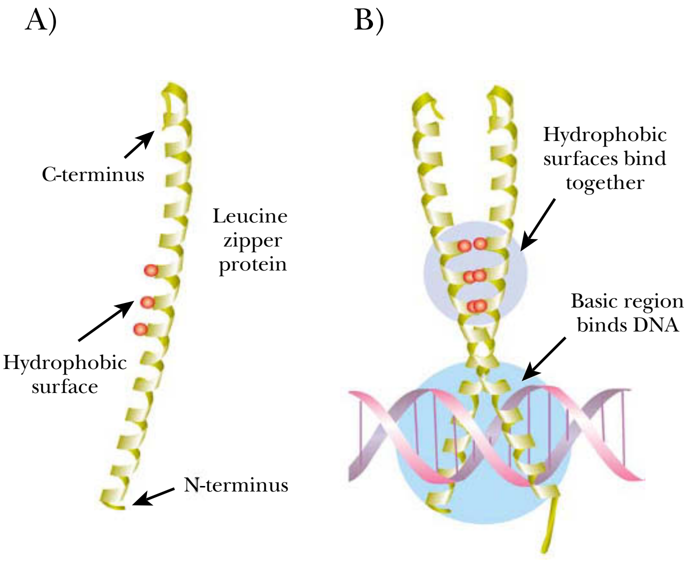


FIGURE 7.48 ***Leucine Zipper Protein Binding DNA***

A) The leucine zipper consists of two α-helixes that have hydrophobic zones and basic ends. B) The helixes of the leucine zipper binds to each other by their hydrophobic regions and to DNA by their basic regions. The basic end region fits into the major groove of the DNA. Because the basic regions are roughly parallel and open up around the DNA, the two helical segments resemble a zipper.

many of them have multiple fingers. The first zinc finger protein discovered was the general eukaryotic transcription factor TFIIIA, from the toad *Xenopus*, which has nine zinc fingers.

Each zinc finger unit usually recognizes three bases in the DNA. Less often, four or five bases are recognized by a single zinc finger. The sequence specificity of each zinc finger depends on the amino acid sequence of the polypeptide chain between the His and Cys residues that bind the zinc. Amino acids in this region make hydrogen bonds with bases in the DNA.

Several modified versions of the zinc finger motif have been found. For example, the **steroid receptor** family of transcription factors has a DNA-binding domain that contains two Zn atoms, each surrounded by four cysteines. A short α-helix that lies between the two zincs binds to the DNA. Several fungal transcription factors, including GAL4 of yeast, contain fingers built around a cluster of two Zn atoms bound to six cysteines.

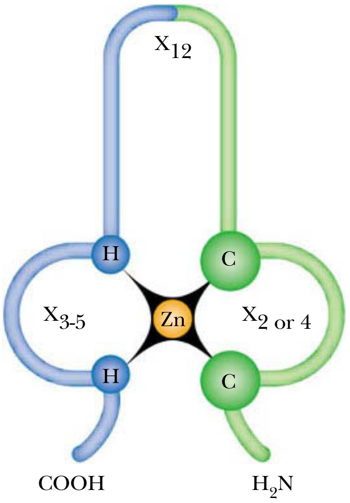


FIGURE 7.49 ***Zinc Finger DNA-Binding Protein***

A central zinc atom is bound to the sulfurs of cysteine (C) and the nitrogens of histidine (H). Chains of amino acids of varying lengths (x = chain length) extend from these binding regions. The zinc finger forms a component of a much larger protein and binds the protein to DNA.

The correct 3D structure of proteins can be destroyed by heating, detergents, acids, bases and certain chemicals.

Denaturation of Proteins

Denaturation is the loss of correct 3-D structure, i.e. the loss of both tertiary and quaternary structure. Denaturation only involves the breaking of non-covalent bonds. Loss of biological activity generally accompanies such structural denaturation. In particular, enzymes are especially sensitive to denaturation. When proteins are denatured they often precipitate out of solution, as happens to the proteins of an egg when it is boiled. Heat, extremes of pH, and a variety of chemical agents destroy non-covalent structure and denature proteins. Even agitation may denature some proteins as when egg whites are whipped to give meringues.

Proteins vary greatly in their stability. Some are very sensitive and even slight changes in pH or temperature may inactivate them. This often causes problems in both the purification of proteins and their biotechnological applications. Consequently, researchers have often searched for natural proteins that are unusually resistant, or have modified proteins to increase their stability. Bacteria that live under extreme conditions are a rich source of such resistant proteins. For example, the Taq polymerase

denaturation Loss of correct 3D structure of proteins or nucleic acids
steroid receptor Protein that binds steroid hormones

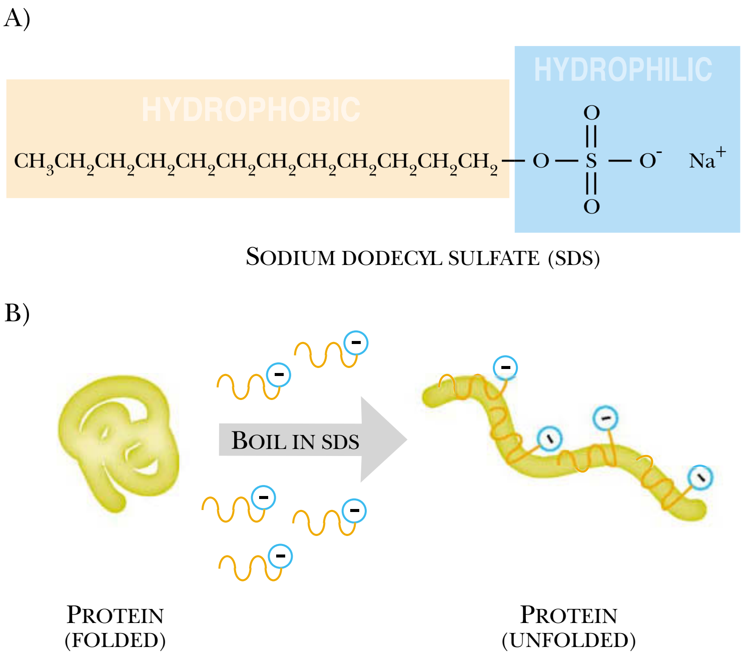

A)
HYDROPHILIC
HYDROPHOBIC
$CH_3CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2—O—S(=O)_2—O^- \ Na^+$
SODIUM DODECYL SULFATE (SDS)
B)
BOIL IN SDS
PROTEIN (FOLDED)
PROTEIN (UNFOLDED)

FIGURE 7.50 ***Structure and Function of Sodium Dodecyl Sulfate (SDS)***

A) SDS contains a strongly hydrophobic carbon chain and a strongly hydrophilic sodium sulfate region. B) When a folded protein is boiled in the presence of SDS, the hydrophobic region winds around the polypeptide backbone and the negatively charged sulfate protrudes. The negative charges repel each other which helps in straightening out the protein.

used in PCR (see Ch. 23) is a highly heat-resistant enzyme made by bacteria that live naturally at temperatures so high they would kill most organisms.

Detergents and chaotropic agents help solubilize hydrophobic groups in water.

The hydrophobic forces responsible for maintaining much of the 3-D structure of proteins are disrupted by **detergents** and **chaotropic agents**. Detergents consist of a hydrophobic tail joined to a highly water-soluble group. They act by directly binding to the hydrophobic regions of other molecules and solubilizing them. In the case of proteins, detergents can bind to the hydrophobic groups that are normally buried deep inside and also to the relatively hydrophobic polypeptide backbone. This destabilizes the 3-D folding of the polypeptide chain.

Sodium dodecyl sulfate is widely used to solubilize proteins before running them on gels.

The detergent **sodium dodecyl sulfate (SDS)** is widely used to solubilize and denature proteins before running them on polyacrylamide gels to separate them by molecular weight (see Ch. 26). SDS has a long hydrocarbon tail that binds to the polypeptide and a negatively charged sulfate group that sticks out into the water and solubilizes the protein/SDS complex (Fig. 7.50). SDS binds to polypeptides along their long axes and converts them to an extended rod-shaped conformation. The precise nature of SDS-polypeptide binding is disputed. One theory is that it binds to the non-polar R-groups of hydrophobic amino acids. However, SDS binds in a ratio of one SDS to every two amino acid residues for the vast majority of proteins and this suggests that it binds to the polypeptide backbone. From a practical viewpoint, it is important that the number of negative charges contributed by bound SDS is proportional to the length of the polypeptide, i.e. to its molecular weight.

Chaotropes alter the structure of water, so allowing hydrophobic groups to dissolve more easily.

Chaotropic agents (e.g. thiocyanate, perchlorate) also destabilize proteins by promoting the exposure of hydrophobic groups, but by an indirect mechanism. When exposed to water, hydrophobic groups induce the formation of regular cages of water molecules around themselves. This decreases the disorder of the water; i.e., it causes an increase in entropy, which is thermodynamically unfavorable. Hydrophobic groups tend to cluster together to avoid contact with water, rather than because of any positive attraction. Chaotropes disrupt water structure and so allow hydrophobic groups to dissolve more readily.

chaotropic agent Chemical compound that disrupts water structure and so helps hydrophobic groups to dissolve
detergent Molecule with both hydrophobic and hydrophilic regions that can solubilize hydrophobic molecules including fats, grease and lipids
sodium dodecyl sulfate (SDS) A detergent widely used to denature and solubilize proteins before separation by electrophoresis

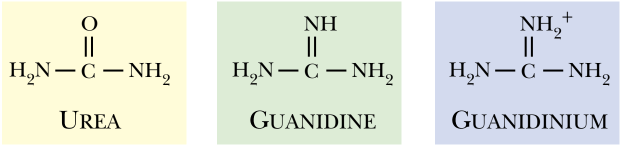


FIGURE 7.51 ***Protein Denaturants Disrupt Hydrogen Bonds***

Urea, guanidine and guanidinium form hydrogen bonds with the peptide groups of proteins. This disrupts the hydrogen bonds within the protein that help maintain secondary structure of the protein. The result is unfolding (denaturation) of the polypeptide chain.

Urea and guanidine disrupt hydrogen bonds.

Protein **denaturants** are molecules that disrupt the hydrogen bonds that maintain secondary structure. Examples are **urea**, **guanidine** and **guanidinium chloride** (Fig. 7.51). They act by forming hydrogen bonds between their own CO or NH_2 groups and all the groups on the protein that can take part in hydrogen bonding. High temperatures and extremes of pH also destroy hydrogen bonding.

Denaturation is also aided by breaking the disulfide bonds of those proteins that rely on them to stabilize their structures. In the laboratory, **β-mercaptoethanol or BME** ($HOCH_2CH_2SH$) is often used to reduce disulfide bonds to two -SH groups, so breaking the linkage.

β-mercaptoethanol (BME) A small molecule with free sulfhydryl groups often used to break disulfide bonds in proteins
denaturant Chemical compound that destroys the 3D structure of proteins, especially by breaking hydrogen bonds
guanidine Non-ionized form of guanidinium
guanidinium chloride A widely used denaturant of proteins
urea A nitrogen waste product of animals; also widely used as a denaturant of proteins

CHAPTER **EIGHT**

Protein Synthesis

Protein Synthesis Follows a Plan

Each protein is made using the genetic information stored in the chromosomes (see Ch. 3 for a brief overview). The genetic information is transmitted in two stages. First the information in the DNA is transcribed into **messenger RNA** (**mRNA**). The next step uses the information carried by the mRNA to give the sequence of amino acids making up a polypeptide chain. This involves converting the nucleic acid "language," the genetic code, to protein "language," and is therefore known as **translation**. This overall flow of information in biological cells from DNA to RNA to protein is known as the central dogma of molecular biology (see Chapter 3, Fig. 3.17) and was first formulated by Sir Francis Crick.

Ribosomes use the information carried by messenger RNA to make proteins.

The decoding of mRNA is carried out by a submicroscopic machine called a **ribosome**, which binds the mRNA and translates it. The ribosome moves along the mRNA reading the message and synthesizing a new polypeptide chain. Bacterial protein synthesis will be discussed first. The process is similar in higher organisms, but some of the details differ and will be considered later.

Proteins Are Gene Products

Gene products include proteins as well as non-coding RNA.

An early rule of molecular biology was Beadle and Tatum's dictum: "one gene—one enzyme" (see Ch. 1). This rule was later broadened to include other proteins in addition to enzymes. Proteins are therefore often referred to as "**gene products**." However, it must be remembered that some RNA molecules (such as tRNA, rRNA, small nuclear RNA) are never translated into protein and are therefore also gene products.

Furthermore, instances are now known where one gene may encode multiple proteins (Fig. 8.01). Two relatively widespread cases of this are known—alternative splicing and polyproteins. In eukaryotic cells, the coding sequences of genes are often interrupted by non-coding regions, the introns. These introns are removed by splicing at the level of messenger RNA. Alternative splicing schemes may generate multiple mRNA molecules and therefore multiple proteins from the same gene. This is especially frequent in higher eukaryotes, in particular vertebrates (see Ch. 12). A set of proteins generated in this manner shares much of their sequence and structure.

gene product End product of gene expression; usually a protein but includes various untranslated RNAs such as rRNA, tRNA, and snRNA
messenger RNA The type of RNA molecule that carries genetic information from the genes to the rest of the cell
ribosome The cell's machinery for making proteins
translation Making a protein using the information provided by messenger RNA

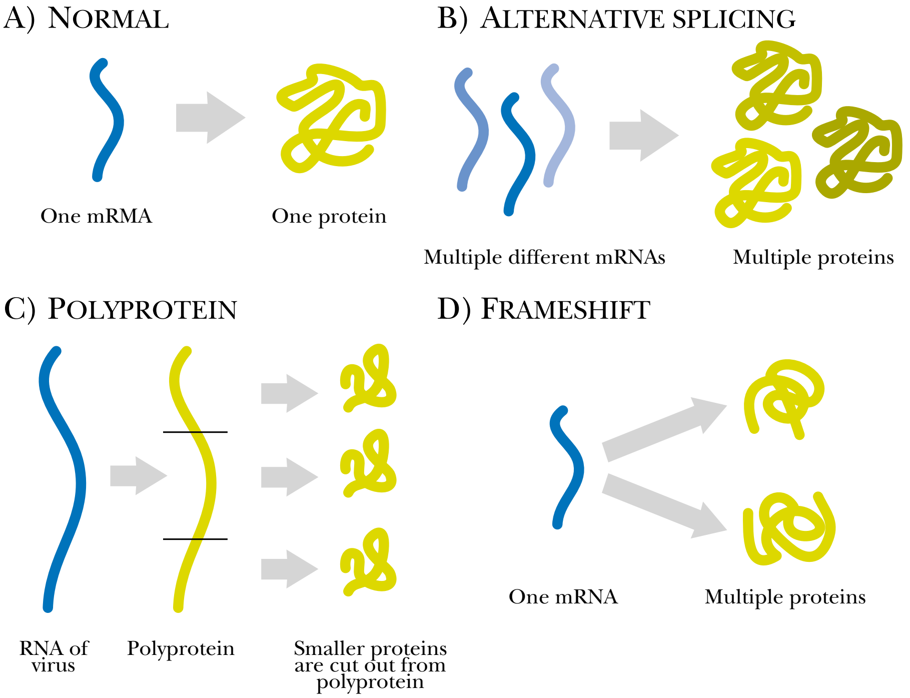


FIGURE 8.01 ***How Many Proteins Per Gene?***

A) Normally each gene is transcribed giving one mRNA and this is translated into a single protein. Variations in the normal theme are B) alternative splicing, C) polyproteins and D) multiple proteins due to the use of different reading frames.

In eukaryotic cells, mRNA only carries information from a single gene and therefore can only be translated into a single protein. This causes problems for certain viruses that infect eukaryotic cells and which have RNA genomes (see Ch. 17). To circumvent the problem, these viruses make a huge "polyprotein" from an extremely long coding sequence in their RNA. This polyprotein is then cut up into several smaller proteins.

Although there are exceptions, most genes give rise to a single protein.

Finally, there are occasional oddities, such as the generation of two proteins from the same gene due to frameshifting (see below). Despite these exceptions, it is still generally true that most genes give rise to a single protein.

Just as the total genetic information of a cell is the genome, so the total number of different proteins that a cell can produce is sometimes known as the **proteome**. In bacteria there is an almost one-for-one correspondence between genes and proteins. However, in higher organisms where alternative splicing is common, there may be an average of two or three final proteins per gene and so the proteome may be significantly larger than the genome.

Decoding the Genetic Code

Each amino acid in a protein is encoded by three bases in the DNA or RNA sequence.

There are 20 amino acids in proteins but only four different bases in the mRNA. So one cannot simply use one base of a nucleic acid to code for a single amino acid when making a protein. During translation, the bases of mRNA are read off in groups of three, which are known as **codons**. Each codon represents a particular amino acid. Four different bases gives 64 possible groups of three bases; that is, 64 different codons in the **genetic code**. Because there are only 20 different amino acids, some are encoded by more than one codon. In addition, three of the codons are used for punctuation. Those are the **stop codons** that signal the end of a polypeptide chain. Figure 8.02 shows nature's genetic code.

To read the codons, a set of adapter molecules that recognize the codon on the mRNA at one end and carry the corresponding amino acid attached to their other end

codon Group of three RNA or DNA bases that encodes a single amino acid
genetic code System for encoding amino acids as groups of three bases (codons) of DNA or RNA
proteome The set of all proteins that an organism can make
stop codon Codon that signals the end of a protein

FIGURE 8.02 ***The Genetic Code***

The 64 codons as found in messenger RNA are shown with their corresponding amino acids. As usual, bases are read from 5′ to 3′ so that the first base is at the 5′ end of the codon. Three codons (UAA, UAG, UGA) have no cognate amino acid but signal stop. AUG (encoding methionine) and, less often, GUG (encoding valine) act as start codons. To locate a codon, find the first base in the vertical column on the left, the second base in the horizontal row at the top and the third base in the vertical column on the right.

	2nd (middle) base				
1st base	U	C	A	G	3rd base
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA stop	UGA stop	A
	UUG Leu	UCG Ser	UAG stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

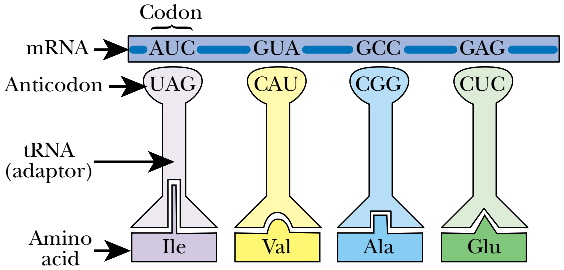


FIGURE 8.03 ***Transfer RNA Recognizes Codons***

Several tRNAs are seen bound to mRNA codons by their anticodons. Each tRNA carries a different amino acid at the end of the adaptor stem. This diagram is intended to show the principle of mRNA decoding. It does NOT illustrate the actual mechanism of protein synthesis. In real life, the codons are contiguous and there are no spacers in between and only two tRNAs are bound at any given time.

The anticodon of tRNA recognizes the codon on mRNA by base pairing.

Each transfer RNA carries one particular amino acid.

is needed. These adapters are small RNA molecules, or **transfer RNA** (**tRNA**). At one end, the tRNA has an **anticodon** consisting of three bases that are complementary to the three bases of the codon on the mRNA. The codon and anticodon recognize each other by base pairing and are held together by hydrogen bonds (Fig. 8.03). At its other end, each tRNA carries the amino acid corresponding to the codon it recognizes.

Transfer RNA Forms a Flat Cloverleaf Shape and a Folded "L" Shape

Transfer RNA molecules are about 80 nucleotides in length. About half the bases are paired to form double helical segments. A typical tRNA has four short base-paired stems and three loops (Fig. 8.04). This is shown best in the **cloverleaf structure**, intended to reveal details of base pairing, which shows the tRNA spread out flat in only two dimensions. (Such a diagram is sometimes called a secondary structure map). The tRNA cloverleaf is folded up further to give an L-shaped 3-D structure, in which the TψC-loop (or T-loop) and the D-loop are pushed together. The anticodon and attached amino acid are located at the two ends of the L-structure. Different tRNA molecules vary considerably in sequence, but they all conform to this same overall structure. Variations in length (from 73 to 93 nucleotides) occur, due mostly to the variable loop.

anticodon Group of three complementary bases on tRNA that recognize and bind to a codon on the mRNA
cloverleaf structure 2-D structure showing base pairing in a tRNA molecule
transfer RNA (tRNA) RNA molecules that carry amino acids to the ribosome

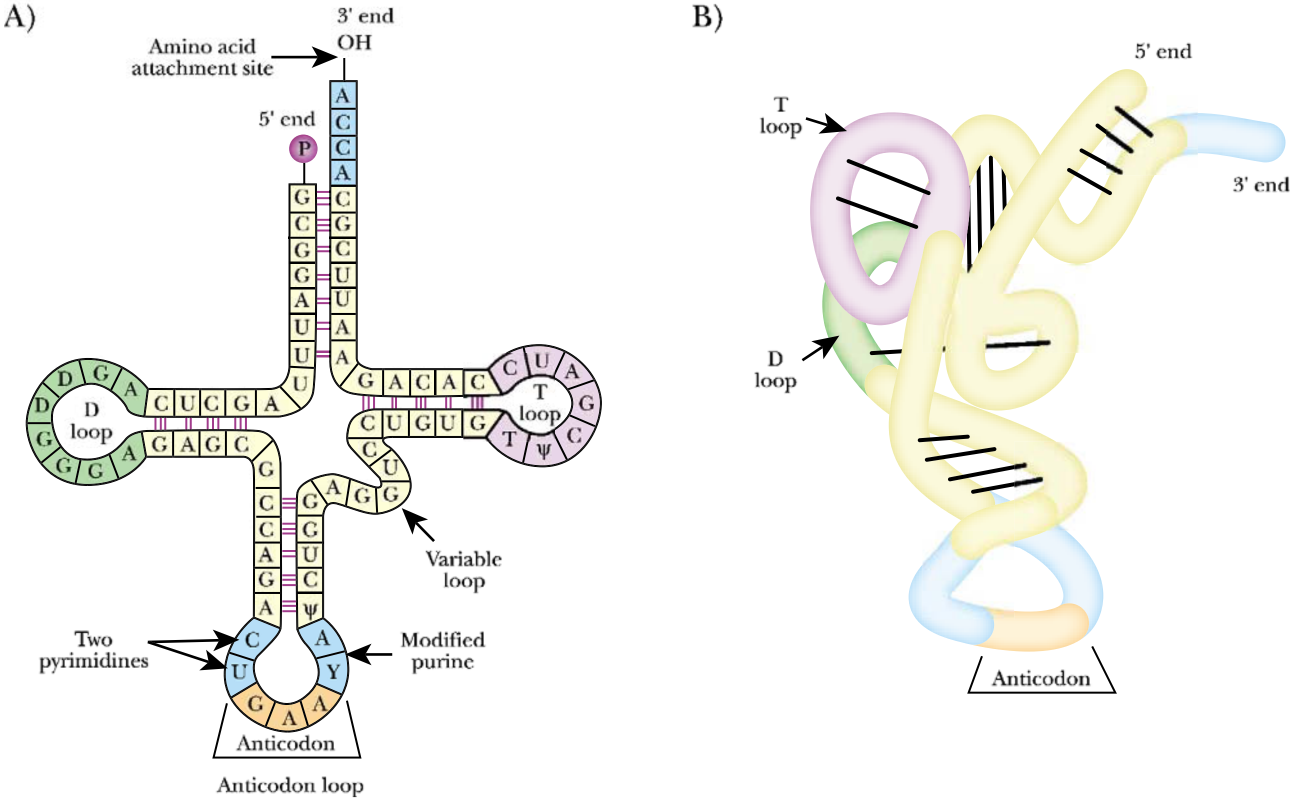


FIGURE 8.04 ***Structure of Transfer RNA***

A) A planar view (secondary structure) of a tRNA shows its cloverleaf structure comprised of the 3′ and 5′ acceptor stem, the T- (or TψC) and D-loops and the anticodon loop. A variable loop, which varies in length in different tRNA molecules is also found. B) The folded (tertiary structure) configuration resembles an "L."

The **acceptor stem** is made by pairing of the 5′-end, which almost always ends in G and is phosphorylated, and the 3′-end, which ends in CCA-OH. The amino acid is bound to the 3′-hydroxyl group of the adenosine at the free 3′-end of the acceptor stem. The anticodon is about halfway round the sequence, in the **anticodon loop**. This consists of seven bases with the three anticodon bases in the middle. The anticodon is always preceded, on the 5′ side, by two pyrimidines and followed by a modified purine (Fig. 8.04).

The other two loops of tRNA are named after **modified bases**. The T ψC-loop contains "ψ" (spelled "psi" but pronounced "sigh"), which stands for pseudouracil; and the D-loop or DHU-loop has "D" for dihydrouracil. These strange bases are required for proper folding and operation of the tRNA. The TψC-loop and the D-loop are needed for binding to the ribosome and other protein factors involved in translation (see below).

Modified Bases Are Present in Transfer RNA

As originally transcribed, RNA contains only the four bases A, U, G and C. However, some RNA molecules contain bases that are altered chemically after the RNA has been made. Some of these are shown in Fig. 8.05. This is especially true for tRNA, in which up to 15 modified bases per molecule may occur.

acceptor stem	Base paired stem of tRNA to which the amino acid is attached
anticodon loop	Loop of tRNA molecule that contains the anticodon
modified base	Nucleic acid base that is chemically altered after the nucleic acid has been synthesized

TABLE 8.01 Wobble Rules for Codon/Anticodon Pairing

	Pairs with Third Codon Base	
First Anticodon Base	**normal**	**by wobble**
G	C	U
U	A	G
I	—	C or U or A
C	G	no wobble
A	U	no wobble

Some bases are chemically modified after RNA has been transcribed. Modified bases are especially common in tRNA.

The most frequent modification is methylation. Methyl or dimethyl versions of A, U, G or C all exist. Methylation prevents pairing of certain bases and also aids binding of ribosomal proteins. Note that thymine (= 5-methyl uracil), which is normally only found in DNA, is also found in the TψC-loop of tRNA, where it is attached to ribose and is made by methylation of uracil after transcription.

In pseudouridine, the uracil itself is not altered, but is attached to ribose by carbon-5 instead of nitrogen-1, as in normal uridine. The base found in the nucleoside **inosine** is actually named hypoxanthine. However, it is written as "I" in sequences and often called I-base to avoid confusion. Similarly, the bases of queuosine and wyosine are referred to as Q-base and Y-base.

Some tRNA Molecules Read More Than One Codon

Each transfer RNA carries only a single amino acid, so at least 20 different tRNAs are needed for the 20 different amino acids. On the other hand (excluding the stop codons), there are 61 codons to be recognized, as some amino acids have more than one codon. In fact, some tRNAs can read more than one codon, though, of course, these must all code for the same amino acid. The minimum set of different tRNA molecules needed to read all 61 codons is 31. The actual number found is usually slightly higher and varies a little from species to species.

Wobbling between the anticodon and codon allows some tRNA molecules to read more than one codon.

Since only complementary bases can pair, how does a tRNA with one anticodon read more than one codon? Remember that the standard base pairing rules apply to bases that form part of a DNA double helix. Since the codon and anticodon do not form a standard double helix, slightly different rules for base pairing apply. The last two bases of the tRNA anticodon, which pair with the first two bases of the mRNA codon, pair strictly according to normal rules. However, the first base of the tRNA anticodon (which pairs with the third base of the mRNA codon) can wobble around a little because it is not squeezed between other bases as in a helix structure. Consequently, the codon/anticodon base pairing rules are known as the **wobble rules** (see Table. 8.01).

If the first anticodon base is G it can pair with C, as usual, or, in wobble mode, with U. For example, tRNA for histidine, with GUG as anticodon, can recognize both the CAC and CAU codons. Similarly, if the first anticodon base is U, it can pair with A or G. Whenever an amino acid is encoded by a pair of codons, the third codon bases are U and C (e.g., histidine, tyrosine) or A and G (e.g., lysine, glutamic acid), but never other combinations. Similarly, those privileged amino acids with four or six codons may be regarded as having two or three such pairs. Due to wobble pairing, only a single tRNA is needed to read each such pair of codons. It is possible for a single tRNA to

inosine An unusual modified nucleoside derived from guanosine
wobble rules Rules allowing less rigid base pairing but only for codon/anticodon pairing

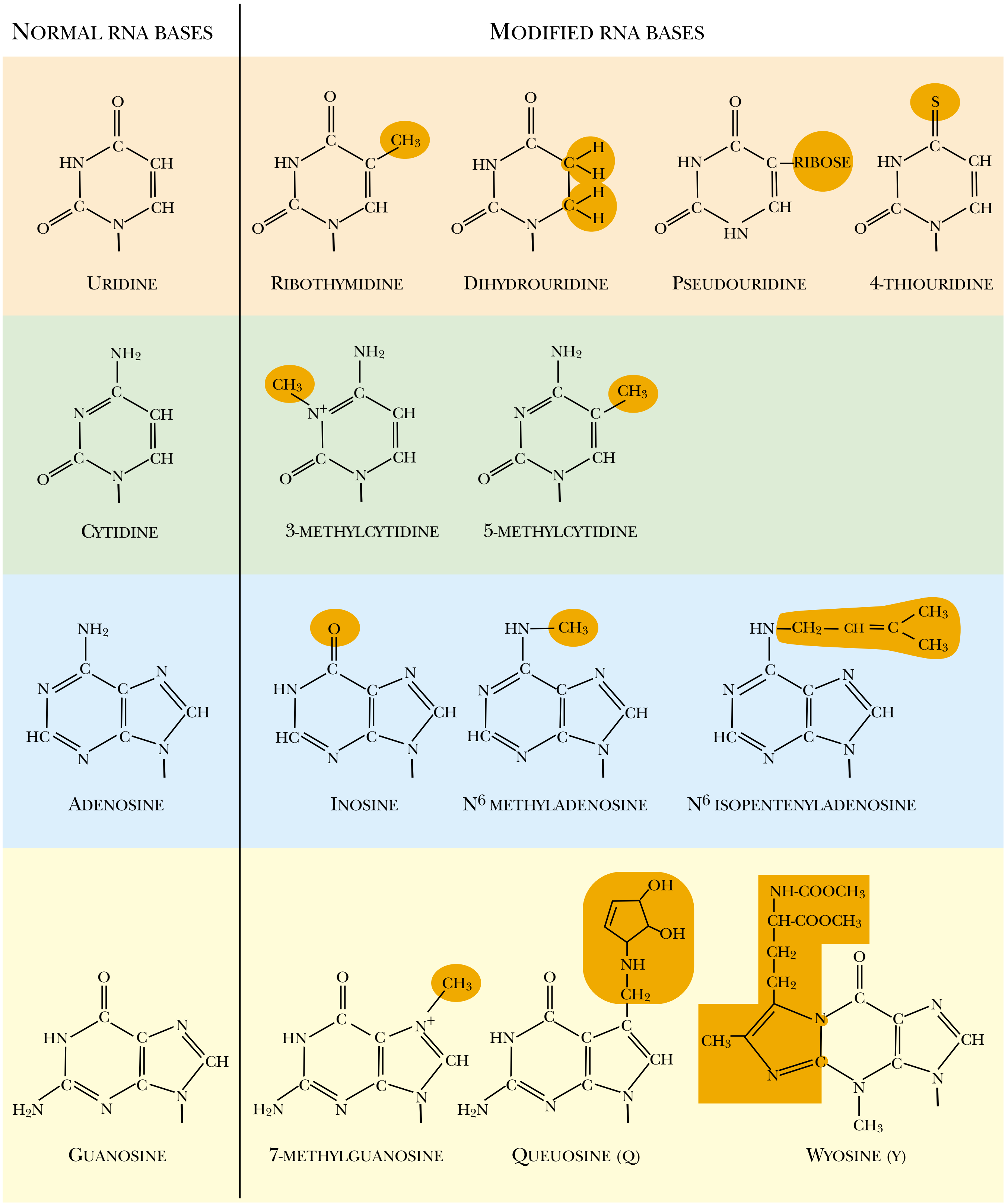


FIGURE 8.05 ***Modified Bases in tRNA***

All four bases normally found in RNA have modified derivatives that may be found in tRNA. The names given are those of the corresponding nucleosides (i.e., base plus ribose).

read three codons by making use of inosine. The I-base is occasionally used as the first anticodon base because it can pair with any of U, C or A.

Charging the tRNA with the Amino Acid

For each tRNA there is a specific enzyme that recognizes both the tRNA and the corresponding amino acid. These enzymes, known as **aminoacyl tRNA synthetases**, attach the amino acid to the tRNA. This is called charging the tRNA. Empty tRNA is known as **uncharged tRNA** while tRNA with its amino acid is **charged tRNA**.

Charging occurs in two steps (Fig. 8.06). First the amino acid reacts with ATP to form aminoacyl-AMP (also known as aminoacyl-adenylate). Next the aminoacyl-group is transferred to the 3′-end of the tRNA.

A specific enzyme attaches the correct amino acid to the correct tRNA.

a) amino acid + ATP → aminoacyl-AMP + PPi

b) aminoacyl-AMP + tRNA → aminoacyl-tRNA + AMP

The amino-acyl tRNA synthetases are highly specific for both the correct amino acid and the correct tRNA. In some cases they recognize the correct tRNA by its anticodon and in others by the sequence of the acceptor stem. Some amino-acyl tRNA synthetases recognize both regions of the tRNA. Figure 8.07 shows an amino-acyl tRNA synthetase bound to its tRNA.

The Ribosome: The Cell's Decoding Machine

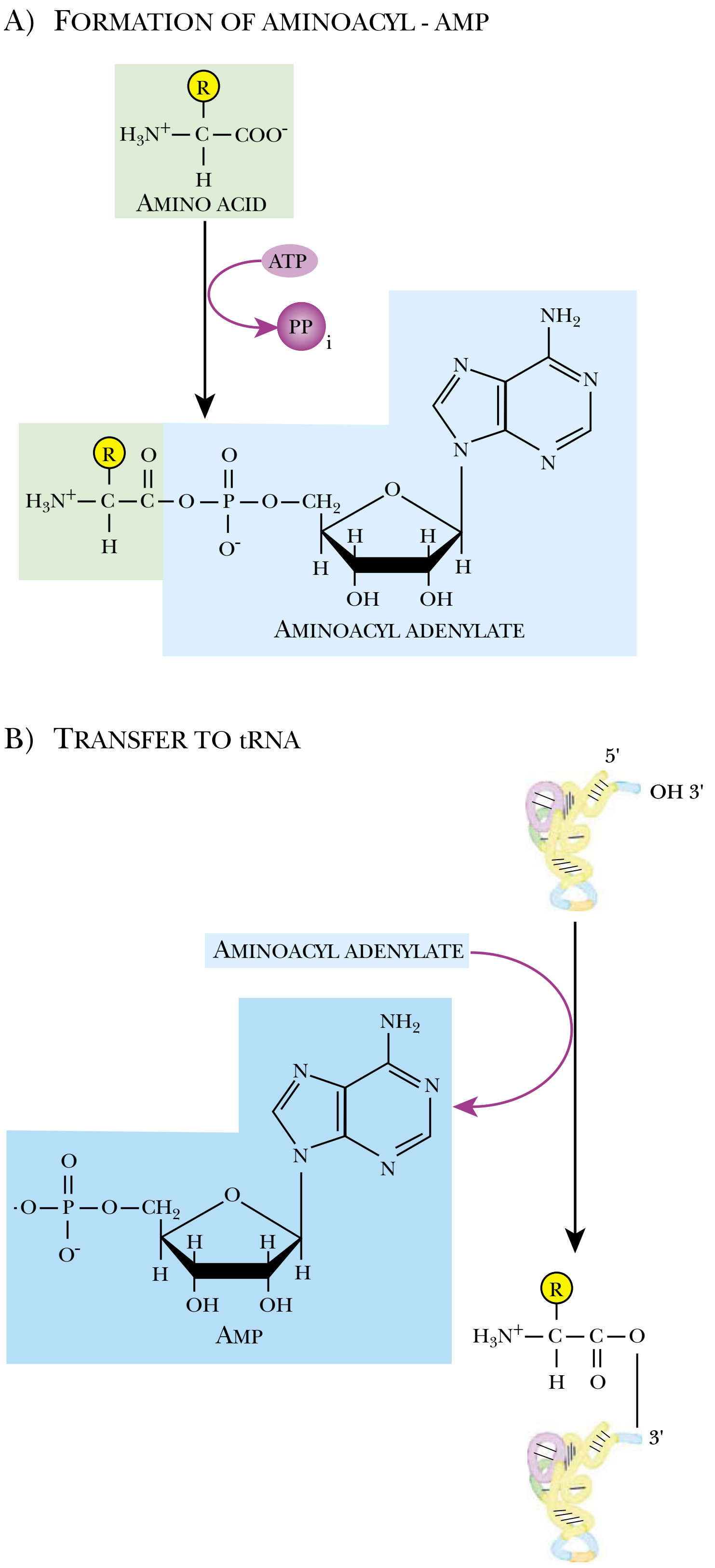
The decoding process is carried out by a submicroscopic machine called a ribosome that binds mRNA and charged tRNA molecules. The mRNA is translated, starting at the 5′-end. After binding to the mRNA, the ribosome moves along it, adding a new amino acid to the growing polypeptide chain each time it reads a codon. Each codon on the mRNA is actually read by an anticodon on the corresponding tRNA and so the information in the mRNA is used to synthesize a polypeptide chain from the amino acids carried by the tRNAs.

Ribosomes consist of proteins plus RNA and their role is to synthesize new proteins.

The ribosome and its components were originally analyzed by ultracentrifugation. Consequently, sizes are referred to in Svedberg units (S-value), which measure sedimentation velocity. Although higher S-values indicate larger particles, the S-value is not directly proportional to molecular weight. The **bacterial (70S) ribosome** consists of two subunits, the **50S** or **large subunit** and the **30S** or **small subunit** (Fig. 8.08). **Eukaryotic (80S) ribosomes** are somewhat larger, consisting of **60S** and **40S subunits** (see below).

The bacterial ribosome contains three ribosomal RNA molecules that make up about two thirds of its weight and about 50 smallish proteins that make up the remaining third. The 30S subunit contains the 16S rRNA and the 50S subunit contains the 5S and 23S rRNA (Fig. 8.08). The 3-D structure of a **70S ribosome** is shown in Figs. 8.09 and 8.10.

30S subunit Small subunit of a 70S ribosome
40S subunit Small subunit of an 80S ribosome
50S subunit Large subunit of a 70S ribosome
60S subunit Large subunit of an 80S ribosome
70S ribosome Type of ribosome found in bacterial cells
80S ribosome Type of ribosome found in cytoplasm of eukaryotic cells
amino-acyl tRNA synthetase Enzyme that attaches an amino acid to tRNA
bacterial (70S) ribosome Type of ribosome found in bacterial cells
charged tRNA tRNA with an amino acid attached
eukaryotic (80S) ribosome Type of ribosome found in cytoplasm of eukaryotic cell and encoded by genes in the nucleus
large subunit The larger of the two ribosomal subunits, 50S in bacteria, 60S in eukaryotes
small subunit The smaller of the two ribosomal subunits, 30S in bacteria, 40S in eukaryotes
uncharged tRNA tRNA without an amino acid attached


A) FORMATION OF AMINOACYL - AMP

B) TRANSFER TO tRNA

FIGURE 8.06 ***Charging Transfer RNA with the Amino Acid***

This two-step procedure begins (A) by attachment of the amino acid to adenosine monophosphate (AMP) to give aminoacyl-AMP or aminoacyl-adenylate. This involves splitting ATP and the release of inorganic pyrophosphate. Then, in the second step (B), the amino acid is transferred to the hydroxyl group of the ribose at the 3′-end of the tRNA, yielding AMP as a byproduct.

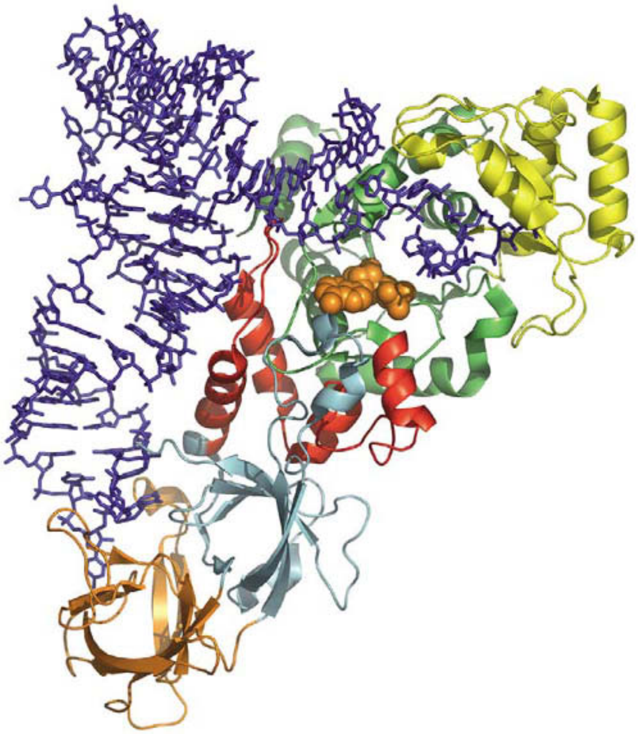


FIGURE 8.07 ***Glutamine tRNA Bound to its Aminoacyl tRNA Synthetase***

Structure of glutaminyl-tRNA synthetase bound to tRNA(Gln) and a glutaminyl adenylate analog. The analog is in orange and is shown in a space-filling representation. The tRNA is depicted in dark blue. Domains of the enzyme are color-coded as follows: active-site Rossman fold, green; acceptor-end binding domain, yellow; connecting helical subdomain, red; proximal beta-barrel, light blue; distal beta-barrel, orange. The image was made in PyMol by John Perona, Department of Chemistry and Biochemistry, University of California at Santa Barbara.

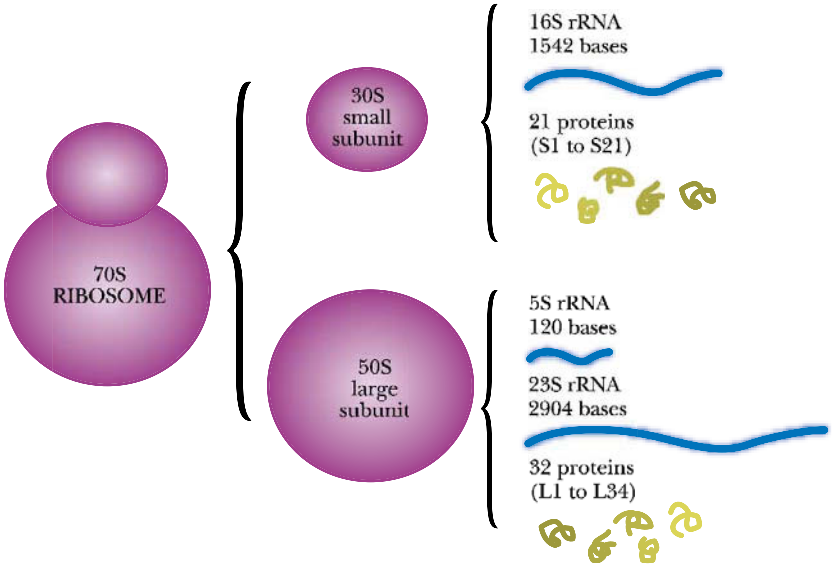


FIGURE 8.08 ***Components of a Bacterial Ribosome***

The ribosome is composed of 30S and 50S subunits. These in turn are composed of ribosomal RNA and numerous proteins. The 30S subunit is built from 16S rRNA together with 21 proteins. The 50S subunit contains 5S and 23S ribosomal RNA plus 34 proteins.

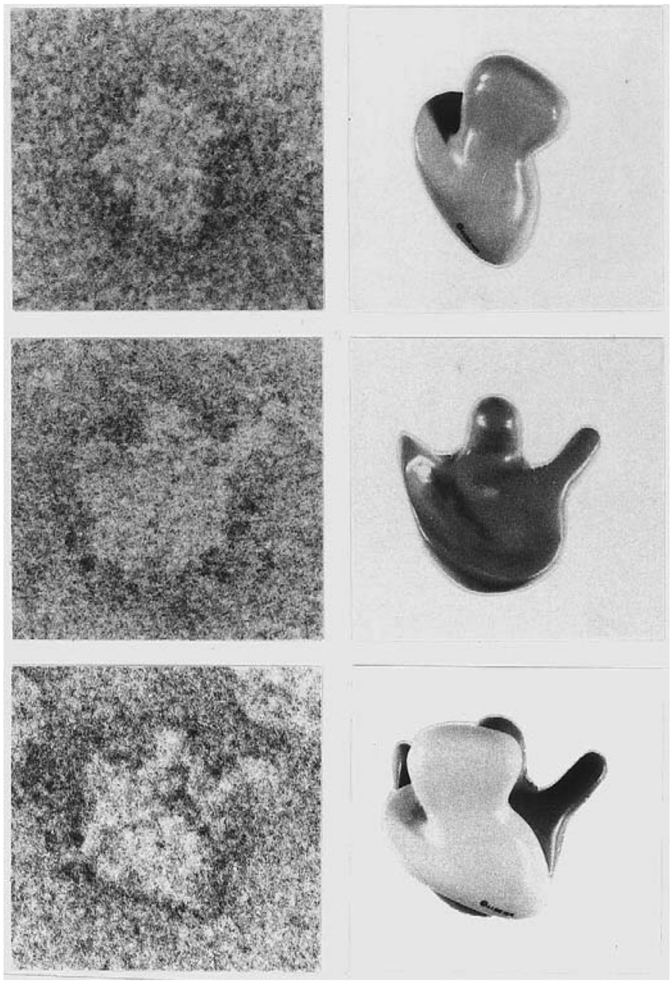
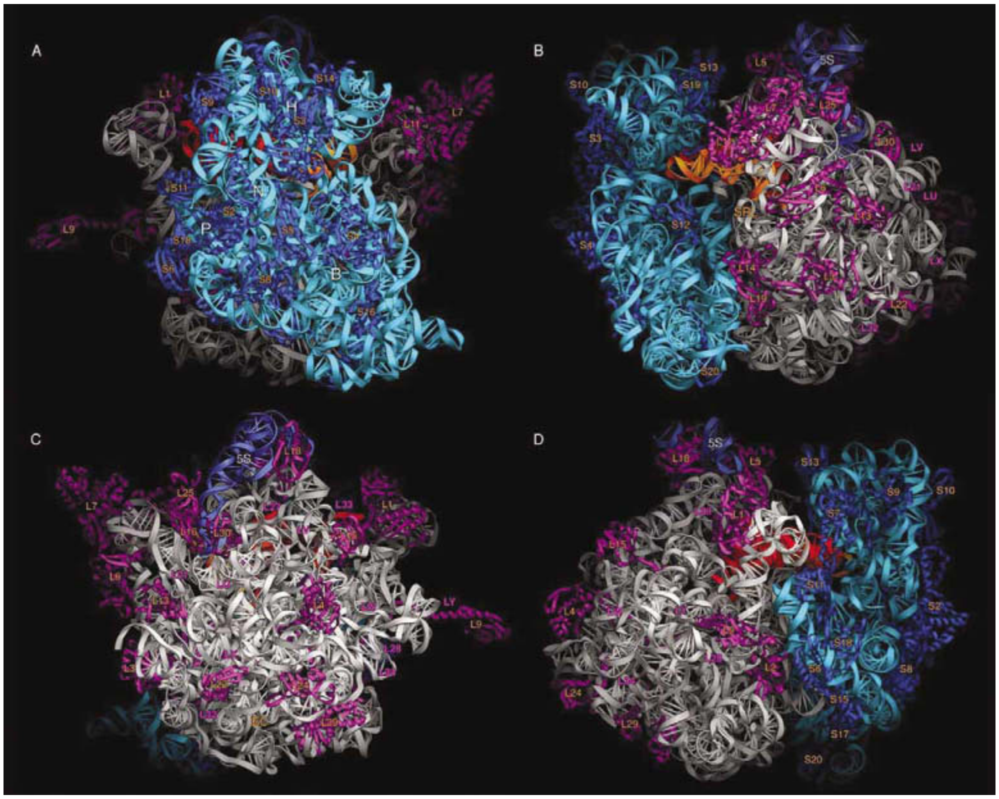


FIGURE 8.09 ***3-D Structure of a Ribosome by EM***

This structure was deduced from negatively stained electron microscope images of a bacterial 70S ribosome.

FIGURE 8.10 ***3-D Structure of a Ribosome by X-Ray***

Views of the structure of the *Thermus thermophilus* 70S ribosome. A. B. C and D are successive 90° rotations about the vertical axis. (A) view from the back of the 30*S* subunit. H, head; P, platform; N, neck; B, body. (B) view from the right-hand side, showing the subunit interface cavity, with the 30*S* subunit on the left and the 50*S* on the right. The anticodon arm of the A-tRNA (gold) is visible in the interface cavity. (C) View from the back of the 50*S* subunit. EC, the end of the polypeptide exit channel. (D) View from the left-hand side, with the 50*S* subunit on the left and the 30*S* on the right. The anticodon arm of the E-tRNA (red) is partly visible. The different molecular components are colored for identification: cyan, 16*S* rRNA; grey, 23*S* rRNA; light blue, 5*S* rRNA; dark blue, 30*S* proteins; magenta, 50*S* proteins. From Yusupov et al., Crystal Structure of the Ribosome at 5.5 Å Resolution. Science 292 (2001) 883-96.



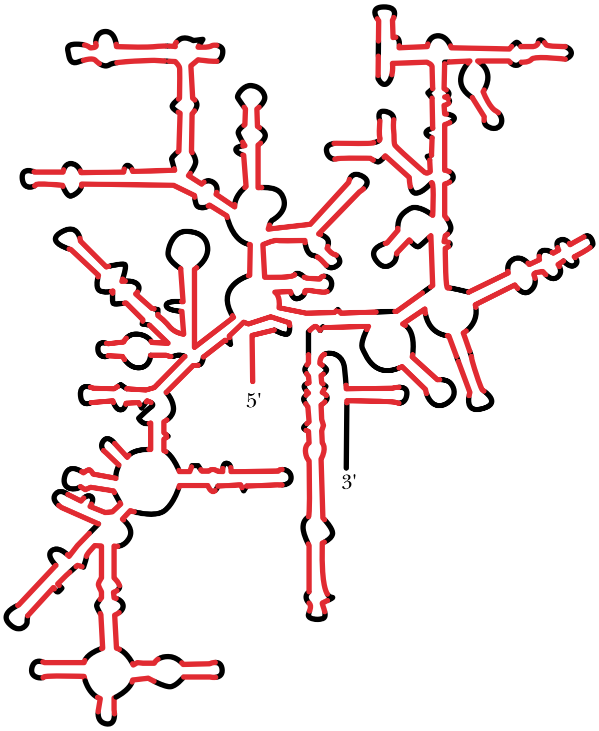


FIGURE 8.11 ***Secondary Structure of a Ribosomal RNA***

The 16S rRNA from the small ribosomal subunit of *E. coli* is complex with extensive secondary structure, forming loops and stems. Red indicates regions of base-pairing.

The peptide bond linking amino acids in the growing protein is made by the largest ribosomal RNA, which acts as a ribozyme.

These rRNA molecules have highly defined secondary structures with many stems and loops (Fig. 8.11). Although it was originally believed to have a largely structural role, recent work indicates that the rRNA is responsible for most of the critical reactions of protein synthesis. In particular, the 23S rRNA of the large subunit is a **ribozyme** that catalyzes the synthesis of the peptide bonds between the amino acids; i.e., it is the **peptidyl transferase**. Indeed, X-ray crystallography of the 50S subunit has shown that no ribosomal proteins are close enough to the catalytic center to take part in the reaction. Alteration by mutation of the catalytic residues in typical ribozymes either abolishes activity completely or reduces it by many-fold. However, the peptidyl-transferase center of 23S rRNA behaves in an atypical manner. Alteration of A2451 or G2447 (*E. coli* numbering) did not greatly reduce catalytic activity, although these residues are present in the catalytic center. These results suggest that the ribosome does not operate via direct chemical catalysis. Rather, the ribosome acts by correctly positioning the two substrates. The activated aminoacyl-tRNA then reacts spontaneously with the end of the growing polypeptide chain.

Three Possible Reading Frames Exist

Since the genetic code is read in groups of three bases, any nucleic acid sequence contains three possible reading frames.

Before mRNA is translated into protein, the issue of **reading frames** must be dealt with. The bases of mRNA are read off in groups of three, with each codon corresponding to one amino acid. How are the base sequences divided into codons? For any given nucleotide sequence there are three alternatives, depending on what is considered the start. Consider the following sequence:

GAAAUGUAUGCAUGCCAAAGGAGGCAUCUAAGG

peptidyl transferase Enzyme activity on the ribosome that makes peptide bonds; actually 23S rRNA (bacterial) or 28S rRNA (eukaryotic)
reading frame One of three alternative ways of dividing up a sequence of bases in DNA or RNA into codons
ribozyme RNA molecule that acts as an enzyme

If we start at base #1 we get the following codons:

GAA | AUG | UAU | GCA | UGC | CAA | AGG | AGG | CAU | CUA | AGG

If translated this would give the following amino acid sequence:

Glu | Met | Tyr | Ala | Cys | Gln | Arg | Arg | His | Leu | Arg

If we start at base #2 we get the following codons:

G | AAA | UGU | AUG | CAU | GCC | AAA | GGA | GGC | AUC | UAA | GG

If translated this would give the following amino acid sequence:

– | Lys | Cys | Met | His | Ala | Lys | Gly | Gly | Ile | Stop | –

And if we start at base #3 we get the following codons:

GA | AAU | GUA | UGC | AUG | CCA | AAG | GAG | GCA | UCU | AAG | G

If translated this would give the following amino acid sequence:

– | Asn | Val | Cys | Met | Pro | Lys | Glu | Ala | Ser | Lys | –

Each set of codons gives a translation completely out of step with each of the others. These three possibilities are known as reading frames. As there are three bases in a codon, there are only three possible reading frames. Changing the reading frame by three (or a multiple of three) provides the same sequence as the first example above.

Any sequence of DNA or RNA, beginning with a start codon, and which can, at least theoretically, be translated into a protein, is known as an **open reading frame**, often abbreviated to (and pronounced!) **ORF**. Since ORFs are derived by examining nucleic acid sequences, deciding whether an ORF is a genuine protein coding sequence requires further information. Any messenger RNA will possess several possible ORFs. The correct one is what matters. Note that the message on an mRNA molecule does not start exactly at the 5′-end. Between the 5′-end and the coding sequence is a short region that is not translated—the **5′-untranslated region** or **5′-UTR**. Hence, the reading frame cannot be defined simply by starting at the front end of the mRNA.

Between the very front of the mRNA and the coding sequence is a short non-translated region.

One way to define the reading frame is by choosing the **start codon.** The first codon is always AUG, encoding for methionine. This will define both the start of translation and the reading frame. In the example considered above, there are three possible start codons (underlined), each of which starts at a slightly different point and gives a different reading frame:

The start codon begins the coding sequence and is read by a special tRNA that carries methionine.

GAAAUGUAUGCAUGCCAAAGGAGGCAUCUAAGGA

5′-untranslated region (5′-UTR) Short sequence at the 5′-end of mRNA that is not translated into protein
open reading frame (ORF) Sequence of mRNA or corresponding region of DNA, that can be translated to give a protein
ORF See open reading frame
start codon The special AUG codon that signals the start of a protein

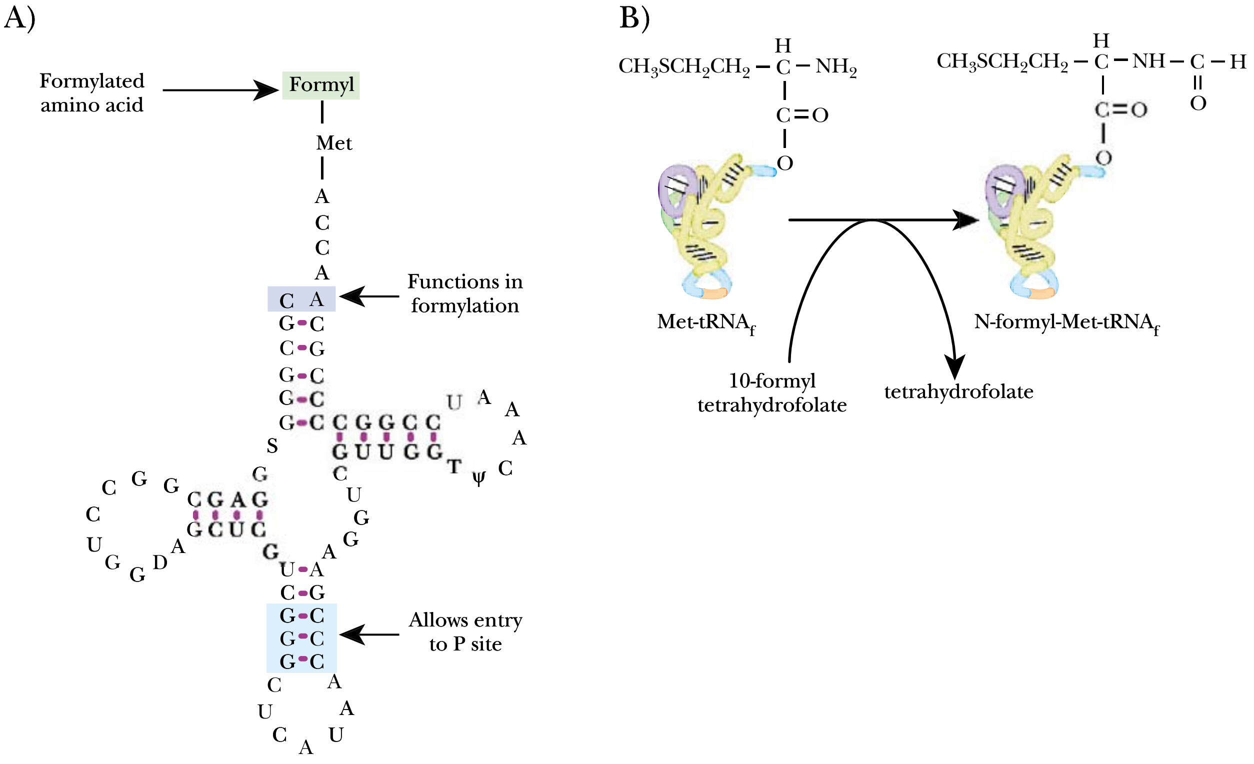


FIGURE 8.12 ***Initiator tRNA Carries N-Formyl-Methionine***

A) The structure of the initiator tRNA, fMet-tRNA, is unique. A CA base pair at the top of the acceptor stem is needed to allow formylation (violet). The initiator tRNA must enter the P-site directly (see below), which requires the three GC base pairs in the anticodon stem (blue). B) The initiator tRNA is first charged with unmodified methionine. Then a formyl group carried by the tetrahydrofolate cofactor is added to the methionine.

The Start Codon Is Chosen

A special tRNA, the **initiator tRNA**, is charged with methionine and binds to the AUG start codon (Fig. 8.12). In prokaryotes, chemically tagged methionine, **N-formyl-methionine (fMet)** is attached to the initiator tRNA whereas in eukaryotes unmodified methionine is used. Consequently all polypeptide chains begin with methionine, at least when first synthesized. Sometimes the initial methionine (in eukaryotes), or N-formyl-methionine (in prokaryotes), is snipped off later, so mature proteins do not always begin with methionine. In bacteria, even when the fMet is not removed as a whole, the N-terminal formyl group is often removed leaving unmodified methionine at the N-terminus of the polypeptide chain.

AUG codons also occur in the middle of messages and result in the incorporation of methionines in the middle of proteins. So how does the ribosome know which AUG codon to start with? Near the front (the 5′-end) of the mRNA of prokaryotes is a special sequence, the **ribosome binding site (RBS)**, often called the **Shine-Dalgarno** or **S-D sequence**, after its two discoverers (Fig. 8.13). The sequence complementary to this, the **anti-Shine-Dalgarno sequence**, is found close to the 3′-end of the 16S riboso-

To choose the correct start codon, the messenger RNA binds to 16S ribosomal RNA at a specific sequence.

anti-Shine-Dalgarno sequence Sequence on 16S rRNA that is complementary to the Shine-Dalgarno sequence of mRNA
initiator tRNA The tRNA that brings the first amino acid to the ribosome when starting a new polypeptide chain
N-formyl-methionine or fMet Modified methionine used as the first amino acid during protein synthesis in bacteria
ribosome binding site (RBS) Same as Shine-Dalgarno sequence; sequence close to the front of mRNA that is recognized by the ribosome; only found in prokaryotic cells
Shine-Dalgarno (S-D) sequence Same as RBS; sequence close to the front of mRNA that is recognized by the ribosome; only found in prokaryotic cells

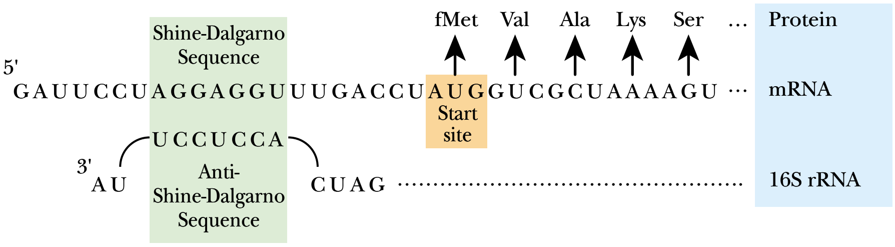


FIGURE 8.13 ***Shine-Dalgarno Sequence of mRNA Binds to 16S rRNA***

The Shine-Dalgarno sequence on the mRNA is recognized by base pairing with the anti-Shine-Dalgarno sequence on the16S rRNA. The first AUG downstream of the S-D/anti-S-D site serves as the start codon.

mal RNA. Consequently, the mRNA and the 16S rRNA bind together by base pairing between these two sequences. The start codon is the next AUG codon after the ribosome binding site. Typically there are about seven bases between the S-D sequence and the start codon. In some cases, the S-D sequence exactly matches the anti-S-D sequence and the mRNAs are translated efficiently. In other cases, the match is poorer and translation is less efficient. [Note that eukaryotes do not use an S-D sequence to locate the start of translation; instead they scan the mRNA starting form the 5′-cap—see below.]

Occasionally, coding sequences even start with GUG (normally encoding valine) instead of AUG. This leads to inefficient initiation and is mostly found for proteins required only in very low amounts, such as regulatory proteins, for example, LacI, the repressor of the *lac* operon (see Ch. 9). Note that when GUG acts as the start codon, the same initiator fMet-tRNA is used as when AUG is the start codon. Consequently, formyl-Met is the first amino acid, even for proteins that start with a GUG codon. This is apparently due to the involvement of the initiation factors, especially IF3—see below.

The Initiation Complexes Must Be Assembled

Before protein synthesis starts, the two subunits of the ribosome are floating around separately. Because the 16S rRNA, with the anti-Shine-Dalgarno sequence, is in the small subunit of the ribosome, the messenger RNA binds to a free small subunit. Next the initiator tRNA, carrying fMet, recognizes the AUG start codon. Assembly of this **30S initiation complex** needs three proteins (IF1, IF2 and IF3), known as **initiation factors**, which help arrange all the components correctly.

Proteins known as initiation factors help the ribosomal subunits, mRNA and tRNA assemble correctly.

IF2 recognizes fMet-tRNA. IF3 is also involved in recognition of the start codon and the matching anticodon end of the initiator tRNA. IF3 prevents the 50S subunit from binding prematurely to the small subunit before the correct initiator tRNA is present. Once the 30S initiation complex has been assembled, IF3 departs and the 50S subunit binds. IF1 and IF2 are now released, resulting in the **70S initiation complex**. This process consumes energy in the form of GTP, which is split by IF2 (Fig. 8.14).

The tRNA Occupies Three Sites During Elongation of the Polypeptide

After the large subunit of the ribosome has arrived, the polypeptide can be made. Amino acids are linked together by the peptidyl transferase reaction, which is catalyzed by the 23S rRNA of the large subunit. The amino acids are carried to the ribosome attached to transfer RNA. The ribosome has three sites for tRNA: the **A (acceptor)**

30S initiation complex Initiation complex for translation that contains only the small subunit of the bacterial ribosome
70S initiation complex Initiation complex for translation that contains both subunits of the bacterial ribosome
A (acceptor) site Binding site on the ribosome for the tRNA that brings in the next amino acid
initiation factors Proteins that are required for the initiation of a new polypeptide chain

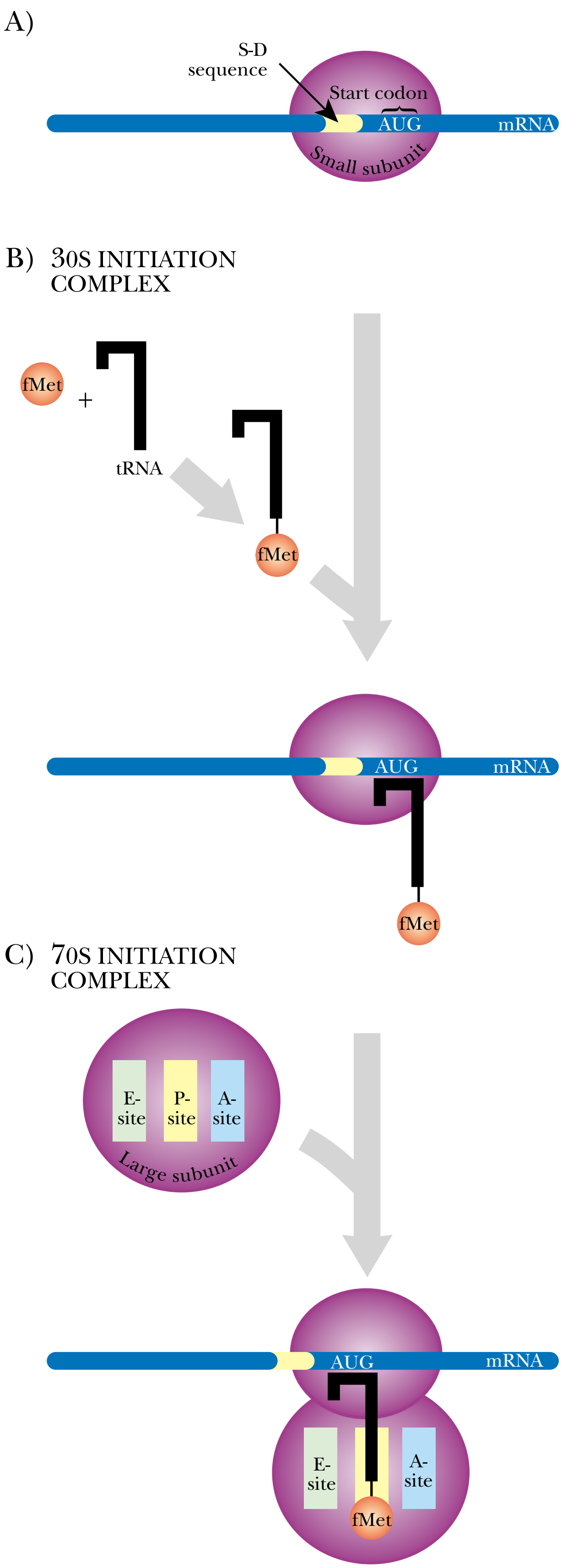


FIGURE 8.14 ***Formation of 30S and 70S Initiation Complexes***

A) The small subunit and the mRNA bind to each other at the Shine-Dalgarno sequence. The start codon, AUG, is just downstream of this site. B) The initiator tRNA becomes tagged with fMet and binds to the AUG codon on the mRNA. C) The large ribosomal subunit joins the small subunit and accommodates the tRNA at the P-site.

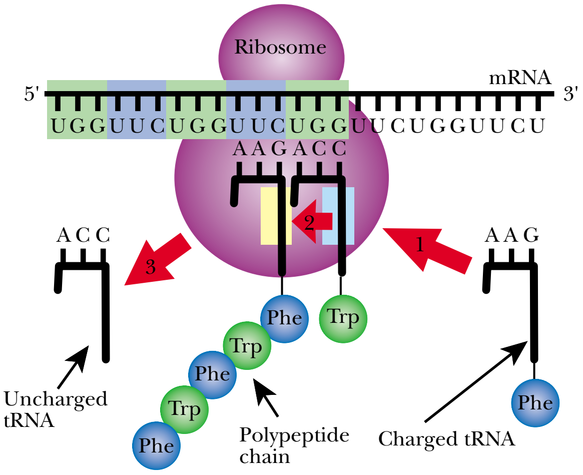


FIGURE 8.15 ***Overview of the Elongation Cycle on the Ribosome***

After the first Met has been added, the incoming charged tRNA first occupies the A-site. The peptide bond is formed between the amino acid at the A-site and the growing polypeptide chain in the P-site. The uncharged tRNA exits the ribosome.

Only two tRNA molecules can occupy the ribosome at any instant.

site, the **P (peptide) site** and the **E (exit) site**. However, only two charged tRNA molecules can be accommodated on the ribosome at any given instant (Fig. 8.15).

After peptide bond formation the tRNA carrying the growing polypeptide chain moves sideways between sites on the ribosome.

The fMet initiator tRNA starts out in the P-site. Another tRNA, carrying the next amino acid, arrives and enters the A-site. The fMet is cut loose from its tRNA and bonded to amino acid # 2, instead. So tRNA #2 now carries two linked amino acids, the beginnings of a growing polypeptide chain. (The enzyme activity that joins two amino acids together is referred to as the peptidyl transferase activity, as the growing peptide chain is transferred from the tRNA carrying it at each step.) After peptide bond formation, the two tRNAs are tilted relative to the A- and P-sites (Fig. 8.16). The tRNA carrying the growing polypeptide chain now occupies part of the A-site on the 30S subunit but part of the P-site on the 50S subunit. This is probably due to movement of the 50S subunit relative to the 30S subunit. The next step is **translocation**, in which the mRNA moves one codon sideways relative to the ribosome (Fig. 8.16). This moves the two tRNAs into the P- and E-sites, leaving the A-site empty.

The empty tRNA leaves the ribosome and a new, charged tRNA enters.

When the next charged tRNA arrives, carrying the third amino acid, it enters the vacant A-site. This triggers release of the tRNA from the E-site. The A- and E-sites cannot be simultaneously occupied. As the peptide chain continues to grow, it is constantly cut off from the tRNA holding it and joined instead to the newest amino acid to be brought by its tRNA into the A-site, hence the name "acceptor" site.

Elongation requires two **elongation factors**, both of which use energy in the form of GTP. EF-T actually consists of a pair of proteins, EF-Tu and EF-Ts. Incoming charged tRNA is delivered to the ribosome and installed into the A-site by elongation factor EF-Tu. This requires energy from the hydrolysis of GTP. EF-Ts is responsible for exchanging the GDP left bound to EF-Tu for a fresh GTP. The second elongation factor, EF-G, oversees the translocation step (Fig. 8.16).

Termination of Protein Synthesis Requires Release Factors

The stop codon is read by a protein, the release factor, not by a tRNA.

Eventually the ribosome reaches the end of the message. This is marked by one of three possible stop codons, UGA, UAG, and UAA. As no tRNA exists to read these three codons, the polypeptide chain can no longer grow. Instead, proteins known

E (exit) site Site on the ribosome that a tRNA occupies just before leaving the ribosome
elongation factors Proteins that are required for the elongation of a growing polypeptide chain
P (peptide) site Binding site on the ribosome for the tRNA that is holding the growing polypeptide chain
translocation a) Transport of a newly made protein across a membrane by means of a translocase; b) Sideways movement of the ribosome on mRNA during translation and c) Removal of a segment of DNA from a chromosome and its reinsertion in a different place

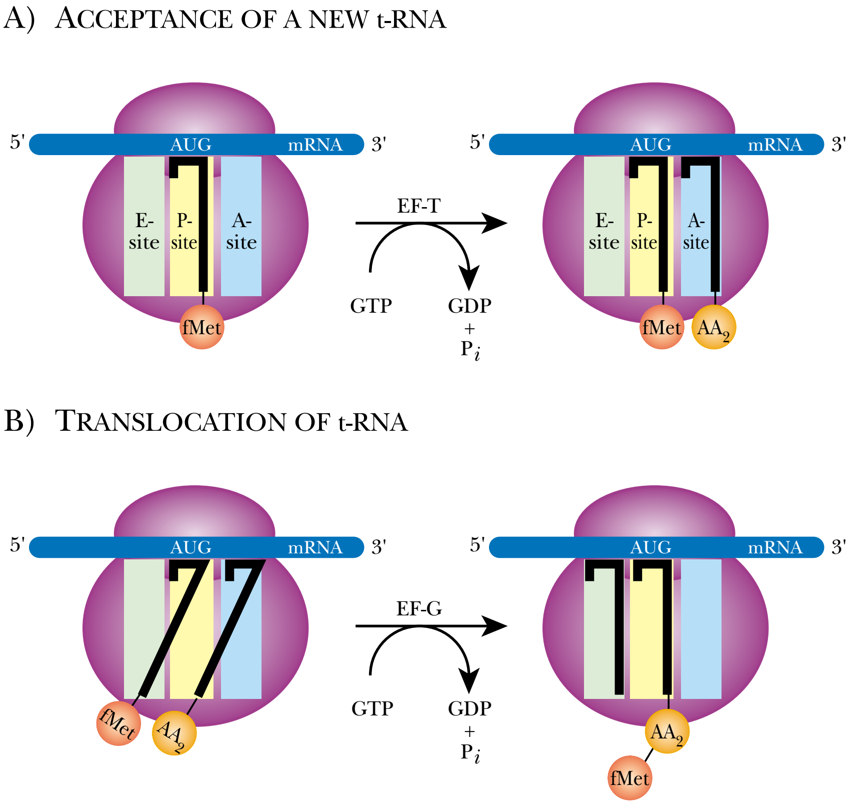


FIGURE 8.16 ***Elongation Factors and Site Occupation***

A) The EF-T factor is important in allowing a new charged tRNA to occupy the A-site. B) The EF-G factor is important in translocation of the tRNAs from the A- and P-sites, to the P- and E-sites, respectively. Note that during translocation the transfer RNA temporarily binds "diagonally" across two sites.

as **release factors (RF)** read the stop signals (Fig. 8.17). RF1 recognizes UAA or UAG and RF2 recognizes UAA or UGA. The completed polypeptide chain is now released from the last tRNA. This is actually done by the peptidyl transferase. Binding of the release factor activates the peptidyl transferase which hydrolyzes the bond between the finished polypeptide chain and the tRNA in the P-site. The polypeptide chain, the tRNA and the mRNA now leave the ribosome, which dissociates into separate subunits. Two further factors aid in dissociation: RF3 releases RF1 or RF2 from the ribosome and **ribosome recycling factor (RRF)** dissociates the large and small subunits.

Several Ribosomes Usually Read the Same Message at Once

Messenger RNA is long enough for several ribosomes to translate it simultaneously.

Once the first ribosome has begun to move, another can associate with the same messenger RNA and travel along behind. In practice, several ribosomes will move along the same mRNA about a hundred bases apart (Fig. 8.18). An mRNA with several attached ribosomes is called a **polysome** (short for polyribosome).

Electron microscope observations have suggested that the polysomes of eukaryotic cells are circular (Fig. 8.19). Apparently, the 3′-end of the mRNA is attached to the 5′-end by protein-protein contact between the poly(A) binding protein (attached to the 3′-poly(A) tail) and the eukaryotic initiation factor, IF4 (attached to the cap at the 5′-end). In prokaryotic cells, such circularization cannot occur as the 3′-end of the mRNA is still being elongated by RNA polymerase while ribosomes have begun translating from the 5′-end.

polysome Group of ribosomes bound to and translating the same mRNA
release factor Protein that recognizes a stop codon and brings about the release of a finished polypeptide chain from the ribosome
ribosome recycling factor (RRF) Protein that dissociates the ribosomal subunits after a polypeptide chain has been finished and released

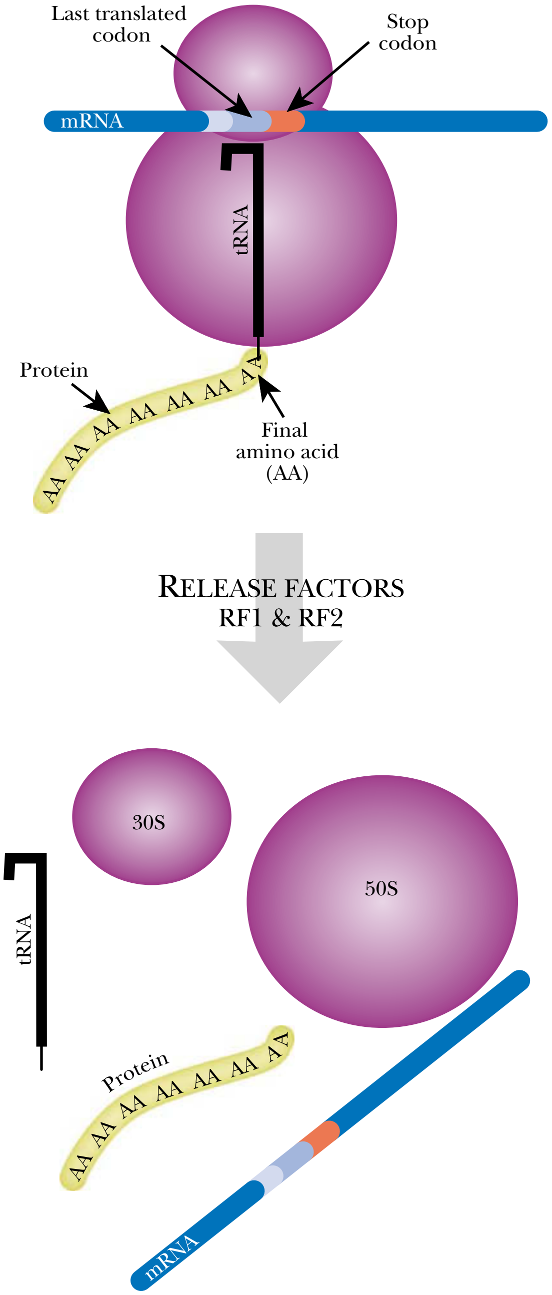


FIGURE 8.17 ***Termination and Release of Finished Polypeptide***

After the ribosome has added the final amino acid, release factors, RF1 and RF2, recognize the stop codon and cause the ribosome complex to dissociate.

Bacterial Messenger RNA Can Code for Several Proteins

Messenger RNA in bacteria often carries several coding sequences.

In bacteria, several genes may be transcribed to give a single messenger RNA. The term **operon** refers to clusters of genes that are co-transcribed. The result is that several proteins may be encoded by the same mRNA. As long as each open reading frame has its own Shine-Dalgarno sequence in front of it, the ribosome will bind and start translating. Open reading frames that are translated into proteins are sometimes known as cistrons; consequently, mRNA which carries several of these is called **polycistronic mRNA** (Fig. 8.20).

operon A cluster of prokaryotic genes that are transcribed together to give a single mRNA (i.e. polycistronic mRNA)
polycistronic mRNA mRNA carrying multiple coding sequences that may be translated to give several different protein molecules; only found in prokaryotic (bacterial) cells

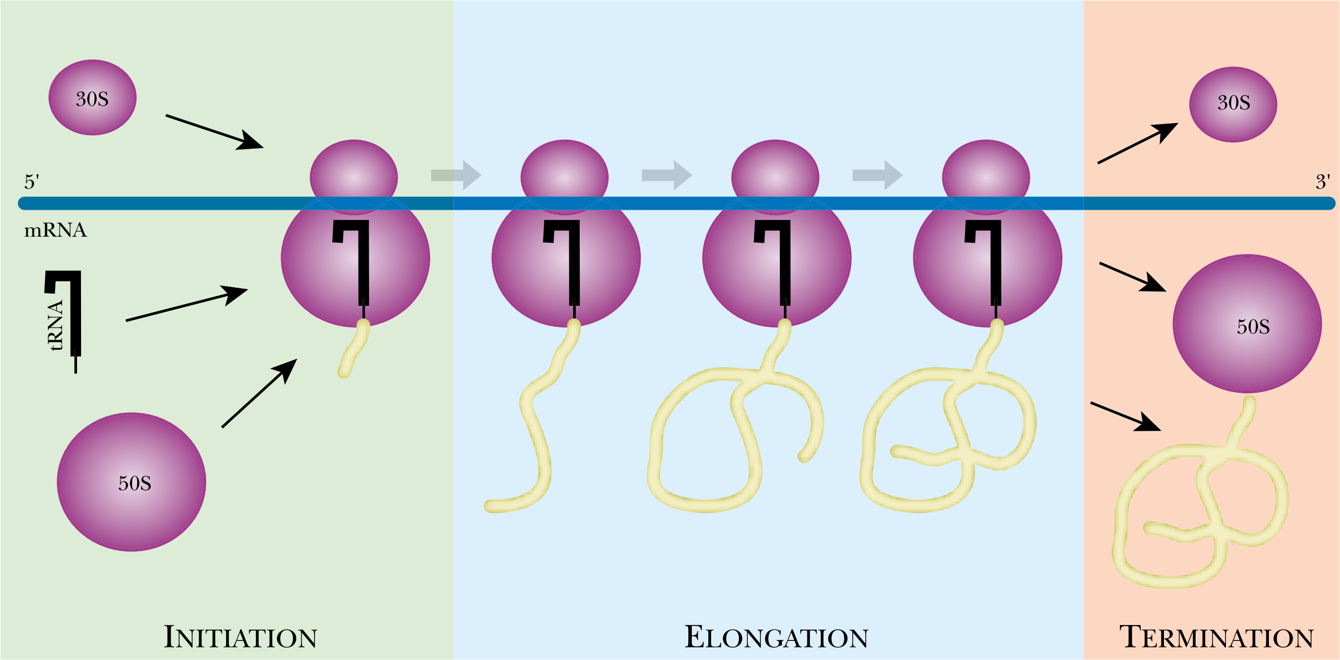


FIGURE 8.18 ***Polysome***

A single mRNA molecule is associated with several ribosomes. At initiation the two subunits assemble; during chain elongation several ribosomes are at different stages in reading the same mRNA message; at termination the ribosome complex disassembles.

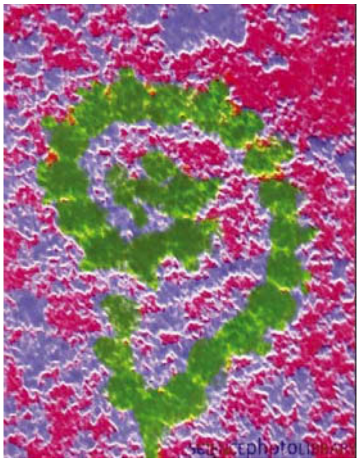


FIGURE 8.19 ***False Color TEM of Polysome***

This false-color transmission electron micrograph (TEM) shows a polysome from a human brain cell. Polysomes consist of several individual ribosomes connected by slender strands of messenger RNA. Magnification: ×240,000. CNRI / Science Photo Library.

Eukaryotic mRNA molecules each code for a single protein.

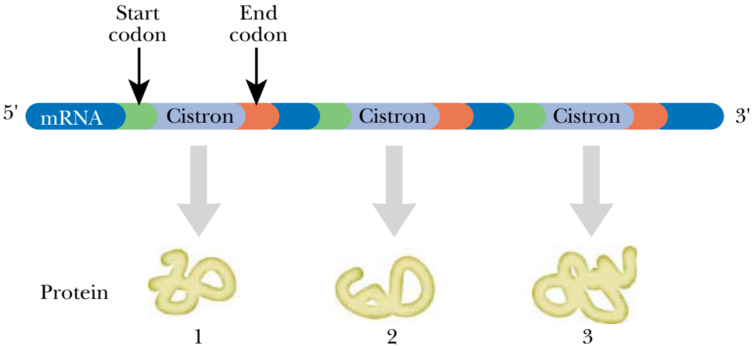


FIGURE 8.20 ***Polycistronic mRNA in Bacteria***

The mRNA contains several cistrons, each of which codes for a protein.

In higher organisms operons are absent and neighboring genes are not co-transcribed. Each individual gene is transcribed separately to give an individual molecule of RNA. Apart from a few exceptional cases, each molecule of eukaryotic mRNA only carries a single protein coding sequence. Furthermore, eukaryotic mRNA does not make use of the Shine-Dalgarno sequence. Instead, the front (5′-end) of the messenger RNA molecule is recognized by its cap structure. Consequently, in eukaryotes only the first open reading frame would normally be translated, even if multiple open reading frames were present (see below for further discussion).

Transcription and Translation Are Coupled in Bacteria

When mRNA is transcribed from the original DNA template, its synthesis starts at the 5′-end. The mRNA is also read by the ribosome starting at the 5′-end. In prokaryotic cells, the chromosome and ribosomes are all in the same single cellular compartment.

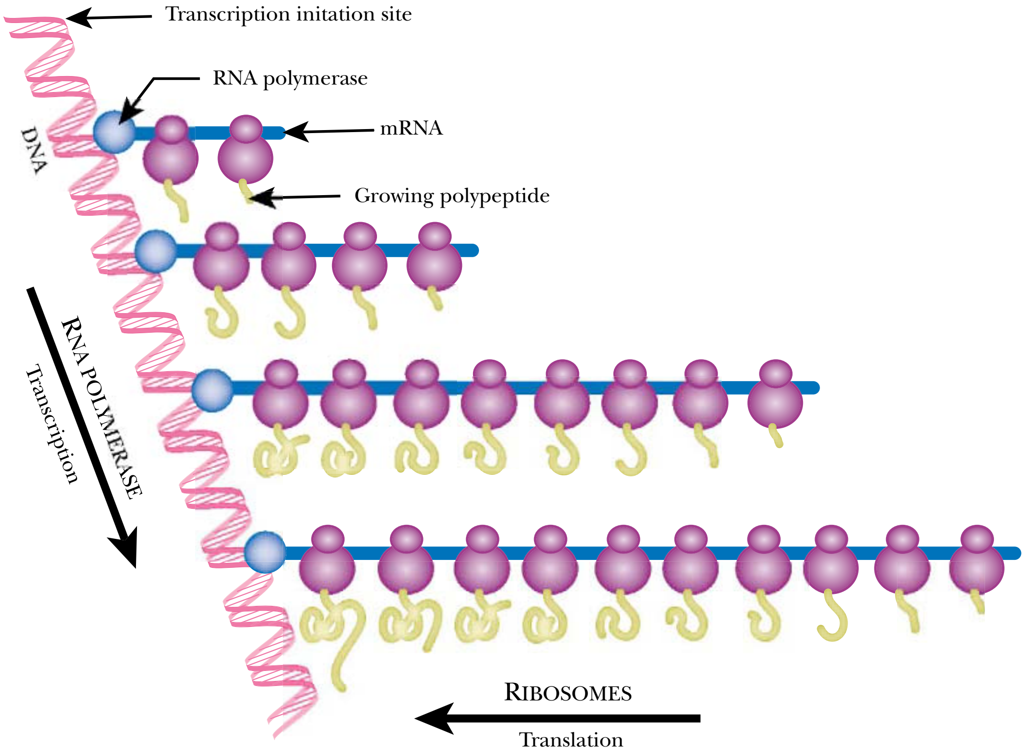


FIGURE 8.21 ***Coupled Transcription-Translation in Bacteria***

Even as the DNA is being transcribed to give mRNA, ribosomes sequentially attach to the growing mRNA and initiate protein synthesis.

In prokaryotes, the ribosomes can begin to translate a message before the RNA polymerase has finished transcribing it.

Therefore the ribosomes can start translating the message before the synthesis of the mRNA molecule has actually been finished. The result is that partly finished mRNA, still attached to the bacterial chromosome via RNA polymerase, may have several ribosomes already moving along it making polypeptide chains. This is known as **coupled transcription-translation** (Fig. 8.21). This is impossible in higher, eukaryotic cells, because the DNA is inside the nucleus and the ribosomes are outside, in the cytoplasm. [Some recent work has suggested that a small amount of translation may occur inside the nucleus of eukaryotes. This issue is presently unresolved. Nonetheless, the majority of eukaryotic translation occurs in the cytoplasm, outside the nucleus.]

Some Ribosomes Become Stalled and Are Rescued

Cellular metabolism is not perfect and cells must allow for errors. One problem ribosomes sometimes run across is defective mRNA that lacks a stop codon. Whether synthesis of the mRNA was never completely finished or whether it was mistakenly snipped short by a ribonuclease, problems ensue. In the normal course of events, a ribosome that is translating a message into protein will, sooner or later, come across a stop codon. Even if an mRNA molecule comes to an abrupt end, ribosomes may be released only by release factor and this in turn needs a stop codon. If the mRNA is defective and there is no stop codon, a ribosome that reaches the end could just sit there forever and the ribosomes behind it will all be stalled, too.

Ribosomes that have stalled due to defective mRNA can be rescued by a special RNA—tmRNA.

Bacterial cells contain a small RNA molecule that rescues stalled ribosomes. This is named **tmRNA** because it acts partly like transfer RNA and partly like messenger RNA. Like a tRNA, the tmRNA carries alanine, an amino acid. When it finds a stalled

coupled transcription-translation When ribosomes of bacteria start translating an mRNA molecule that is still being transcribed from the DNA
tmRNA Specialized RNA used to terminate protein synthesis when a ribosome is stalled by a damaged mRNA

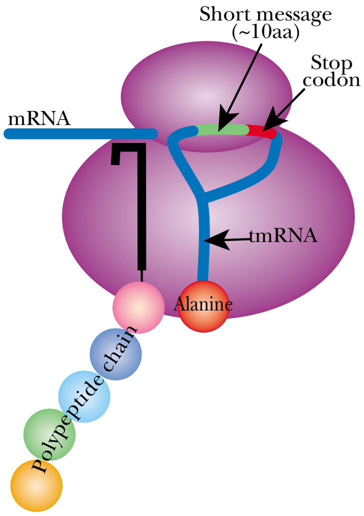


FIGURE 8.22 ***Stalled Ribosome Liberated by tmRNA***

Binding of a tmRNA carrying alanine allows the translation of a damaged message to continue. First alanine is added, then a short sequence of about 10 amino acids encoded by the tmRNA. Finally the stop codon of the tmRNA allows proper termination of the polypeptide chain.

ribosome, it binds beside the defective mRNA (Fig. 8.22). Protein synthesis now continues, first using the alanine carried by tmRNA, and then continuing on to translate the short stretch of message that is also part of the tmRNA. Finally, the tmRNA provides a proper stop codon so that release factor can disassemble the ribosome and free it for it for continued protein synthesis. The tRNA domain of tmRNA lacks an anticodon loop and a D-loop. A protein known as SmpB (not shown in Fig. 8.22 for clarity) binds to the tRNA domain and makes contacts to the ribosome that would normally be made by the missing D-loop.

Clearly, the protein that has just been made is defective and should be degraded. As might be supposed, the tmRNA has signaled that the protein that was made is defective. The short stretch of 11 amino acids specified by the message part of tmRNA and added to the end of the defective protein acts as a signal, known as the ssrA tag. [SsrA stands for small stable RNA A, a name used for tmRNA before its function was elucidated.] The ssrA tag is recognized by several proteases (originally referred to as "**tail specific protease**") which degrade all proteins carrying this signal. These include the Clp proteases and the HflB protease involved in the heat shock response (see Ch. 9). Eukaryotic cells lack tmRNA. Since they do encounter stalled ribosomes, they presumably have some presently unknown mechanism to deal with this situation.

Differences between Eukaryotic and Prokaryotic Protein Synthesis

Eukaryotic ribosomes are larger than those of prokaryotes.

The overall scheme of protein synthesis is similar in all living cells. However, there are significant differences between bacteria and eukaryotes. These are summarized in Table 8.02 and discussed in the following sections. Note that eukaryotic cells contain mitochondria and chloroplasts, which have their own DNA and their own ribosomes. The ribosomes of these organelles operate similarly to those of bacteria and will be considered separately below. In eukaryotic protein synthesis, it is usually the cytoplasmic ribosomes that translate nuclear genes. Several aspects of eukaryotic protein synthesis are more complex. The ribosomes of eukaryotic cells are larger and contain more rRNA and protein molecules than those of prokaryotes. In addition, eukaryotes have more initiation factors and a more complex initiation procedure.

A few aspects of protein synthesis are actually less complex in eukaryotes. In prokaryotes, mRNA is polycistronic and may carry several genes that are translated to give several proteins. In eukaryotes, each mRNA is monocistronic and carries only a single gene, which is translated into a single protein. In prokaryotes, the genome and the ribosomes are both in the cytoplasm, whereas in eukaryotes the genome is in the nucleus. Consequently, coupled transcription and translation is not possible for eukaryotes (except for their organelles; see below).

Both prokaryotes and eukaryotes have a special initiator tRNA that recognizes the start codon and inserts methionine as the first amino acid. In prokaryotes, this first methionine has a formyl group on its amino group (i.e., it is N-formyl-methionine) but in eukaryotes unmodified methionine is used.

Initiation of Protein Synthesis in Eukaryotes

Eukaryotic mRNA is recognized by its cap structure (not by base pairing to rRNA).

Initiation of protein synthesis differs significantly between prokaryotes and eukaryotes. Eukaryotic mRNA has no ribosome binding site (RBS). Instead recognition and binding to the ribosome rely on a component that is lacking in prokaryotes. The cap structure, at the 5′-end, is added to eukaryotic mRNA before it leaves the nucleus (see Ch. 12). Cap binding protein (one of the subunits of eIF4) binds to the cap of the mRNA (Fig. 8.23). Eukaryotes also have more initiation factors than prokaryotes and

tail specific protease Enzyme that destroys mis-made proteins by degrading them tail first, i.e., from the carboxyl end

TABLE 8.02 Comparison of Protein Synthesis

Prokaryotes	Eukaryotes (cytoplasm)
Polycistronic mRNA	Monocistronic mRNA
Coupled transcription and translation	No coupled transcription and translation for nuclear genes
Linear polyribosomes	Circular polyribosomes
No cap on mRNA	5′-End of mRNA is recognized by cap
Start codon is next AUG after ribosome binding site	No ribosome binding site so first AUG in mRNA is used
First amino acid is formyl-Met	First Met is unmodified
70S ribosomes made of:	80S ribosomes made of:
30S and 50S subunits	40S and 60S subunits
Small 30S subunit: 16S rRNA 21 proteins	Small 40S subunit: 18S rRNA 33 proteins
Large 50S subunit: 23S and 5S rRNA 31 proteins	Large 60S subunit: 28S, 5.8S and 5S rRNA 49 proteins
Elongation factors EF-T (2 subunits) and EF-G	Elongation factors eEF1 (3 subunits) and eEF2
Three initiation factors IF1, IF2 and IF3	Multiple initiation factors eIF2 (3 subunits), eIF3, eIF4 (4 subunits), eIF5
Shut-off by dimerization of ribosomes in non-growing cells	Control via IF sequestration

Internal Ribosome Entry Sites

Although most eukaryotic mRNA is scanned by the 40S subunit to find the first AUG, exceptions do occur. Sequences known as internal ribosome entry sites (IRES) are found in a few mRNA molecules. As the name indicates, these allow ribosomes to initiate translation internally, rather than at the 5′-end of the mRNA. IRES sequences were first found in certain viruses that have polycistronic mRNA despite infecting eukaryotic cells. In this case, the presence of IRES sequences in front of each coding sequence allows a single mRNA to be translated to give multiple proteins. The best known examples are members of the Picornavirus family, which includes poliovirus (causative agent of polio) and rhinovirus (one of the agents of common cold).

More recently, it has been found that a few special mRNA molecules encoded by eukaryotic cells themselves also possess IRES sequences. During major stress situations, such as heat shock or energy deficit, synthesis of the majority of proteins is greatly decreased. Much of this regulation occurs at the initiation stage of translation (see below). However, a few proteins are exempted from this down-regulation as they are needed under stress conditions. The mRNAs encoding these proteins often contain an IRES sequence. In these cases the mRNA carries only a single coding sequence and the IRES is located in the 5′-UTR, between the 5′-end of the mRNA and the start of the coding sequence. This allows translation to be initiated at the IRES even in the absence of the standard initiation/scanning procedure.

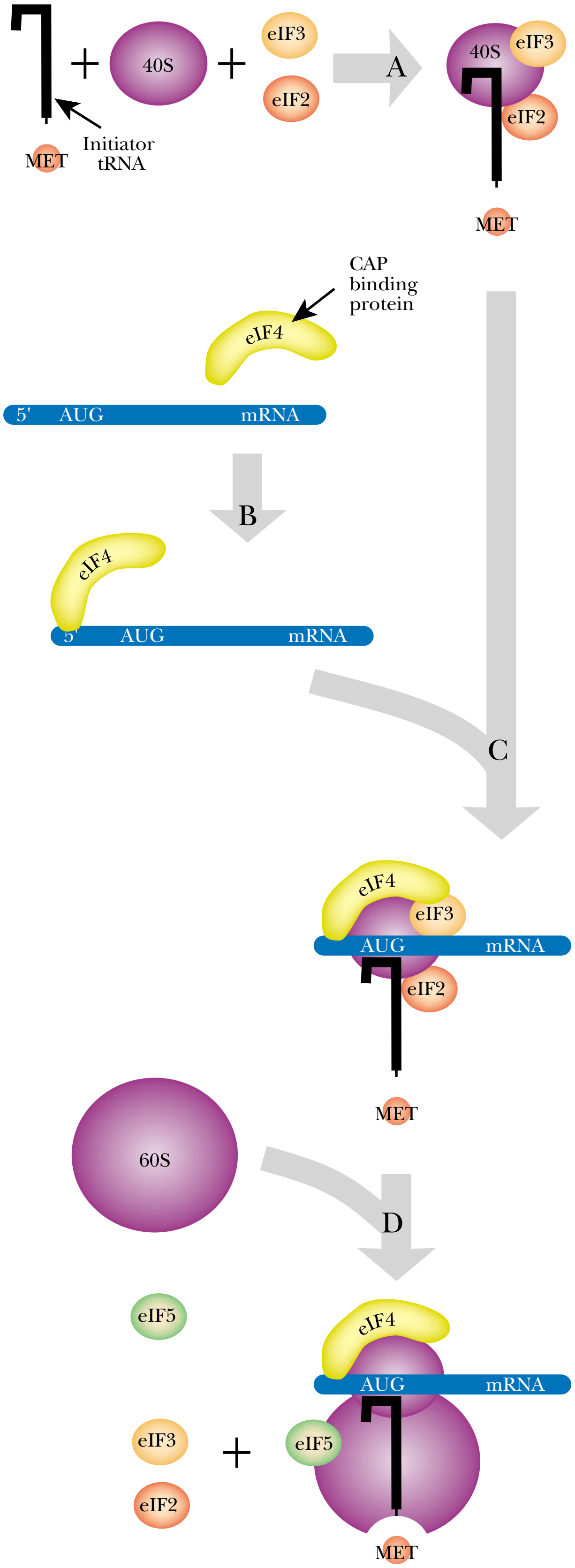


FIGURE 8.23 ***Assembly of the Eukaryotic Initiation Complex***

A) Assembly of the small subunit plus initiator Met-tRNA involves the binding of factors eIF3 and eIF2. B) The cap binding protein of eIF4 attaches to the mRNA before it joins the small subunit. C) The mRNA binds to the small subunit via cap binding protein and the 40S initiation complex is assembled. D) Assembly of the large subunit requires factor eIF5. After assembly, eIF2 and eIF3 depart.

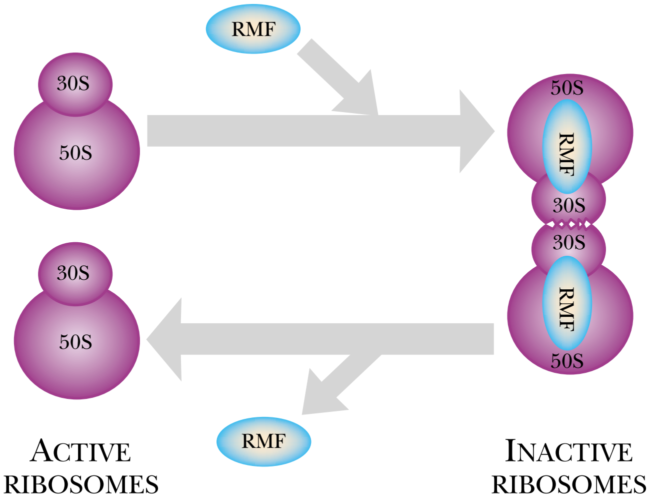


FIGURE 8.24 ***Bacterial Ribosomes on Standby During Bad Conditions***

Active bacterial ribosomes can become inactive when the RMF protein binds to them. The ribosomes form dimers with the 30S subunits attached to one another. When conditions are favorable, dissociation occurs.

the order of assembly of the initiation complex is different. Factor eIF2 binds to the initiator Met-tRNA, factor eIF3 binds to the small (40S) subunit of the ribosome and factor eIF4 binds to the mRNA (via Cap binding protein). These components then assemble to form the initiation complex (Fig. 8.23).

The 40S subunit then moves along the mRNA, starting from the 5′-end, until it finds a start codon. This process is referred to as scanning. Normally the first AUG to be found is used as the start codon, although the sequence surrounding the AUG is important. The consensus is GCCRCC<u>AUG</u>G (R = A or G). If its surrounding sequence is too far from consensus an AUG may be skipped. Once a suitable AUG has been located, eIF5 is needed to allow the 60S subunit to join and for eIF2 and eIF3 to depart.

Protein Synthesis Is Halted When Resources Are Scarce

Proteins make up about two-thirds of the organic matter in a cell and their synthesis consumes a major part of the cell's energy and raw materials. Clearly, when cells run low on nutrients or energy they cannot continue to synthesize proteins at the normal rate. In bacteria, ribosomes are taken out of service during stationary phase or periods of slow growth. A small basic protein, **ribosome modulation factor (RMF)** binds to ribosomes and inactivates them (Fig. 8.24). The inactive ribosomes exist as dimers. When favorable conditions return, the inactive dimers are disassembled and the ribosomes are reactivated.

Higher organisms also stop protein synthesis when nutrients or energy run low. However, they do so by inactivating the initiation factors rather than the ribosomes (Fig. 8.25). Initiation factor eIF2 uses energy by hydrolyzing GTP to GDP. After initiation is over, it is released from the ribosome with the GDP still bound. It then binds to eIF2B, which exchanges GDP for GTP, so recycling the eIF2. In times of stress, a kinase phosphorylates eIF2 and prevents the removal of GDP. The GDP bound form of eIF2 cannot initiate translation and protein synthesis is halted.

A Signal Sequence Marks a Protein for Export from the Cell

Exported proteins have a signal sequence at the front.

Once a protein has been made, it must find its correct location. Although cytoplasmic proteins are made in the cell compartment where they belong, other proteins, which do not reside in the cytoplasm, must be transported. Proteins destined to be exported

ribosome modulation factor (RMF) Protein that inactivates surplus ribosomes during slow growth or stationary phase in bacteria

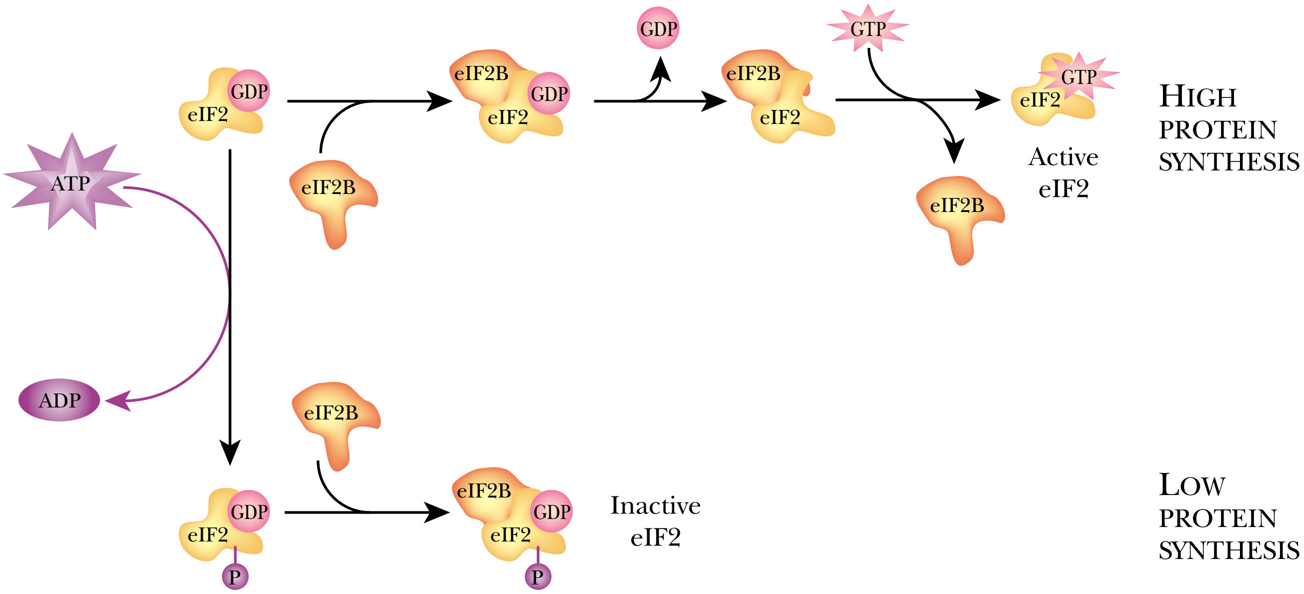


FIGURE 8.25 ***Recycling of Initiation Factor eIF2 is Controlled***

When eukaryotes down-regulate the level of protein synthesis, a protein kinase phosphorylates eIF2/GDP. This prevents eIF2B from removing the GDP and eIF2/GDP stays locked in an inactive complex with eIF2B. Absence of active eIF2 decreases the rate of initiation

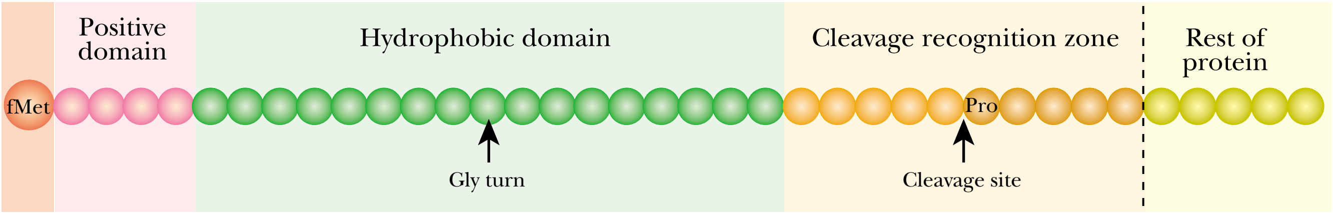


FIGURE 8.26 ***Standard Signal Sequence for Exported Proteins***

The signal sequence contains a positive charged domain (containing lysine and/or arginine), an α-helical hydrophobic domain (rich in alanine, leucine and valine) and a cleavage site preceded by a glycine or serine and followed by a proline. A reverse turn due to glycine is found approximately half way through the hydrophobic domain.

to the exterior of the cell must be exported through the cell membrane. Similar systems exist in bacterial and eukaryotic cells. Proteins destined for export are tagged at the N-terminus with a **signal sequence**. This is cut off after export, by proteases attached to the outside of the membrane, and is therefore not present in the mature protein. The signal sequence consists of approximately 20 amino acids that form an α-helix. There is little specific sequence homology between signal sequences from different exported proteins. A positively charged, basic N-terminus of two to eight amino acids is followed by a long stretch of hydrophobic amino acids. The amino acid just before the cleavage site has a short side chain (Fig. 8.26).

A polypeptide destined for export is recognized by its signal sequence. In bacteria, the signal recognition protein (SecA) binds the signal sequence and guides it to the **translocase** complex in the cell membrane. The rest of the protein being exported

After export of a protein across the cell membrane via the translocase, the signal sequence is cut off.

signal sequence Short, largely hydrophobic sequence of amino acids at the front of a protein that label it for export
translocase Enzyme complex that transports proteins across membranes

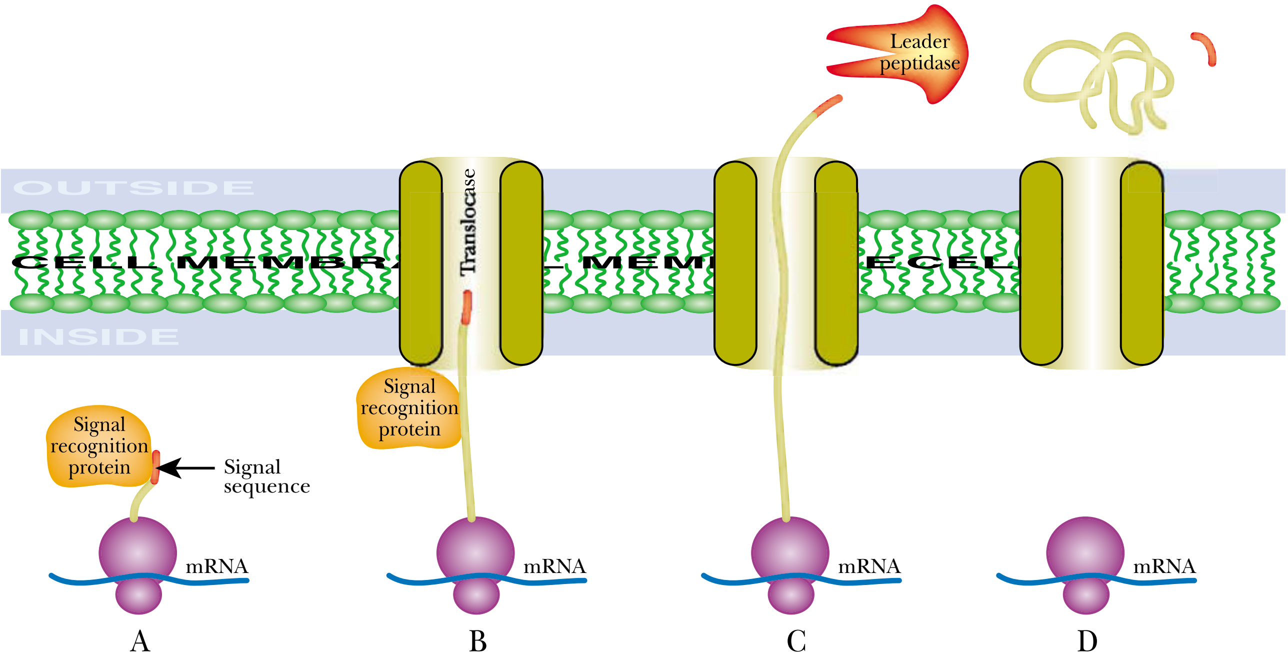


FIGURE 8.27 ***Cotranslational Export of Proteins***

A) The ribosome making the polypeptide chain approaches the cell membrane. The polypeptide with its signal sequence binds to the signal recognition protein. B) The signal recognition protein recognizes the translocase and binds to it, allowing the polypeptide chain to begin its journey through the membrane. C) After the signal sequence exits the translocase, leader peptidase cuts the polypeptide chain, liberating the signal peptide. D) Final folding of the protein occurs outside the cell.

is synthesized and follows the signal sequence into and through the membrane via the translocase. This is known as **cotranslational export**, since the protein is exported as it is made. The signal sequence is cut off by the **leader peptidase** after translocation (Fig. 8.27).

There are approximately 500 translocases per *E. coli* cell. Each cell exports about 1 × 10^6 proteins from the cytoplasm prior to dividing. In a cell that doubles in 20 minutes, 100 proteins are exported per minute per translocase. Protein export is 10-fold faster than protein synthesis. So the demand for a growing protein chain will allow the translocase to be ready for a new chain as fast as the ribosome can make it. Note that in gram-negative bacteria such as *E. coli*, most of these exported proteins are structural components of the outer membrane that are being made constantly, rather than enzymes being excreted outside the cell for digestive purposes.

In eukaryotes, cotranslational export occurs across the membranes of the endoplasmic reticulum. In multi-cellular eukaryotes proteins involved in digestion, such as amylases and proteases, must be exported. So must proteins located in blood and other body fluids, such as antibodies, albumins and circulating peptide hormones. When the animal genes for preproinsulin or ovalbumin are put into *E. coli*, correct export across the cell membrane occurs and cleavage of the signal sequence by the *E. coli* leader peptidase happens at the correct position. Conversely, yeast cells correctly process and excrete bacterial β-lactamase. Thus, the export machinery is highly conserved between diverse organisms.

cotranslational export Export of a protein across a membrane while it is still being synthesized by a ribosome
leader peptidase Enzyme that removes the leader sequence after protein export

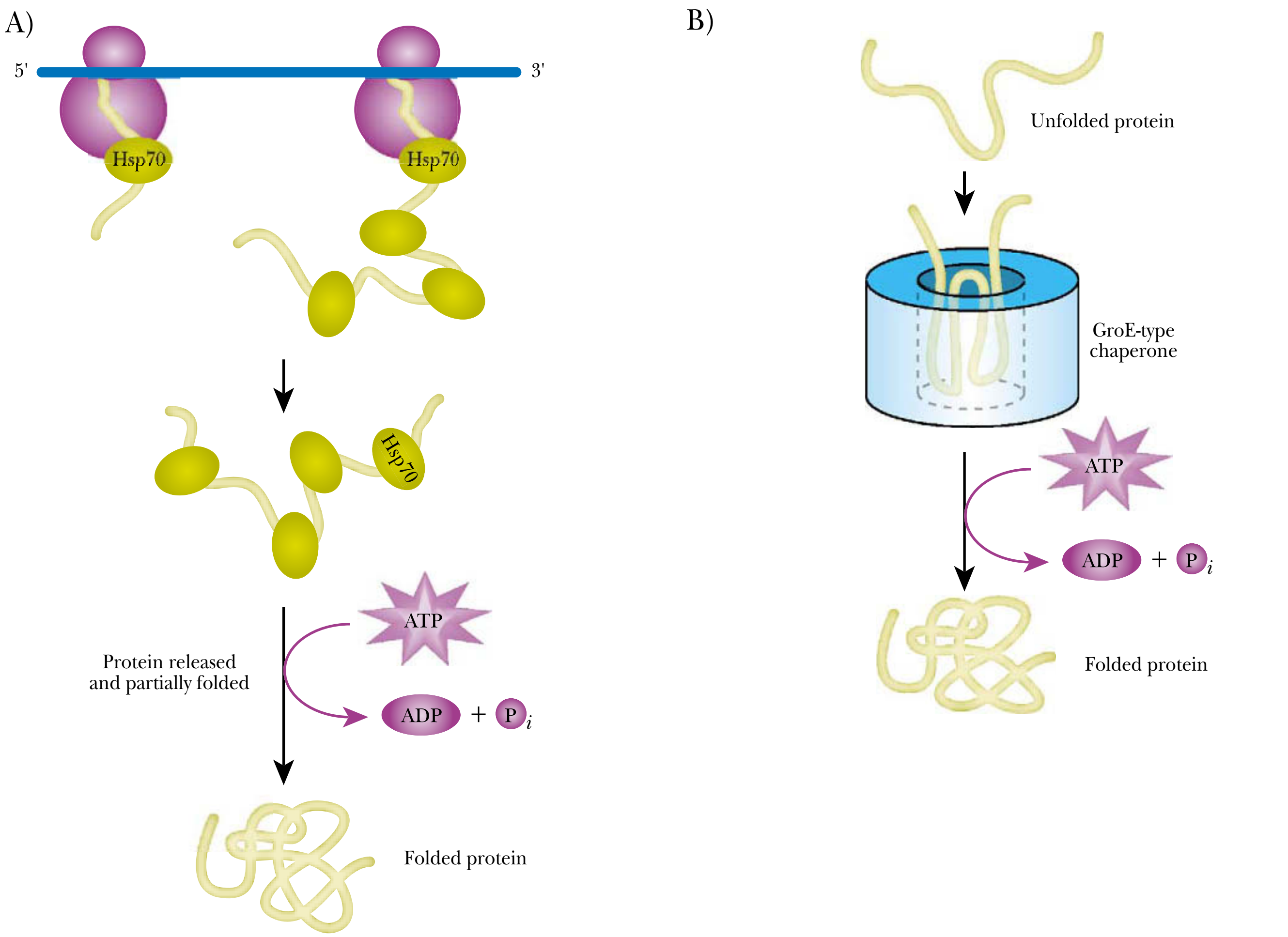


FIGURE 8.28 ***Chaperonins Act by Two General Mechanisms***

A) Chaperonins of the Hsp70 type act during protein formation by binding to hydrophobic patches of the protein. Once chaperonins are released, the protein automatically folds. B) Large chaperonins, such as GroE, act after translation by sequestering misfolded protein in a central cavity. Freed from the influences of other molecules in the cytoplasm, the protein will fold correctly.

Molecular Chaperones Oversee Protein Folding

Molecular **chaperones**, or **chaperonins**, are proteins that oversee the correct folding of other proteins. Many chaperonins are called **heat shock proteins (HSPs)**, as their levels increase at high temperature (see Ch. 9). Chaperonins may be divided into two main classes: those that prevent premature folding and those that attempt to rectify misfolding. Obviously, chaperonins cannot "know" the correct 3-D structure for several other proteins. Mechanistically, they act to prevent incorrect folding, rather than actively creating a correct structure.

During bacterial protein export, the secretory chaperonin SecB keeps the polypeptide chain from folding up prematurely. Secreted proteins must travel through a narrow translocase channel and so must remain unfolded until they reach the other side of the membrane. The Hsp70 set of chaperonins tends to bind to newly made or highly uncoiled proteins (Fig. 8.28).

Chaperonins are proteins that promote the correct folding of other proteins.

chaperone Sometimes "molecular chaperone"; same as chaperonin
chaperonin Protein that oversees the correct folding of other proteins
heat shock protein (HSP) Protein induced in response to high temperature. Many heat shock proteins are chaperonins

The more complex GroE (= Hsp60/Hsp10) chaperonin machine is involved in attempting to refold damaged or misfolded proteins. When polypeptide chains unfold, they expose hydrophobic regions that are normally clustered in the center of the folded protein. Left to themselves, many proteins could refold. However, inside a cell, there is a high concentration of protein. Consequently, exposed hydrophobic regions from multiple proteins bind to each other and the proteins aggregate together. The GroE chaperonin machine forms a cavity in which a single polypeptide can refold on its own, protected from interactions with other polypeptide chains.

Protein Synthesis Occurs in Mitochondria and Chloroplasts

Protein synthesis in mitochondria and chloroplasts resembles that of bacteria in many respects.

Mitochondria and chloroplasts are thought to be of prokaryotic origin. The symbiotic hypothesis of organelle origins argues that symbiotic prokaryotes evolved into organelles by specializing in energy production and progressively losing their genetic independence (see Ch. 20 for further details). Both mitochondria and chloroplasts contain circular DNA that encodes some of their own genes and they divide by binary fission. They contain their own ribosomes and make some of their own proteins. Organelle ribosomes resemble the ribosomes of bacteria rather than the ribosomes of the eukaryotic cytoplasm. The initiation and elongation factors of organelles are also bacterial in nature. Nonetheless, there are differences in composition between organelle and bacterial ribosomes, as shown in Table 8.03.

TABLE 8.03 Components of Cytoplasmic, Organelle and Bacterial Ribosomes

Location	Subunits	Ribosomal RNA	Proteins
Animal Cytoplasm	40S	18S	33
	60S	28S, 5.8S, 5S	49
Animal Mitochondria	28S	12S	31
	39S	16S	48
Plant Cytoplasm	40S	18S	~35
	60S	28S, 5.8S, 5S	~50
Plant Chloroplast	30S	16S	22–31
	50S	23S, 5S, 4.5S	32–36
Plant Mitochondria	30S	18S	>25
	50S	26S, 5S	>30
Bacterial	30S	16S	21
	50S	23S, 5S	31
Archeal	30S	16S	26–27
	50S	23S, 5S	30–31

Although they are larger, the cytoplasmic/nucleus-encoded ribosomes of eukaryotes resemble those from archaebacteria in the way they operate. A similar relation holds for the associated initiation and elongation factors. In fact, it is possible to make hybrid ribosomes containing one subunit from yeast and one from *Sulfolobus* (an archaebacterium). These still make protein, albeit less efficiently than native ribosomes. In contrast, hybrid ribosomes made by mixing subunits from yeast and *E. coli* are totally functionless.

Proteins Are Imported into Mitochondria and Chloroplasts by Translocases

The size of organelle genomes varies considerably from organism to organism. Generally, the more advanced eukaryotes have smaller organelle genomes. The mitochondria of mammals make only around 10 proteins and in higher plants the chloroplasts make approximately 50 proteins. As has been described, the other organelle proteins are encoded by nuclear genes and made on the cytoplasmic ribosomes. They are then transported into the organelles.

Proteins for import into mitochondria have a leader sequence at the N-terminus. This consists of 20 or more amino acids with a positively charged lysine or arginine every three or four residues and no negatively charged residues. The leader forms an α-helix with a positively charged face and a hydrophobic face. This is recognized by a receptor on the mitochondrial surface. The protein is imported successively through two translocase complexes known as TOM (translocase, outer mitochondrial) and TIM (translocase, inner mitochondrial) that lie in the outer and inner membranes of the mitochondria, respectively. After importing the protein, its leader sequence is trimmed off.

Many mitochondrial and chloroplast proteins are made in the eukaryotic cytoplasm and enter the organelle after synthesis.

Plant cells are more complex than animal cells as they possess not only mitochondria but also chloroplasts. The principle of protein import is similar. The leader sequences for chloroplast proteins resemble those for mitochondria, and in fact only plant cells can tell them apart. Thus, the mitochondria of fungi will import chloroplast proteins if genes encoding these are artificially introduced into the fungal cell. It is still unclear how plants decide between chloroplast and mitochondrial leader sequences; however, the leaders are longer in plants than in other organisms and may have additional complexities. The chloroplast contains two translocases equivalent to TIM and TOM, which are known as TIC and TOC (C for chloroplast).

Protein import by organelles also needs chaperonins on both sides of the membrane. An imported protein must travel through the narrow translocase channel in an uncoiled conformation. To avoid premature folding, newly synthesized organelle proteins are kept in a loosely folded conformation by chaperonins. Later, when the imported protein emerges from the translocase into the inside of the organelle, it is bound by another set of chaperonins. In particular, an Hsp70-type chaperonin is responsible for hauling in the incoming protein. The Hsp70 acts as a ratchet, binding to successive segments of unfolded polypeptide chain. Each binding and release of Hsp70 consumes energy in the form of ATP.

Mistranslation Usually Results in Mistakes in Protein Synthesis

Ribosomes are not perfect and make occasional mistakes. Perhaps 1 in 10,000 codons is misread and results in the wrong amino acid being incorporated. Two amino acids whose codons differ by only one base are most likely to be confused. Other possible errors are shifts in the reading frame ("**frameshift**") or reading through stop codons. These assorted errors are collectively known as **mistranslation**.

Although such events are rare, a few weird genes actually require such errors for proper expression. For example, the *pol* gene of retroviruses (see Ch. 17) is only translated if the ribosome frameshifts or reads through a stop codon while translating the preceding *gag* gene. In Chapter 5, it was noted that the *dnaX* gene of *E. coli* gives rise to two proteins, tau and gamma, both subunits of DNA polymerase. The gamma protein is made only as a result of frameshifting. Release factor 2 (RF2) of *E. coli* also requires a frameshift for its synthesis.

frameshift Alteration in the reading frame during polypeptide synthesis
mistranslation Errors made during translation

TABLE 8.04 Exceptions to the Universal Genetic Code

Exceptions in the Chromosomal Genome

Codon	**Universal**	***Mycoplasma***	***Paramecium***	***Euplotes***	***Candida***
UGA	Stop	**Trp**	Stop	**Cys**	Stop
UAA/UAG	Stop	Stop	**Gln**	Stop	Stop
CUG	Leu	Leu	Leu	Leu	Ser

Exceptions in the Mitochondrial Genome

Codon	**Universal**	**Fungi**	**Protozoa**	**Mammals**	**Flatworm**
UGA	Stop	**Trp**	**Trp**	**Trp**	**Trp**
UAA	Stop	Stop	Stop	Stop	Tyr
AUA	Ile	**Met**	**Met**	**Met**	Ile
AGA/AGG	Arg	Arg	Arg	**Stop**	**Ser**
AAA	Lys	Lys	Lys	Lys	Asn
CUA	Leu	**Thr**	Leu	Leu	Leu

The Genetic Code Is Not "Universal"

Minor variations in the genetic code are found in mitochondria and certain microorganisms.

The genetic code is not quite universal. Despite this, the term "**universal genetic code**" is used to refer to the codon table shown above (Fig. 8.02), which applies to almost all organisms. Rare exceptions are found in some protozoans and mycoplasmas and in the mitochondrial genome of animals and fungi (Table 8.04). Mycoplasmas are parasitic bacteria with unusually small genomes. *Paramecium* and *Euplotes* are ciliated protozoans and *Candida* is a yeast.

Note that there is no general mitochondrial genetic code. Although fungal and animal mitochondria share similarities (e.g., UGA = Trp), there are also differences. However, plant mitochondria and chloroplasts use the universal genetic code.

Unusual Amino Acids are Made in Proteins by Post-Translational Modifications

Although the genetic code has codons for only 20 amino acids, many other amino acids are occasionally found in proteins. Apart from selenocysteine and pyrrolysine (see below), these extra amino acids are made by modifying genetically encoded amino acids after the polypeptide chain has been assembled. This is known as **post-translational modification**.

An example of medical importance is **diphthamide**, which is derived from histidine by post-translational modification (Fig. 8.29). It is found only in elongation factor eEF2 of eukaryotes and archaebacteria, in a region of the amino acid sequence that is highly conserved. The corresponding bacterial factor, EF-G, does not contain diphthamide.

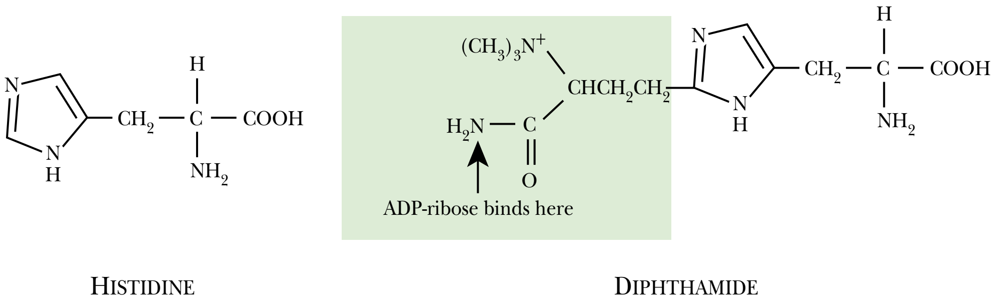
Selenocysteine: The 21st Amino Acid

Selenocysteine (Sec) is not one of the standard 20 amino acids and yet it is incorporated into a few rare proteins during translation of the mRNA by the ribosome. This occurs both in bacteria and in eukaryotes, including humans. Sequencing of the genes

diphthamide Modified amino acid found only in eukaryotic elongation factor eEF2 that is the target for diphtheria toxin
post-translational modification Modification of a protein or its constituent amino acids after translation is finished
selenocysteine (Sec) Amino acid resembling cysteine but containing selenium instead of sulfur
universal genetic code Version of the genetic code used by almost all organisms

FIGURE 8.29 *Histidine and Diphthamide*

After the polypeptide chain of eEF2 has been made on the ribosome one specific histidine residue is converted to diphthamide. Diphtheria toxin stops protein synthesis by adding an ADP-ribose group to the side chain of diphthamide. This inhibits the action of eEF2.



Diphthamide was named after diphtheria, an infectious disease caused by *Corynebacterium diptheriae*. Diphtheria toxin attaches an ADP-ribose fragment to elongation factor eEF2 and this inhibits protein synthesis and kills the target cells. eEF2 normally splits GTP and uses the energy released to move the peptidyl-tRNA from the A-site to the P-site. ADP-ribosylated eEF2 still binds GTP but cannot hydrolyze it or translocate the peptidyl-tRNA.

Very rarely, the stop codon UGA is read as the unusual amino acid selenocysteine.

and proteins involved has shown that selenocysteine is encoded by UGA. However, UGA is one of the stop codons. Apparently, UGA is normally read as "stop" but is occasionally translated to give selenocysteine, which therefore has the honor of being the 21st genetically encoded amino acid. The choice between "stop" and selenocysteine depends on a special recognition sequence in the following part of the gene—the **selenocysteine insertion sequence (SECIS element)**. Selenocysteine has its own tRNA and a special protein factor to escort charged tRNA-Sec to the ribosome. In fact, selenocysteine-tRNA is initially charged with serine. Then the attached serine is enzymatically modified to form selenocysteine.

When bacteria use selenocysteine, the selenocysteine insertion sequence forms a stem and loop structure in the mRNA molecule just after the UGA. SelB protein recognizes both charged tRNA-Sec and this stem and loop. Thus selenocysteine bound to tRNA is delivered to the right place (Fig. 8.30A). [In bacteria, the stem and loop form temporarily from part of the coding sequence, and this section of the mRNA is therefore translated after insertion of the selenocysteine.] In mammals, the stem and loop structure is found beyond the end of the coding sequence, in the 3′-untranslated region—not next to the critical UGA codon! A pair of proteins is responsible for binding the tRNA-Sec and recognizing stem and loop. Somehow, they deliver the tRNA-Sec to the correct position for insertion (Fig. 8.30B).

Selenocysteine is an analog of cysteine, but has selenium instead of sulfur (Fig. 8.31). Selenium is more susceptible to oxidation than sulfur and so proteins that contain it must be protected from oxygen. Examples are the formate dehydrogenases found in many bacteria. These contain selenocysteine in their active sites and function in anaerobic metabolism. They are inactivated by oxygen and are normally made only in the absence of air. Higher organisms contain about 20 proteins that contain selenocysteine. Zebrafish selenoprotein P contains 17 Sec residues, the largest number in any known protein.

The stop codon UAG is occasionally translated as the rare amino acid, pyrrolysine.

Pyrrolysine: The 22nd Amino Acid

In 2002, a 22nd genetically encoded amino acid was discovered—pyrrolysine, a derivative of lysine with an attached pyrroline ring (Fig. 8.32). This is found in a few archaebacteria where it is encoded by the stop codon UAG in occasional proteins. Pyrrolysine was first discovered in the active site of methylamine methyl-transferases found in methane producing archaebacteria of the genus *Methanosarcina*. An unusual

selenocyteine insertion sequence (SECIS element) Recognition sequence that signals for insertion of selenocysteine at a UGA stop codon

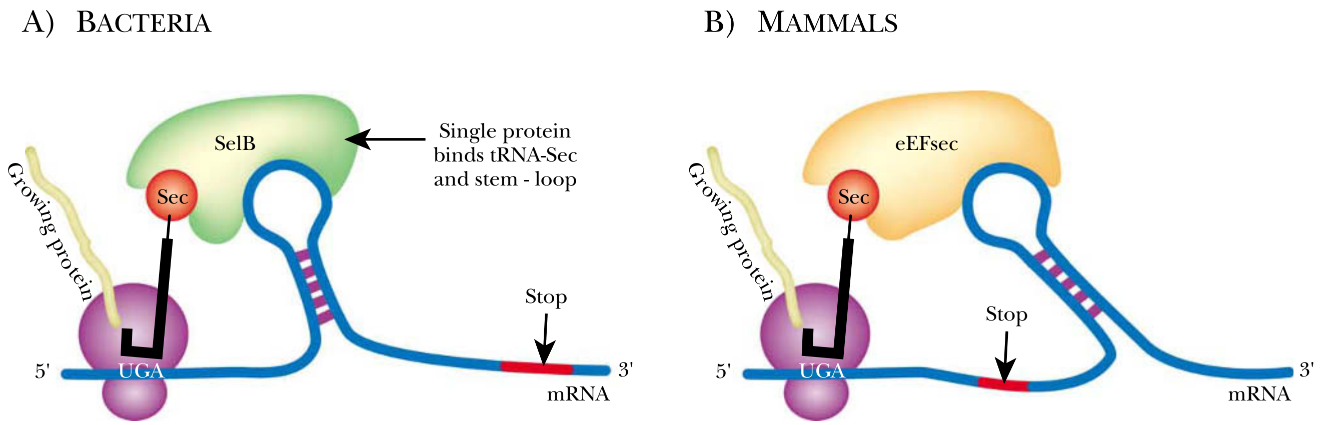


FIGURE 8.30 ***Delivery of tRNA with Selenocysteine to an Internal UGA Stop Codon***

A) In bacteria, the tRNA carrying selenocysteine (Sec) first binds to SelB and the complex then binds to a stem and loop in the mRNA. This aligns the tRNASec with a UGA codon within the coding sequence on the mRNA. Selenocysteine is then inserted as part of the growing polypeptide. Only the fully bound complex is shown. B) In mammals, the protein that binds the stem and loop and the tRNASec is called eEFsec. In addition, the stem and loop are more distant, being found after the stop codon.

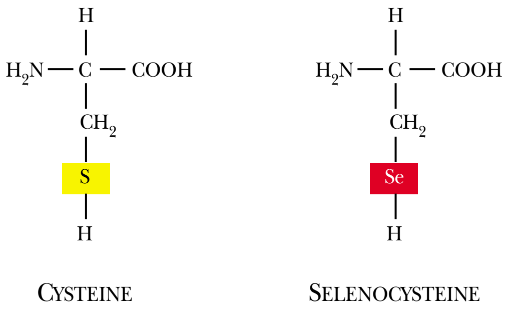


FIGURE 8.31 ***Selenocysteine and Cysteine***

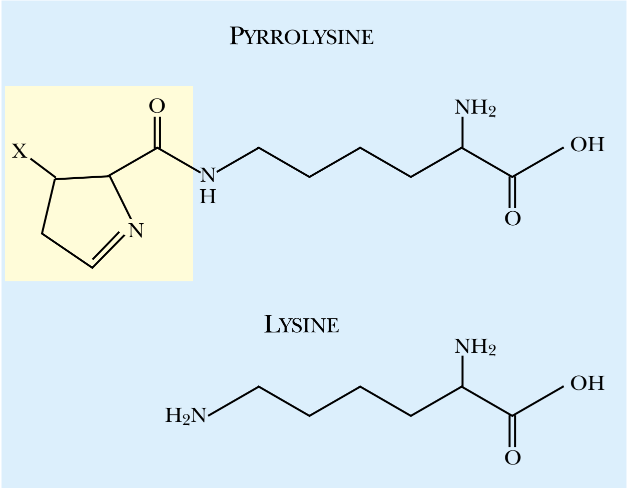


FIGURE 8.32 ***Pyrrolysine and Lysine***

Pyrrolysine is (5*R*,5*R*)-4-substituted-pyrroline carboxylate. The 4-substituent (shown as X) is thought most likely to be a methyl group, but this is not certain.

Why Selenocysteine?

One plausible hypothesis to explain why UGA is both a stop codon and a selenocysteine codon is evolutionarily based. When oxygen first appeared with the evolution of the photosynthetic process, selenocysteine was replaced by the more stable cysteine in all but a few proteins. This allowed UGA to be reassigned as a stop codon. There are some problems with this hypothesis. Many groups of organisms had already diverged before photosynthesis appeared. That they all independently dropped selenocysteine and reallocated UGA to mean "stop" seems highly unlikely.

aminoacyl-tRNA synthase, a special tRNA and genes for three accessory proteins are also found in organisms with pyrrolysine. By analogy with selenocysteine, it is believed that the pyrrolysine-tRNA is first charged with lysine, which is then modified to form pyrrolysine. However, the mechanism of pyyrolysine synthesis and insertion remains to be elucidated in detail. In particular, it is unknown how UAG codons for pyrrolysine are distinguished from those still meaning stop. Genome sequence analysis has found genes homologous to those for the pyrrolysine system in occasional eubacteria suggesting that pyrrolysine may be present. However, pyrrolysine itself has not yet been identified directly in these organisms.

Many Antibiotics Work by Inhibiting Protein Synthesis

Many well-known antibiotics work by inhibiting protein synthesis. Most of these are specific for prokaryotic ribosomes. However, very high concentrations of these agents will inhibit the ribosomes of mitochondria and chloroplasts, which are of prokaryotic ancestry.

Streptomycin and related antibiotics bind to rRNA in the small subunit of the bacterial ribosome.

Aminoglycoside antibiotics bind to the 30S subunit. **Streptomycin** binds to the 16S rRNA near where the two ribosomal subunits touch. The presence of streptomycin distorts the A-site and hinders binding of incoming charged tRNA. In particular, binding of initiator tRNA-Met is inhibited and so initiation of translation is prevented. Streptomycin-resistant mutants have alterations in nucleotide 523 of 16s rRNA or in ribosomal protein S12 (RpsL), which assists antibiotic binding. Many of the other aminoglycosides, such as *gentamycin* and *kanamycin*, bind to multiple sites on the 30S subunit and mainly inhibit the translocation step of protein synthesis. Streptomycin and other aminoglycosides also cause misreading of the mRNA.

Tetracycline binds rRNA in the small subunit of both prokaryotic and eukaryotic ribosomes.

Tetracyclines inhibit both bacterial and eukaryotic ribosomes. They bind to the 16S (or 18S) rRNA of the small subunit and block the attachment of charged tRNA. Despite inhibiting both types of ribosome, tetracyclines inhibit bacteria preferentially due to the fact that bacteria actively take them up whereas eukaryotic cells actively export them.

Chloramphenicol binds to 23S rRNA and prevents peptide bond formation.

Chloramphenicol binds to the 50S subunit, to the loop of 23S rRNA that interacts with the acceptor stem of the tRNA, and inhibits the peptidyl transferase. **Cycloheximide** binds to the 60S subunit of eukaryotic ribosomes and inhibits the peptidyl transferase. **Erythromycin** and related macrolide antibiotics bind to the 23S rRNA of bacterial ribosomes and inhibit the translocation step.

Fusidic acid is a steroid derivative that binds to prokaryotic elongation factor EF-G. In the presence of fusidic acid, EF-G, with its bound GDP, is frozen in place

aminoglycosides Class of antibiotics that inhibits protein synthesis; includes streptomycin, neomycin, kanamycin, amikacin and gentamycin
chloramphenicol An antibiotic that inhibits bacterial protein synthesis
cycloheximide An antibiotic that inhibits eukaryotic protein synthesis
erythromycin An antibiotic that inhibits bacterial protein synthesis
Fusidic acid An antibiotic that inhibits protein synthesis
streptomycin An antibiotic of the aminoglycoside family that inhibits protein synthesis
tetracyclines Family of antibiotics that inhibit protein synthesis

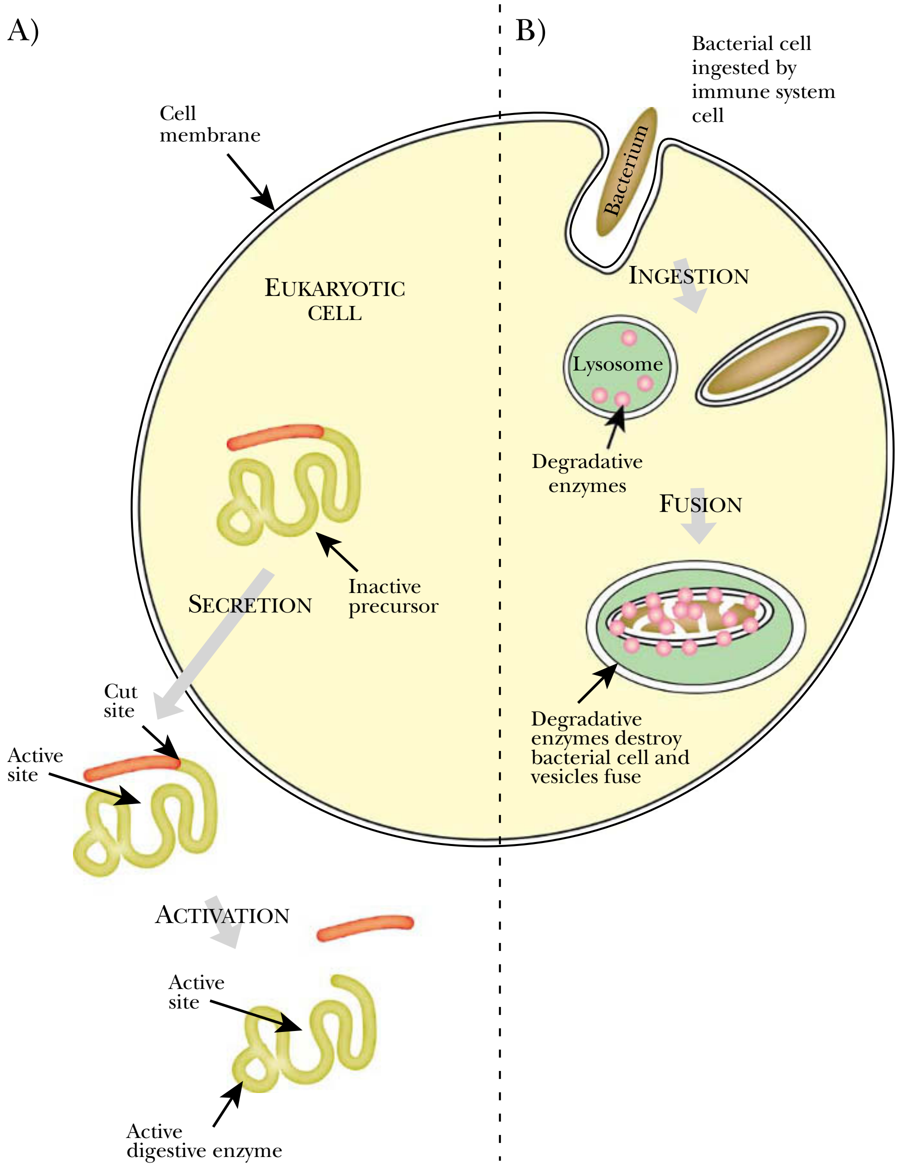


FIGURE 8.33 ***Digestive Enzymes Are Activated on Location***

A) Proteases destined for export are made as precursors and are cleaved to form the active protease once safely outside the cell. B) Proteases in the membrane-bound lysosome degrade ingested material.

on the ribosome. Fusidic acid also inhibits the corresponding eukaryotic elongation factor EF-2; however, in practice, animal cells are unaffected as they do not take up the antibiotic.

Degradation of Proteins

Living cells not only synthesize proteins, they also degrade them. Although protein degradation is nowhere near as complex as synthesis, it is nonetheless carefully controlled and often highly specific. **Proteases** are enzymes that degrade proteins. They are therefore potentially dangerous to the organism that makes them and must be carefully controlled. Proteases are often located in separate compartments where they can act without endangering other components of the organism. Alternatively, proteases may be designed so that they only accept specifically tagged proteins for degradation.

Proteases are found in three main locations. Animals secrete proteases into their digestive tracts. These enzymes are usually synthesized as inactive precursors and only activated once they are safely outside the cells of the animal that made them (Fig. 8.33A).

protease Same as proteinase; an enzyme that degrades proteins

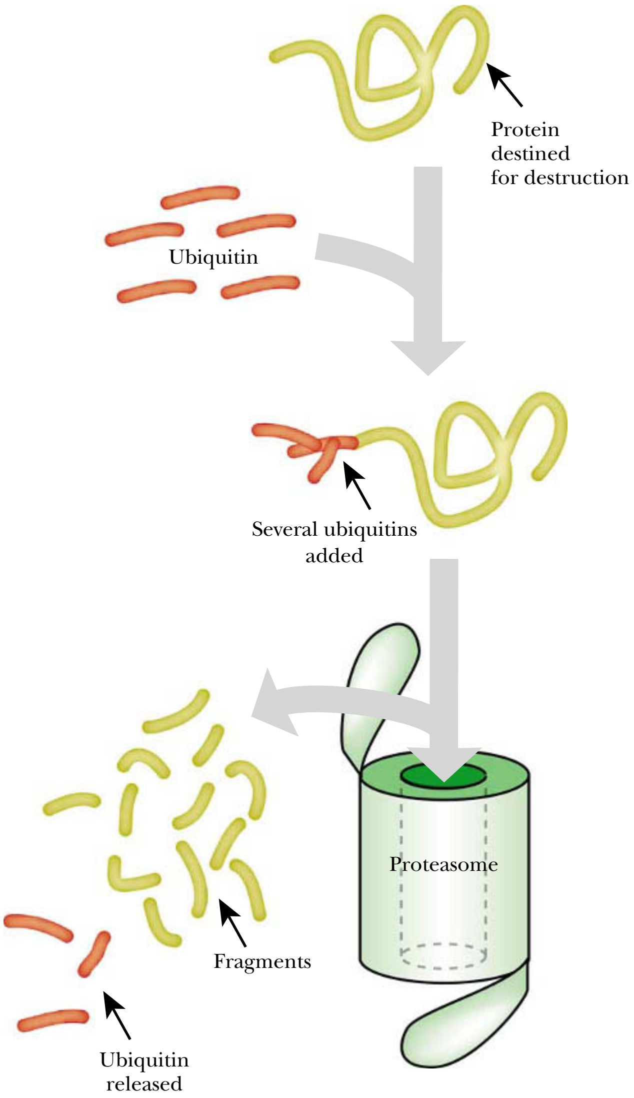


FIGURE 8.34 ***Operation of Proteasome***

Ubiquitin tags damaged proteins and is recognized by the cylindrical proteasome. After degradation, the polypeptide fragments and ubiquitin are extruded.

Enzymes that degrade proteins are dangerous. They are usually kept in separate compartments and often made as inactive precursors.

Examples are trypsin (and its precursor trypsinogen) and pepsin (and its precursor pepsinogen). Plants that catch insects, fungi that trap nematodes and bacteria that live in rotting animal or plant tissue also secrete proteases. As with animals, these proteases are generally secreted as inactive precursors and only activated once outside the cells of the producer organism.

Lysosomes are membrane-bound organelles found in eukaryotic cells. They contain a variety of digestive enzymes, including proteases, and function in self-defense. When cells of the immune system have engulfed bacteria or virus particles, the vesicle containing the invader is merged with lysosomes and the infectious agent is, hopefully, digested (Fig. 8.33B). Bacteria do not always cooperate—for example, many pathogenic strains of *Salmonella* can survive the toxins and digestive enzymes inside lysosomes.

Proteases located in the cytoplasm itself must be very carefully controlled. Nonetheless, the cell needs some internal proteases to degrade damaged or mis-folded proteins. The proteases found inside bacterial cells tend to form rings, with the dangerous active site on the inside of the ring. Proteins slated for destruction are ferried to the protease ring and pushed into its center by accessory proteins. The number of

lysosome Membrane bound organelle of eukaryotic cells that contains degradative enzymes mis-folded proteins and consequently the level of protein degradation increases greatly under certain conditions, in particular when cells are exposed to uncomfortably high temperatures that tend to disrupt protein structure. This induces the heat shock response described in more detail in Ch. 9.

Eukaryotes have more sophisticated structures, known as **proteasomes**. These are cylindrical, with the protease active sites inside. The top and bottom of the cylinder are covered by protein complexes that recognize and bind damaged or unwanted proteins. Proteins destined for degradation are recognized because they are tagged with **ubiquitin**. This is a small protein that is fixed to damaged or mis-folded proteins and also to certain proteins that are needed only for a brief period (Fig. 8.34). The ubiquitin tagged proteins are unfolded and then fed into the barrel of the proteasome where they are degraded into short peptides. The ubiquitin tags themselves are cleaved off and recycled.

proteasome Protein assembly found in eukaryotic cells that degrades proteins
ubiquitin Small protein attached to other proteins as a signal that they should be degraded; used by eukaryotic cells, not bacteria

CHAPTER **NINE**

Regulation of Transcription in Prokaryotes

Gene Regulation Ensures a Physiological Response

Regulation at the Level of Transcription Involves Several Steps

Alternative Sigma Factors in Prokaryotes Recognize Different Sets of Genes

Heat Shock Sigma Factors in Prokaryotes Are Regulated by Temperature

Cascades of Alternative Sigma Factors Occur in *Bacillus* Spore Formation

Anti-sigma Factors Inactivate Sigma; Anti-anti-sigma Factors Free It to Act

Activators and Repressors Participate in Positive and Negative Regulation

The Operon Model of Gene Regulation

Some Proteins May Act as Both Repressors and Activators

Nature of the Signal Molecule

Activators and Repressors May Be Covalently Modified

Two-Component Regulatory Systems

Phosphorelay Systems

Specific Versus Global Control

Crp Protein Is an Example of a Global Control Protein

Accessory Factors and Nucleoid Binding Proteins

Action at a Distance and DNA Looping

Anti-termination as a Control Mechanism

Gene Regulation Ensures a Physiological Response

The functioning of living creatures depends on the *regulated expression of genetic information*. Most genes encode proteins although a minority of genes encode non-translated RNA molecules that function without being translated into proteins. In this chapter, we will focus on the expression of genes that encode proteins and how this is regulated at the level of transcription. Much of cell growth and metabolism is due to the functioning of the proteins thus produced. The mere presence of a protein is not sufficient to ensure a correct physiological response and the activity of many proteins is regulated after they are made as discussed in Chapter 7.

Cells respond actively to their environments by switching genes on or off.

In bacteria such as *Escherichia coli*, about 1,000 of the 4,000 genes are expressed at any given time. If conditions change, some genes are turned off and others are switched on (Fig. 9.01). A major change of growth conditions, such as a shift in temperature, may result in altered expression of 50 to 100 genes.

Single-celled organisms regulate their genes in response both to changes in the environment (such as temperature, osmotic pressure or availability of nutrients) and to the internal state of the cell (such as readiness for cell division). In multicellular organisms, one must also consider near and far communication between cells and the developmental progression of the organism as a whole.

Gene expression, to yield a functional protein, may be regulated at a number of different stages.

—*Transcription of the gene* to give the primary transcript

—*Processing of the primary transcript* to give mRNA

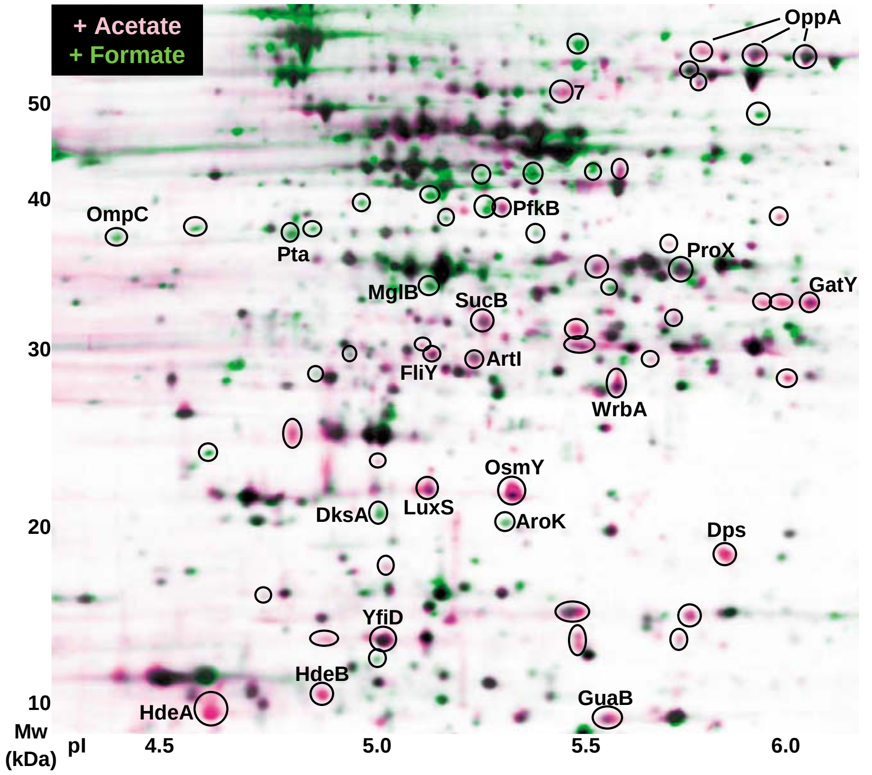


FIGURE 9.01 ***2D Protein Gels of* E. coli *under Different Conditions***

Escherichia coli was grown in the presence of 50 mM acetate or with 20 mM formate. Three gels were run for each condition and the figure shows a layered view of two three-gel composites. The pink and green spots are proteins induced in acetate or formate, respectively. The circled spots are those that were statistically validated, based on a pair-wise comparison of all the individual gels. Copyright Joan L. Slonczewski and Christopher Kirkpatrick, Kenyon College, Gambier, Ohio.

—*Stability of mRNA* to degradation

—*Translation of mRNA* to give polypeptide chains

—*Processing and assembly of polypeptide chains* and any necessary cofactors to give a functional protein

—*Control of activity* of an enzyme or other protein

—*Degradation of protein*

Each of the steps listed above may be regulated, but not with equal frequency. For example, transcription is more frequently regulated than is translation. Each of the major steps listed above may be subdivided further. For example, transcription involves the following steps: access of transcription apparatus to DNA, recognition of promoter sequences, initiation of RNA synthesis, elongation of RNA and termination. Any of the subdivided processes may be regulated, but in practice, regulation of certain steps is much more common than the regulation of others, For example, the initiation of transcription is more often controlled than its elongation phase.

Efficient regulation: control mRNA synthesis. Rapid regulation: control protein activity.

Nature has evolved strategies to optimize regulation. Clearly, it is less wasteful to control the initial synthesis of mRNA rather than wait until the mRNA has been made before deciding whether or not to translate it into protein. However, factors other than efficiency may intrude. If a rapid response is critical to the survival and propagation of a species, making an inactive enzyme and holding it on standby is a good strategy. In general, regulation of transcription is most efficient for the organism, but control of enzyme activity allows for a fast response (Fig. 9.02). Not surprisingly, these are the most common targets of regulation. Nonetheless, examples of regulation may be found for virtually every step between DNA and the final active gene product.

Genes may be regulated positively or negatively.

Both positive and negative regulation control mechanisms exist. In **positive regulation**, a gene is incapable of expression unless it receives a positive signal of some sort. In **negative regulation**, a gene is inherently active but is prevented from expressing itself unless certain inhibitory factors are removed. Some genes are regulated positively, others negatively, and still others by multiple regulators, including both types.

In higher organisms with complex development patterns, genes tend to be controlled by the interaction of multiple positive factors. A gene may be expressed only if it receives several positive signals indicating that certain requirements are satisfied. For example, fetal hemoglobin is only expressed at a specific stage of fetal development and in red blood cells. Negative regulation is relatively rare in higher organisms. In bacteria negative control is more common.

Regulation at the Level of Transcription Involves Several Steps

In Ch. 6, the basic requirements for transcribing a gene were described, including the need for activator proteins to help RNA polymerase bind to the promoter and the possibility of repressor proteins blocking access of RNA polymerase. Here, the regulatory aspects of transcription will be examined in more detail. It is useful to subdivide regulation into sub-processes such as *access*, *recognition*, *initiation*, *elongation* and *termination*.

Access to coding DNA: Coding DNA may be inaccessible under some circumstances. This is especially true of eukaryotes (see next chapter) where DNA is often condensed into **heterochromatin** while not being transcribed. In prokaryotes, access is a relatively minor issue.

heterochromatin A highly condensed form of chromatin that cannot be transcribed because it cannot be accessed by RNA polymerase
negative regulation Control by a repressor that prevents expression of a gene unless it is somehow removed
positive regulation Control by an activator that promotes gene expression when it binds

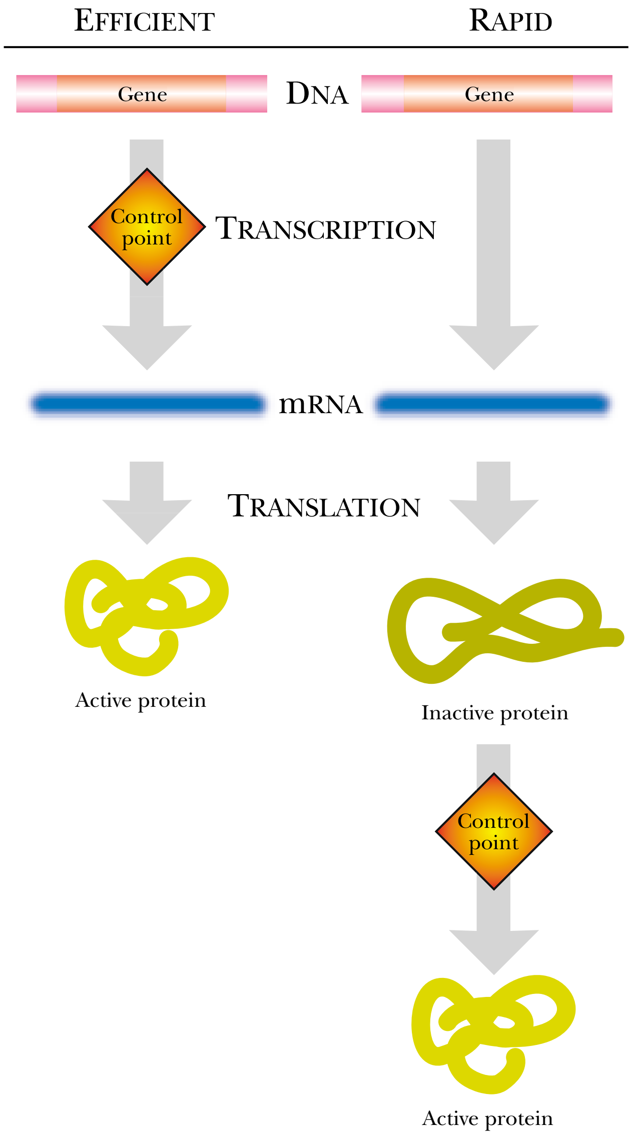


FIGURE 9.02 ***Efficiency Versus Rapid Response***

Two common regulatory mechanisms serve two separate needs. For an efficient response—in other words, one requiring less energy—the gene is often regulated at the level of transcription. When a rapid response is needed, a precursor protein is produced in advance and then rapidly converted to an active protein when the conditions warrant.

Recognition: Obviously many regulatory proteins recognize binding sites on the DNA. Here we are referring to the recognition of the promoter by RNA polymerase itself. In eukaryotes, there are three different RNA polymerases that recognize and transcribe different categories of genes. In bacteria, there is a single RNA polymerase, but there are multiple different sigma factors.

Initiation: Even if recognition is successful, RNA polymerase may be unable to initiate RNA synthesis. In some cases, activator proteins that bind upstream of RNA polymerase may be needed. In other cases, a repressor that blocks movement of RNA polymerase prevents transcription.

Elongation: Once transcription has been initiated, it usually continues without interruption. Regulatory effects at the stage of elongation are uncommon. They may be subdivided into *slowing down of the elongation rate* and *premature termination*.

Termination: Normally, RNA polymerase stops at terminator sites. However, in a few rare cases, termination may be over-ridden by **anti-terminator proteins**. This allows for the expression of those genes downstream of the terminator and their regulation.

anti-terminator protein Protein that allows transcription to continue through a transcription terminator

TABLE 9.01 Alternative Sigma Factors of *Escherichia coli*

Sigma factor	Name		Consensus Sequence		
			–35	Spacing	–10
Housekeeping	σ70	RpoD	TTGACA	16–18	TATAAT
Stationary phase	σ38	RpoS	CCGGCG	16–18	CTATACT
Nitrogen control	σ54	RpoN	TTGGNA	6	TTGCA
Flagellar motion	σ28	FliA	CTAAA	15	GCCGATAA
Heat shock	σ32	RpoH	CTTGAA	13–15	CCCCATNT
Extracytoplasmic heat shock	σ24	RpoE	GAACTT	16	TCTGAT

(N = any base; Y = any pyrimidine; R = any purine)

Alternative Sigma Factors in Prokaryotes Recognize Different Sets of Genes

Different sigma factors recognize different groups of genes.

In bacteria, the sigma subunit of RNA polymerase recognizes the promoter. However, several major groups of genes have promoters lacking the –10 and –35 recognition sequences to which the standard sigma factor binds (see Fig. 6.05 of Chapter 6). These genes have distinct promoter recognition sequences that are recognized by **alternative sigma factors**.

The various sigma factors are named either as σ followed by the molecular weight in kd or as "RpoX," where Rpo refers to RNA polymerase and X signifies the function. The standard or "housekeeping" sigma factor is σ70 or RpoD protein. Each alternative sigma factor is needed for the expression of a large suite of genes (typically 50 or more) that is required under particular conditions such as stationary phase, nitrogen starvation, etc. Some examples of alternative sigma factors found in *Escherichia coli* are shown in Table 9.01.

Heat Shock Sigma Factors in Prokaryotes Are Regulated by Temperature

Almost all organisms that have been studied respond to **heat shock**, so this environmental condition can work well as an example. Many of the genes expressed at high temperatures are highly conserved throughout evolution. However, the regulatory circuits that control the heat shock genes vary greatly from one group of organisms to another. In bacteria such as *E. coli*, two alternative sigma factors, RpoH and RpoE, control the heat shock response.

At increasingly high temperatures, the 3-D structure of proteins begins to unravel. Unfolded or misfolded proteins not only lose their own activity, but they may also bind to other functional proteins and create insoluble aggregates. Thus, the heat shock response is primarily concerned with protecting the cell from damaged and/or misfolded proteins.

Overheating provokes a protective response.

E. coli is optimized for growth at body temperature (37°C). It grows happily up to about 43°C but almost stops growing at 46°C. At 46°C, about 30 percent of all the proteins made by *E. coli* are **heat shock proteins**. Most of these fall into two categories: some are **chaperonins** that help other proteins fold correctly and prevent aggregation,

alternative sigma factor A nonstandard sigma factor needed to recognize a specialized subset of genes
chaperonin A protein that helps other proteins fold correctly
heat shock proteins A set of proteins that protect the cell against damage caused by high temperatures
heat shock response Response to high temperature by expressing a set of genes that encode heat shock proteins

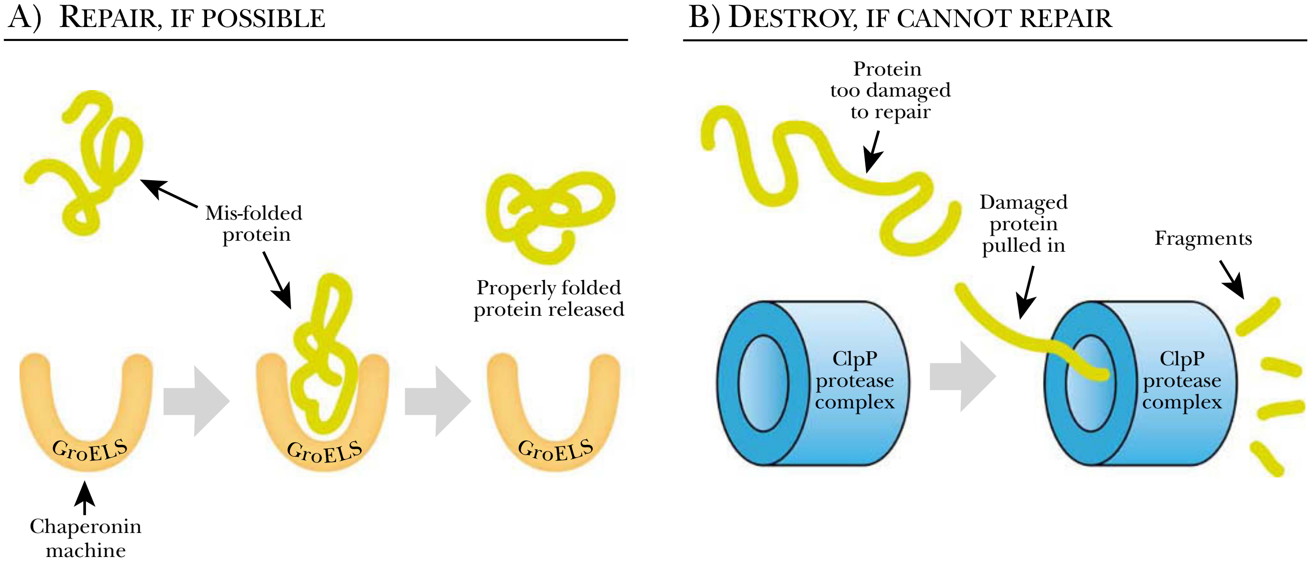


FIGURE 9.03 ***Heat Shock Response in* E. coli**

The cell responds to heat shock either by repairing misfolded proteins or by degrading them.

while others are **proteases** that degrade heat-damaged proteins that are past rescue (Fig. 9.03).

When the environmental temperature of cells is increased from 30°C to 43°C, the level of RpoH rises. This results in elevated expression of the heat shock genes that depend on RpoH for transcription. Control of the level of RpoH itself is very complex and modulated by a variety of minor factors; however, the main signal is the level of misfolded proteins present in the cell. This is monitored by two heat shock proteins, DnaK (a chaperonin) and HflB (a protease). When the level of misfolded proteins is low, DnaK and HflB are free and they apparently bind to RpoH and degrade it. In addition, they bind to partly synthesized RpoH protein, even before it is finished by the ribosome, and block further translation. When the level of misfolded proteins rises, DnaK and HflB bind to these and are unable to affect RpoH levels (Fig. 9.04).

Transcription of the *rpoH* gene from its main promoter requires the standard sigma factor σ70 or RpoD. At temperatures above 50°C, σ70 is inactivated and synthesis of RpoH would come to a halt, thus undermining the heat shock response. This is prevented by the presence of a second promoter for the *rpoH* gene that can be recognized by the RpoE sigma factor. Transcription can continue until 57°C, when the core enzyme of RNA polymerase is inactivated.

The level of RpoE (E for extra-cytoplasmic) is controlled in response to the level of misfolded proteins in the outer membrane and periplasmic space, rather than in the cytoplasm, as in the case of RpoH. In addition to *rpoH*, another group of a dozen or so heat shock genes requires RpoE for their transcription.

Cascades of Alternative Sigma Factors Occur in *Bacillus* Spore Formation

Some bacteria survive hard times by making spores.

The requirement of RpoE for transcription of the *rpoH* gene illustrates, at a simple level, that the expression of one sigma factor may depend upon another. Indeed, in some complex processes, a series of alternative sigma factors may depend on each other. The classic case of a cascade of alternative sigma factors is the regulation of

protease Enzyme that degrades proteins

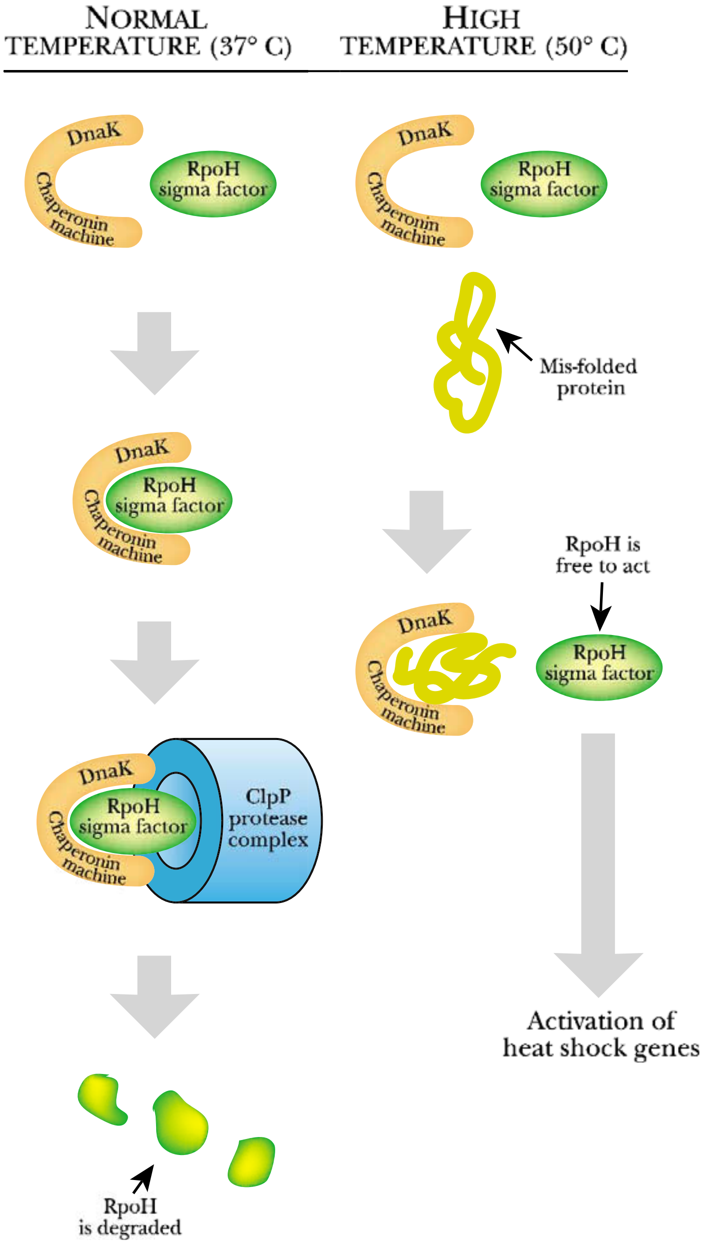


FIGURE 9.04 ***Regulation of the Heat Shock Response***

At normal temperatures, the RpoH sigma factor binds the chaperonin and is presented to the protease for digestion. At high or extreme temperatures, the chaperonin binds instead to misfolded proteins, leaving RpoH free to activate the heat shock genes.

spore formation in the gram-positive bacterium, *Bacillus*. When nutrients are scarce, *Bacillus* forms spores designed to survive bad times. Spores are formed by an asymmetric division that gives a full-sized mother cell and a much smaller spore. The spore is surrounded by the mother cell until it is fully developed. The mother cell then bursts, releasing the spore. This represents cellular differentiation at its most primitive level (Fig. 9.05 and Fig. 9.06).

Spore formation is controlled by four alternative sigma factors, σE and σK in the mother cell, and, once sporulation has started, σF and σG in the developing spore. Two of these, σE and σK, are first synthesized as inactive precursor proteins—pre-σE and pre-σK that must be activated by specific proteases. First, an environmental signal activates the synthesis of pre-σE in the mother cell and σF in the spore. The presence of σF allows transcription of early sporulation genes in the spore. These include the gene for the sigma factor, σG, as well as protein sporulation factors that move into the mother cell and split pre-σE protein to give active σE. The activated σE switches on

spore A cell specialized for survival under adverse conditions and/or designed for distribution

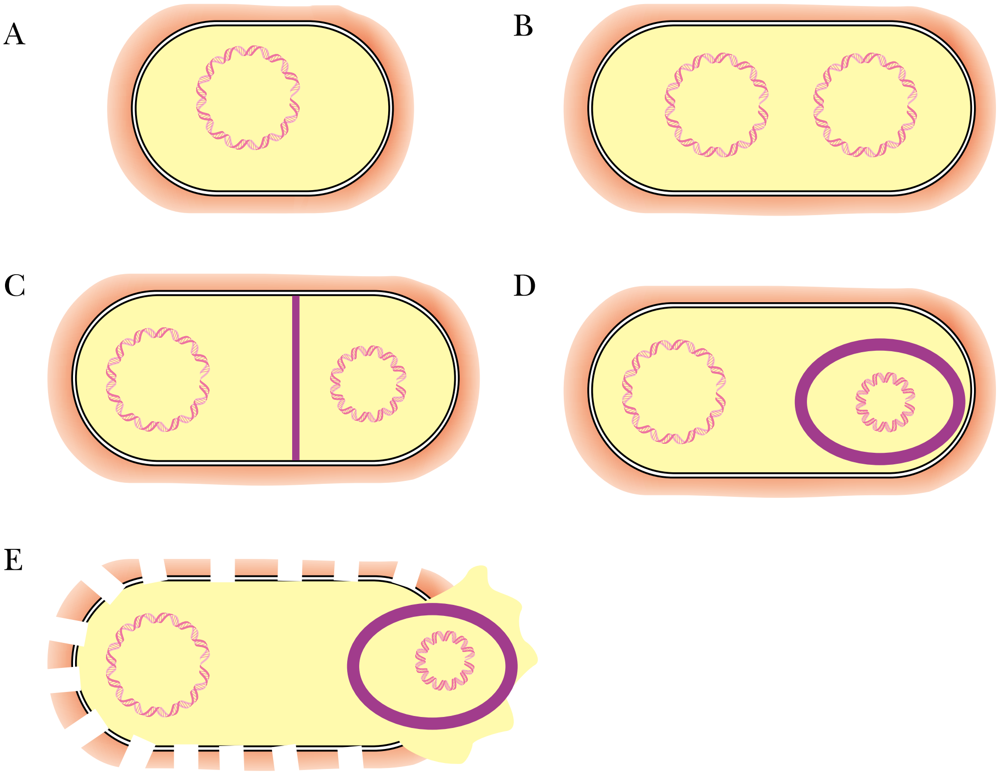


FIGURE 9.05 ***Spore Formation in* Bacillus**

Bacillus (A) first duplicates its DNA (B), then walls off the new DNA into a spore that lies within the cell (C and D). The spore is released from the original mother cell as it bursts and dies (E).

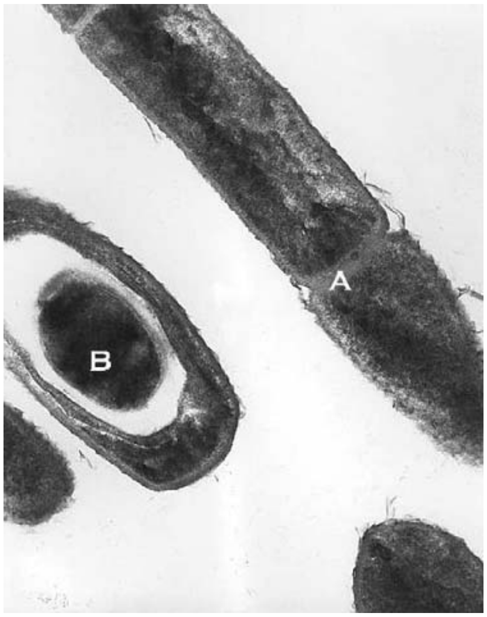


FIGURE 9.06 ***Spore Formation in* Bacillus anthracis**

Transmission electron micrographic image of *Bacillus anthracis* from an anthrax culture, showing cell division (A), and spores (B). Public Health Image Library (CDC) by Dr. Sherif Zaki and Elizabeth White.

several genes in the mother cell, including the gene that encodes pre-σK. As a result, pre-σK accumulates in the mother cell. The presence of σG in the spore allows transcription of late sporulation genes in the spore, including factors that move into the mother cell and activate pre-σK to active σK (Fig. 9.07). Each of these sigma factors is responsible for the transcription of a group of genes needed for successive stages in the development and release of the spore.

The cascade of regulators seen in spore formation ensures that steps in a complex series of events follow each other in the correct sequence. Each regulator controls one stage in the process and also controls the regulator for the next stage. In addition,

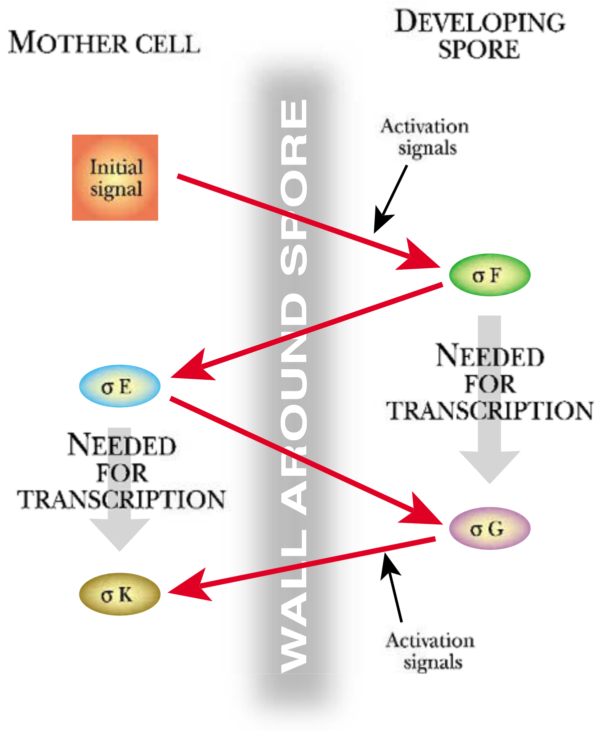


FIGURE 9.07 ***Outline of Spore Formation Regulatory Cascade***

A cascade of four sigma factors is involved in the stepwise development of the spore in *Bacillus*. An external signal activates synthesis of σF in the spore. This is required for transcription of the gene for σG (inside spore) and for factors that cross into the mother cell (red arrow) and convert pre-σE into active σE (in mother cell). Active σE is required for synthesis of the precursor, pre-σK. Finally, σG allows synthesis of factors that cross into the mother cell (red arrow) and convert pre-σK into active σK (in mother cell).

regulatory signals are exchanged between more than one cell. Thus, the developing spore and mother cell cross-regulate each other. The development and differentiation of higher organisms use much the same principles as spore formation, but the regulatory schemes are vastly more complex.

Anti-sigma Factors Inactivate Sigma; Anti-anti-sigma Factors Free It to Act

Positive regulatory factors are often opposed by negative factors.

Sigma factors may be inhibited by proteins known as **anti-sigma factors**. These bind to specific alternative sigma factors and prevent them from associating with RNA polymerase (Fig. 9.08). When σF is first made in the developing spore, it is inactive. Unlike σE and σK, which need to be activated by the proteolysis of an inactive precursor protein, σF is kept inactive by an anti-sigma factor (SpoIIAB). This anti-sigma factor is, in turn, displaced from σF by an **anti-anti-sigma factor** (SpoIIAA). This event triggers the cascade of gene activation described above.

Regulation of alternative sigma factors by binding to anti-sigma factors and their release by anti-anti-sigma factors is not especially common, but several examples are known. Assembly of flagella in *E. coli* and *Salmonella* is under control of the FliA sigma factor and the FlgM anti-sigma factor. A clinically important example is the production of mucus by *Pseudomonas aeruginosa*. This bacterium often infects the lungs of cystic fibrosis patients, where it switches on the genes for mucus production. These are under control of the alternative sigma factor AlgU and the anti-sigma factor MucA. The bacterial mucus clogs the patient's airways and is a major contributor to the symptoms of the disease. Strictly speaking, the material made by *Pseudomonas* is alginate, an acidic polysaccharide that is a repeating polymer of mannuronic and glucuronic acids. This is chemically distinct from the true mucus made by animal cells, which con-

anti-anti-sigma factor Protein that binds to an anti-sigma factor and so prevents the anti-sigma factor from binding to and inhibiting a sigma factor
anti-sigma factor Protein that binds to a sigma factor and blocks its role in the initiation of transcription

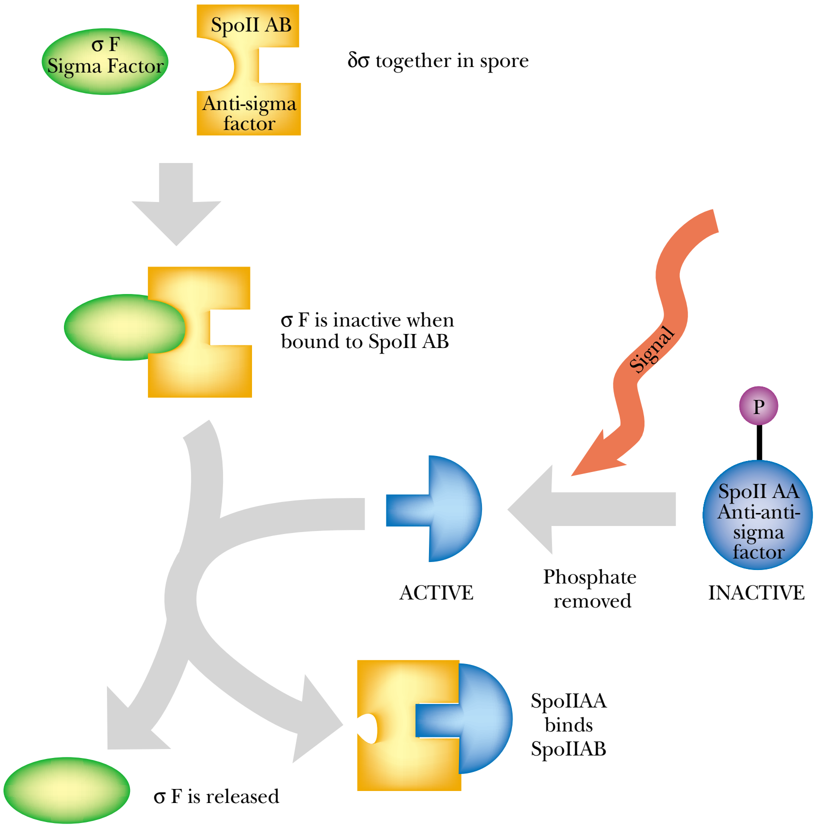


FIGURE 9.08 ***Anti-Sigma Factors***

The anti-sigma factor SpoIIAB binds to σF and inactivates it. When the cell receives an external signal, the phosphorylated form of SpoIAA, an anti-anti-sigma factor, loses its phosphate and engages SpoIIAB. This releases σF, which is then free to activate the sporulation cascade shown above in Fig 9.07.

sists of glycoproteins with many short side chains of galactose, N-acetyl-galactosamine and N-acetyl-neuraminic acid.

Activators and Repressors Participate in Positive and Negative Regulation

Activator proteins turn genes on. Repressor proteins turn genes off.

Some promoters, both in lower and higher organisms (see Ch. 6) function poorly or not at all in the absence of extra proteins known as gene activator proteins, or transcription factors. In addition, there is a class of gene regulator proteins known as *repressors* that act to turn genes off.

In **positive control**, an activator is required to turn a gene on, in response to a signal of some kind. In **negative control**, a gene is switched off by a repressor and is only expressed in the presence of a signal that removes the repressor from the gene. Positive and negative control may be exerted at the level of transcription or at later stages in gene expression. Furthermore, although most activators and repressors are proteins, cases are known in which regulation is due to regulatory RNA or even small molecules.

In both positive and negative control, a small **signal molecule**, the **inducer**, typically binds to the regulatory protein and induces gene expression. In the standard model of positive regulation, an inactive activator protein binds the signal molecule and is converted to its DNA-binding form, which then turns on the gene (Fig. 9.09). Similarly, in typical negative regulation, the DNA-binding form of a repressor protein is converted to its inactive form by binding the signal molecule.

inducer A signal molecule that turns on a gene by binding to a regulatory protein
negative control or regulation Regulatory mode in which a repressor keeps a gene switched off until it is removed
positive control or regulation Control by an activator that promotes gene expression when it binds
signal molecule A small molecule that triggers a regulatory response by binding to a regulatory protein

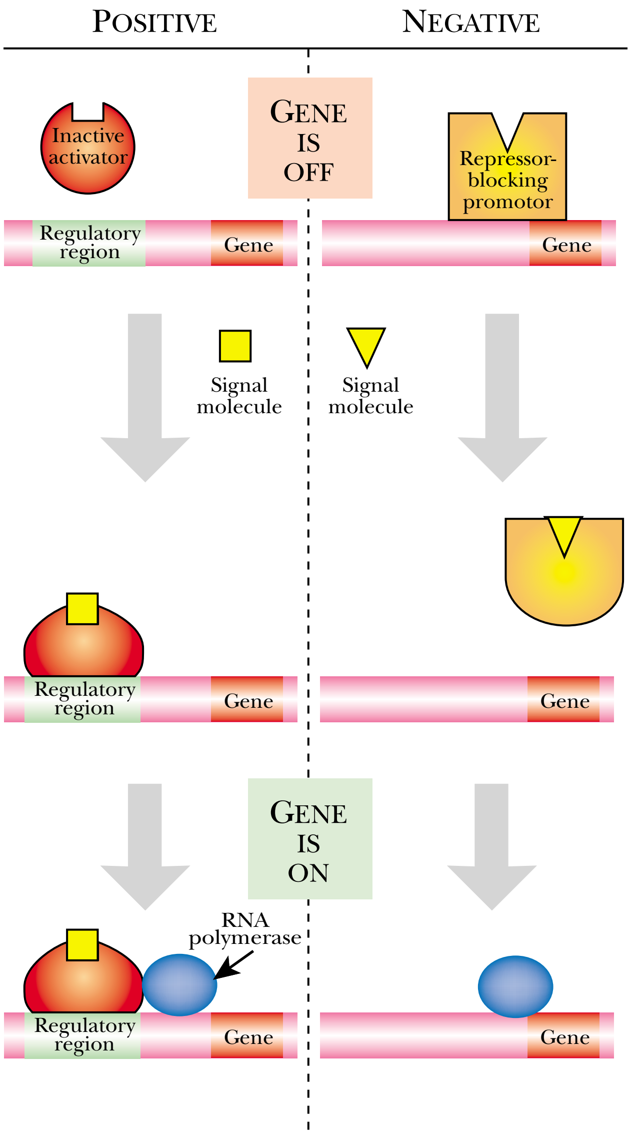


FIGURE 9.09 ***Principle of Positive and Negative Regulation***

In positive regulation, a signal changes the conformation of an inactive regulator, which then becomes active and binds to the regulatory region of a gene. Its presence aids the binding of the RNA polymerase and helps switch on the gene. In negative regulation, a repressor molecule blocks the promoter of the gene. A signal changes the conformation of the repressor, releasing it from the gene and allowing the RNA polymerase to bind.

Operons are clusters of genes that are controlled as a unit.

The Operon Model of Gene Regulation

An **operon** is a cluster of genes that are transcribed together to give a single messenger RNA molecule, which therefore encodes multiple proteins (Fig. 9.10). Such **polycistronic mRNA** is only found in prokaryotes. The genes in an operon are often related functionally, so it makes good sense to regulate them as a group. For example, an operon may encode several enzymes that take part in the same biochemical pathway. Some operons have only a single gene; most have two to half a dozen and a few have more. Despite having more genes than bacteria, higher organisms do not have operons; their genes are transcribed one at a time.

The operon model for regulating bacterial genes was first proposed by François Jacob and Jaques Monod, using the negatively regulated lactose genes of *E. coli* as an example. Since then a vast number of bacterial genes, including those with activators as well as those with repressors, have been fitted to this model or variants of it. The lactose operon, like many bacterial operons, is controlled at two levels. **Specific regu-**

operon A cluster of prokaryotic genes that are transcribed together to give a single mRNA (i.e. polycistronic mRNA)
polycistronic mRNA mRNA that carries several structural genes or cistrons
specific regulation Regulation that applies to a single gene or operon or to a very small number of related genes

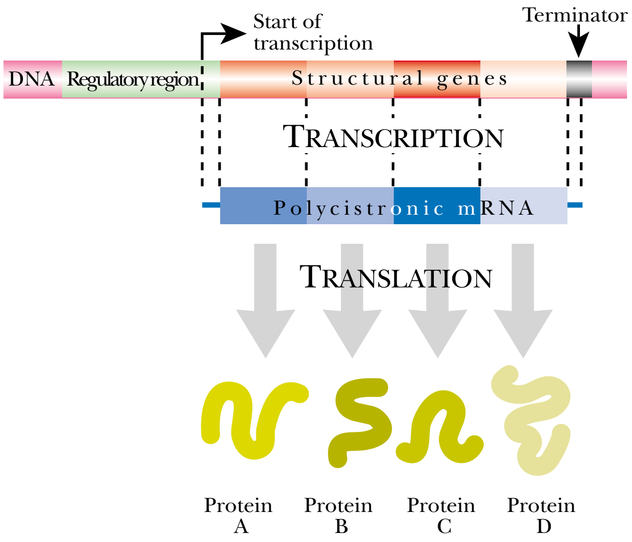


FIGURE 9.10 ***Transcription of an Operon to Give Polycistronic mRNA***

Several structural genes are located side by side and are transcribed into a single length of mRNA. During translation, several proteins, usually functionally related, are made using the single mRNA.

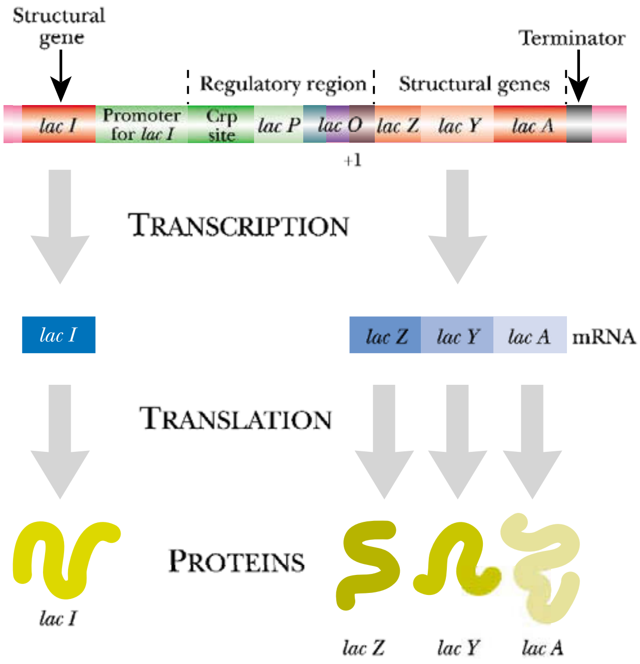


FIGURE 9.11 ***Components of the lac Operon***

The *lac* operon consists of three structural genes, *lacZYA*, all transcribed from a single promoter, designated *lacP*. The promoter is regulated by binding of the repressor at the operator, *lacO*, and of Crp protein at the Crp site. Note that in reality the operator partly overlaps both the promoter and the structural gene. The single *lac* mRNA is translated to produce the LacZ, LacY, and LacA proteins. The *lacI* gene that encodes the LacI repressor has its own promoter and is transcribed in the direction opposite to the *lacZYA operon*.

lation refers to regulation in response to factors specific for a particular operon, in this case the availability of lactose. **Global regulation**, discussed later, is regulation in response to more general conditions, such as the overall carbon and energy supply of the cell.

The lac operon includes genes for lactose uptake and metabolism.

The lactose or *lac* operon consists of three structural genes, *lacZ*, *lacY*, and *lacA*, together with an upstream regulatory region (Fig. 9.11). The *lacZ* structural gene encodes **β-galactosidase**, the enzyme that degrades lactose. The *lacY* gene encodes **lactose permease** (a transport protein) and *lacA* encodes lactose acetylase, whose role is not known (it is not needed for growth on lactose by *E. coli)*. The *lac* operon is regulated by the **LacI** repressor protein, which is encoded by the *lacI* gene. This lies upstream of *lacZYA* and is transcribed in the opposite direction.

beta-galactosidase or β-galactosidase (LacZ) Enzyme that splits lactose and related molecules to release galactose
global regulation Regulation of a large group of genes in response to the same stimulus
LacI The lactose repressor protein
lactose permease (LacY) The transport protein for lactose

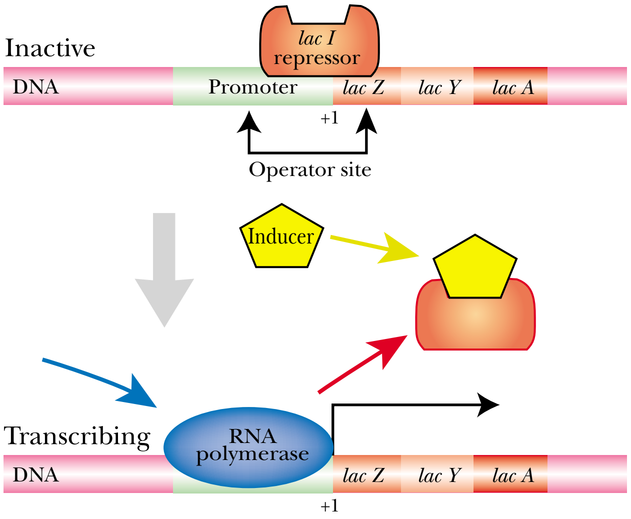


FIGURE 9.12 ***Specific Regulation of the* lac *Operon***

When LacI binds to the operator site, no transcription takes place. The presence of the inducer can remove LacI from the operator, allowing RNA polymerase to bind and transcribe the operon. The inducer may be *allo*-lactose, derived from lactose or an artificial compound, such as IPTG.

The upstream region contains a recognition sequence for the repressor protein, known as the **operator** (*lacO* in Fig. 9.11). If no inducer is present, LacI protein binds to the operator. This blocks the binding of RNA polymerase to the promoter. When lactose is present, it induces the *lac* operon. The actual inducer is not lactose itself, but *allo*-lactose, an isomer of lactose that is made from lactose by β-galactosidase. The LacI repressor binds inducer and changes shape to its inactive form, which cannot bind DNA. RNA polymerase is now able to bind to the promoter and transcribe the *lac* operon (Fig. 9.12). LacI protein exists as a tetramer that can, in fact, bind two DNA recognition sites if these are available in the promoter region. This will result in looping of the DNA—see below. [Unlike many allosteric proteins, the LacI protein does not alternate between monomer and tetramer forms, but exists as a tetramer of unusual stability under all physiological conditions.]

IPTG is an artificial inducer of the lactose operon.

In the laboratory, the *lac* operon is often induced by the compound **IPTG** (***iso*-propyl-thiogalactoside**) (Fig. 9.13). This artificial compound is known as a **gratuitous inducer** because it is not metabolized by the products of the genes it induces. In this particular case, although IPTG induces the *lacZYA* genes, it is not broken down by β-galactosidase, the enzyme that degrades lactose. Consequently, IPTG continues to induce the *lac* operon long-term, whereas, natural inducers only induce for a short period of time before they are broken down.

Some Proteins May Act as Both Repressors and Activators

Activators generally bind upstream of the promoter and help RNA polymerase to bind. Conversely, repressors bind downstream of the promoter and either block the binding of RNA polymerase or prevent it from moving forward and transcribing the gene. Not surprisingly, the same DNA-binding protein can act as an activator for one gene and a repressor for another if it binds at different locations in the two cases (Fig. 9.14).

Regulatory proteins can alternate between two different forms, both of which bind DNA. This is somewhat different than the binding already discussed, which alternates between an active, DNA-binding form and an inactive, nonbinding form. Here, the two forms of the protein act as an activator and a repressor that both bind DNA, but at

gratuitous inducer A molecule (usually artificial) that induces a gene but is not metabolized like the natural substrate; the best known example is the induction of the *lac* operon by IPTG
IPTG (*iso*-propyl-thiogalactoside) A gratuitous inducer of the *lac* operon
operator The site on DNA where a repressor binds

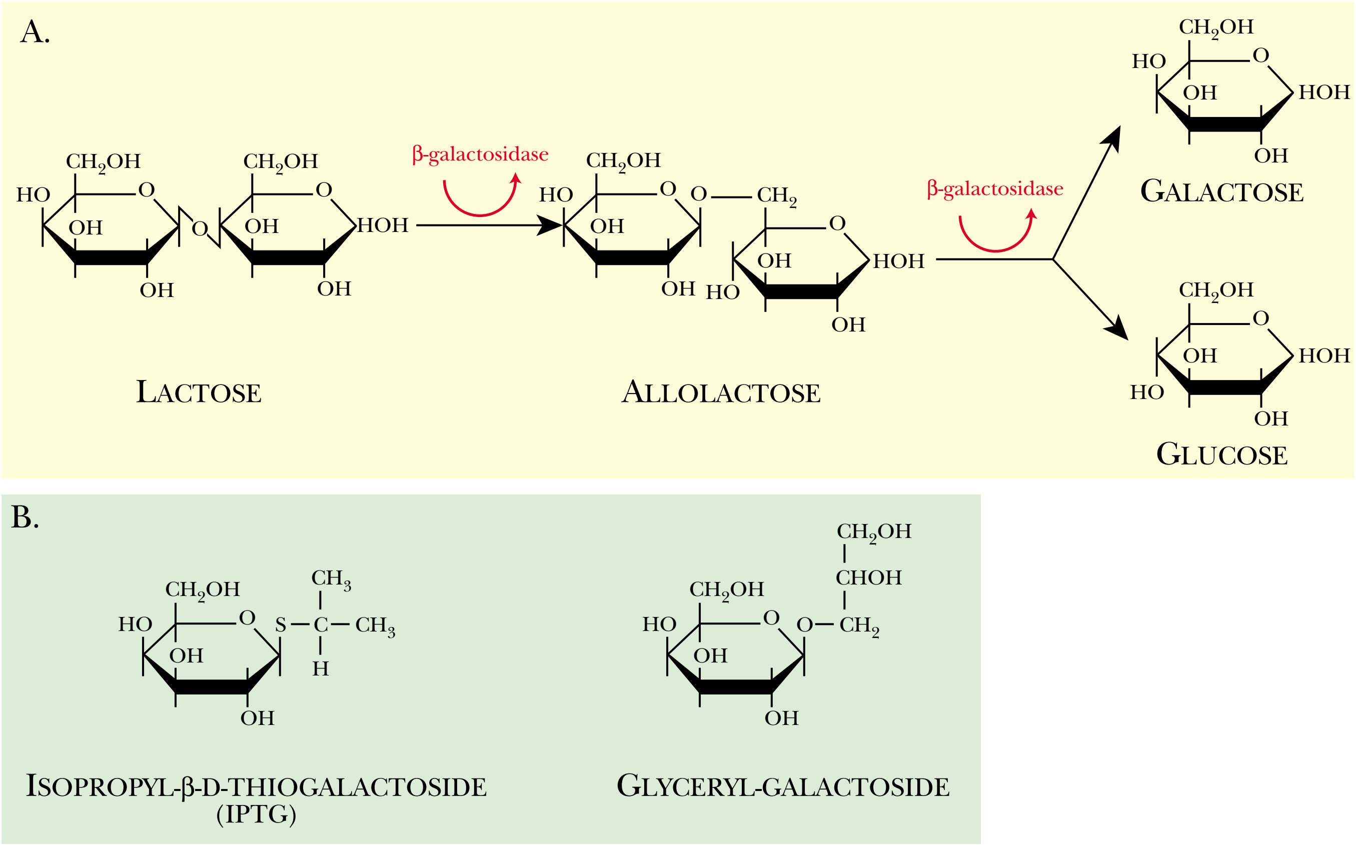


FIGURE 9.13 ***Lactose and Related Galactoside Derivatives***

A) The enzyme β-galactosidase splits lactose into galactose plus glucose. β-Galactosidase can also interconvert lactose with its isomer, *allo*-lactose, which is the true inducer of the *lac* operon. B) The structures of the gratuitous inducer, IPTG, and of the most likely natural substrate for β-galactosidase, glyceryl-galactoside, are also shown.

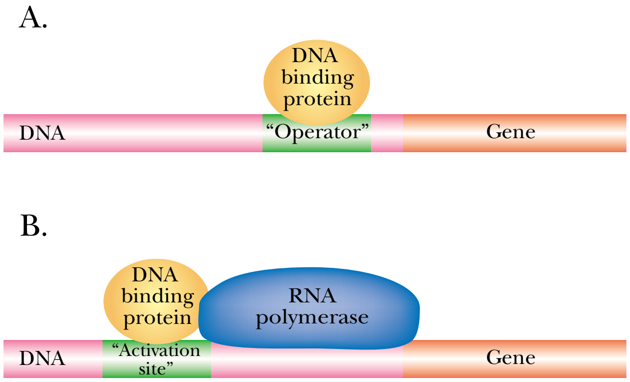


FIGURE 9.14 ***Binding Site Determines Action: Repressor or Activator***

A) The DNA binding protein shown in orange acts as a repressor when it binds to the operator region of the promoter, thus preventing binding of the RNA polymerase. B) The same protein may also bind to an activation site on the DNA of another operon, thus facilitating the binding of the RNA polymerase and promoting gene transcription. The recognition sequences for the DNA binding protein are identical; only their position relative to the RNA polymerase has changed.

The *lac* operon is actually an intruder into the *E. coli* genome.

The Lactose Operon Is Not Really Typical

Although the lactose genes were the first whose regulation was characterized in detail, and although they are often cited as a typical example, they are aberrant in several ways. Curiously, lactose itself is not the inducer. Lactose, which consists of glucose linked to galactose, is converted to ***allo*-lactose**, an isomer in which the same two sugars are linked differently. This transformation is carried out by β-galactosidase, which normally splits lactose, but makes a small amount of *allo*-lactose as a side reaction. It is *allo*-lactose that actually binds to the LacI protein and acts as an inducer.

Lactose is present in the milk consumed by babies and children, but adult diets often contain very little. Moreover, lactose is almost all absorbed in the small intestine, so in nature *E. coli*, which lives in the large intestine, will never get any lactose. In reality, the lactose genes are probably intended to digest glyceryl-galactoside, a compound derived from breakdown of the lipids of animal cells. This is released into the large intestine as the cells lining it are sloughed off and disintegrate. Glyceryl-galactoside is both a genuine inducer, which binds to LacI protein, and a substrate for β-galactosidase, which splits it into glycerol plus galactose.

Furthermore, the segment of DNA containing the lactose genes is missing from *Salmonella* and several other bacteria of the enteric family that are close relatives of *E. coli*. It seems likely that this segment of DNA is a relative newcomer to the *E. coli* genome, and came originally from some source outside the enteric bacteria.

In retrospect, it was both fortuitous and fortunate that Jacob and Monod chose a rather anomalous gene rather than a typical one. The regulation of the *lac* operon is simpler than that of many genes that are more fully integrated into the central metabolism of *E. coli*.

different recognition sites. The AraC regulatory protein controls the transport and metabolism of the five-carbon sugar **arabinose**. When arabinose binds to AraC, it converts it from a repressor to an activator. The ***araBAD* operon** for arabinose metabolism and the *araFG* operon for arabinose uptake are repressed by AraC in the absence of arabinose and activated by AraC plus arabinose (Fig. 9.16).

Quite often regulatory proteins control their own production. This is known as **autogenous regulation**, or auto-regulation. For example, the AraC protein represses the *araC* gene and the Mlc protein (see below) represses transcription of the *mlc* gene.

Biological signals are often carried by small molecules. These signals are detected by proteins that bind to them.

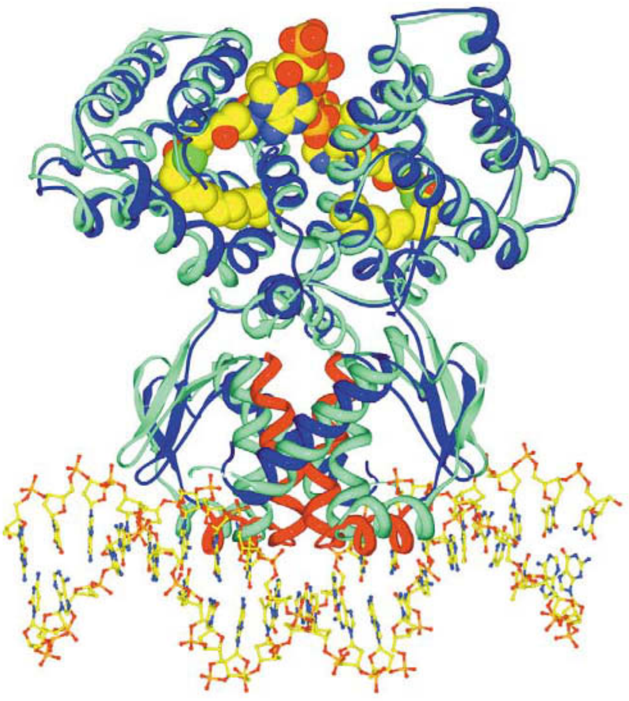
Nature of the Signal Molecule

The substrate specificity and the inducer specificity of an operon need not be identical. One is determined by which molecules fit the active site(s) of the enzyme(s) of the pathway and the other by which molecules fit the binding site on the regulatory protein. In the case of the lactose operon *allo*-lactose, glyceryl-galactoside and IPTG are true inducers (i.e. they bind to the LacI protein). Lactose itself is only an apparent inducer, as it does not bind directly to LacI and must first be converted to *allo*-

***araBAD* operon** Operon that encodes proteins involved in metabolism of the sugar arabinose
arabinose A five-carbon sugar often found in plant cell wall material that can be used as a carbon source by many bacteria
***allo*-lactose** An isomer of lactose that is the true inducer of the *lac* operon
autogenous regulation Self regulation, i.e. when a DNA-binding protein regulates the expression of its own gene

FadR—An Example of a Repressor and an Activator

An example of differential binding of a DNA-binding protein is the FadR protein of *E. coli*. FadR represses the genes for fatty acid breakdown, but also activates certain genes involved in fatty acid synthesis. FadR responds to the availability of long-chain fatty acids in the growth medium. The cell can incorporate pre-made fatty acids into its lipids and can also break them down for energy. Fatty acids are taken up as coenzyme A derivatives, not free fatty acids; hence the signal molecule recognized by FadR is a long-chain acyl-CoA. In the absence of acyl-CoA, FadR represses the operons for fatty acid degradation and also activates *fabA*, a gene involved in fatty acid biosynthesis. When the FadR protein binds acyl-CoA, it no longer binds to DNA. Fatty acid degradation is induced and in addition the level of expression of *fabA* decreases, so fewer fatty acids are manufactured.



Current Opinion in Structural Biology

FIGURE 9.15 ***FadR Structure and Binding***

Overlay of FadR bound to DNA (in blue) and to myristoyl-CoA (in green). Atoms of myristoyl-CoA are shown as spheres. The HTH motif is colored red in the DNA-bound structures. From Huffman & Brennan, Current Opinion in Structural Biology 12 (2002) 98–106.

The same regulatory protein can sometimes turn genes on or off depending on where it binds on the DNA.

lactose. However, lactose, *allo*-lactose, and glyceryl-galactoside are all substrates of β-galactosidase whereas IPTG is not (Fig. 9.13, above).

For induction of the *lac* operon by lactose, low levels of both LacY (transport protein) and LacZ (β-galactosidase) proteins are necessary. The inducer must be transported into the cell before it can bind to LacI and β-galactosidase must convert some of the lactose to *allo*-lactose. In practice, when a gene is "switched off," it is not utterly inactive. Even when the *lac* operon is not induced, occasional mRNA molecules are made and a few molecules of LacY and LacZ are present. The maltose system allows transport and metabolism of maltose and longer oligosaccharides also made of glucose

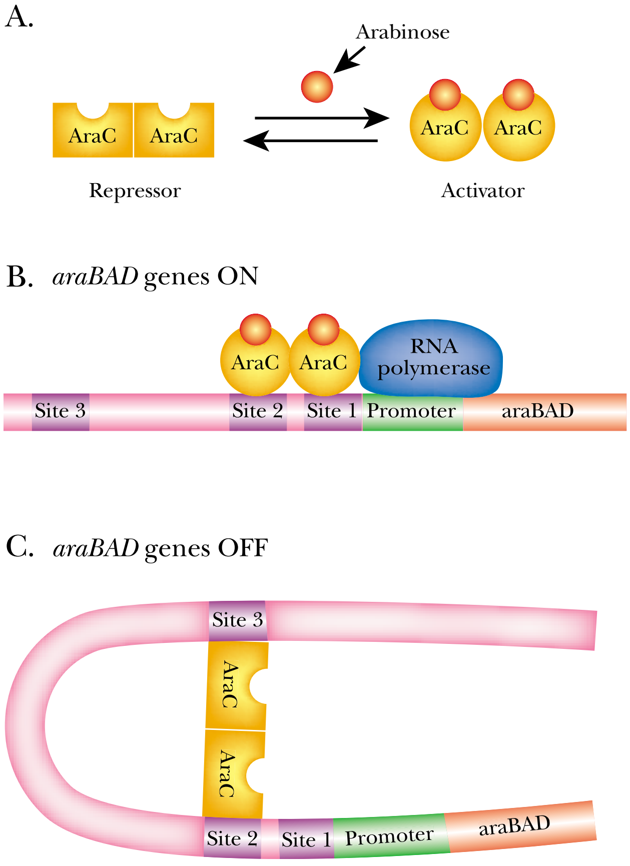


FIGURE 9.16 *AraC Repressor and Activator*

A) AraC dimers are either activators or repressors depending on whether arabinose is bound or not. B) When AraC binds arabinose, the dimer changes configuration and binds to DNA at sites 1 and 2. Here it acts as an activator, allowing the RNA polymerase to bind. C) When the repressor form of AraC binds DNA, it occupies sites 2 and 3, forming a loop in the DNA and causing gene inactivation.

subunits. Again, maltose itself is not the true inducer. The MalT protein actually binds maltotriose, a trisaccharide consisting of three glucose residues.

Some repressors are only active when they bind a small signal molecule called a **co-repressor**. This situation is often found when regulating biosynthetic pathways. If an amino acid, such as tryptophan, is present in the culture medium, then the cell does not need to make it. On the other hand, if the amino acid is not present in sufficient amounts, the pathway for synthesis needs to be turned on. In general, the cell should turn biosynthetic pathways off when their products are present in the medium or have been synthesized in sufficient amounts. Thus biosynthetic pathways respond to the corresponding nutrient. An example is the ArgR repressor of *E. coli*, which binds the amino acid arginine as its co-repressor (Fig. 9.17).

The signal molecule itself is not always small. Sometimes repressors or activators bind other proteins, rather than small metabolites. For example, the Mlc repressor regulates glucose transport and a variety of other genes involved in the uptake and metabolism of monosaccharides. The Mlc protein does not bind glucose, yet responds to its presence indirectly. When glucose enters the cell, it is converted to glucose-6-phosphate by the PtsG transporter. When glucose is absent, phosphate groups accumulate on PtsG protein. Conversely, when glucose is present, PtsG rapidly transfers the phosphate to glucose and most PtsG protein is therefore non-phosphorylated. This form of PtsG binds to Mlc and prevents it from binding to DNA. The inactive Mlc protein is thus found attached to the cell membrane where PtsG is located (Fig. 9.18).

co-repressor In prokaryotes—a small signal molecule needed for some repressor proteins to bind to DNA; in eukaryotes—an accessory protein, often a histone deacetylase, involved in gene repression

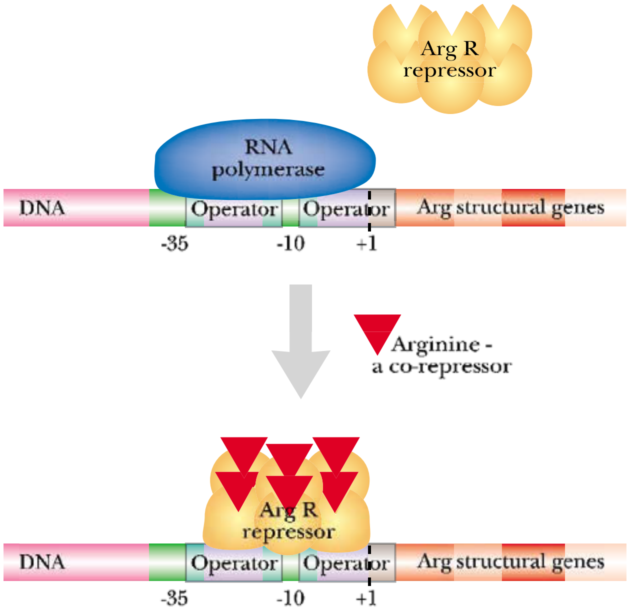


FIGURE 9.17 *ArgR Repressor Uses Arginine as a Co-repressor*

In the absence of high levels of arginine the ArgR repressor cannot bind to DNA. Therefore RNA polymerase is active in transcribing the genes for the synthesis of arginine. When sufficient arginine is present, the arginine acts as a co-repressor by binding to ArgR. The complex then binds to the double operator sites and represses the genes for the synthesis of arginine.

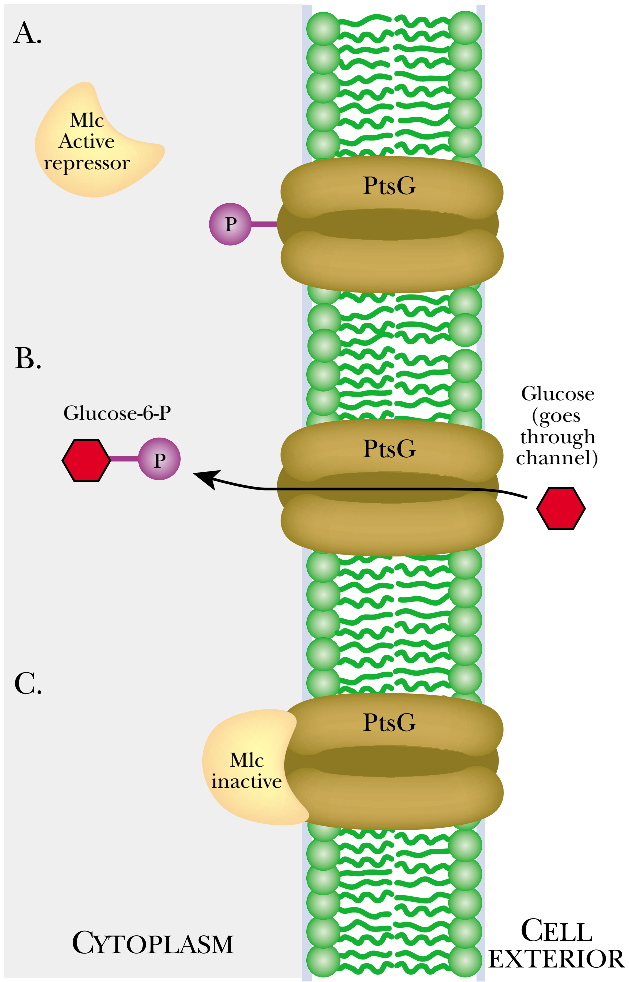


FIGURE 9.18 *Sequestration of Mlc by PtsG Protein*

A) The membrane transporter PtsG has an open channel for glucose when it is phosphorylated (P).
B) Glucose enters the cell and is converted into glucose-6-phosphate so removing the phosphate.
C) The dephosphorylated PtsG protein binds the Mlc repressor. This inactivates the Mlc repressor as it can no longer bind DNA when trapped by PtsG. If the supply of glucose runs out, PtsG will be able to retain its phosphate and Mlc is released.

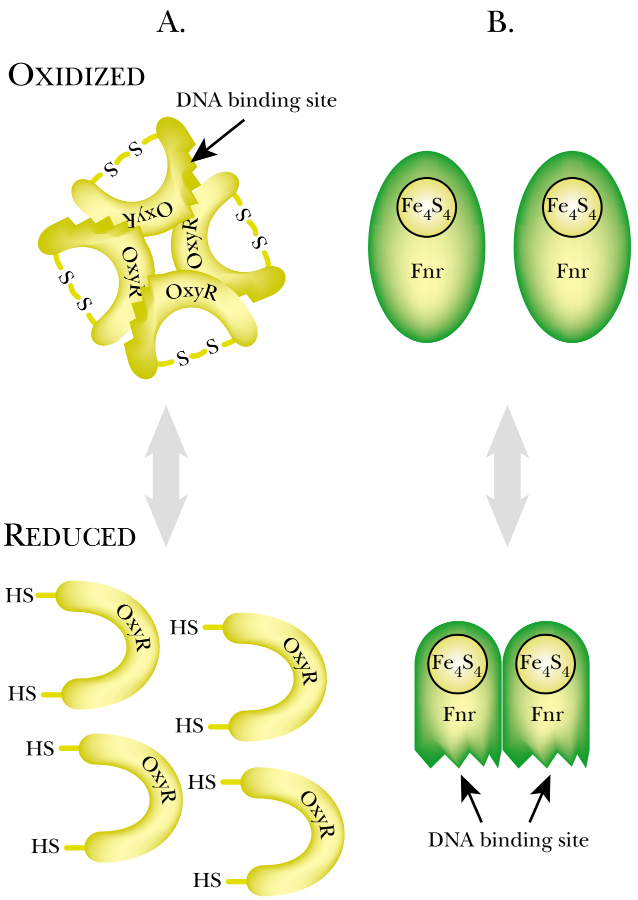


FIGURE 9.19 ***Regulatory Proteins That Respond to Oxidation or Reduction***

A) OxyR changes from the DNA-binding disulfide form to the inactive sulfhydryl form. B) The reduced state of Fnr actively binds DNA, whereas the oxidized state is inactive. Note that the disassembled form of the protein is inactive in both instances.

Removal of Mlc results in derepression of several genes for glucose uptake and metabolism, including *ptsG* itself.

Activators and Repressors May Be Covalently Modified

Some signals consist of chemical alterations to protein molecules.

Some regulatory proteins do not bind a separate independent signal molecule. Instead, some activators and repressors are chemically modified. Most often this is done by the attachment of a chemical group, usually phosphate (see below). Less commonly, the regulatory protein is altered chemically in some other way, for example, by oxidation or reduction.

Examples of bacterial regulatory proteins that are altered by oxidation or reduction are the activators OxyR and Fnr. OxyR is converted to its active form by hydrogen peroxide or related oxidizing agents that oxidize sulfhydryl groups to disulfides (Fig. 9.19A). It then activates a set of genes involved in protecting bacterial cells against oxidative damage.

In contrast, Fnr is inactive when oxidized and becomes an activator when reduced. In this case, an Fe_4S_4 **iron sulfur cluster** in the N-terminal domain of Fnr is reduced under anaerobic conditions. This results in the formation of dimers and a change in shape of the C-terminal DNA-binding domain (Fig. 9.19B). The Fnr activator then activates genes involved in **anaerobic respiration**, such as those for nitrate reductase, fumarate reductase and formate dehydrogenase.

anaerobic respiration Respiration using other oxidizing agents (e.g. nitrate) instead of oxygen
iron sulfur cluster Group of iron and sulfur atoms found in proteins and involved in oxidation/reduction reactions

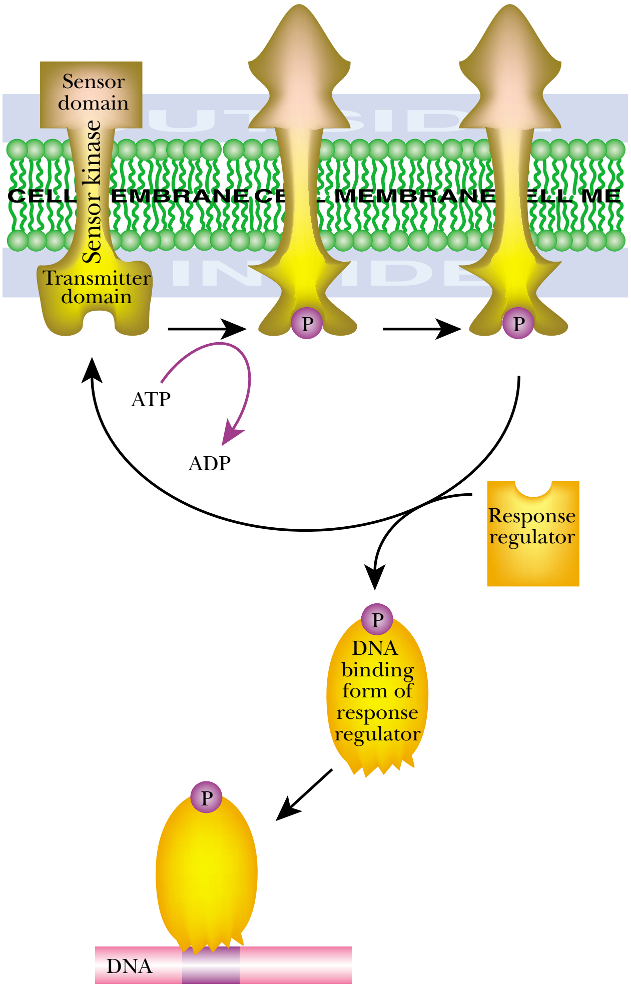


FIGURE 9.20 *Model of Two-Component Regulatory System*

The two-component regulatory system includes a membrane component and a cytoplasmic component. Outside the cell, the sensor domain of the kinase detects an environmental change, which leads to phosphorylation of the transmitter domain. The response regulator protein receives the phosphate group, and as a consequence, changes configuration so as to bind the DNA.

A sensor protein plus a regulator protein often act together as a two component regulatory system.

Two-Component Regulatory Systems

Covalent addition of a chemical group (as opposed to binding an entire signal molecule) is widely used to control the activity of both enzymes and DNA-binding proteins. Phosphate is the most common group used, but methyl, acetyl, AMP and ADP-ribose moieties may also be used.

One large class of regulatory systems that use a phosphate group are the **two-component regulatory systems**. Although often regarded as characteristic of bacteria, they have also been found in lower eukaryotes, including yeast and slime molds. As the name implies, two-component regulatory systems consist of two proteins that co-operate to regulate gene expression. The first component is a DNA-binding regulator protein that only binds DNA when phosphorylated. The second is a trans-membrane **sensor kinase** that senses a change in the environment and changes shape. This causes the sensor kinase to phosphorylate itself using ATP and then transfer the phosphate group to the DNA-binding regulator (Fig. 9.20).

There are many different two-component regulatory systems in *E. coli* (see Table 9.02 for examples). Usually the sensor kinase is a membrane protein that senses either physical conditions of some sort (e.g., aeration, osmotic pressure) or a nutrient (e.g., phosphate, nitrate). The DNA-binding form of the regulator may act as an activator or a repressor. The ArcAB system senses aerobic versus anaerobic conditions. Under

sensor kinase A protein that phosphorylates itself when it senses a specific signal (often an environmental stimulus, but sometimes an internal signal)
two-component regulatory system A regulatory system consisting of two proteins, a sensor kinase and a DNA-binding regulator

TABLE 9.02 Two-Component Regulatory Systems in *E. coli*

Stimulus/Function	Sensor	Regulator
Lack of oxygen	ArcB	ArcA
Osmolarity, envelope proteins	EnvZ	OmpR
Osmolarity, potassium transport	KdpD	KdpE
Phosphate deprivation	PhoR	PhoB
Nitrogen metabolism	NtrB	NtrC
Nitrate respiration	NarX	NarL
Nitrate and nitrite respiration	NarQ	NarP

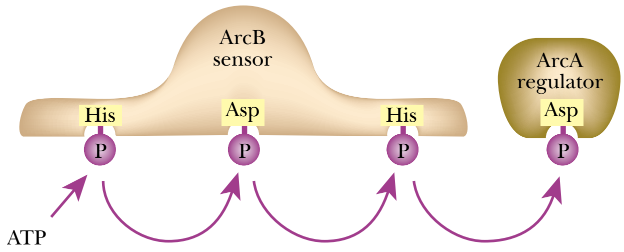


FIGURE 9.21 ***Four Domain Phosphorelay***

The ArcAB two-component regulatory system consists of a membrane sensor, ArcB, and a DNA-binding regulator protein, ArcA. The direction of movement of the phosphate group along the ArcB sensor protein is shown. The phosphate is finally passed from the Arc B sensor to the ArcA regulator.

anaerobic conditions, ArcB phosphorylates itself and then phosphorylates ArcA. The ArcA-P regulator then represses about 20 genes that are only required for aerobic metabolism and activates half a dozen genes needed when oxygen is absent or very low.

Signals are often passed on by adding or removing phosphate groups.

Phosphorelay Systems

The pathway of phosphate transfer in two-component regulatory systems actually involves four protein domains. These domains are highly conserved among different regulatory proteins and are of two types, those where the phosphate is attached to a histidine residue and those where it is carried by an aspartate. In a typical phosphorelay, the phosphate passes from His to Asp to His to Asp. In the case of the ArcAB system, the first three sites are on the ArcB protein and the fourth is on ArcA (Fig. 9.21).

In addition to the two-component regulatory systems, other control systems use phosphorelays. The number of proteins, the total number of phosphate binding domains, and their arrangement varies in different regulatory systems. The regulator at the end of the line may bind DNA upon being phosphorylated or it may activate/deactivate one or more enzymes. In eukaryotic cells, especially in multi-cellular organisms, there are many highly complex signal transmission pathways, which often include one or more phosphorelays.

Global regulators control large families of genes.

Specific versus Global Control

Specific regulation refers to control by a signal specific for a small group of genes. Thus, lactose induces the *lac* operon, maltose induces the *mal* operon, etc. **Global regulators** control large numbers of genes in response to a more general signal or stimulus. Most

global regulator A regulator that controls a large group of genes, generally in response to some stimulus or developmental stage
specific regulation Regulation that applies to a single gene or operon or to a very small number of related genes

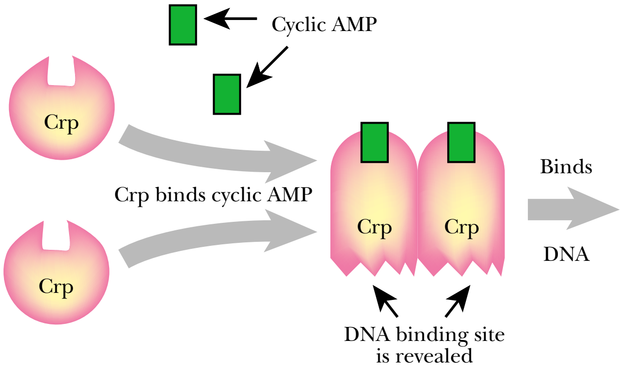


FIGURE 9.22 *Cyclic AMP and the Crp Global Regulator*

Individual Crp units bind cyclic AMP to form a dimer that has the ability to bind DNA.

genes respond to both specific and global signals. Thus, in addition to specific control by the *lac* repressor, the *lac* operon is regulated by the global activator protein, **Crp** (**Cyclic AMP Receptor Protein**). The maltose genes are also regulated by Crp, which regulates the selection among different sugars.

Many bacteria can grow on a wide range of sugars, such as fructose (fruit sugar), lactose (milk sugar) and maltose (from starch breakdown), as well as glucose. When a preferred sugar, such as glucose, is present, less favored sugars, such as fructose, lactose or maltose, are not used. Only when glucose runs out will the other sugars be consumed. In molecular terms, this means the genes for using these other sugars are switched off when glucose is available.

A **regulon** is a group of several genes or operons that are turned on or off in response to the same signal by the same regulatory protein. The members of a regulon have separate promoters and are widely separated on the chromosome. Two examples of regulons in *E. coli* are the genes for using maltose and the genes for the synthesis of arginine. The arginine regulon consists of a dozen genes for biosynthesis and transport scattered over nine locations on the chromosome. They are controlled by a repressor, ArgR, which binds arginine as co-repressor, and is unlinked to any of the genes it controls.

Crp Protein Is an Example of a Global Control Protein

The global regulator Crp binds the signal molecule, cyclic AMP.

Crp is a global activator that is required for switching on the genes for using maltose, lactose and other nutrients less favored than glucose. The Crp protein is allosteric. In order to bind DNA and activate genes, it must first bind its signal molecule, cyclic AMP. When Crp binds cyclic AMP, it forms dimers and these can bind to a recognition site in the DNA upstream of the promoter. The presence of Crp helps RNA polymerase bind to the promoter (Fig. 9.22)

Cyclic AMP is a global signal that the bacterial cell has run out of glucose, its favorite energy source. Only when this has occurred can the genes for using less favored nutrients be switched on. Consequently, in order to switch on genes for using any individual sugar, say, lactose, both an individual signal (the availability of lactose) and a global signal which indicates the need for nutrition (cyclic AMP) are required. Because of this role, Crp has also been called CAP, for catabolite activator protein.

Whether or not the *lac* operon is switched on or off thus depends on the two regulator proteins, LacI and Crp. The various possibilities are illustrated in Figure 9.23. Only when the repressor, LacI, is absent and the Crp protein is present to help it bind can RNA polymerase bind to the promoter and make mRNA.

CRP (cyclic AMP receptor protein) Bacterial protein that binds cyclic AMP and then binds to DNA
regulon A set of genes or operons that are regulated by the same regulatory protein even though they are at different locations on the chromosome

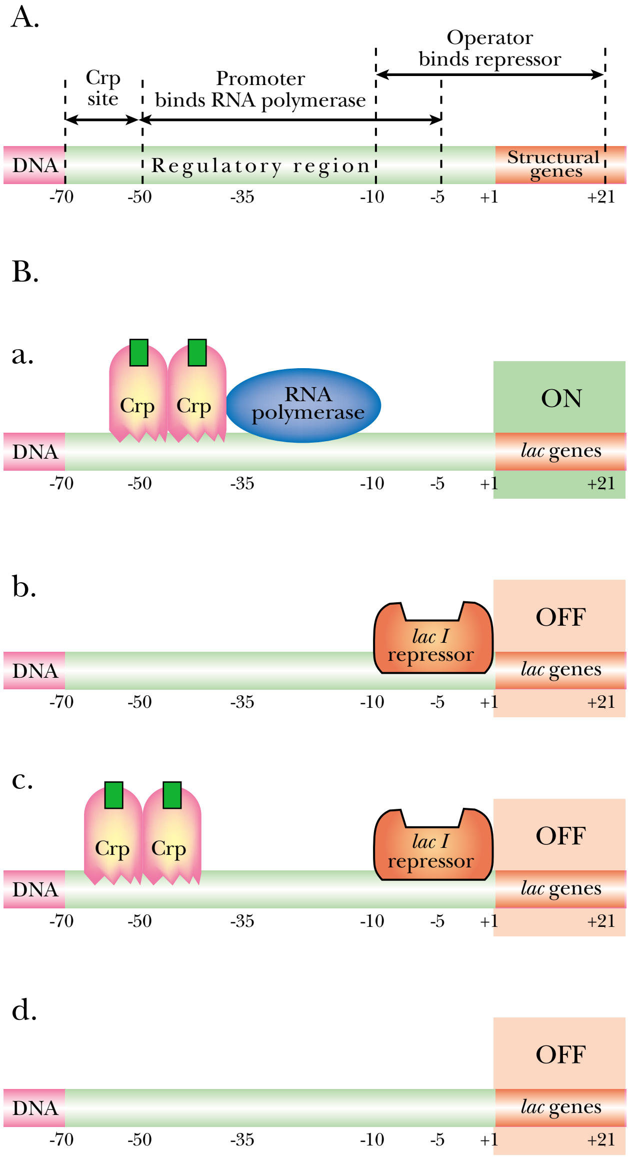


FIGURE 9.23 ***Overall Regulation of the* lac *Operon***

The drawing summarizes the conditions that regulate the *lac* genes. In A, the binding sites for various regulatory components are shown. In B, only the first (a) of the four conditions shown allows RNA polymerase to bind and ensures expression of the genes; the others do not allow gene expression. In scenario (a) glucose is absent, hence CRP (plus cyclic AMP) binds and lactose is present, hence the LacI repressor is removed from the DNA by binding the inducer. In (b) glucose is present, hence CRP is absent and lactose is also absent, hence LacI is still bound. In (c) both glucose and lactose are absent and so although CRP is present, the LacI repressor still blocks transcription. In (d) lactose is present, hence the LacI repressor is removed from the DNA. However, glucose is present, hence CRP is absent and RNA polymerase still cannot bind and transcribe the genes.

Accessory Factors and Nucleoid Binding Proteins

In bacteria such as *E. coli*, there are several rather nonspecific DNA-binding proteins that are sometimes referred to as accessory factors or **histone-like proteins.** They function partly in affecting chromosome structure and partly in gene regulation. Some tend

histone-like protein Bacterial protein that binds nonspecifically to DNA and participates in maintaining the structure of the nucleoid; they do not actually have much in common with true histones

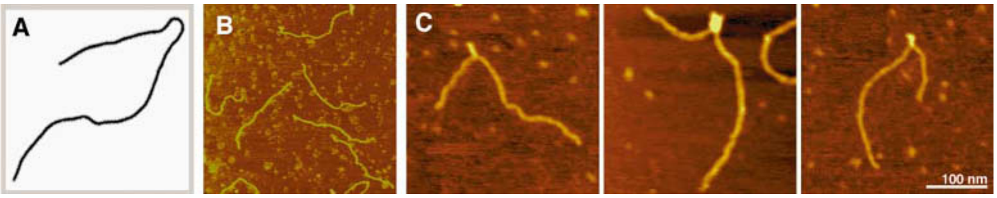


FIGURE 9.24 ***H-NS Binds Preferentially to Curved DNA***

A. Structure of the linear DNA fragment used as predicted by the CURVATURE program. The fragment is shown in the plane of its intrinsic curvature. **B**. Atomic force microscope (AFM) image of naked linear DNA molecules with the curved region at one-third of their length. The image shows an area 900 × 900 nm. **C**. AFM image of DNA after incubation with H-NS (1 monomer per 20 bp). H-NS-DNA complexes are specifically formed at the position of the curved region only. The image shows a 300 × 300 nm surface area. The color scale ranges from 0.0 to 3.0 nm (from dark to bright). From: Structural basis for preferential binding of H-NS to curved DNA Dame RT, Wyman C, Goosen N, Biochimie 83 (2001) 231–234.

to have negative effects on gene expression (e.g., H-NS, StpA), whereas, others usually act in a positive manner (e.g., HU, IHF). However, these regulatory effects tend to be indirect and rather nonspecific.

Bacterial DNA is covered with non-specific binding proteins.

The primary role of the **H-NS (h̲istone-like n̲ucleoid s̲tructuring)** protein of *E. coli* and related bacteria is to maintain the structure of the bacterial nucleoid. The term nucleoid refers to the compact structure formed by the bacterial chromosome together with its accessory proteins. H-NS protein binds in a relatively nonspecific manner, although it prefers regions of bent DNA as illustrated in Fig. 9.24. H-NS consists of two domains, one for DNA-binding and one for protein-protein interaction. These domains are joined by a linker region. H-NS binds to DNA and then the H-NS proteins bind to each other, forming aggregates of four or more H-NS units, thus helping the condensation of the DNA into the nucleoid.

In addition, H-NS binds with higher affinity to the regulatory regions of a wide range of genes scattered throughout the bacterial chromosome. Most of these genes respond to some sort of environmental conditions, but apart from this they are unrelated. The presence of H-NS represses these genes. In contrast to genuine global regulators, H-NS does not control a specific response nor does it respond to any particular signal; therefore, this effect is referred to as **silencing**. Induction of these genes requires specific transcriptional activators to overcome the silencing.

The StpA protein is very similar to H-NS but is present in smaller amounts and appears predominantly under certain stressful conditions (e.g., high temperature or high osmotic pressure). The protein-binding domains of StpA and H-NS bind to each other so that mixed aggregates are formed. StpA silences fewer genes than H-NS. In particular, it allows expression of genes induced by stress. For example, the *proU* gene, expressed at high osmotic pressure, is silenced by H-NS but not by StpA.

Action at a Distance and DNA Looping

The **HU** (**heat-unstable nucleoid protein**) and **IHF** (**integration host factor**) proteins are often required as positive factors in gene expression. Both are **heterodimers**, consisting of two different subunits. The four subunits (two from each protein) are all similar in sequence and 3-D structure and HU and IHF can, to some extent, substitute for each other. HU is relatively nonspecific whereas IHF is more specific. Both HU

H-NS protein (h̲istone-like n̲ucleoid s̲tructuring protein) A bacterial protein that binds nonspecifically to DNA and helps maintain the higher level structure of the nucleoid
heterodimer Dimer composed of two different subunits
HU protein (heat-unstable nucleoid protein) A bacterial protein that binds to DNA with low specificity and is involved in bending of DNA
IHF (integration host factor) A bacterial protein that bends DNA so helping the initiation of transcription of certain genes; named after its role in helping the integration of bacteriophage lambda into the chromosome of *E. coli*
silencing In genetic terminology, refers to switching off genes in a relatively nonspecific manner

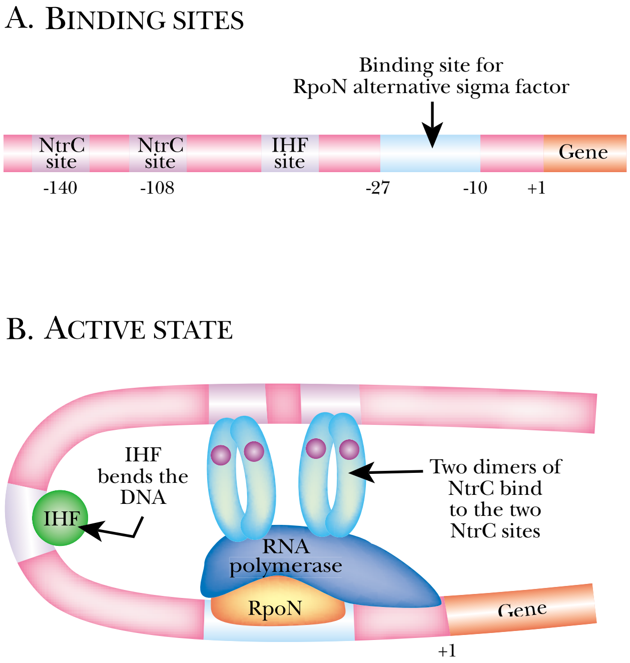


FIGURE 9.25 ***Looping of DNA in RpoN-dependent Promoter***

The sites for binding of transcription factors and the alternative sigma factor, RpoN, are shown in A. The IHF protein induces a bend in the DNA, which brings the NtrC sites close to the binding site for the alternative sigma factor RpoN. Because the DNA loops around, the RNA polymerase can be bound by two sets of NtrC dimers as well as by the RpoN protein.

and IHF are examples of proteins that are involved in bending DNA (as opposed to H-NS, which binds to DNA already bent as a result of its sequence; see Ch. 4). They help in integration, inversion and recombination events by bending DNA into the appropriate conformation (see Ch. 14). They also affect the expression of certain genes that need the DNA in their upstream regions to be bent, in order to be transcribed. A variety of other accessory proteins, activator proteins, and repressor proteins are also involved in the looping of DNA.

DNA may be bent into a loop by some regulatory proteins.

Many genes for nitrogen metabolism require the alternative sigma factor RpoN (= NtrA = σ54) for their transcription. In addition, they are regulated by activator proteins that bind far upstream. Most activator proteins, at least in bacteria, bind just upstream of the promoter and make direct contact with RNA polymerase, so helping it bind to the promoter. For the activators of RpoN-dependent promoters to touch the RNA polymerase, the DNA must be bent around, forming a loop. The bend results from IHF binding between the promoter and activator sites (Fig. 9.25).

Genes for using many alternative nitrogen sources are regulated by the NtrBC two-component regulatory system; although NtrB is not a membrane protein as is typically the case (see above). In the absence of ammonia, the NtrB protein phosphorylates NtrC protein. NtrC-P then binds to the upstream region of nitrogen-source genes and activates transcription. Similarly, the genes for nitrogen fixation, in *Klebsiella* and related bacteria, require the RpoN sigma factor and the activator NifA. In both cases, IHF must bend the DNA into a loop for activation to work.

The enhancers that activate eukaryotic promoters also act at a distance and depend on looping of DNA (see Ch. 10). Because of this similarity, the activator sites of RpoN-dependent promoters have sometimes been called bacteria enhancers. However, the bacterial activator sites are only about 100 bp upstream and occupy a fixed position. In contrast, eukaryotic enhancers may lie several kilobases upstream or downstream of their target genes, and can work in either orientation.

In genetics there are examples of anti-everything—including anti-termination.

Anti-Termination as a Control Mechanism

Anti-termination factors are proteins that prevent termination at specific sites. The RNA polymerase therefore continues on its way and transcribes the region of DNA

anti-termination factor Protein that allows transcription to continue through a transcription terminator

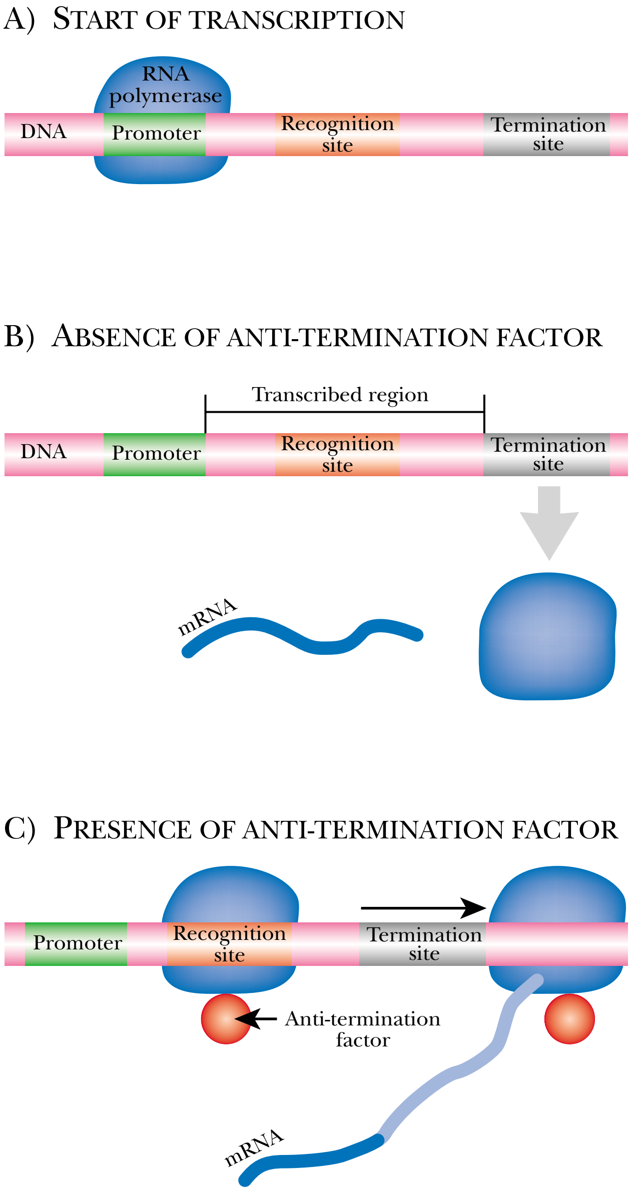


FIGURE 9.26 ***Operation of Anti-Termination Factor***

A) Transcription of the DNA begins with the RNA polymerase bound to the promoter. B) In the absence of an anti-terminator factor, the RNA polymerase reaches the terminator region and drops off, having transcribed a short RNA. C) In the presence of an anti-termination factor, the RNA polymerase binds the factor when it reaches the recognition site. The factor allows the RNA polymerase to transcribe through the termination site.

beyond the terminator (Fig. 9.26). This mechanism for controlling gene expression is common in bacteriophages but is also found for a few bacterial genes.

Anti-termination factors attach themselves to the RNA polymerase before it reaches the terminator. The recognition sequences for anti-termination are found in the DNA well upstream of the terminator. As the RNA polymerase passes by, the anti-termination factors are loaded on. They remain attached and allow the RNA polymerase to travel past the stem and loop region of the terminator without pausing. Consequently, termination is suppressed (Fig. 9.26).

Anti-termination in *E. coli* involves several **Nus proteins**. **NusA protein** is probably attached to the core RNA polymerase shortly after the sigma factor is lost just after initiation. NusA itself actually promotes termination, apparently by increasing the duration of pauses at hairpin structures. NusA and sigma cannot both bind to the core enzyme simultaneously. As long as RNA polymerase is attached to DNA, the Nus proteins cannot be dislodged. However, addition of sigma displaces NusA from free RNA

Nus proteins A family of bacterial proteins involved in termination of transcription and/or in anti-termination
NusA protein A bacterial protein involved in termination of transcription

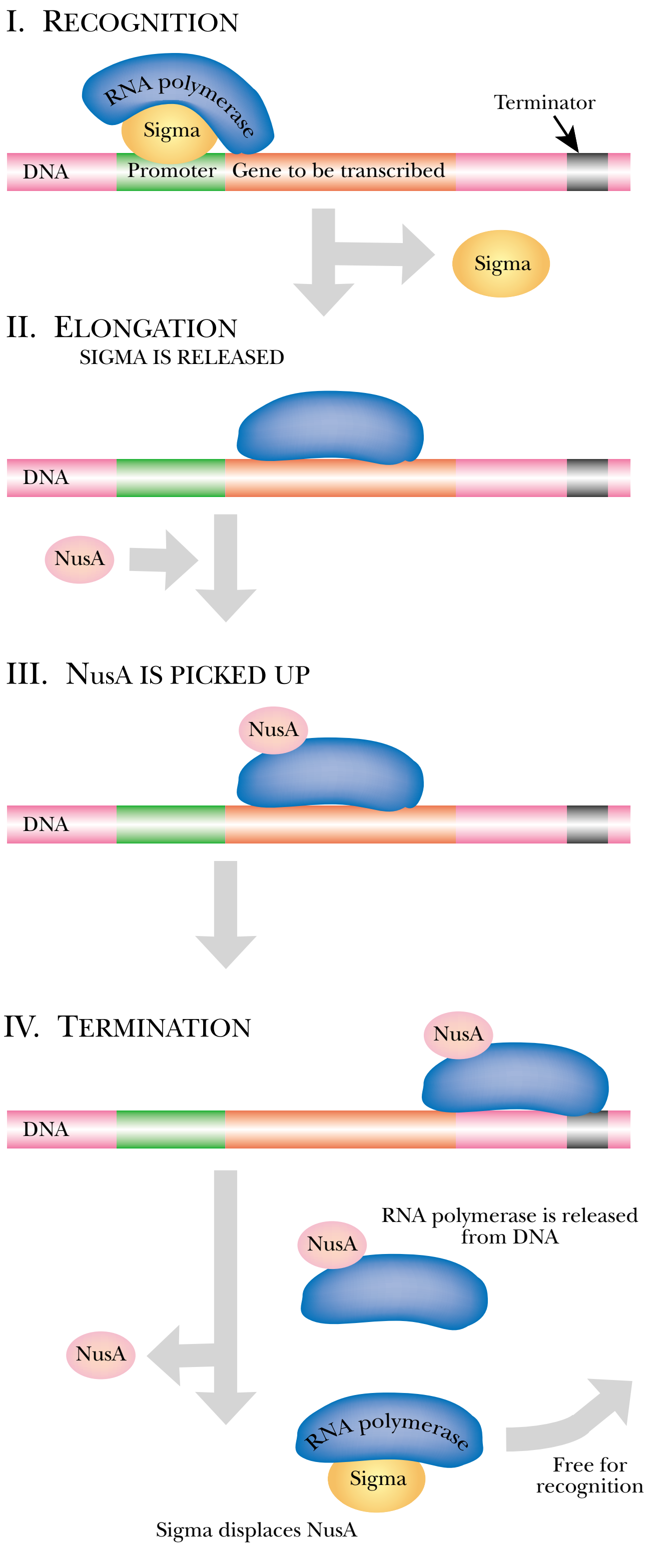


FIGURE 9.27 ***Sigma NusA Cycle of RNA Polymerase***

RNA polymerase needs the sigma subunit to recognize and bind to the promoter. Once RNA polymerase moves forward and starts transcribing, it releases sigma and picks up NusA protein. After termination, NusA protein is displaced by sigma.

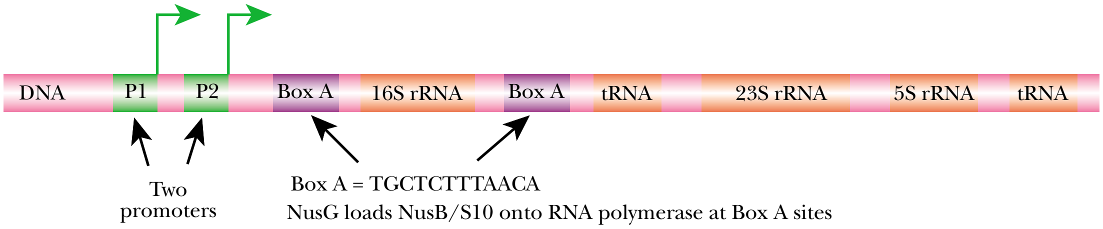


FIGURE 9.28 ***Operation of Anti-Termination in* E. coli rrn *Genes***

As RNA polymerase passes the *boxA* sequences, NusG protein loads NusB plus NusE (= S10 = RpsJ) onto the polymerase. These Nus proteins prevent premature termination.

polymerase (Fig. 9.27). Thus, RNA polymerase cycles between initiation mode (with sigma bound) and termination mode (with NusA bound).

The other Nus proteins are involved in anti-termination. Two of these proteins, NusB plus RpsJ (= NusE), are attached to the RNA polymerase as it passes a "*boxA*" anti-termination sequence (Fig. 9.28). RpsJ (S10) is also found in the small subunit of the ribosome. The connection between its two roles is still obscure. NusG protein probably helps in loading. The presence of NusA is required for NusB plus RpsJ to bind to RNA polymerase. The best known genes in *E. coli* that show anti-termination are the *rrn* genes for synthesis of ribosomal RNA.

CHAPTER TEN

Regulation of Transcription in Eukaryotes

Transcriptional Regulation in Eukaryotes Is More Complex Than in Prokaryotes

Controlling gene expression in eukaryotes is complicated by the high number of genes and the segregation of the chromosomes in the nucleus.

The same principles of transcriptional regulation apply to both prokaryotes and eukaryotes. It is assumed that the concepts in the previous chapter have been understood before continuing into eukaryotic transcriptional regulation. However, the regulation of transcription in eukaryotes, especially multicellular organisms, is more complex than that in prokaryotes. Higher eukaryotes have many more genes than bacteria and regulate their expression differently in different tissues of the body and at different stages of development. For transcription to occur, the DNA must first be exposed. In general, expression of a eukaryotic gene requires the presence of several activators. These may bind to the upstream region of the promoter or to enhancer sequences that may be several kilobases away from the promoter, as described briefly in Ch. 6. Furthermore, eukaryotic enhancers may lie downstream of their target genes and work in either orientation.

Eukaryotic genes are sequestered in the nucleus. Since transcription factors are proteins, they are made by ribosomes in the cytoplasm, but to act they must enter the nucleus. Although both bacterial and eukaryotic DNA are condensed and covered with protein this is much more pronounced in eukaryotic cells. Here the DNA highly condensed into nucleosomes and covered with histones. Long sections of the DNA are frequently folded tightly into heterochromatin and are therefore inaccessible to RNA polymerase (Fig. 10.01). This makes access to the DNA difficult, both for RNA polymerase and transcription factors.

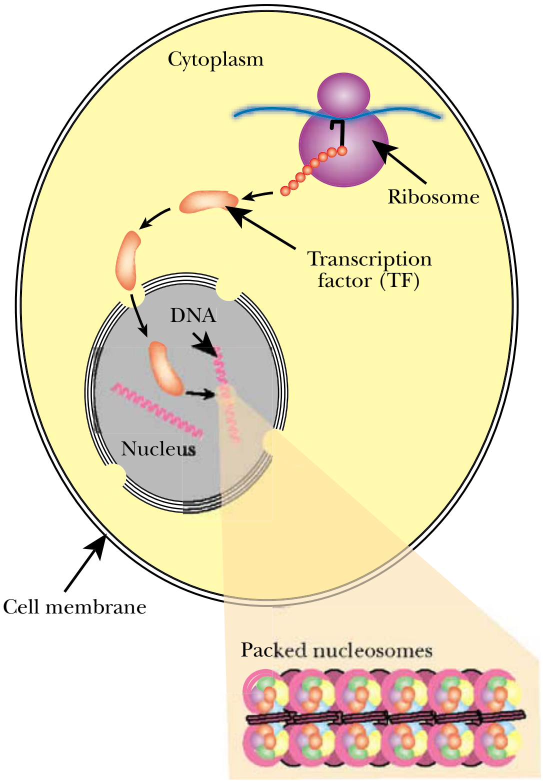


FIGURE 10.01 ***Eukaryotic Genes Are Difficult to Access***

Compartmentalization of the eukaryotic cell into nucleus and cytoplasm means that the transcription factors must be made in the cytoplasm and transported to the nucleus. The DNA in the nucleus is often highly condensed and difficult to access.

Specific Transcription Factors Regulate Protein Encoding Genes

This section deals with the regulation of genes that encode proteins and that are transcribed by RNA polymerase II. As already discussed in Ch. 6, several general transcription factors are required for expression of these genes. Expression also requires specific transcription factors that only affect certain genes in response to specific stimuli or signals. Transcription factors may bind to upstream elements in the promoter region or to enhancer elements that lie far away from the promoter.

Typical specific transcription factors share four general properties:

1. They respond to a stimulus which signals that one or more genes should be turned on.
2. Unlike most proteins, transcription factors are capable of entering the nucleus where the genes reside.
3. They recognize and bind to a specific sequence on the DNA.
4. They also make contact with the transcription apparatus, either directly or indirectly.

Protein coding genes of eukaryotes are regulated by transcription factors that respond to specific signals.

Some DNA-binding proteins may respond directly to a stimulus, something that is often the case in prokaryotes. In contrast, the transcription factors of higher organisms are often separated from the original signal by several intervening steps. Here, the DNA-binding proteins and their effects on transcription will be considered.

Transcription factors usually have at least two domains, one that binds to DNA and another that interacts with the transcription apparatus. This may be illustrated by using artificial hybrid proteins consisting of the DNA-binding domain from one protein plus the activation domain of another (Fig. 10.02). A hybrid with the DNA-binding domain from a bacterial protein plus the activation domain of yeast GAL4 activator will no longer activate transcription from the original yeast promoter. However, it will activate transcription from an artificial promoter into which the recognition sequence for the bacterial protein has been inserted. The bacterial DNA-binding protein used in these experiments was LexA, which is actually a repressor.

The Mediator Complex Transmits Information to RNA Polymerase

The mediator complex combines multiple signals to regulate transcription of genes by RNA polymerase II.

In prokaryotes, sigma factors recognize the promoter and activators generally help RNA polymerase to bind to the promoter. In eukaryotes, recognition and binding to the promoter are both functions of the general transcription factors. Activators in eukaryotes may be viewed as granting RNA polymerase permission to proceed forward from the promoter. Some eukaryotic activators make contact with the general transcription factors TFIIB, TFIID, and TFIIH. However, this is not sufficient to initiate transcription.

The **mediator** is a protein complex that sits on top of RNA polymerase II and provides a site of contact for activators, especially those that are bound at enhancer sequences (Fig. 10.03). The mediator consists of about 20 protein subunits and receives signals from activators. Apparently, it combines the signals from multiple activators and/or repressors and sends the final result to the RNA polymerase II enzyme. Some subunits of the mediator act in a positive manner while others act in a negative manner.

mediator A protein complex that transmits the signal from transcription factors to the RNA polymerase in eukaryotic cells

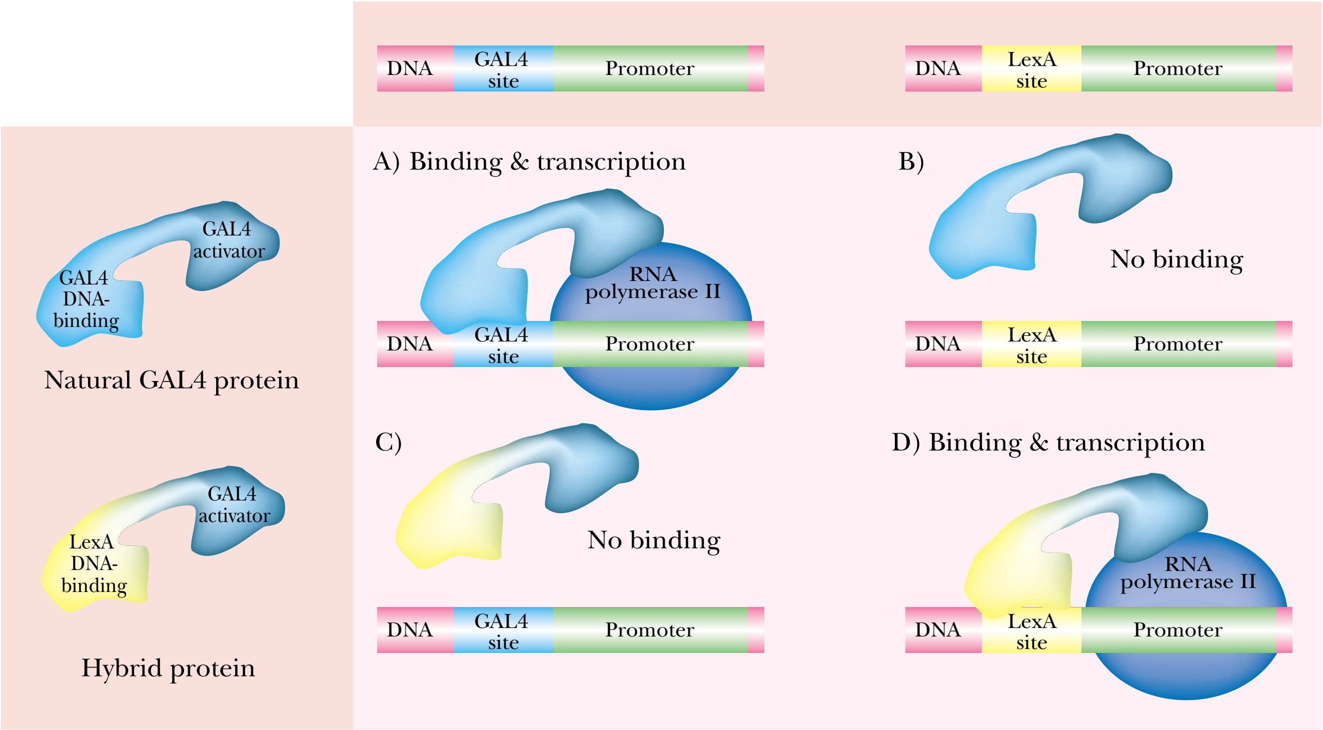


FIGURE 10.02 ***Transcription Factors Have Two Independent Domains***

A) One domain of a GAL4 transcription factor normally binds to the GAL4 DNA recognition sequence and another binds to the transcription apparatus. B) If the LexA sequence is substituted for the GAL4 site, the transcription factor does not recognize the DNA and no binding occurs. C) An artificial protein made by combining a LexA binding domain with a GAL4 activator domain will not recognize the GAL4 binding site, but D) will bind to the LexA recognition sequence and activate transcription. Thus, the GAL4 activator domain acts independently of any particular recognition sequence. It works as long as it is held in close contact to the DNA.

The mediator consists of a constant core that is similar in yeast and higher eukaryotes. Attached to this are other subunits that vary between organisms and also between different tissues within the same organism. Many individual mediator proteins were identified as "co-activator" proteins before it was realized that they belong together in a complex.

Enhancers and Insulator Sequences Segregate DNA Functionally

Enhancer sequences loop around to contact the transcription apparatus.

Enhancers may be found up to several kilobases distant and either upstream or downstream from the promoters they control. This is done by looping of DNA around so that the activator proteins bound at the enhancer can make contact with the transcription apparatus via the mediator complex as discussed above (Fig. 10.03).

Insulator sequences prevent enhancers from interfering with the wrong genes.

This looping mechanism allows a single enhancer to control several genes in its vicinity. But how is an enhancer prevented from activating genes further along the chromosome, that are supposed to be under control of another, closer enhancer? It appears that chromosomes are divided into regulatory neighborhoods by special sequences known as "boundary elements," or **insulators** (Fig. 10.04). An enhancer is prevented from controlling a gene if an insulator sequence lies between them on the chromosome.

insulator A DNA sequence that shields promoters from the action of enhancers and also prevents the spread of heterochromatin

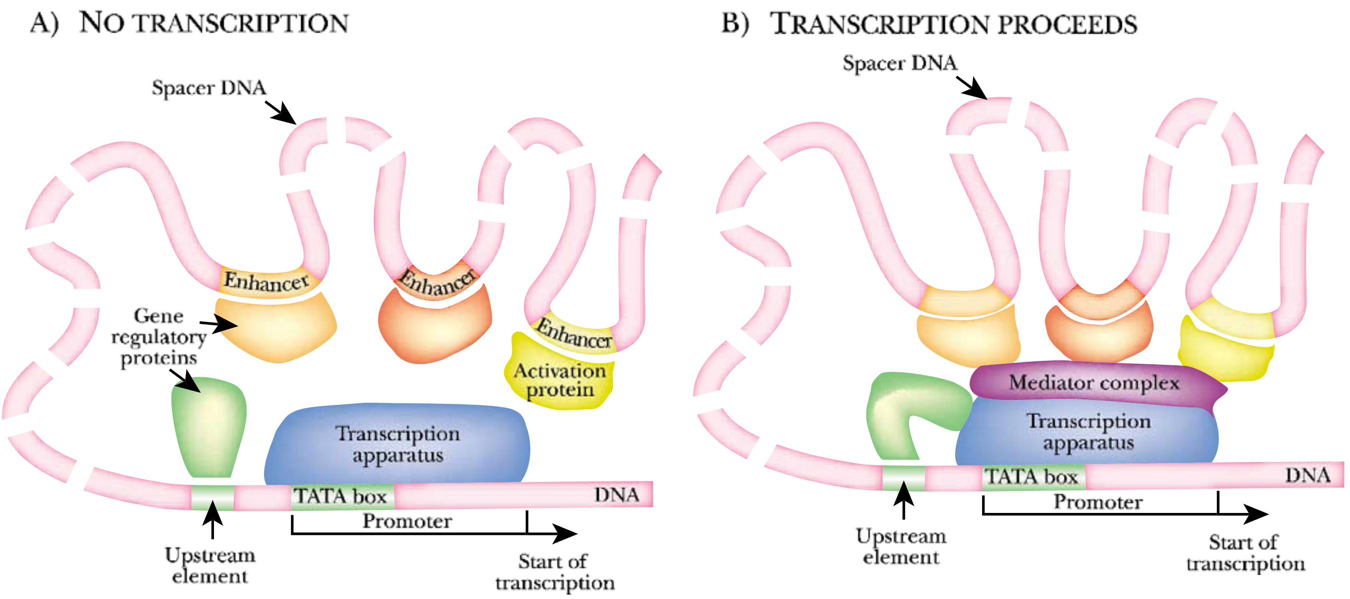


FIGURE 10.03 ***Activator Proteins and the Mediator***

A) The folding of the DNA allows numerous activators that are bound to enhancer sequences to approach the transcription apparatus. B) The mediator complex allows contact of the activators and/or repressors with the DNA polymerase.

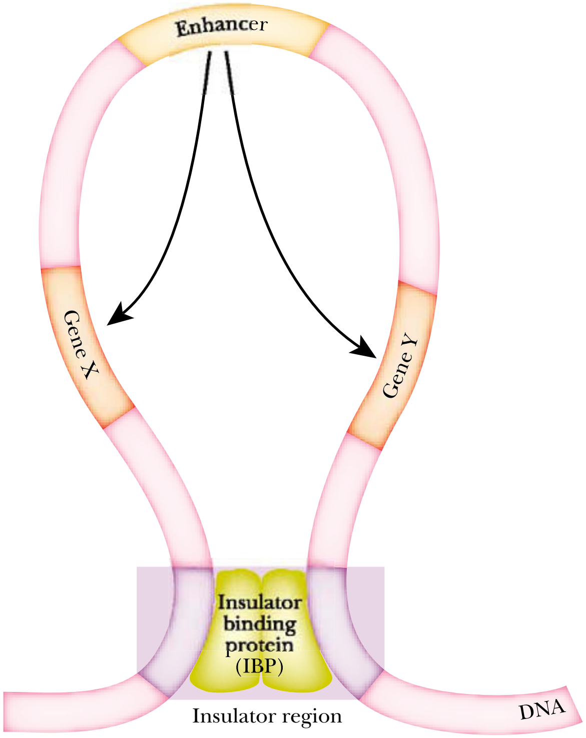


FIGURE 10.04 ***Insulator Sequences Restrict the Range of Enhancer Action***

A large loop of DNA is shown with an enhancer that may interact with Gene X or Gene Y. Insulator sequences at the base of the loop are recognized by an IBP (insulator binding protein) which prevents the enhancer from acting outside of the loop.

Insulators are regions of DNA consisting of clusters of sequences that bind multiple copies of special zinc-finger proteins known as **insulator binding proteins** (**IBPs**). In vertebrates, the best known of these is CTCF (CCCTC-binding Factor). These must bind to the insulator sequences to block enhancer action. [In some organ-

insulator binding protein (IBP) Protein that binds to insulator sequence and is necessary for the insulator to function

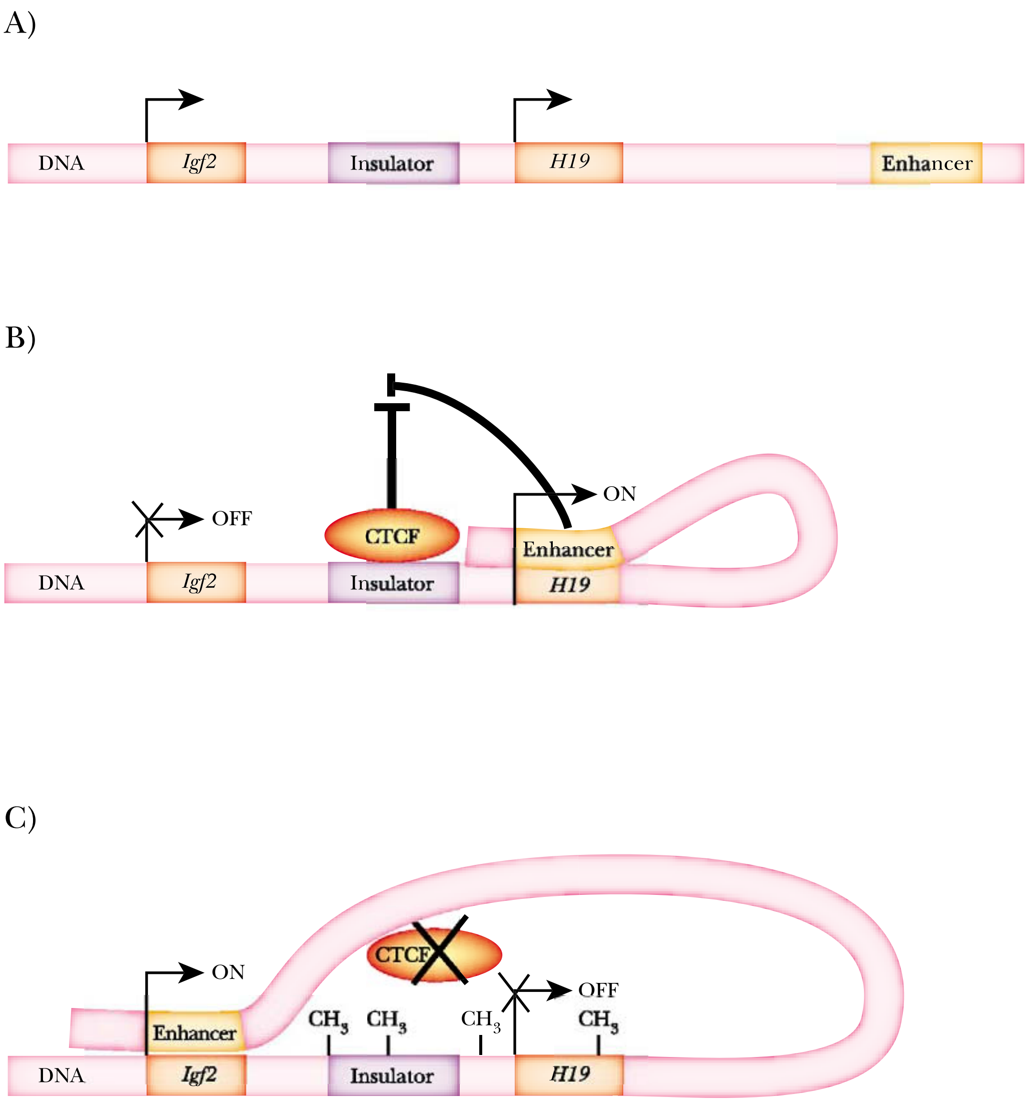


FIGURE 10.05 ***Methylation of Insulator Sequences and Binding***

A) The *Igf2* (insulin-like growth factor-II) gene is distant from an enhancer element. B) When the insulator is not methylated, the CTCF protein binds to the insulator and the enhancer can only affect the *H19* gene. C) When the insulator and the *H19* gene are methylated, the CTCF protein does not bind, allowing the enhancer to activate the *Igf2* gene.

isms, such as the fruit-fly *Drosophila*, there are not only fixed insulator sequences, but also mobile ones. An example is the co-called gypsy element, which is a retrotransposon and can move from place to place within the genome (see Ch. 15 for transposons).]

Insulators can be inactivated by methylating their CG sequences.

In some cases, at least in vertebrates, insulators may be converted between operational or nonoperational forms. Insulators are GC-rich and, as described below, CG sequences may be methylated. When the insulator element is methylated, it no longer binds the CTCF protein and no longer functions (Fig. 10.05). The *Igf2* gene (encoding insulin-like growth factor-II) and the *H19* gene are close together and face in the same direction. The maternal copy of the *Igf2* gene is normally silenced but the paternal copy is active. Conversely, the maternal *H19* gene is normally active whereas the paternal copy is silenced. This is due to differing methylation patterns on the maternal and paternal chromosomes (i.e. an imprinting mechanism—see below).

The reason insulators were also called "boundary elements" is because they form boundaries to regions of heterochromatin. In addition to blocking the action of

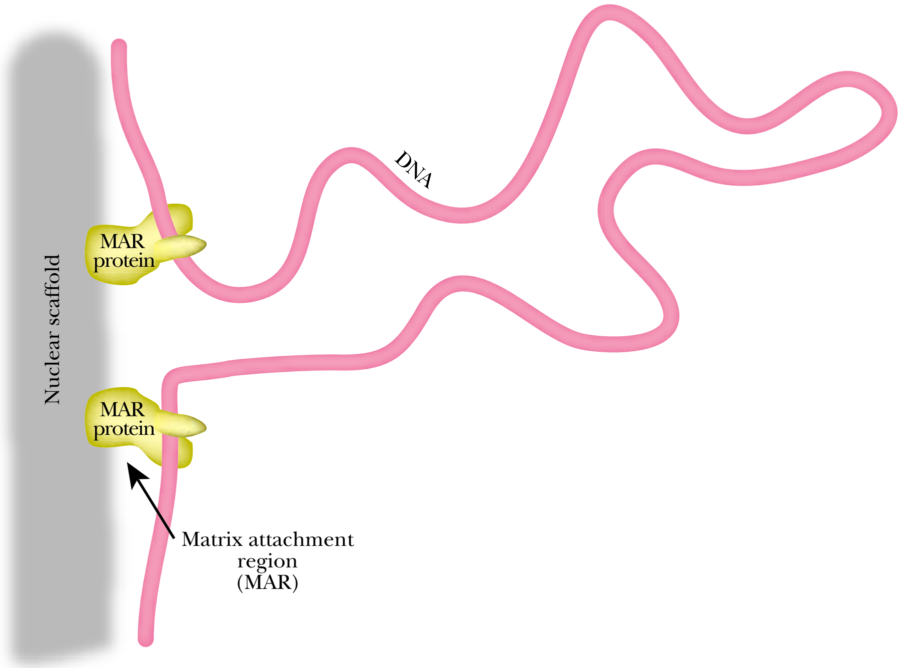


FIGURE 10.06 ***Looped Domains between MAR Sites***

Although many loops are present, only a single loop of histone-free DNA is drawn coming from a region of the nuclear scaffold. The matrix attachment regions contain matrix attachment proteins (MAR protein) that anchor the DNA to the scaffold.

enhancers, insulators also prevent the spread of heterochromatin and the resultant silencing of genes (see below).

Matrix Attachment Regions Allow DNA Looping

DNA forms giant loops that are attached to scaffold proteins by special AT-rich sequences.

Both in bacteria and in eukaryotes, the DNA is arranged in giant loops attached at intervals to the chromosomal scaffold (see Ch. 4). In bacteria the loops consist of about 40 kbp of DNA, whereas the eukaryotic loops are somewhat longer, about 60 kbp.

During interphase, a filamentous web of proteins, the **nuclear matrix**, appears just on the inside of the nuclear membrane. DNA is attached to the proteins of the matrix by sites known as **matrix attachment regions**, or **MARs**. Because the same DNA sites are used for attachment to the chromosomal scaffold during replication as for attachment to the nuclear matrix during interphase, they are sometimes also called **SARs** (**scaffold attachment regions**).

These MAR/SAR sites are 200–1000 bp long and AT-rich (70% AT) but otherwise share no obvious consensus. DNA with multiple runs of A's is inherently bent (see Ch. 4) and the nuclear proteins that bind the MAR sites recognize the bent DNA rather than a specific sequence. Topoisomerase II recognition sites are often found next to MAR sites, implying that the supercoiling of each giant loop is adjusted independently. Enhancers and other regulatory elements are often associated with MAR sites, and, at least in some cases, chromatin remodeling (see below) starts from a MAR site and affects the whole of the chromatin loop (Fig. 10.06).

In transgenic animals and plants, the efficient expression of the transgene is helped by making sure that it lies between two MAR sites. This region of chromatin is then more likely to be opened up for transcription.

matrix attachment region (MAR) Site on eukaryotic DNA that binds to proteins of the nuclear matrix or of the chromosomal scaffold—same as SAR sites
nuclear matrix A mesh of filamentous proteins found on the inside of the nuclear membrane and used in anchoring DNA
scaffold attachment region (SAR) Site on eukaryotic DNA that binds to proteins of the chromosomal scaffold or of the nuclear matrix—same as MAR sites

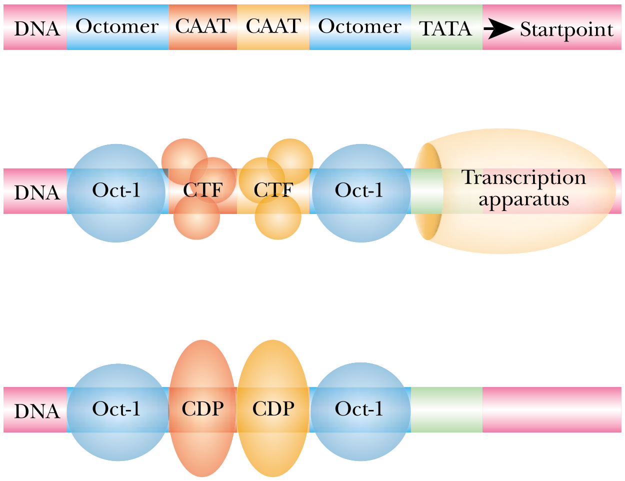


FIGURE 10.07 ***Blocking the CAAT Box in Sea Urchins***

A) Several elements must bind the appropriate transcription factors before transcription occurs.
B) Transcription can be prevented if CAAT displacement protein (CDP) binds to the site that CAAT-binding factor (CTF) normally fills. This prevents assembly of the transcription apparatus and so stops gene expression.

Negative Regulation of Transcription Occurs in Eukaryotes

Simple repressors are common in prokaryotes, but rarely found in eukaryotes. Those examples of repressors that do exist are usually found in simpler, single-celled eukaryotes, such as yeast. Although repressors are rare in eukaryotes, this does not mean that **negative regulation** itself is uncommon. On the contrary, some form of **negative control** is vital to most of the complex regulatory circuits found in higher organisms. Generally, negative signals act by hindering activator proteins in some manner.

Negative regulators in eukaryotes often act by interfering with activators (rather than by obstructing the movement of RNA polymerase).

One obvious way to obstruct an activator is to occupy its recognition site on the DNA and so prevent the activator from binding. An example of this concerns the **CAAT box**, often found in eukaryotic promoters. Activation of the sea urchin *H2B* gene occurs in the testis only and requires, among other things, the binding of the activator protein CTF to the CAAT sequence. However, the CAAT-displacement protein (CDP) may also occupy the CAAT box and prevent binding of the activator. This occurs in embryonic tissue and prevents premature expression of testis-specific genes. The presence of CDP prevents assembly of the transcriptional apparatus. Note, however, that CDP does not block the binding site for RNA polymerase as a classical bacterial repressor would do (Fig. 10.07).

Another example of negative regulation involves the **MyoD** transcription factor, which induces a set of genes specifically needed for formation of muscle cells. MyoD is produced only in cells destined to differentiate into muscle tissue. It is a member of the large class of basic helix-loop-helix (bHLH) proteins. As discussed in Ch. 7, the helix-loop-helix is a widespread motif found in DNA-binding proteins. The basic HLH proteins share a stretch of basic amino acids, located next to the first helix, which helps in binding DNA.

The activity of some transcription factors is controlled by forming mixed dimers with different partners.

HLH proteins bind to DNA as dimers. If both partners have a basic region, the dimer can bind to DNA. Basic HLH proteins usually function as **heterodimers** consisting of a tissue-specific bHLH protein plus one of the widely expressed bHLH proteins known as E-proteins. By itself, MyoD dimerizes poorly. In order to bind DNA, MyoD must form mixed dimers with E12 or E47. These are also basic HLH proteins that are alternative splicing products from the same gene, *E2A*. They are similar in shape and structure to MyoD, but unlike MyoD, they are expressed in all tissues. The

CAAT box A sequence often found in the upstream region of eukaryotic promoters and which binds transcription factors
heterodimer Dimer composed of two different subunits
MyoD A eukaryotic transcription factor that takes part in muscle cell differentiation
negative control See negative regulation
negative regulation Control by a repressor that prevents expression of a gene unless it is somehow removed

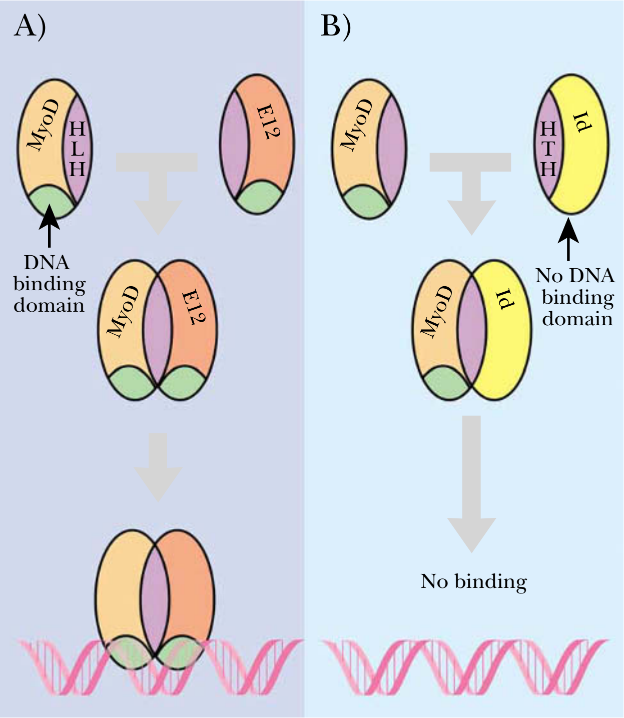


FIGURE 10.08 ***MyoD and Its Alternative Partners***

A) MyoD and E12 both possess basic DNA binding domains. When MyoD dimerizes with E12 the dimer therefore binds to DNA. B) In contrast, Id protein lacks a basic region. When MyoD dimerizes with Id, this dimer cannot bind DNA.

MyoD/E12 or MyoD/E47 heterodimer binds to the DNA sequence CAAATG and activates muscle-specific genes.

Other HLH proteins lack the basic region and cannot bind DNA. An example is the Id protein. This binds to MyoD and E12 or E47. The heterodimers formed by Id cannot bind DNA (Fig. 10.08). Thus, the presence of Id protein *i*nhibits *d*ifferentiation. Id therefore plays a negative role, but without binding to DNA like a true repressor. Id protein is present in large amounts in precursor myoblasts, where it plays a role in restraining MyoD activity. During myoblast differentiation, the level of Id falls and this allows the activation of MyoD.

Heterochromatin Causes Difficulty for Access to DNA in Eukaryotes

In bacteria, the DNA is freely accessible to RNA polymerase and regulatory proteins. However, in eukaryotes, the DNA is coiled around the histones, forming nucleosomes, as discussed in Ch. 4. The nucleosomes are wound into a helix and held close together largely by interactions due to the histone proteins. Densely packaged DNA is referred to as **heterochromatin** and cannot be transcribed, because the RNA polymerase cannot gain access to the promoters. Generally, visible DNA in electron micrographs is heterochromatin, whereas DNA that is not dense is called euchromatin (Fig. 10.09).

The linker histone, H1, has two arms extending from its central spherical domain. The central part of H1 binds to its own nucleosome and the two arms are thought to bind to the nucleosomes on either side, however the exact arrangement is uncertain. The histones of the nucleosome core (H2A, H2B, H3 and H4) have a body of about 80 amino acids and a tail of 20 amino acids at the N-terminal end that faces outwards from the core. Interactions due to these tails are believed to be important in nucleosome aggregation and higher level folding of the chromatin.

Acetylation of histones controls access of regulatory proteins to the DNA.

The histone tails contain several lysine residues that may have acetyl groups added or removed. All four of the core histones may be acetylated, although H3 and H4 are

heterochromatin A highly condensed form of chromatin that cannot be transcribed because it cannot be accessed by RNA polymerase

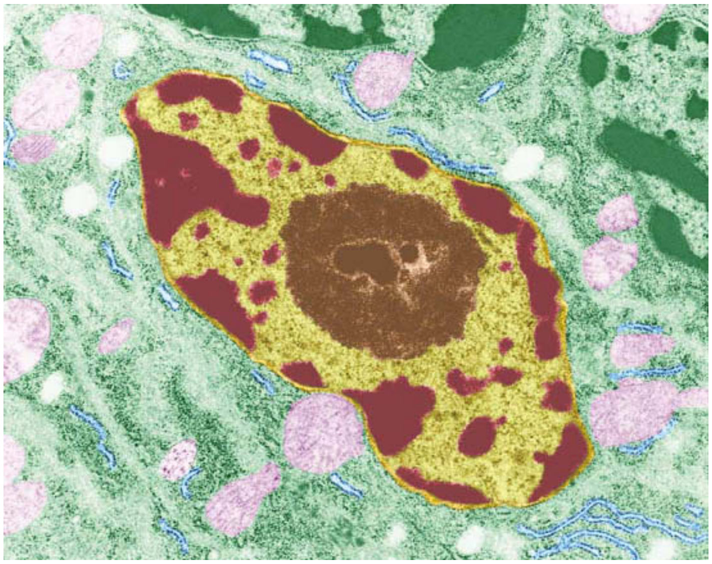


FIGURE 10.09
Heterochromatin Versus Euchromatin

The nucleus shown contains regions of densely packed heterochromatin and less densely packed euchromatin regions. This electron micrograph is of a neuroglial cell nucleus from an insect nervous system, magnified by 3980. Note the nucleolus (brown) and condensed DNA (dark red) in the nucleus. Mitochondria (pink) can be seen in the surrounding cytoplasm. Copyright Dennis Kunkel.

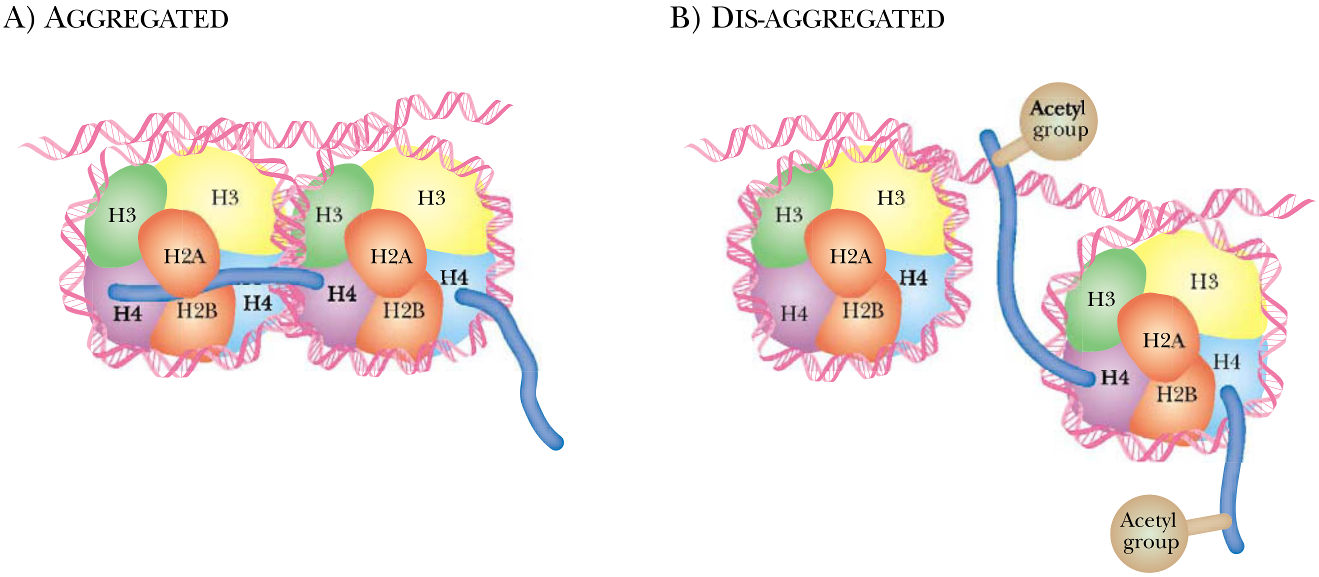


FIGURE 10.10 ***Acetylation of Histone Tails Disaggregates Nucleosomes***

A) Closely packed nucleosomes are stabilized by binding of histone tails to histones in the next nucleosome. B) When the tail of H4 is acetylated, it no longer binds to histones in an adjacent nucleosome. This promotes disaggregation of neighboring nucleosomes. [Histone H1 binds to the linker DNA between the nucleosomes but is not shown in this figure for the sake of clarity.]

most often modified. The degree of **acetylation** affects the state of nucleosome aggregation and therefore of gene expression. Non-acetylated histones form highly condensed heterochromatin, whereas acetylated histones form less condensed chromatin. Note that the nucleosomes themselves are not disassembled by acetylation, but their clustering is loosened up (Fig. 10.10).

Enzymes known as **histone acetyl transferases** (**HATs**) add acetyl groups and **histone deacetylases** (**HDACs**) remove them. Several proteins previously known as co-activators are actually HATs. Examples include the human CBP and p300 proteins

acetylation Addition of an acetyl (CH_3CO) group
histone acetyl transferase (HAT) Enzyme that adds acetyl groups to histones
histone deacetylase (HDAC) Enzyme that removes acetyl groups from histones

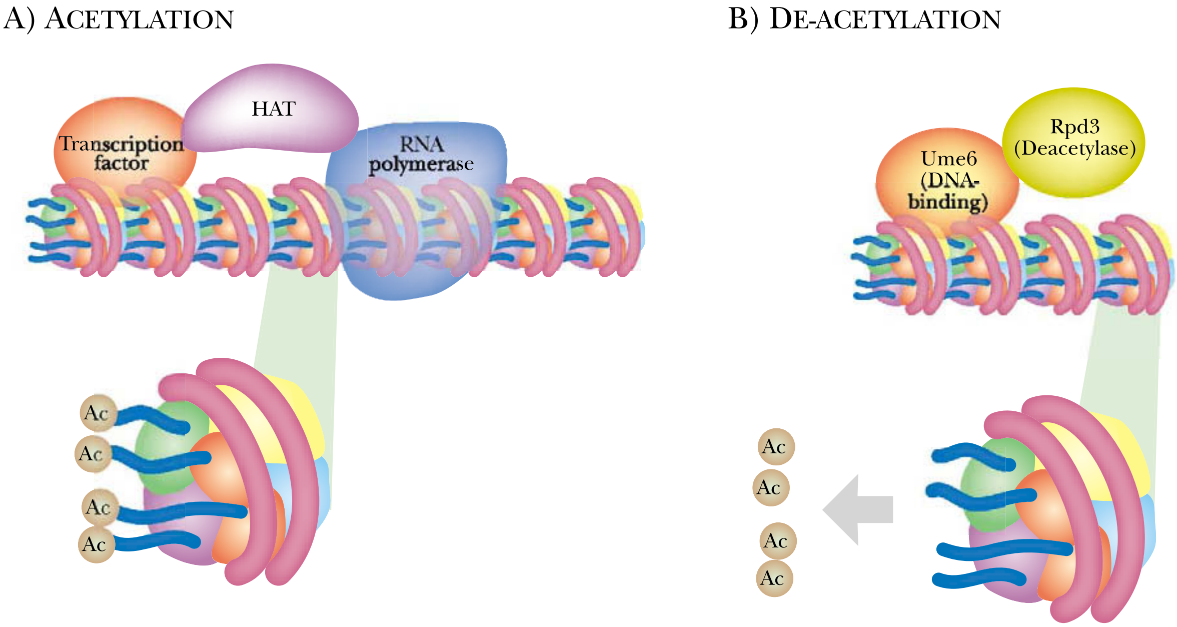


FIGURE 10.11 ***Acetylation and Deacetylation of Histones***

A) Acetylation of histone tails is performed by co-activators known as histone acetyl transferases (HATs). B) Deacetylation of histone tails is due to a repressor complex containing both a DNA-binding subunit and a deacetylase.

involved in cell cycle control and differentiation. Similarly, several so-called **co-repressor** proteins are histone deacetylases. Co-activators and co-repressors do not bind to the DNA, itself, but bind to transcription factors that have already bound to the DNA (Fig. 10.11).

Access to eukaryotic DNA involves moving or restructuring the nucleosomes.

In addition to disaggregating the nucleosomes by acetylation, a further step is needed to provide access to the DNA itself. This is performed by **chromatin remodeling complexes**. These carry out two main types of remodeling. Firstly, they can slide nucleosomes along a DNA molecule, so exposing sequences for transcription. Secondly, they are able to rearrange the histones, so remodeling nucleosomes into a looser structure that allows access to the DNA. ATP is used to provide energy for this remodeling.

There are two families of chromatin remodeling complexes. The larger **Swi/Snf** ("switch sniff") complexes consist of eight to 12 proteins and bind to DNA strongly (Fig. 10.12). Swi/Snf can both slide and remodel nucleosomes. Apparently, Swi/Snf merges two nucleosomes into a new, looser structure. The smaller **ISWI** ("imitation switch") complexes contain two to six polypeptides and can slide nucleosomes but cannot rearrange them. They bind to histones rather than to DNA (Fig. 10.12). (The Swi factors were named after the switching of mating type; Snf factors refer to sucrose nonfermenting mutants. Both were found first in yeast.)

Binding of the chromatin remodeling complexes by transcription factors targets them to the stretch of DNA that needs opening up. The precise order in which transcription factors, histone acetyl transferases and chromatin remodeling complexes bind appears to vary from promoter to promoter. A generalized overall sequence of events for the activation of a eukaryotic gene is as follows:

chromatin remodeling complex A protein assembly that rearranges the histones of chromatin in order to allow transcription
co-repressor In prokaryotes—a small signal molecule needed for some repressor proteins to bind to DNA; in eukaryotes—an accessory protein, often a histone deacetylase, involved in gene repression
ISWI ("imitation switch") complex Smaller type of chromatin remodeling complex
Swi/Snf ("switch sniff") complex Larger type of chromatin remodeling complex

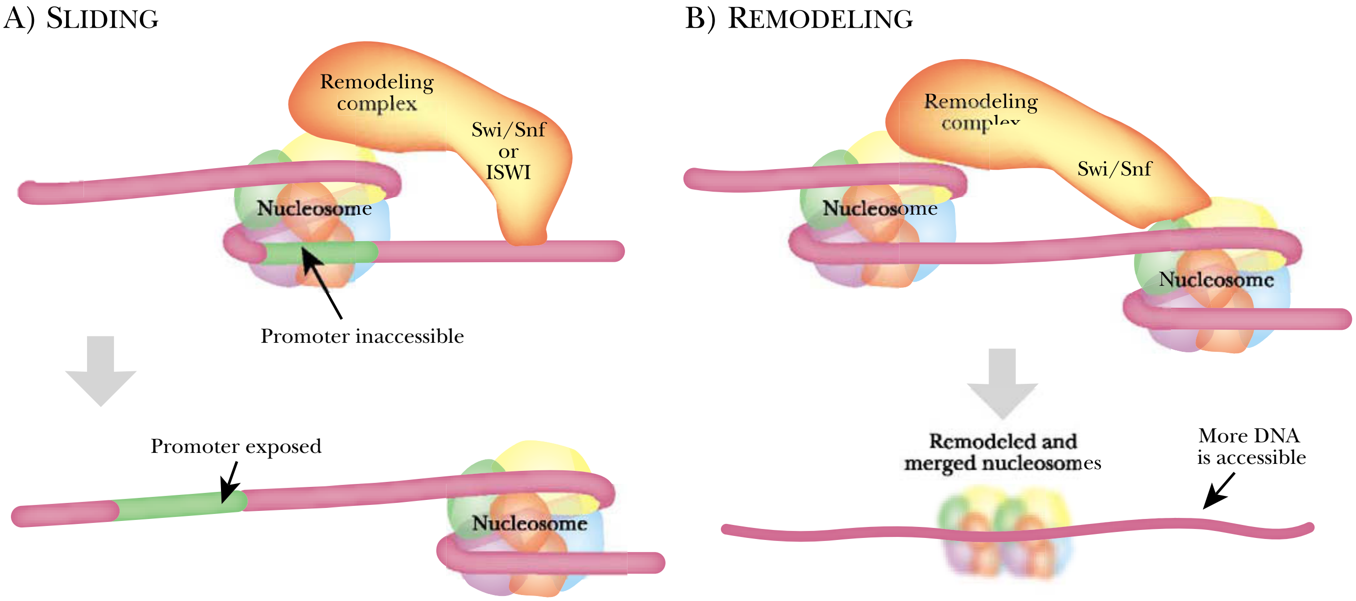


FIGURE 10.12 ***Sliding and Remodeling of Nucleosomes***

A) Sliding of the nucleosome relative to the DNA exposes a previously inaccessible promoter. B) A remodeling complex such as Swi/Snf can merge two nucleosomes and loosen the winding of DNA, making more of the DNA accessible.

1. A transcription factor binds to the DNA.
2. A histone acetyl transferase binds to the transcription factor.
3. The HAT acetylates the histones in the vicinity and the association of the nucleosomes is loosened.
4. The chromatin remodeling complex slides or rearranges the nucleosomes, allowing binding access to the DNA.
5. Further transcription factors bind.
6. RNA polymerase binds to the DNA.
7. Initiation requires a positive signal to be transmitted via the mediator complex from one or more specific transcription factors.

Consider, for example, the yeast *HO* gene, which encodes an endonuclease required for the switching of mating type in yeast. First, the Swi5p transcription factor binds. Then the Swi/Snf complex binds to Swi5p. Next to arrive is SAGA, a histone acetyl transferase, which depends for its binding on the presence of Swi/Snf. The histones in the promoter region are then acetylated. This allows another transcription factor, SBF, to bind. This then allows the general transcription factors to bind, followed by the RNA polymerase (Fig. 10.13).

Methylation of DNA in Eukaryotes Controls Gene Expression

Methylation of DNA is often used to control gene expression during development of higher organisms.

Methylation of bases in DNA occurs in both prokaryotes and eukaryotes, although the purpose is generally quite different. [Prokaryotes use methylation to distinguish newly synthesized DNA as discussed in Ch. 14. In eukaryotes, newly synthesized DNA is recognized by other means that are still unclear.] Nonetheless, many eukaryotes do methylate their DNA as a marker for regulating gene expression.

Methylation of DNA is rare in lower eukaryotes. Higher animals methylate up to 10 percent of their cytosines, and higher plants methylate up to 30 percent. In these

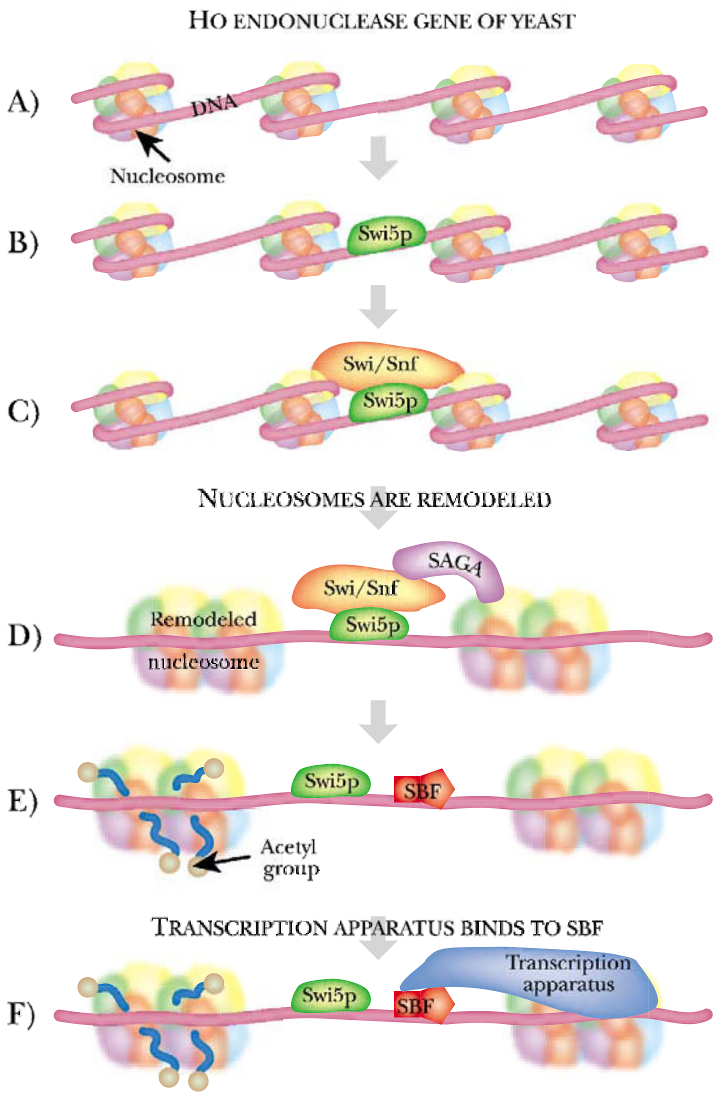


FIGURE 10.13 ***Sequence of Events at the HO Promoter***

A) The HO endonuclease gene of yeast is covered by nucleosomes. B) The transcription factor Swi5p binds to the DNA. C) This is followed by binding of the Swi/Snf complex to Swi5p. D) The remodeled nucleosomes allow binding of an acetyl transferase, SAGA, to the Swi/Snf complex and to a nucleosome. E) As the acetylated histones become less compact, SBF, a transcription factor, binds. F) The transcription apparatus binds.

multi-cellular organisms DNA methylation is often used as a marker for genes whose expression is involved in tissue differentiation. The recognition sequences are extremely short; typically CG for animals and CNG for plants. There are two types of methylases. **Maintenance methylases** add methyl groups to newly made DNA at locations opposite methyl groups on the old, parental DNA strand. This ensures that the pattern of methylation is inherited during chromosome division. Changing the pattern of methylation involves ***de novo*** **methylases** to add new methyl groups and **demethylases** to remove methyl groups (Fig. 10.14).

Methylation in eukaryotes silences gene expression, as discussed below. In animals, about half the genes are located close to **CG-islands** (i.e., clusters of CG sequences). **Housekeeping genes**, which are expressed in all tissues, possess non-methylated CG-islands. In contrast, the CG-islands of tissue specific genes are only non-methylated in those particular tissues where the genes are expressed. The maintenance of the pattern of methylation therefore makes sure that the pattern of gene expression stays constant among the cells of a particular tissue. In plants, certain transposable elements may also be inactivated by methylation.

CG-islands Region of DNA in eukaryotes that contains many clustered CG sequences that are used as targets for cytosine methylation
de novo **methylase** An enzyme that adds methyl groups to wholly nonmethylated sites
demethylase An enzyme that removes methyl groups
housekeeping genes Genes that are switched on all the time because they are needed for essential life functions
maintenance methylase Enzyme that adds a second methyl group to the other DNA strand of half-methylated sites

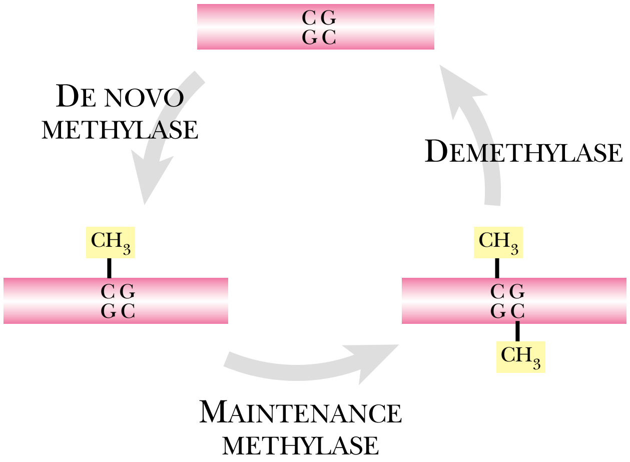


FIGURE 10.14 ***Control of Methylation of DNA in Eukaryotes***

Three enzymes control methylation of DNA. *De novo* methylase adds methyl groups to non-methylated CG-islands. Maintenance methylase adds a second methyl group on the opposite strand of hemi-methylated sites. Demethylase removes methyl groups.

Silencing of Genes Is Caused by DNA Methylation

Silencing is a somewhat vague term that refers to repression of a large number of genes in a relatively nonspecific manner. Limited silencing is known in bacteria (see Ch. 9). However, silencing is widespread in eukaryotes, where it involves the covalent modification of both DNA and of the histones. Silencing may affect a single gene, a cluster of genes, a substantial region of a chromosome or even a whole chromosome. The most spectacular example is the almost total silencing of a whole X-chromosome in mammalian cells (see below).

Genes are silenced by methylation of the DNA followed by removal of acetyl groups from the histones.

Silencing involves the methylation of the cytosine in 5′-CG-3′ sequences of eukaryotic DNA. Occasionally, 5′-CNG-3′ sequences are also methylated. The methyl groups project into the major groove of the DNA and thus hinder the binding of most transcription factors. In addition, methylated CG sequences are recognized by **methylcytosine binding proteins** (**MeCPs**). Bound MeCPs are, in turn, recognized by other proteins that remove acetyl groups from the histones (especially H4). This results in the condensation of the DNA to form heterochromatin that is no longer accessible for transcription (Fig. 10.15). The genes in such regions are said to be "**silenced**."

DNA methylation patterns are reprogrammed when a new zygote is formed.

The pattern of methylation in adult, differentiated cells is duplicated after each round of cell division. This is done by the maintenance methylases that add methyl groups to newly made DNA at locations opposite methyl groups on the old, parental DNA strand. *De novo* methylases and demethylases change the pattern of methylation when necessary, especially during development. Just after fertilization of an egg to form a zygote, the methylation patterns of most of the DNA are erased. New methylation patterns are then laid down in a tissue-specific manner by processes still poorly understood. Those genes whose promoter regions are methylated are silenced.

Genetic Imprinting in Eukaryotes Has Its Basis in DNA Methylation Patterns

A few special genes retain their methylation patterns through fertilization.

Imprinting occurs when methylation patterns from the gametes survive the formation of the zygote and affect gene expression in the new organism. This only applies to very few genes, although examples are known from mammals, fungi and plants. Imprinting

imprinting When the expression of a particular allele depends on whether it originally came from the father or the mother; (imprinting is a rare exception to the normal rules of genetic dominance)
methylcytosine binding protein (MeCP) Proteins in eukaryotes that recognize methylated CG-islands
silencing In genetic terminology, refers to switching off genes in a relatively nonspecific manner

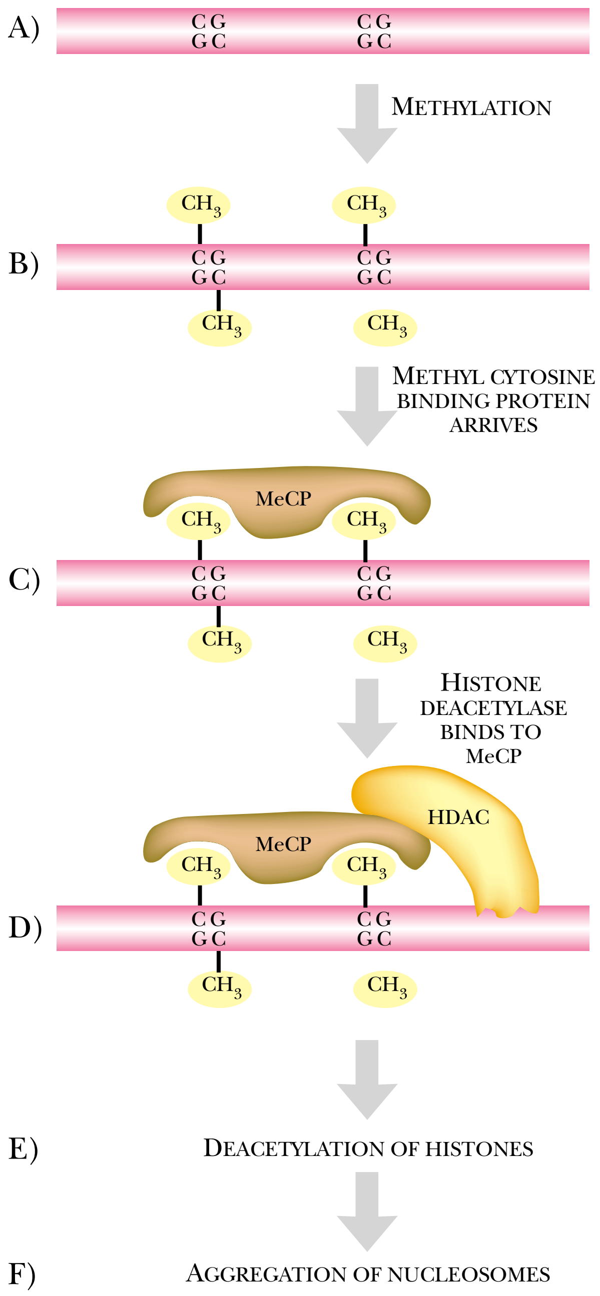

A)
C G
G C
C G
G C
METHYLATION
B)
CH_3
CH_3
CH_3
CH_3
METHYL CYTOSINE BINDING PROTEIN ARRIVES
C)
MeCP
HISTONE DEACETYLASE BINDS TO MeCP
D)
HDAC
E)
DEACETYLATION OF HISTONES
F)
AGGREGATION OF NUCLEOSOMES

FIGURE 10.15 ***Silencing Starts with Methylation***

A) CG regions on the DNA are B) methylated. C) A methyl cytosine binding protein (MeCP) is attracted to the methylated sites and D) histone deacetylase (HDAC) is bound to both MeCP and DNA. E and F) The consequence is deacetylation of the histone tails and aggregation of nucleosomes to form heterochromatin.

is a mechanism to ensure that only one of a pair of genes in a diploid cell is expressed. The second copy is silenced by methylation. The choice as to which allele of a gene to express depends on its parental origin.

Methylation patterns are set up during the formation of the gametes. Most genes in eggs and sperm cells are inactive and therefore silenced by methylation. However, some genes remain active and there are a few differences in methylation patterns between male and female gametes. This causes an initial difference in expression of alleles inherited from the father and mother. As the embryo develops, previous methylation patterns are erased and the genome is reprogrammed for development (Fig. 10.16).

Only about 20 imprinted genes are known in the mouse and these tend to be clustered. For example, IGF-II (insulin-like growth factor II) from the father is expressed, whereas the maternal allele is not. Usually, it is the paternal allele that is expressed in the zygote, but not always. An example of the expression of the maternal allele is IGF-IIR, the receptor for IGF-II.

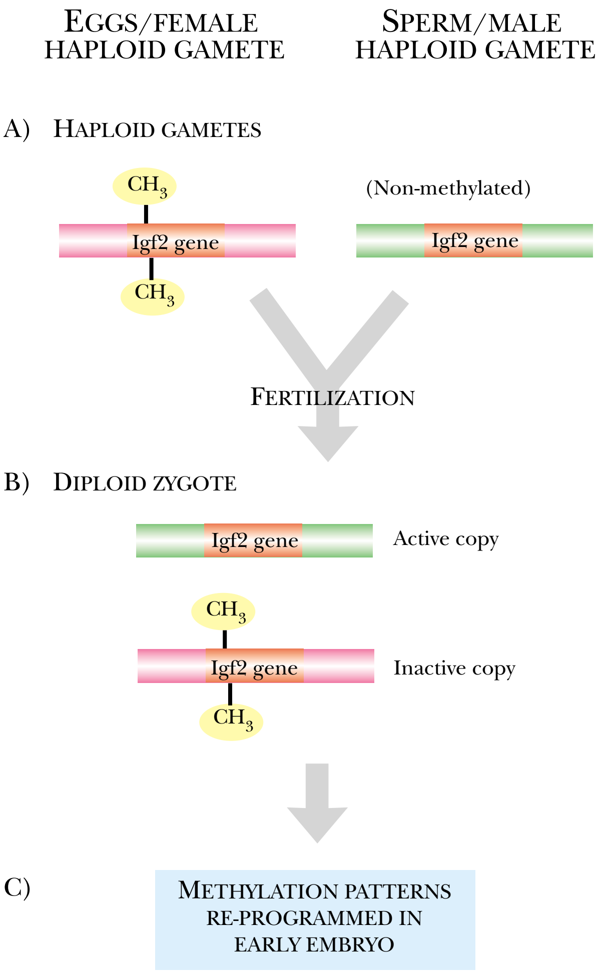


FIGURE 10.16 ***Imprinting and Development***

The two copies of the same gene inherited from each parent are not always methylated in an identical manner. A) Here the *Igf2* gene from the mother has methyl groups and is inactive, whereas the non-methylated gene from the father is active. B) After fertilization, the embryo has one active and one inactive *Igf2* gene. C) Most methylation patterns are reprogrammed in the early embryo but a few survive, resulting in imprinting.

X-Chromosome Inactivation Occurs in Female XX Animals

Only one of the two X chromosomes in a female mammalian cell is active and expresses its genes.

X-inactivation is a special form of imprinting found in animals. Females possess two X chromosomes, whereas males have one X chromosome plus a much shorter Y chromosome. Consequently, females have two copies of most genes carried on the X chromosomes, whereas males only have one. Evolution has developed a variety of mechanisms for gene dosage compensation in order to avoid different levels of gene expression in male and female.

In nematodes, such as *C. elegans*, the expression level of genes on both X chromosomes is halved. Conversely, in insects, such as *Drosophila*, the expression of genes on the single X chromosome in the male is doubled. In mammals, one of the pair of X chromosomes in each female cell is silenced (except for a few loci that are exempted and are referred to as “pseudo-autosomal” regions). In worms and insects, protein complexes that bind specifically to the X chromosomes are responsible for decreasing (worms) or increasing (insects) transcription from genes located on the X chromosomes. In female mammals, a mechanism involving non-coding RNA inactivates just one of the X chromosomes (see below).

In *C. elegans*, there is no Y chromosome and males have a single unpaired X chromosome (this situation is designated XO). Furthermore, XX animals are actually her-

X-inactivation The condensation and complete shutting down of gene expression of one of the two X-chromosomes in cells of female mammals maphrodites and possess both male and female sex organs. Dosage compensation relies on a protein, Sdc2, that is only expressed in XX animals. The Sdc2 protein binds to specific sites on the X chromosomes and the dosage compensation complex, consisting of half a dozen proteins, assembles on Sdc2 and decreases gene expression. The mechanism used by *Drosophila* is essentially a mirror image of that in *C. elegans*. In flies, a protein, Msl2, that is only expressed in XY animals binds to specific sites on the X chromosome. The dosage compensation complex assembles around Msl2 and increases gene expression. The dosage compensation complex of *Drosophila* includes two non-coding RNAs as well as several proteins.

Worms and flies use different mechanisms from mammals to regulate X chromosome expression.

In mammals, X-inactivation is controlled by methylation of the ***Xist* gene**, which is itself located on the X-chromosome. The *Xist* gene of the X-chromosome that remains active is inactivated by methylation. Once established, this methylation pattern is inherited at cell division; thus, the same one of the pair of X-chromosomes will remain active in the daughter cells. The expression of the *Xist* gene is in turn regulated by the antisense RNA, Tsix, which is transcribed from the *Xist* locus, but in the reverse direction. However, the timing of expression of Tsix RNA varies among different mammals and although Tsix appears to be involved in choosing which X chromosome to inactivate, its role is presently unclear.

The DNA of inactivated X chromosomes is highly condensed.

Expression of the *Xist* gene causes the inactivation of the X chromosome that carries it. A long non-translated RNA is transcribed from the *Xist* gene. This Xist RNA coats the inactive X-chromosome. Starting from the *Xist* gene and proceeding along the X chromosome in both directions, the DNA is converted into heterochromatin, a condensed form of DNA that cannot be transcribed (Fig. 10.17). Highly condensed X chromosomes are visible in the cells of female mammals and are called **Barr bodies**, after Murray Barr, who discovered them in 1948. The presence or absence of Barr bodies has sometimes been used to check whether female Olympic athletes are genetic females.

If an active *Xist* gene is inserted into another chromosome, this is only partly inactivated. So another factor(s) is needed to explain X-inactivation. [The X chromosome has twice as many LINE-1 elements (see Ch. 4) per unit length than other chromosomes and it has been suggested that these may somehow promote the binding of Xist RNA. The converse theory argues that more LINE-1 elements have accumulated on the X chromosomes precisely because they are often inactivated!]

Xist RNA is involved in silencing X chromosomes.

The mechanism of Xist-induced silencing only partly understood. After Xist RNA binds, it recruits proteins that are responsible for the actual transcriptional silencing and heterochromatin formation. Changes occur in the histones of the inactive X chromosome. First, histone 3 becomes methylated on Lys7 instead of on Lys4 as in active chromatin. Next, histone H4 loses most of its acetyl groups. Somewhat later, an unusual histone, macroH2A, an H2A variant with an extra C-terminal domain, is found solely on the inactive X chromosome. Finally, methylation of CpG islands occurs along the chromosome. Once silencing has been established, the Xist RNA is no longer required for its maintenance.

In those rare cases where three or more X chromosomes are present in female mammals, only one remains active. Moreover, mice with a single X chromosome (and no Y-chromosome) are healthy and fertile, implying that the second X chromosome is not even necessary. In marsupials, the X-chromosome from the father is always inactivated. In other mammals, the choice is random. Furthermore, which X chromosome is active varies in different cell lines. Consequently, female mammals consist of a genetic mosaic, in which different alleles of genes borne on the X-chromosome are expressed in different regions of tissues. This is illustrated by the variegated coat color seen in female mice that are heterozygous for a coat color mutation in an X-linked gene (Fig. 10.18).

Barr body Inactive and highly condensed X-chromosome as seen in the light microscope
Xist gene A gene that causes the inactivation of the X-chromosome that carries it

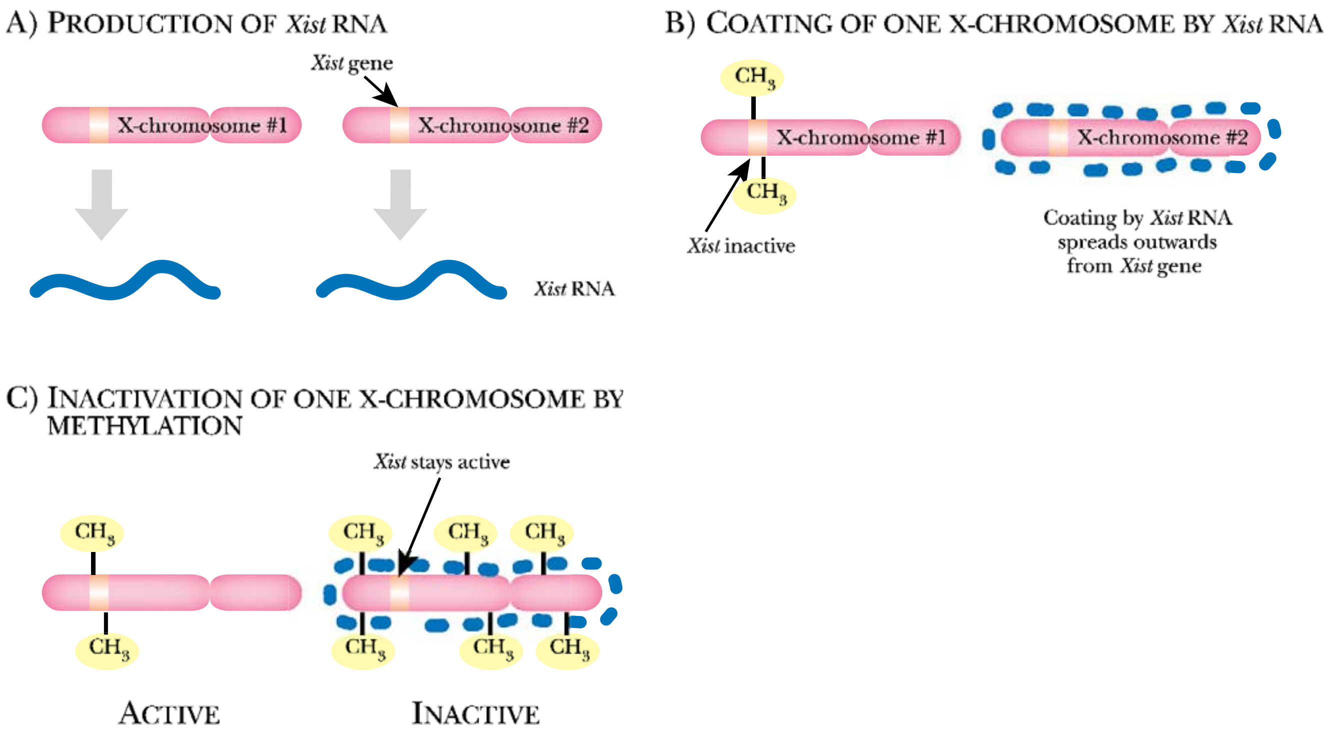

A) PRODUCTION OF Xist RNA
Xist gene
X-chromosome #1
X-chromosome #2
Xist RNA
B) COATING OF ONE X-CHROMOSOME BY Xist RNA
CH3
X-chromosome #1
CH3
Xist inactive
X-chromosome #2
Coating by Xist RNA spreads outwards from Xist gene
C) INACTIVATION OF ONE X-CHROMOSOME BY METHYLATION
Xist stays active
CH3
ACTIVE
INACTIVE

FIGURE 10.17 ***X-inactivation Involves the Xist Gene and Xist RNA***

A) Originally, both X chromosomes transcribe Xist RNA from the *Xist* gene. B) The X chromosome that will remain active is methylated in the *Xist* gene region, which inactivates the *Xist* gene. The Xist RNA coats the other X chromosomes and inactivates it. C) The inactive X chromosome is almost entirely methylated, except for the *Xist* gene (together with a few aberrant loci that are exempted from X-inactivation and are not shown here). This causes the X chromosome to transform largely into heterochromatin. Only the *Xist* gene itself remains active.

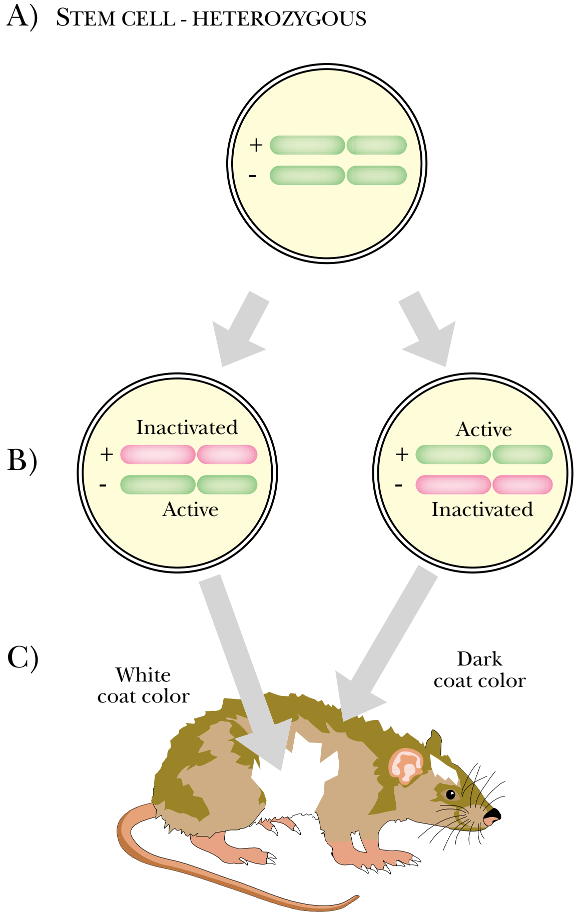


FIGURE 10.18 ***X-Inactivation Causes Skin Patterning in Mice***

A female mouse is shown that is heterozygous for an X-linked gene involved in hair pigmentation. A) All cells contains two X chromosomes, one with a functional copy (+) of the hair color gene and the other with a defective copy (–). B) During development, different X chromosomes are inactivated at random in different ancestral cells. Each ancestral cell divides and gives rise to a patch of cells on the body surface. C) The result is a mixture of zones of different skin colors. The white (mutant) zones appear when the active X chromosome carries the defective hair color gene. The dark zones appear when the active X chromosome carries a wild-type hair color gene.

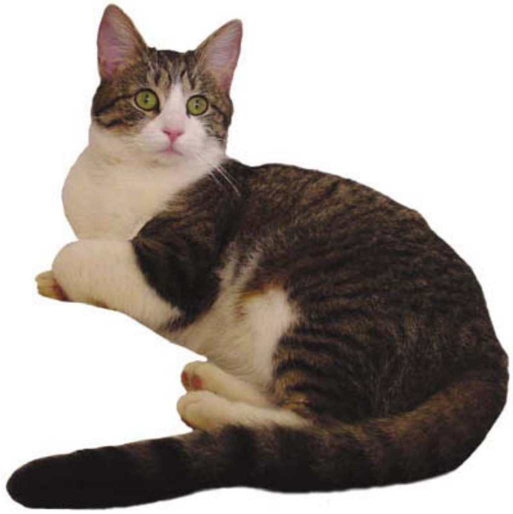


FIGURE 10.19 ***A Calico Cat***

Calico coloration is seen only in female cats. It occurs because genes for fur color are carried on the X chromosomes. If a female cat is heterozygous for mutant and wild-type alleles, random inactivation of the two X chromosomes in different regions creates the pattern. White patches are due to cells where the activated X chromosome contains the mutant allele.

In any given cell, only one X chromosome is active and so only one of the two alleles will be expressed. The descendents of a particular ancestral cell stay together and form regions of skin with the same color. Hence, some regions of the coat are wild-type and others show the mutant color. A similar effect is seen in calico cats (Fig. 10.19).

CHAPTER ELEVEN

Regulation at the RNA Level

Regulation at the Level of RNA

Regulation at the level of translation is rare in bacteria but more common in higher organisms.

Regulation of gene expression at the level of transcription is most efficient in conserving materials and energy whereas regulation of enzyme activity provides the most rapid response. Since regulation at the level of translation is neither the most efficient nor the most rapid, it is consequently less frequent than these other forms of regulation. Generally, once mRNA has been made, it quickly moves to the ribosome where it is translated. This is especially true in prokaryotic organisms where there is no nuclear membrane restricting access between the transcription machinery and the ribosomes. It used to be thought that translational regulation was very rare. Partly this was due to the greater difficulty of measuring translation and associated phenomena such as mRNA stability rather than assaying transcription or protein levels. Consequently, more recent work has revealed a growing number of cases of regulation at the level of translation, especially in eukaryotes. In particular, plants seem to favor translational regulation via the use of small regulatory RNA molecules. Nonetheless, there are still significantly fewer known cases of translational than of transcriptional regulation.

In addition to controlling the translation of mRNA after it has been made, there is also the possibility of aborting the synthesis of messenger RNA after transcription has been initiated and only a short stretch of RNA has been made. This somewhat ambiguous mechanism is referred to as transcriptional attenuation. It is sometimes classified as a form of transcriptional regulation. However, it has been included in this chapter as it is closely related to true translational regulation in the sense that alternative structures of messenger RNA are involved in both cases.

Many of the known cases of translational regulation occur as extra steps in highly complex regulatory cascades that also include regulation at the level of transcription and of protein activity. Examples include the heat shock response in both bacteria and animals and the control of cell growth and differentiation in higher animals. In this chapter, we have attempted to illustrate regulation at the RNA level using examples where other regulatory mechanisms do not overly complicate the issue.

Given a pre-made mRNA molecule, there are several ways in which the translation of the message may be regulated:

1. Control over the rate of degradation of the mRNA
2. Modification of untranslatable RNA to a form that can be translated
3. Control of mRNA translation by regulatory proteins
4. Binding of anti-sense RNA to mRNA to prevent its translation
5. Control of mRNA translation by riboswitches
6. Preferential translation of certain classes of mRNA due to alteration of the ribosome

Binding of Proteins to mRNA Controls The Rate of Degradation

Controlling the rate of messenger RNA destruction is occasionally used to regulate gene expression.

Unlike the DNA that constitutes the cell's genome, mRNA is relatively short-lived. All cells contain a series of **ribonucleases** whose role is to remove mRNA once it has served its function. The half-life of a typical mRNA in bacteria such as *E. coli* is 2–3 minutes. The susceptibility of mRNA to degradation depends on its secondary structure and thus some mRNA molecules are inherently more stable than others. Here we are concerned with cases where the susceptibility of mRNA to degradation is altered in response to regulatory signals. The degradation of mRNA may be hindered or has-

ribonuclease An enzyme that degrades RNA tened by the binding of RNA-specific regulatory proteins. Interpretation of the precise mechanism is often difficult because binding to ribosomes protects mRNA from degradation. Thus mRNA that is not being translated will be degraded faster. There are two main ways in which the binding of a protein might affect mRNA stability. First, it could directly alter susceptibility to ribonuclease attack. Second, the protein might help or hinder binding to the ribosome, which would alter the rate of translation, and affect mRNA stability indirectly.

RNA binding proteins may change the conformation and stability of messenger RNA.

The CsrAB regulatory system of *E. coli* consists of an RNA-binding protein, CsrA, and a **non-coding RNA** molecule, CsrB, which acts as a dock for CsrA. The CsrA protein may either bind to those mRNA molecules that it regulates or to CsrB RNA. Each CsrB RNA can accommodate around 18 CsrA proteins. The CsrAB regulatory system controls the balance between sugar storage as glycogen, which accumulates in bacteria, especially in stationary phase, and the breakdown of sugars by glycolysis (Csr = carbohydrate storage regulator). Overall, CsrA activates glycolysis and represses glucose synthesis and glycogen synthesis.

The CsrAB system activates the *flhDC* operon by stabilizing the mRNA. The primary event is debatable. Binding of CsrA to the mRNA might directly activate translation. Alternatively protection of the mRNA from degradation by ribonucleases would indirectly increase translation. CsrA protein also binds to mRNA carrying genes involved in glycogen synthesis (Fig. 11.01). The binding of CsrA hastens the decay of *glg* mRNA so preventing its translation. Again, there are two ways in which the binding of CsrA might affect mRNA stability. It could directly promote ribonuclease attack or might first inhibit translation, in which case the mRNA will not be protected by the ribosome and will be degraded faster. Deciding experimentally between these two alternatives is not easy.

Some mRNA Molecules Must Be Cleaved Before Translation

In rare cases, the messenger RNA of bacteria must be cleaved to reveal the ribosome binding site.

Eukaryotic mRNA is the result of processing that removes the introns and adds a cap and tail to the **primary transcript** (see Ch. 12 for details). Generally, prokaryotic mRNA is not processed in this manner. The primary transcript, as made by RNA polymerase, is generally the messenger RNA in prokaryotes. However, in a few rare cases, further processing of a prokaryotic mRNA is needed before it can be translated.

In *Escherichia coli*, **ribonuclease III** is one of several ribonucleases that are mostly involved in processing the precursors to tRNA and rRNA. In addition, a few mRNA molecules must be processed by RNase III before they can be translated. The mRNA from the *speF* gene that encodes ornithine decarboxylase is translated about four-fold better if cut by RNase III. However, mRNA from the *adhE* gene, encoding alcohol dehydrogenase, has an absolute requirement for processing by RNase III. In such cases, the original mRNA molecule is folded so that the ribosome binding site and start codon are inaccessible. Cleavage of the mRNA upstream of the ribosome binding site by RNase III frees these sequences for recognition by the ribosome (Fig. 11.02). In *rnc* mutants, which lack RNase III, the *adhE* mRNA cannot be translated and the cells are unable to grow anaerobically by fermentation.

The breakdown of *adhE* message is also controlled by ribonuclease activity. The mature *adhE* mRNA is specifically degraded by ribonuclease G, another ribonuclease normally involved in processing rRNA. In *rng* mutants, which lack RNase G, the half-life of *adhE* mRNA increases from 4 min to 10 min, levels of the mRNA rise and AdhE protein is over-produced.

non-coding RNA Any RNA molecule that is not translated to give protein
primary transcript The original RNA molecule obtained by transcription from a DNA template, before any processing or modification has occurred
ribonuclease III A ribonuclease of bacteria whose main function is processing rRNA and tRNA precursors

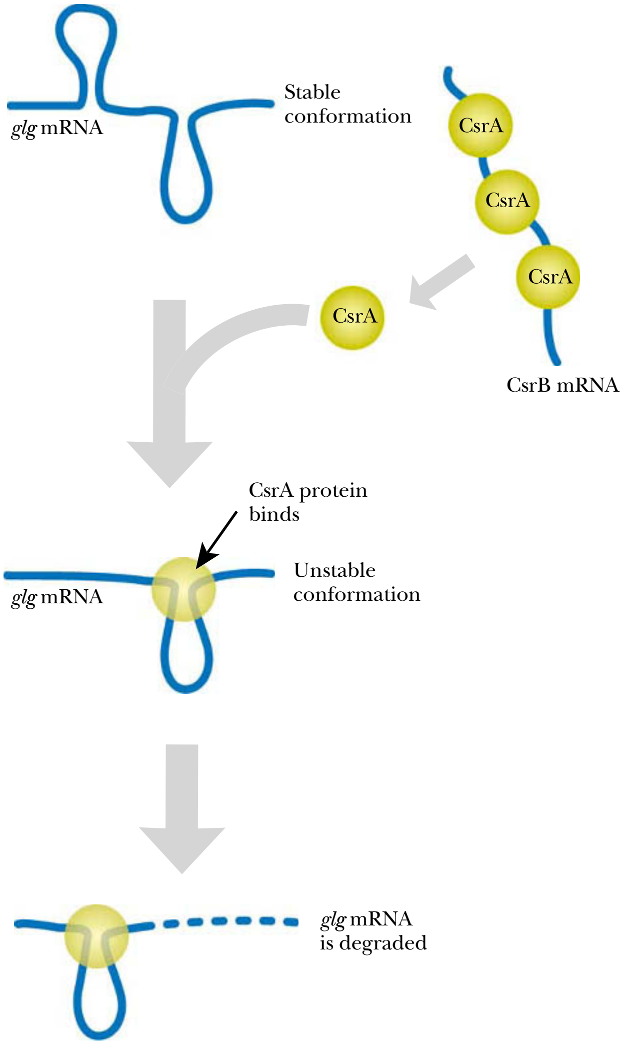


FIGURE 11.01 ***Control of mRNA Degradation by CsrA***

Stability of *glg* mRNA is regulated by CsrA protein. While idle, the RNA-binding protein CsrA is bound to CsrB RNA. When CsrA protein binds to *glg* mRNA, the configuration of the *glg* mRNA is altered to a form that is much more susceptible to degradation.

Some Regulatory Proteins May Cause Translational Repression

Just as regulatory proteins may bind to DNA and either promote or hinder transcription, proteins may also bind to specific sequences on a mRNA and regulate its translation. Examples of both positive and negative translational regulation are known. The response of mRNA to iron is used below as an example of translational repression.

Iron is an essential nutrient and is the cofactor for many proteins, such as cytochromes and hemoglobin. However, free iron generates toxic free radicals and is therefore dangerous. Consequently, surplus iron atoms are stored by **ferritin** and its prokaryotic counterpart, **bacterioferritin**. Ferritin is a hollow spherical protein consisting of 24 subunits. Up to 5,000 iron atoms may be stored as a hydroxyphosphate complex inside this sphere.

The level of ferritin is regulated in response to the iron supply. In plants, ferritin levels are regulated at the level of transcription. However, in both animals and bacteria, ferritin levels depend on translational regulation. When iron is scarce, translation

bacterioferritin The bacterial analog of ferritin, an iron storage protein
ferritin An iron storage protein
translational repression Form of control in which the translation of a messenger RNA is prevented

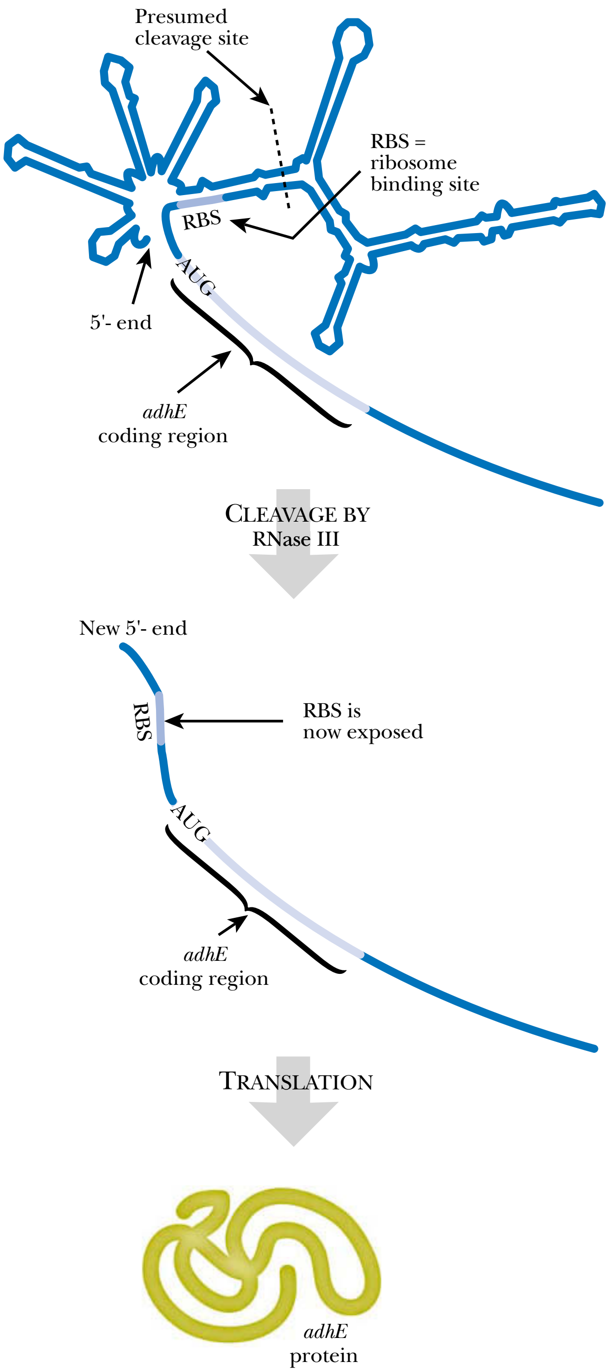


FIGURE 11.02 ***Cleavage of adhE mRNA by RNase III***

The ribosome binding site of *adh*E messenger RNA can not bind to the ribosome due to folding of the pre-mRNA. RNAase III cleaves the *adh*E pre-mRNA so exposing the ribosome binding site.

of ferritin mRNA is reduced. When iron is plentiful, more ferritin is made. In animals, an RNA-binding protein is responsible, whereas in bacteria control is by antisense RNA (see below).

The **5′-untranslated region (5′-UTR)** of ferritin mRNA of animals contains a special recognition sequence known as an **iron-responsive element (IRE)**, which forms a stem loop structure. When iron is scarce, an **iron regulatory protein (IRP)** binds to the IRE stem and loop and prevents translation. Surplus iron results in the detachment of IRP from the ferritin mRNA, which can then be translated (Fig. 11.03).

In animals, translation of genes involved in iron uptake and storage is controlled by an iron regulatory protein that binds to the mRNA.

5′-untranslated region (5′-UTR) The untranslated sequence between the 5′-end of an mRNA and the start codon
iron regulatory protein (IRP) Translational regulator that controls expression of mRNA in animals in response to the level of iron
iron-responsive element (IRE) Site on mRNA where the IRP binds

FIGURE 11.03 ***Regulation of Ferritin mRNA Translation by IRP***

The ferritin mRNA has a loop structure (the iron-responsive element or IRE) in the 5′ untranslated region. When iron is scarce, the iron regulatory protein (IRP) binds preventing the small subunit of the ribosome from binding to the cap structure of the mRNA to initiate translation. When iron is abundant no IRP is bound and translation occurs.

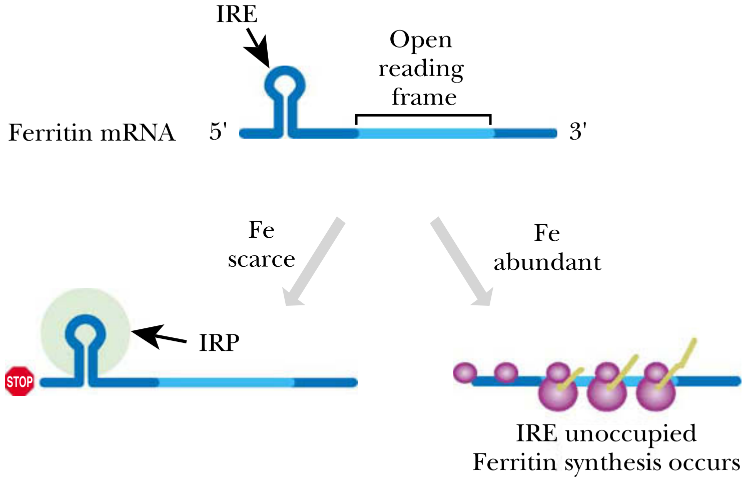
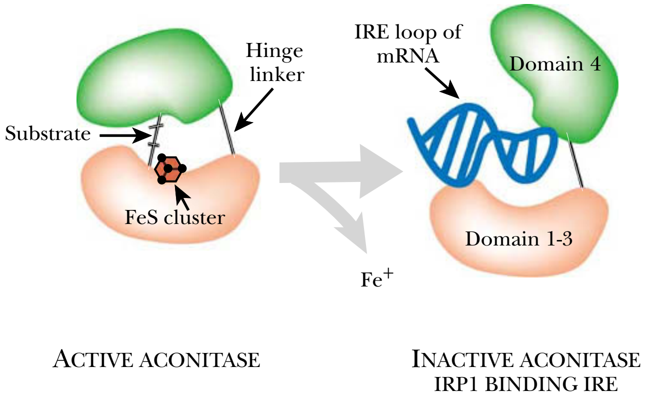


FIGURE 11.04 ***Aconitase Activity versus RNA-binding of IRP1***

The IRP1 protein acts as an enzyme (aconitase) when iron is plentiful. Upon losing an iron atom from the central Fe_4S_4 cluster, during situations of iron scarcity, the two major domains open, allowing binding to RNA. The consequence of binding of IRP1 to mRNA is the blockage of transferrin synthesis.



In animals, the enzyme δ-aminolevulinic acid synthase catalyses the rate limiting step in the pathway for synthesizing the iron-containing cofactor heme. Not surprisingly, its mRNA contains an iron-responsive element and is also under IRP translational control. The receptor for the iron transport protein transferrin, which is found in blood, is also regulated by IRP binding to an IRE on the mRNA, but in this case mRNA stability is affected.

The free iron concentration in the cytoplasm is directly monitored by the iron regulatory proteins. Although there are several IRPs, the major one, IRP1, is identical to the cytoplasmic enzyme, **aconitase**. This enzyme has an **Fe_4S_4 cluster** that is needed for enzyme activity. When iron is scarce, one of the four iron atoms is lost from the Fe_4S_4 cluster. The aconitase/IRP1 then loses enzyme activity and changes conformation, exposing its RNA-binding site. Thus when iron is plentiful, aconitase/IRP1 acts as aconitase (an enzyme in the Krebs Cycle that converts citrate to isocitrate) and when iron is scarce, it acts as an RNA-binding **translational repressor** (Fig. 11.04).

Ribosomal proteins of bacteria regulate their own synthesis by binding to their own messenger RNA.

Regulation by translational repression is also used to control the synthesis of ribosomal proteins in bacteria such as *E. coli*. The ribosomal proteins are grouped in several operons. For each operon, one of the encoded ribosomal proteins binds to the mRNA and so auto-regulates synthesis of all proteins in the operon. These ribosomal proteins bind to rRNA preferentially and only bind to their own mRNA when there is no rRNA available. This mechanism ensures that the amount of rRNA and ribosomal proteins is balanced. If there is an excess of ribosomal protein over rRNA then translation will be decreased (Fig. 11.05).

aconitase An enzyme of the Krebs cycle that, in animals, also acts as an iron regulatory protein
Fe_4S_4 cluster A group of inorganic iron and sulfur atoms found as a cofactor in several proteins
translational repressor A protein that binds to mRNA and prevents its translation

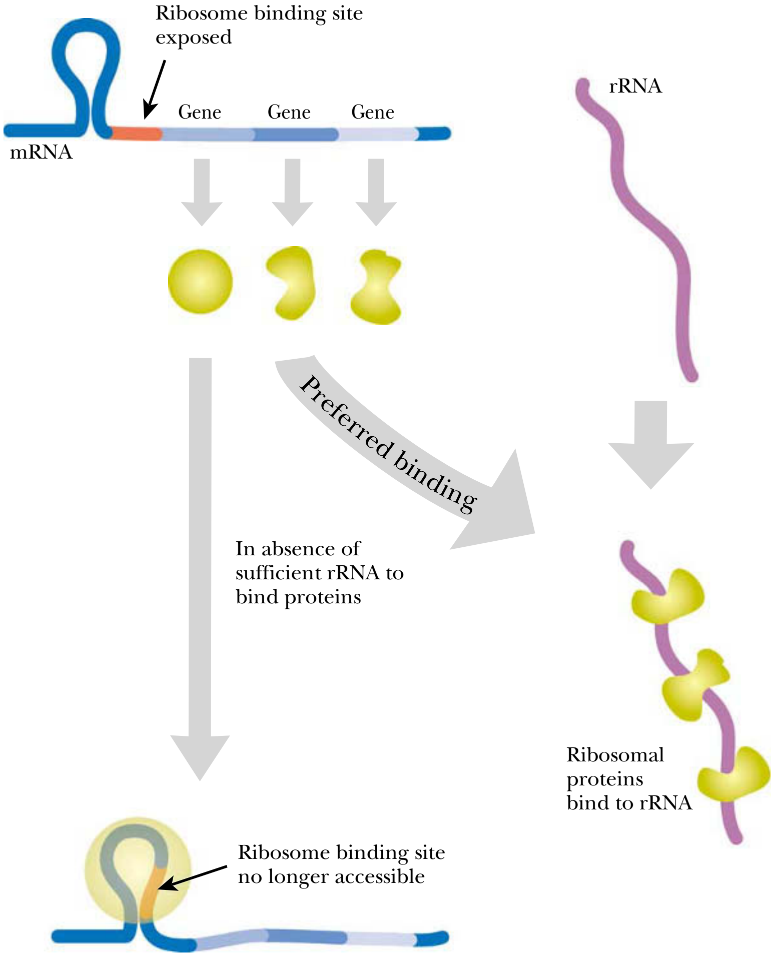

Ribosome binding site exposed
Gene
Gene
Gene
mRNA
rRNA
Preferred binding
In absence of sufficient rRNA to bind proteins
Ribosomal proteins bind to rRNA
Ribosome binding site no longer accessible

FIGURE 11.05 ***Regulation of Synthesis of Ribosomal Proteins in Bacteria***

The mRNA shown at the top of the figure contains a loop, a ribosome binding site (RBS) and structural genes coding for several ribosomal proteins. These proteins prefer to bind rRNA. In situations where there is insufficient rRNA to bind the ribosomal proteins, one of them binds to its own mRNA and alters its configuration so blocking the ribosomal binding site. This prevents synthesis of excess ribosomal proteins.

Some Regulatory Proteins Can Activate Translation

Positive regulation of translation is used to control protein synthesis in chloroplasts after light stimulation. Synthesis of many chloroplast proteins is induced as much as a hundred-fold by light. The levels of some of these proteins are controlled by transcription, others by translation and others by protein degradation.

For example, the mRNA for the large subunit of **Rubisco** accumulates in developing chloroplasts even in the dark. [Rubisco is ribulose bisphosphate carboxylase, a critical enzyme in the fixation of carbon dioxide during photosynthesis. It is the most abundant protein on earth.] Another example is PsbA (= D1 protein) a component of photosystem II. Translation of these mRNAs is controlled by proteins encoded by the nucleus that act as **translational activators**. These proteins bind to an adenine rich region in the 5′-UTR of the mRNA. The activators bind to the mRNA in the light and allow translation. In the dark, they do not bind to the mRNA which cannot be translated due to its unfavorable secondary structure (Fig. 11.06).

The translational activator **cPABP (chloroplast polyadenylate binding protein)** exists in two conformations, only one of which can bind RNA. The interconversion of the two forms of cPABP is controlled by light. Energized electrons from photosystem I are passed down a short electron transport chain to cPABP. The electrons reduce the disulfide form of cPABP to the sulfhydryl form. The reduced sulfhydryl form can bind to RNA and activate translation, whereas, the disulfide form cannot (Fig. 11.07).

cPABP (chloroplast polyadenylate binding protein) A translational activator protein that controls expression of chloroplast mRNA
Rubisco (ribulose bisphosphate carboxylase) A critical enzyme in the fixation of carbon dioxide during photosynthesis
translational activator A protein that binds to mRNA and promotes its translation

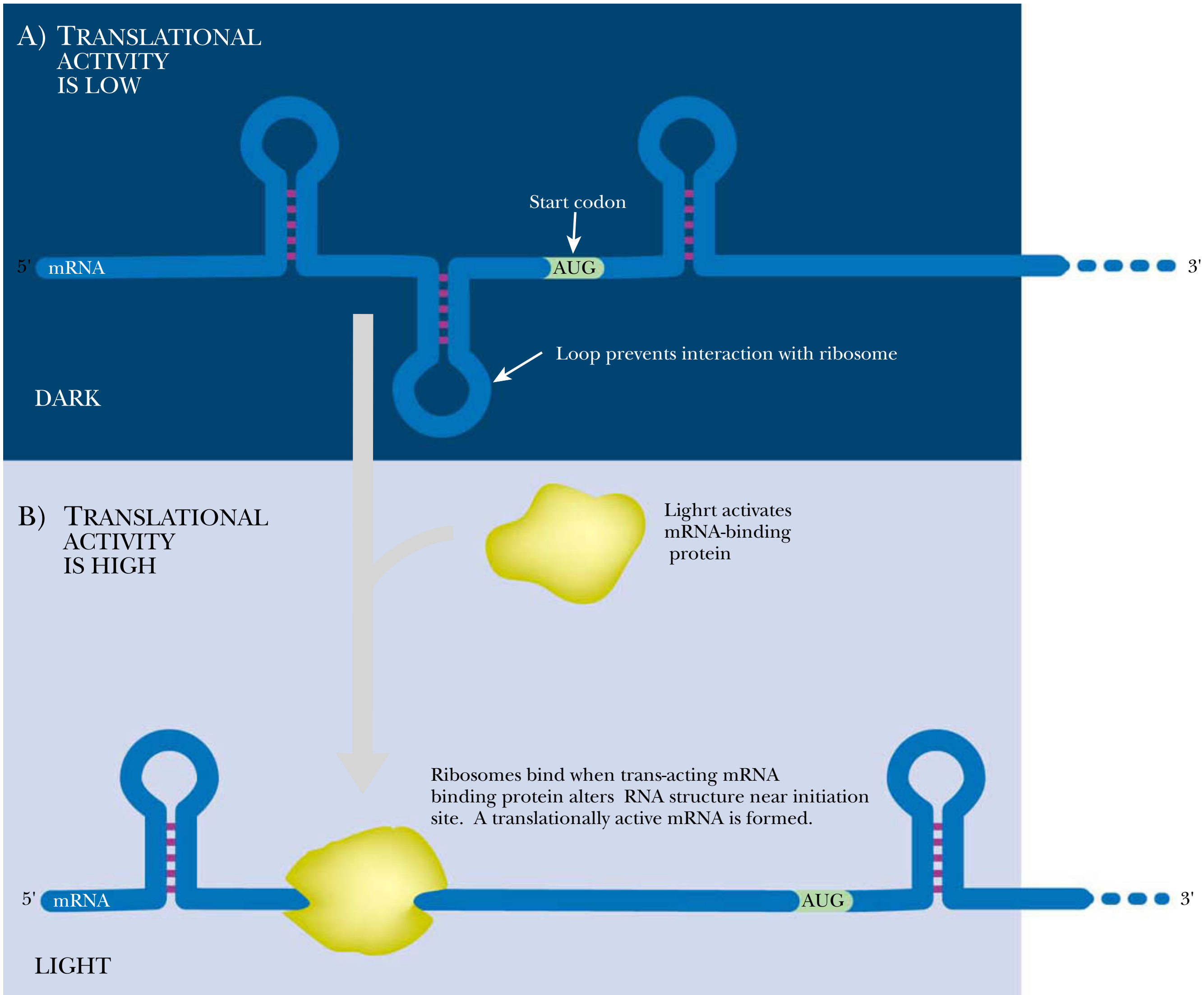

A) TRANSLATIONAL ACTIVITY IS LOW
Start codon
5'
mRNA
AUG
3'
Loop prevents interaction with ribosome
DARK
B) TRANSLATIONAL ACTIVITY IS HIGH
Lighrt activates mRNA-binding protein
Ribosomes bind when trans-acting mRNA binding protein alters RNA structure near initiation site. A translationally active mRNA is formed.
5'
mRNA
AUG
3'
LIGHT

FIGURE 11.06 ***Translational Activation of Chloroplast mRNA***

In the dark, the central loop in the mRNA prevents the mRNA from binding to the ribosome. When there is sufficient light, an mRNA binding protein straightens out the RNA to allow ribosome binding.

Translation May Be Regulated by Antisense RNA

Antisense RNA binds to messenger RNA and prevents its translation.

Messenger RNA is transcribed using only one DNA strand as the template. This is referred to variously as the template strand, non-coding strand or antisense strand. The mRNA produced is consequently **sense RNA**. The other strand of DNA (the coding strand or sense strand) is not normally used as a template for transcription. If RNA were transcribed using the coding strand as template we would produce an RNA molecule complementary in sequence to the mRNA. This is known as **antisense RNA** and can base pair with its complementary mRNA, just as the two strands of DNA in the original gene base pair with each other (Fig. 11.08). [Note that uracil pairs with adenine in duplex RNA].

Antisense RNA is occasionally used in gene regulation both by bacteria and eukaryotes. If antisense RNA is made, it will base pair with the mRNA and prevent

antisense RNA An RNA molecule that is complementary to mRNA
sense RNA Normal RNA that has been produced from the non-coding strand of DNA

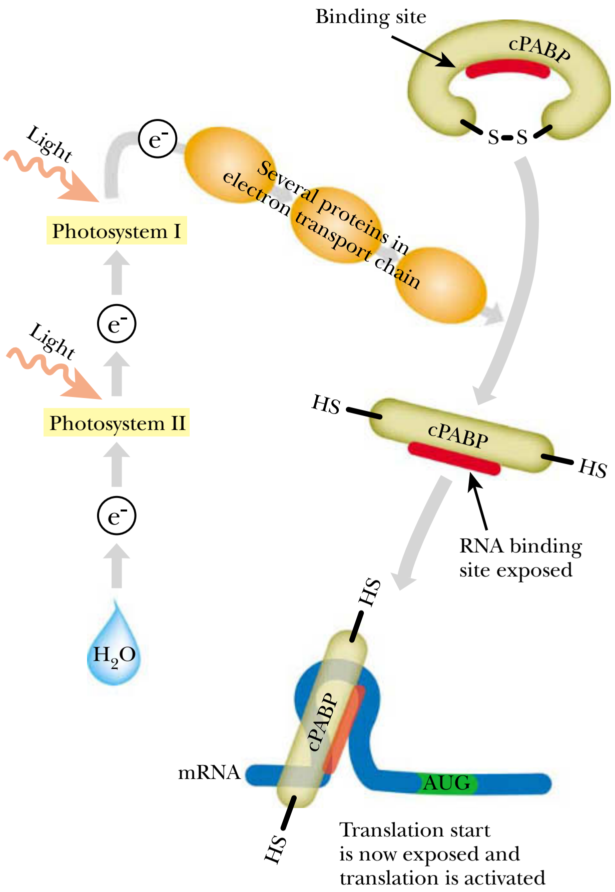


FIGURE 11.07 ***The Conformation of Translational Activator cPABP Responds to Light Intensity***

Light initiates a chain reaction in photosystem II (PS II) and I (PS I) of plant chloroplasts whereby an electron transfer takes place through the electron transport chain. The electron reduces the disulfide bond of cPABP so changing its conformation and allowing it to bind mRNA and activate translation.

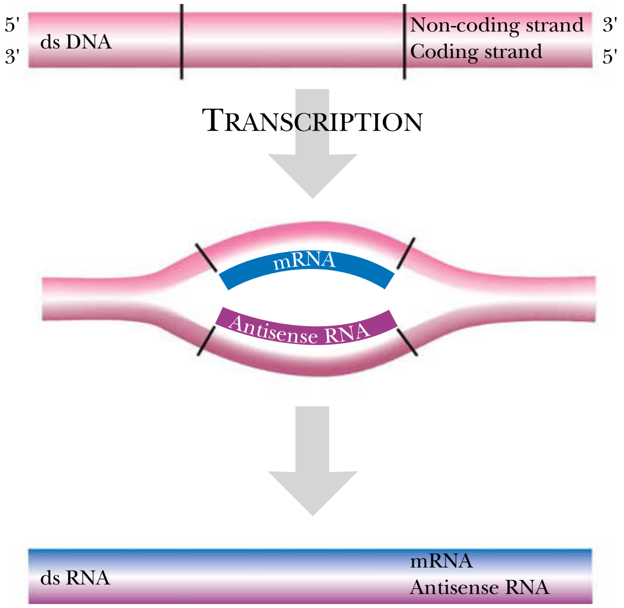


FIGURE 11.08 ***Antisense RNA can Base Pair with mRNA***

mRNA is normally made using the non-coding strand of DNA as a template. Such mRNA is also known as sense RNA. If RNA is made using the coding strand as a template, it will be complementary in sequence to mRNA and is known as anti-sense RNA. The sense and antisense strands of RNA can base pair.

it from binding to the ribosome and being translated (Fig. 11.09). In practice, antisense RNA is not made by transcribing the non-template strand of the same gene that gives the mRNA. Another, quite distinct "anti-gene" is used for making the antisense RNA.

Bacterioferritin is the protein used by bacteria to store surplus iron atoms. The *bfr* gene encodes bacterioferritin itself and the anti-*bfr* gene encodes the antisense RNA. Since only a relatively short piece of antisense RNA is needed to block the mRNA,

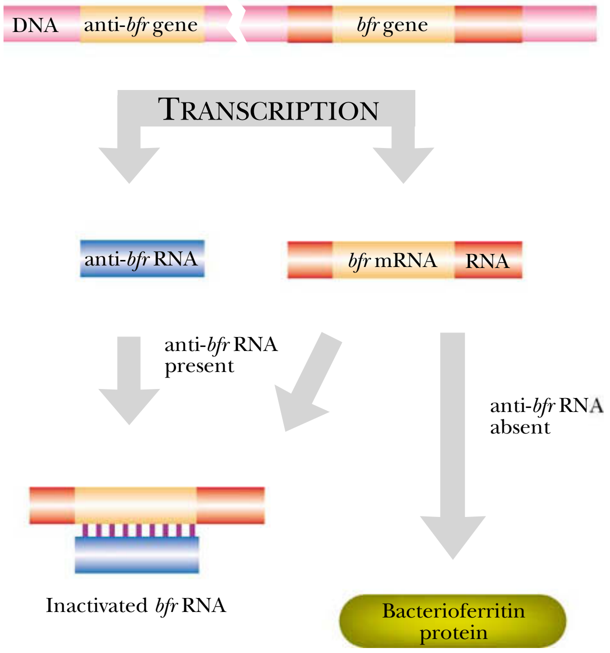


FIGURE 11.09 *Antisense RNA Regulates Bacterioferritin Synthesis*

The bacterial chromosome contains genes for both *bfr* mRNA and anti-*bfr* RNA. If both RNA molecules are transcribed the anti-*bfr* RNA pairs with the *bfr* mRNA and prevents it from being translated. When iron is plentiful the anti-*bfr* gene is not expressed and only the *bfr* mRNA is produced. Under these conditions translation of the *bfr* mRNA to give bacterioferritin can take place.

the anti-gene is similar in sequence but shorter than the original gene. When the iron concentration in the culture medium is low, bacterioferritin is not needed, but it is made if the iron level goes up. The *bfr* gene itself is transcribed to give mRNA in both conditions. However, the anti-*bfr* gene is controlled by a regulatory protein known as **Fur (Ferric Uptake Regulator)**, which senses iron levels.

When plenty of iron is present, Fur acts as a repressor and turns off the transcription of a dozen or more operons needed for adapting the cell to iron scarcity. These include genes for several iron uptake systems designed to capture trace levels of this essential nutrient. In addition, Fur plus iron turns off the anti-*bfr* gene, which turns on the production of bacterioferritin (Fig. 11.09). In low iron the anti-*bfr* gene is transcribed to give antisense RNA. This prevents synthesis of the bacterioferritin protein when iron is scarce. Thus, by using antisense RNA, one gene can be regulated the opposite way to a group of others although all respond to the same stimulus.

Artificially synthesized antisense RNA will interfere with gene expression or any other cell process involving RNA. For example, antisense RNA is being tested experimentally to suppress cancer by stopping chromosome division.

Regulation of Translation by Alterations to the Ribosome

The same ribosomes have to translate many different messages, expressed under many different conditions. Not surprisingly, the ribosomes themselves are rarely modified, except is such general cases as putting whole ribosomes on standby to reduce the overall rate of protein synthesis in non-growing cells, as discussed in Chapter 8.

One rare exception is the phosphorylation of the S6 protein of the small subunit of the ribosome in the cells of mammals. The S6 protein may be phosphorylated up to five times on a cluster of serine residues close to its C-terminus. [Lower eukaryotes, such as yeast lack these phosphorylation sites.] The phosphorylation occurs after cells in a growing tissue receive a signal to proliferate. The modified ribosomes preferentially translate mRNA molecules that possess **5′-terminal oligopyrimidine (5′-TOP)**

5′-terminal oligopyrimidine tract (5′-TOP) Long pyrimidine-rich tracts located between the 5′-end of mRNA and the start codon.
Fur (ferric uptake regulator) Global regulatory protein that senses iron levels in bacteria

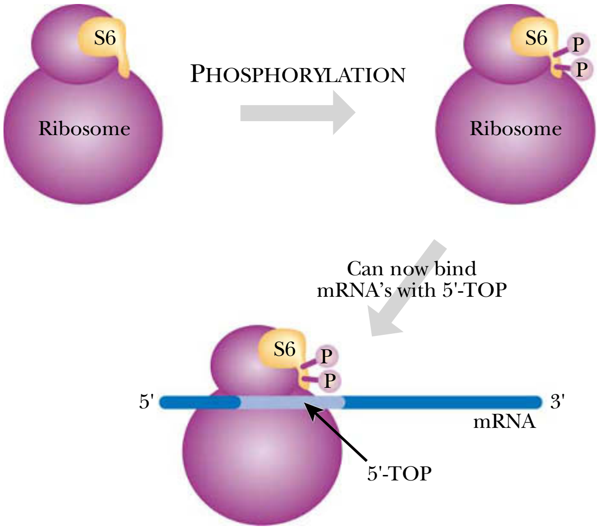


FIGURE 11.10
Phosphorylation of S6 Favors mRNA with 5′-TOP

The S6 protein of the small ribosome subunit can be phosphorylated by an activated protein kinase. Those mRNA possessing a 5′-TOP tract then bind better to the ribosome and commence translation.

tracts (Fig. 11.10). These are long pyrimidine-rich sequences located at the 5′-end of the mRNA, just upstream of the start codon.

It turns out that most of the favored mRNAs encode ribosomal proteins and elongation factors. In other words, phosphorylation of S6 protein stimulates the ribosome to make parts for more new ribosomes—an evolved response to an increased rate of cell growth and division.

RNA Interference (RNAi)

RNA interference (RNAi) is a mechanism for gene silencing that is induced by double-stranded RNA (dsRNA). It is sequence specific and involves the degradation of both dsRNA and single-stranded RNA molecules—usually mRNA—that are homologous in sequence to the dsRNA that triggered the response.

Double stranded RNA is normally destroyed by living cells of all organisms.

It is thought that RNAi originated as a defense mechanism against viruses. During the normal course of events, cells contain dsDNA and ssRNA but double-stranded RNA is not found. However, during infection by most RNA viruses, the virus genome passes through a double-stranded RNA intermediate (the replicative intermediate). This is true both for viruses that carry their genomes as ssRNA in the virus particle and those that use dsRNA (see Ch. 17 Viruses). Consequently, dsRNA is seen as a signal for infection and triggers an anti-viral response.

RNA interference destroys messenger RNA that has the same sequence as double-stranded RNA detected in the cell.

RNA interference is triggered by dsRNA that is fully base-paired and is at least 21–23 base pairs in length. Longer molecules of dsRNA are cleaved into fragments of 21–23 bp by a nuclease known as "**Dicer**" (Fig. 11.11). These RNA fragments are referred to as **siRNA (short interfering RNA)** and are bound by proteins of the **RNA-induced silencing complex (RISC)**. The RISC complex recognizes and degrades single-stranded RNA that corresponds in sequence to the siRNA. This involves unwinding and strand separation of the siRNA within the RISC complex and subsequent base pairing to the target RNA, as shown in Fig. 11.11. The nuclease activity of the RISC complex, sometimes referred to as "**Slicer**", then degrades the target RNA.

Dicer Ribonuclease that cleaves double-stranded RNA into segments of 21–23 bp
RNA-induced silencing complex (RISC) Protein complex induced by siRNA that degrades single-stranded RNA corresponding in sequence to the siRNA
RNA interference Response that is triggered by the presence of double-stranded RNA and results in the degradation of mRNA or other RNA transcripts homologous to the inducing dsRNA
short interfering RNA (siRNA) Double-stranded RNA molecules of 21–22 nucleotides involved in triggering RNA interference in eukaryotes
Slicer Ribonuclease activity of the RISC complex

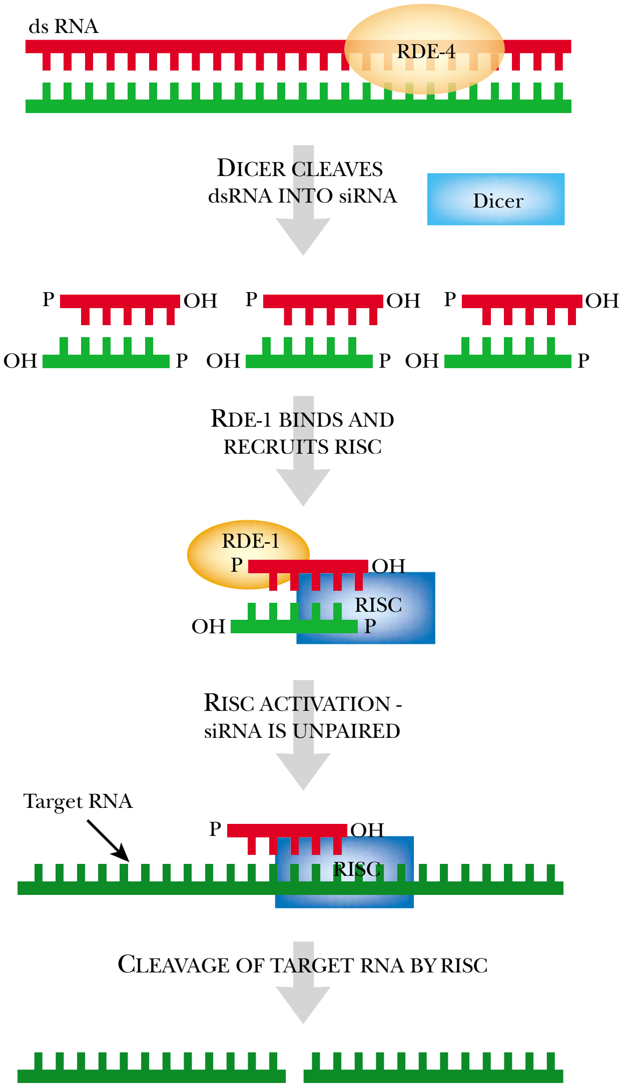

ds RNA
RDE-4
DICER CLEAVES dsRNA INTO siRNA
Dicer
P
OH
OH
P
RDE-1 BINDS AND RECRUITS RISC
RDE-1
RISC
RISC ACTIVATION - siRNA IS UNPAIRED
Target RNA
RISC
CLEAVAGE OF TARGET RNA BY RISC

FIGURE 11.11 ***Mechanism of RNA Interference***

Intruding double-stranded RNA (dsRNA) is recognized (by RDE-4 and other proteins in *Caenorhabditis*). Dicer cleaves the dsRNA into segments of 21 or 22 nucleotides with one or two base overhangs—short interfering RNA (siRNA). This is recognized by RDE-1 which recruits the RNA-induced silencing complex (RISC). The strands of the siRNA are separated during RISC activation. Finally, RISC cleaves target RNA that corresponds to the siRNA.

RNA interference operates in a wide range of eukaryotes, including protozoa, invertebrates, mammals and plants. It is not found in prokaryotes. [In bacteria, ribonuclease III rapidly degrades dsRNA molecules as short as 12 bp.] Recent investigations have complicated the simple idea of RNAi operating solely through RNA degradation. RNAi-related mechanisms can silence transcription of targeted genes by altering chromatin structure and promoting DNA methylation.

Amplification and Spread of RNAi

RNA interference can be amplified and can spread throughout an organism, after being triggered in a localized zone.

RNAi is remarkably potent. Thus less than 50 molecules of siRNA can silence target RNA that is present in thousands of copies per cell. This results from amplification of the siRNA via an **RNA-dependent RNA polymerase (RdRP)** that does not need a primer (Fig. 11.12). Cutting of the target mRNA by Slicer gives two aberrant and unstable RNA molecules, one capped but without the poly(A) tail and the other with a tail but no cap. One or both of these aberrant RNA molecules are apparently used as template by RdRP to generate dsRNA. The dsRNA then acts as a substrate for Dicer, which generates more siRNA so amplifying the RNA interference effect.

RNA-dependent RNA polymerase (RdRP) RNA polymerase that uses RNA as a template and is involved in the amplification of the RNAi response

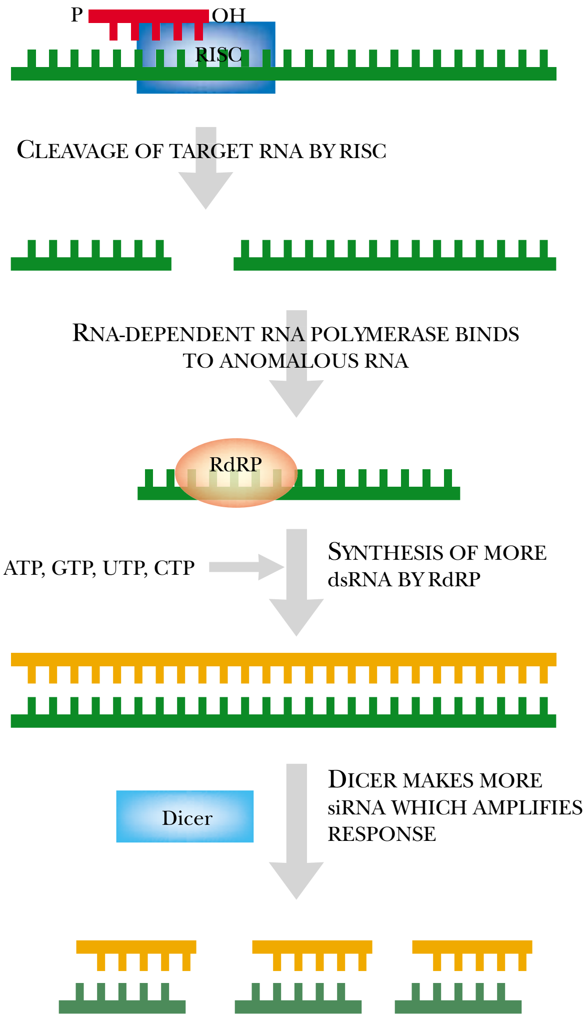

P
OH
RISC
CLEAVAGE OF TARGET RNA BY RISC
RNA-DEPENDENT RNA POLYMERASE BINDS TO ANOMALOUS RNA
RdRP
ATP, GTP, UTP, CTP
SYNTHESIS OF MORE dsRNA BY RdRP
Dicer
DICER MAKES MORE siRNA WHICH AMPLIFIES RESPONSE

FIGURE 11.12
Amplification of RNA Interference by RdRP

Anomalous RNA generated by RISC-mediated cleavage is used as a template by RNA-dependent RNA polymerase (RdRP). This generates more dsRNA which in turn is converted into more siRNA by Dicer.

In addition to amplification, the RNAi effect is capable of spreading from cell to cell, and may travel considerable distances through an organism. This effect is especially noticeable in plants (see below). Spreading of the siRNA signal throughout the body is also seen in animals. Remarkably, in *C. elegans* the RNAi effect is passed on for several generations (without alterations in the genomic DNA sequence of the targeted gene occurring). Mammals do not possess the RdRP responsible for RNAi amplification, hence RNAi remains relatively localized. It is thought that the development of the specific immune system has made RNAi less important in mammals.

Experimental Administration of siRNA

RNA interference is now being widely used to investigate gene function in animals and plants.

RNAi is widely used in laboratory research to prevent expression of cellular genes (as opposed to virus genetic information). RNAi in animals was first seen with the roundworm *Caenorhabditis elegans*. Injection of dsRNA into *C. elegans* causes destruction of the mRNA of corresponding sequence and hence stops gene expression at the post-transcriptional level. RNAi is now being used to survey gene function in *C. elegans*, since the complete genome sequence is available and it is therefore possible to synthesize siRNA corresponding to any gene of interest.

Using RNAi allows investigation of gene function without the need to make mutants with altered or inactivated versions of a particular gene. This is especially useful for eukaryotes, most of which are diploid and where classic genetic analysis therefore requires introducing mutations into both copies. However, RNAi by its very nature, prevents expression of all copies of the target gene. Indeed, RNAi may be used in organisms where multiple gene copies are present. For example, protozoa such as *Paramecium* contain two types of nuclei, a germline micronucleus and a highly polyploid somatic nucleus (see Ch. 19). This situation renders them refractory to standard genetic analysis.

Experimentally, RNAi may be induced by providing long molecules of dsRNA that are cut into siRNA by the Dicer enzyme (Fig. 11.11). Single-stranded antisense RNA against cellular genes may also trigger RNAi by base pairing with the corresponding plus RNA strand. This generates dsRNA inside the cell. Alternatively short dsRNA molecules of 21–23 nucleotides in length may be administered directly and will act as siRNA.

A variety of technical tricks are used to generate double-stranded RNA inside the experimental organism.

Although dsRNA may be given as such, it is often more convenient to provide a DNA construct that generates dsRNA *in vivo*. This may be done by three main variations (Fig. 11.13):

i. A single DNA segment transcribed from a single promoter that generates a stem and loop structure. Here the plus and minus strands are in tandem but separated by a short stretch of DNA that remains unpaired and forms the loop.

ii. A DNA segment flanked by two opposing promoters. Consequently one promoter transcribes the plus and the other transcribes the minus strand from the same dsDNA segment.

iii. Two DNA segments, one being the inverse of the other and both having separate promoters. Consequently one promoter transcribes the plus strand from the sense version of the DNA and the other transcribes the minus strand from the other, inverted, antisense, DNA segment.

The dsRNA generated from the DNA constructs may be long (and rely on Dicer to generate siRNA) or short, giving siRNA directly.

In the case of *Caenorhabditis elegans*, dsRNA may be injected into the worm. Alternatively the worm may be fed bacteria, such as *E. coli*, carrying plasmids with DNA constructs as described above that generate dsRNA in vivo. [Conveniently enough, bacteria form the natural diet of *C. elegans*.]

PTGS in Plants and Quelling in Fungi

Although discovered independently, it seems that **post-transcriptional gene silencing (PTGS)** in plants, quelling in *Neurospora* and RNAi in animals share a conserved mechanism. This system is also found in protozoa such as *Paramecium* and trypanosomes. In both PTGS and RNAi the presence of dsRNA triggers the Dicer/RISC system for RNA degradation. PTSG in plants may be propagated throughout the whole organism from a local initiation site. This has been illustrated by grafting experiments in which silencing is transmitted from grafted tissue to the stock receiving the graft. The siRNA signal is amplified by RdRP (see above) and then spreads from cell to cell via the plasmodesmata until they reach the vascular system, which distributes the siRNA throughout the whole plant.

RNA interference works well in plants and the response is distributed around the whole plant via its vascular system.

PTGS was actually observed in plants several years before RNAi was discovered in animals. PTGS was noticed when extra copies of plant genes were introduced into plant cells by genetic engineering. Instead of increased levels of gene expression, as intended, sometimes the result was a massive reduction in gene expression. This proved to be due to destruction of mRNA corresponding to the sequences introduced and occurs by a mechanism closely related to RNAi as described above for animals. Quelling is a similar phenomenon seen in fungi such as *Neurospora*. Mutants of the plant *Arabidopsis* that are defective in PTGS show increased sensitivity to infection

post-transcriptional gene silencing (PTGS) Plant version of the RNA interference response to double-stranded RNA that results in the degradation of mRNA or other RNA transcripts homologous to the inducing dsRNA

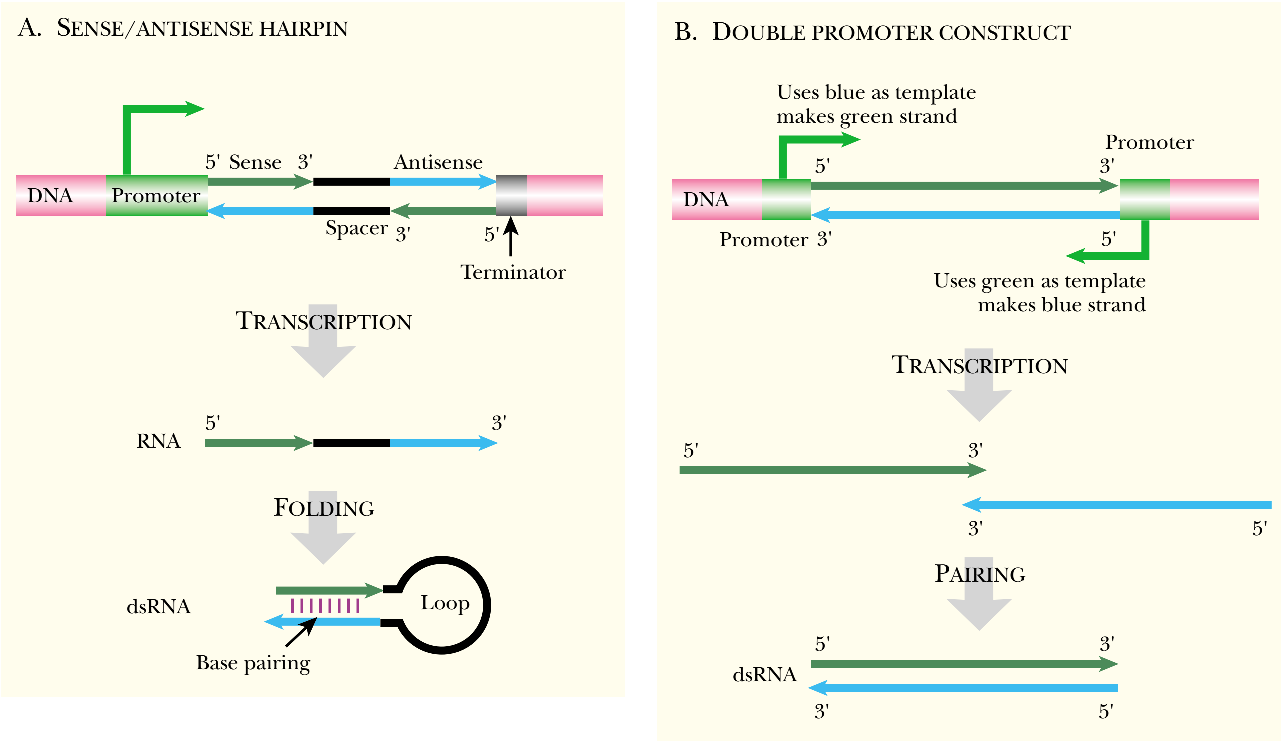


FIGURE 11.13 ***Experimental Induction of RNA Interference***

RNA interference occurs when both the sense and antisense RNA of a gene are present and form dsRNA. Two constructs are shown that direct the synthesis of a dsRNA molecule. The first construct (A) has a sense region and an antisense region that base pair. A spacer separates the sense and antisense regions and forms a loop at the end of the hairpin. The double promoter construct (B) has a promoter that directs the transcription of the sense strand, and another promoter for the antisense strand. The two resulting RNA molecules are complementary and form a dsRNA molecule.

by certain plant RNA viruses. This again suggests that the natural role of PTGS/RNAi is in protection against virus infection.

Like RNAi, PTGS requires the formation of dsRNA. Indeed, PTGS may be induced in plants by administration of small segments of pre-made dsRNA, just as RNAi is induced in animals. However, the main difference between PTGS and RNAi is that dsRNA may accumulate in plants after transcription of an introduced transgene. The mechanism for this is not fully understood. In some cases transcription of both strands of the introduced DNA constructs by opposing promoters may lead to the formation of RNA duplexes. In other cases, base pairing may occur internally between regions that are complementary in sequence. Nonetheless, PTGS may also be induced by single transgenes that highly transcribed. It is thought that aberrant mRNAs may sometimes be made and that these act as templates for an RNA-dependent RNA polymerase (RdRP). This then generates the dsRNA.

Micro RNA—A Class of Small Regulatory RNA

Micro RNA (miRNA) molecules are short RNA molecules that share several properties in common with siRNA. However, microRNA molecules regulate gene expression by blocking translation of mRNA, not by promoting degradation of the mRNA. They block translation by binding to mRNA, often in the 3′ untranslated region, less commonly in the coding region.

micro RNA (miRNA) Small regulatory RNA molecules of eukaryotic cells

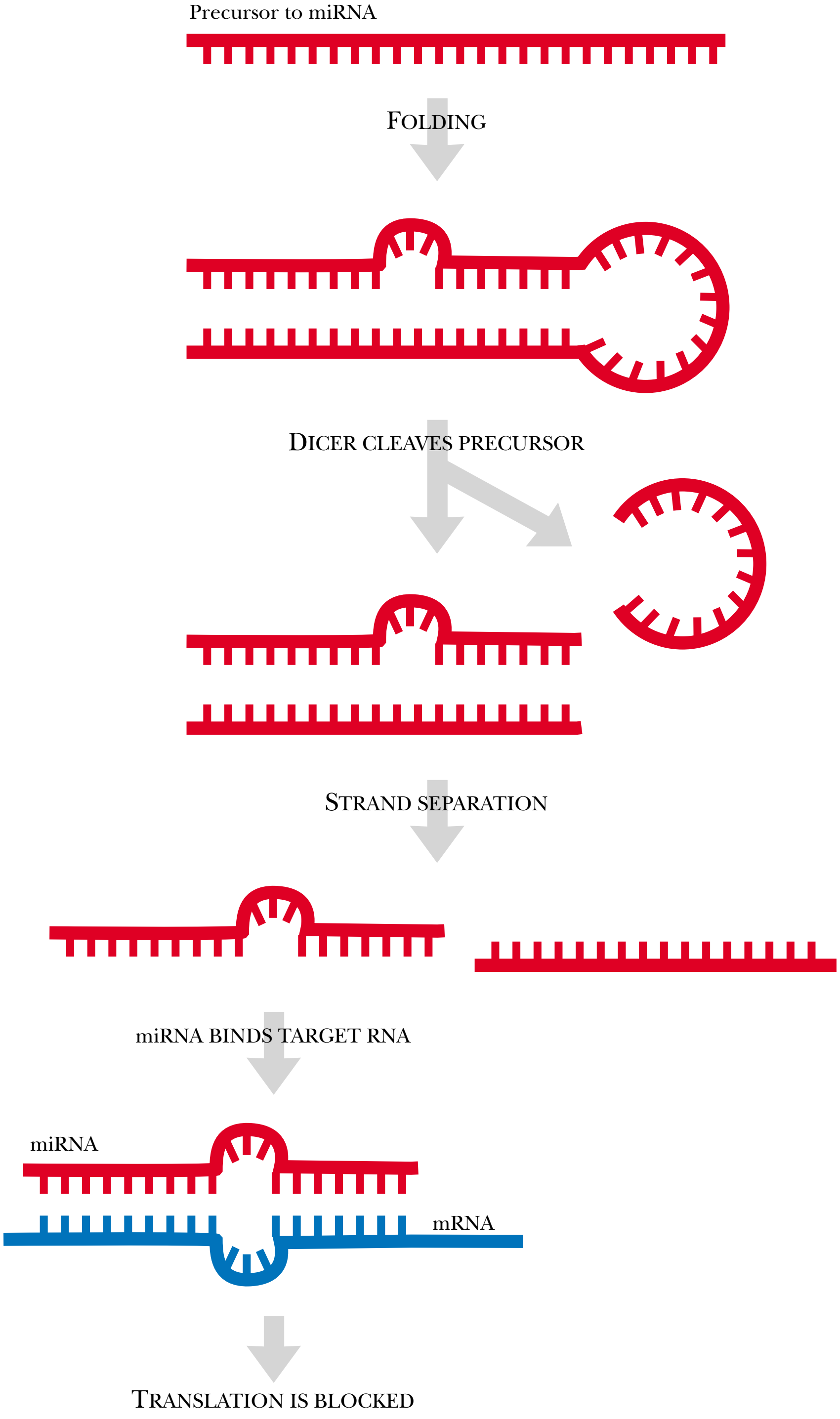


FIGURE 11.14 ***Micro RNA***

Micro RNA (miRNA) is made by processing a longer precursor that folds into a stem and loop. Dicer, the same nuclease that makes siRNA, is responsible for this processing. After strand separation, one strand of the miRNA binds to the target mRNA and prevents translation.

Micro RNA is a kind of short regulatory RNA that blocks translation of mRNA.

Micro RNA molecules are produced from longer RNA precursors of approximately 70 nucleotides that are transcribed from chromosomal genes. These precursor RNA molecules fold into stem-loop structures (Fig. 11.14). The double-stranded stem region is then cut by Dicer, the same nuclease that generates siRNA. The miRNA molecules have 1–3 unpaired bases in the middle and thus differ from siRNA, which is completely base paired. Instead of Rde1, which recognizes siRNA, the miRNA is bound by the related proteins, Alg1 and Alg2 (in *C. elegans*). It is uncertain whether the RISC complex then binds to the miRNA and separates the strands (as for siRNA). An alternative is that the strands separate spontaneously since miRNA is imperfectly base-paired. In any case, one strand of the miRNA then binds to the target mRNA. However, base pairing is not usually perfect (as it would be with siRNA). The result

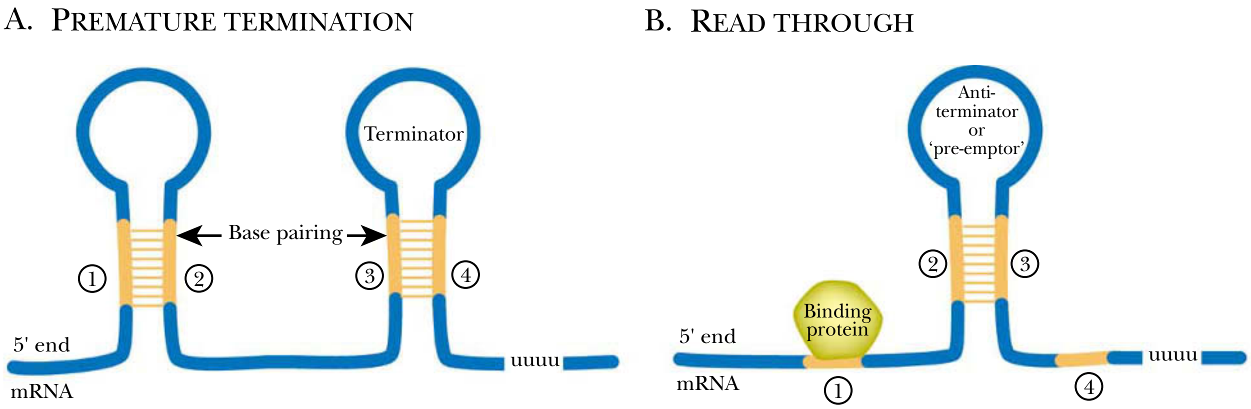


FIGURE 11.15 ***Alternative Secondary Structures of mRNA Leader Region***

The sequences in the leader region of the messenger RNA designated by 1, 2, 3 and 4 can base pair in two alternative ways. A) The structure for premature termination of mRNA is due to base pairing that forms two stem and loop structures in the mRNA. The second loop, 3 plus 4, causes termination. B) Termination can be prevented if a protein binds to the mRNA at site 1, allowing sites 2 and 3 to pair off. This creates the "pre-emptor" and prevents the terminator from forming.

is that translation of the mRNA is blocked, but the mRNA is not degraded. The precise mechanism of inhibition of translation is presently uncertain.

Micro RNA is found in worms, insects, mammals and plants—i.e. the same organisms that display RNA interference. The first miRNAs were found in *C. elegans* where they were called small temporal RNA (stRNA) because they regulate the timing of worm development during the conversion of the larva into the adult. Many miRNAs appear to be involved in the regulation of development and many of the targets for miRNA are mRNAs that encode transcription factors, which in turn regulate the expression of other genes.

Premature Termination Causes Attenuation of RNA Transcription

Transcriptional **attenuation** is a regulatory mechanism that involves premature termination of mRNA synthesis. The basic principle of attenuation is that the first part of the mRNA to be made, the **leader region**, can fold up into two alternative secondary structures. One of these allows continued transcription but the other secondary structure causes premature termination. The status of attenuation is somewhat ambiguous. It is often viewed as regulation at the level of transcription. However, it does involve mRNA that has already been partly transcribed. Consequently it is sometimes regarded as "post-transcriptional". Since attenuation is closely related to other mechanisms based on alternative RNA stem and loop structures, I have chosen to include it along with other forms of RNA based regulation.

Regulation by attenuation involves alternative stem and loop structures in the mRNA.

Typically, the leader region contains four sub-regions (sequences 1 through 4) that may base pair in two different ways. When no other factors intervene, sequence 1 pairs with 2 and sequence 3 pairs with 4, so forming two stem and loop structures (Fig. 11.15). The second of these stem and loop structures, containing paired sequences 3 and 4, acts as a terminator. However, sequence 2 may instead pair with 3. For this to happen, a protein must bind to sequence 1 and remove it from play. The net result is that the terminator loop, normally consisting of sequences 3 and 4, no longer forms and transcription of the mRNA can continue.

attenuation Type of transcriptional regulation that works by premature termination and depends on alternative stem and loop structures in the leader region of the mRNA

leader region The region of an mRNA molecule in front of the structural genes, especially when involved in regulation by the attenuation mechanism

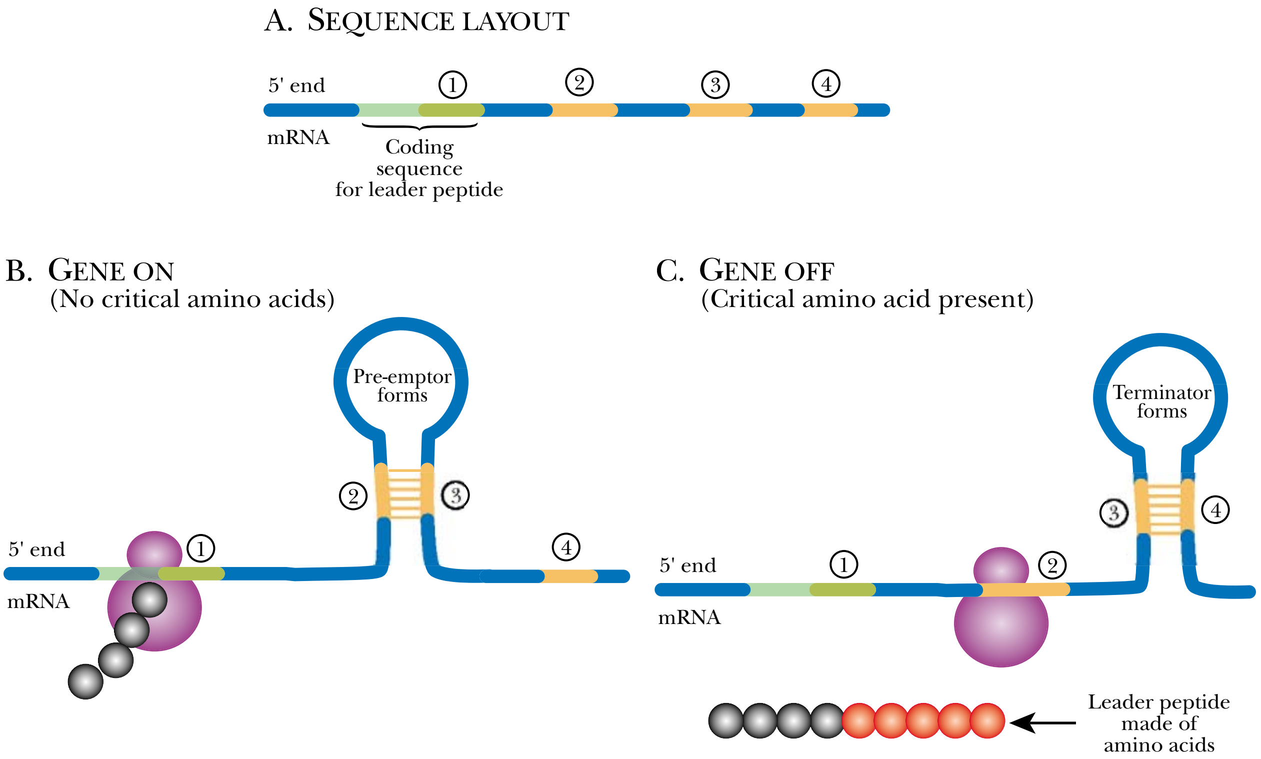

A. SEQUENCE LAYOUT
5' end
1
2
3
4
mRNA
Coding sequence for leader peptide
B. GENE ON
(No critical amino acids)
Pre-emptor forms
2
3
5' end
1
4
mRNA
C. GENE OFF
(Critical amino acid present)
Terminator forms
3
4
5' end
1
2
mRNA
Leader peptide made of amino acids

FIGURE 11.16 ***Stalled Ribosome Prevents Formation of Terminator Loop***

Attenuation controls whether synthesis of mRNA is completed or aborted. [The RNA polymerase is not shown in this figure, just the attenuation mechanism.] (A) The leader region of the mRNA contains the coding sequence for the leader peptide and four specific sequences (1, 2, 3 and 4) that can base pair to form stem and loop structures. (B) When there is a shortage of the corresponding amino acid, the ribosome slows down at region 1, allowing the pre-emptor structure to form and transcription of the mRNA by RNA polymerase to continue. Note that the leader peptide is not completed. (C) When there is an abundance of the corresponding amino acid, the leader peptide is made and the ribosome quickly moves to region 2, allowing regions 3 and 4 to form the terminator loop. This prevents further elongation of the mRNA.

Attenuation is used to regulate the genes for biosynthesis of amino acids in both gram-negative bacteria, such as *E. coli*, and gram-positive bacteria, such as *Bacillus*. If the supply of amino acid is plentiful, then the genes for its biosynthesis should be turned off. Conversely, if the level of the amino acid is low, the biosynthetic genes should be transcribed.

In *E. coli*, attenuation is complicated and usually involves the binding of ribosomes to the mRNA leader region where they translate a **leader peptide**. The leader peptide is encoded by a short open reading frame and only consists of 14 or 15 amino acids. It lies close to the 5′-end of the mRNA, upstream of the structural genes for the enzymes of the biosynthetic pathway (Fig. 11.16A).

The leader peptide contains several tandem codons for the amino acid in question. For example, in the leader peptide of the *his* operon of *E. coli* there are seven codons for histidine in a row. In the leader peptide of the *thr* operon, there are 11 clustered codons for threonine and isoleucine. Since threonine is the precursor for isoleucine, codons for both of these amino acids are included in attenuation control. When an amino acid is in short supply, the ribosome has difficulty finding a charged tRNA carrying that particular amino acid and it slows down. When several codons for a scarce amino acid follow each other, the ribosome grinds to a halt. The stalled ribosome covers sequence 1, loop 2/3 forms and the terminator loop is not made (Fig. 11.16B). The RNA polymerase carries on, transcribing the rest of the mRNA.

In *Bacillus*, ribosomes are not involved and there is no leader peptide. Nonetheless, the leader region of the mRNA possesses four sequences that can pair up

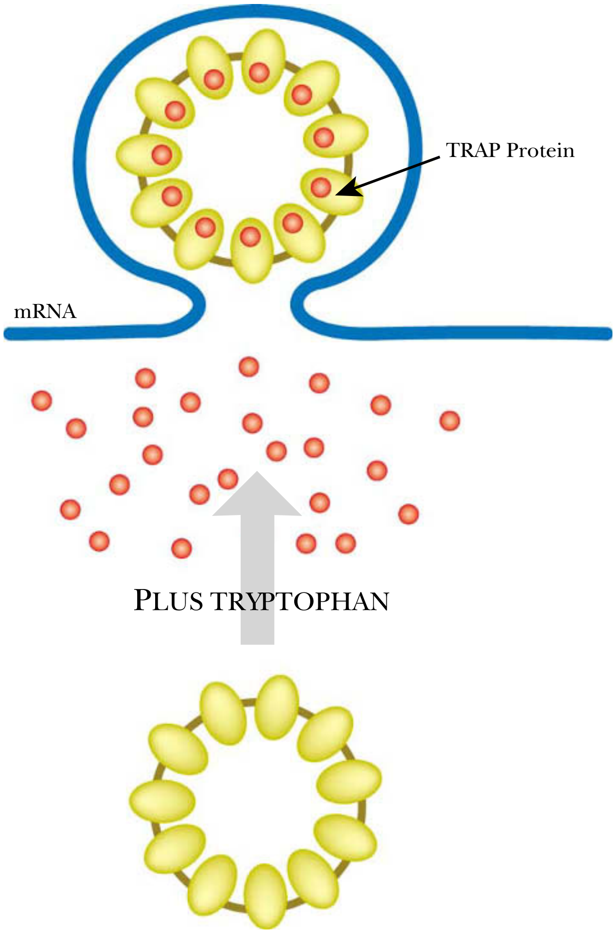


FIGURE 11.17 ***Attenuation by RNA-Binding Protein***

When tryptophan is present, it binds avidly to the tryptophan attenuation protein. This binds to the RNA and alters its structure. The alternative RNA structure possesses a stem and loop that causes premature termination.

to give two alternative structures. An **attenuation protein** binds the amino acid in question. In the presence of the amino acid, the attenuation protein binds to the mRNA leader region and promotes termination. In *Bacillus*, the 5′-region of the leader of *trp* mRNA contains a run of eleven UAG or GAG triplets separated from each other by two or three other bases. Eleven subunits of the tryptophan attenuation protein (TRAP) bind to these, forming an eleven-membered ring (Fig. 11.17). This allows formation of the terminator stem and loop and so causes premature termination of transcription.

Riboswitches—RNA Acting Directly as a Control Mechanism

A few messenger RNA molecules can control their own translation via riboswitch domains that bind small molecules.

One of the most fascinating recent stories in molecular biology has been the discovery that RNA can carry out many of the functions that were previously believed to need proteins. Ribozymes, that is to say catalytically active RNA, are involved in RNA splicing, protein synthesis and viroid replication. Antisense RNA and a variety of small regulatory RNA molecules, such as siRNA are involved in gene regulation. Most recently, it has been found that RNA domains at the front of messenger RNA, referred to as **riboswitches**, can directly interact with small molecules and can control gene expression. The vast majority of riboswitches have been found in bacteria and so far it is only in bacteria that experimental evidence for riboswitch operation exists.

attenuation protein Regulatory protein involved in attenuation and that binds to the leader region of mRNA
riboswitch Domain of messenger RNA that directly senses a signal and controls translation by alternating between two structures

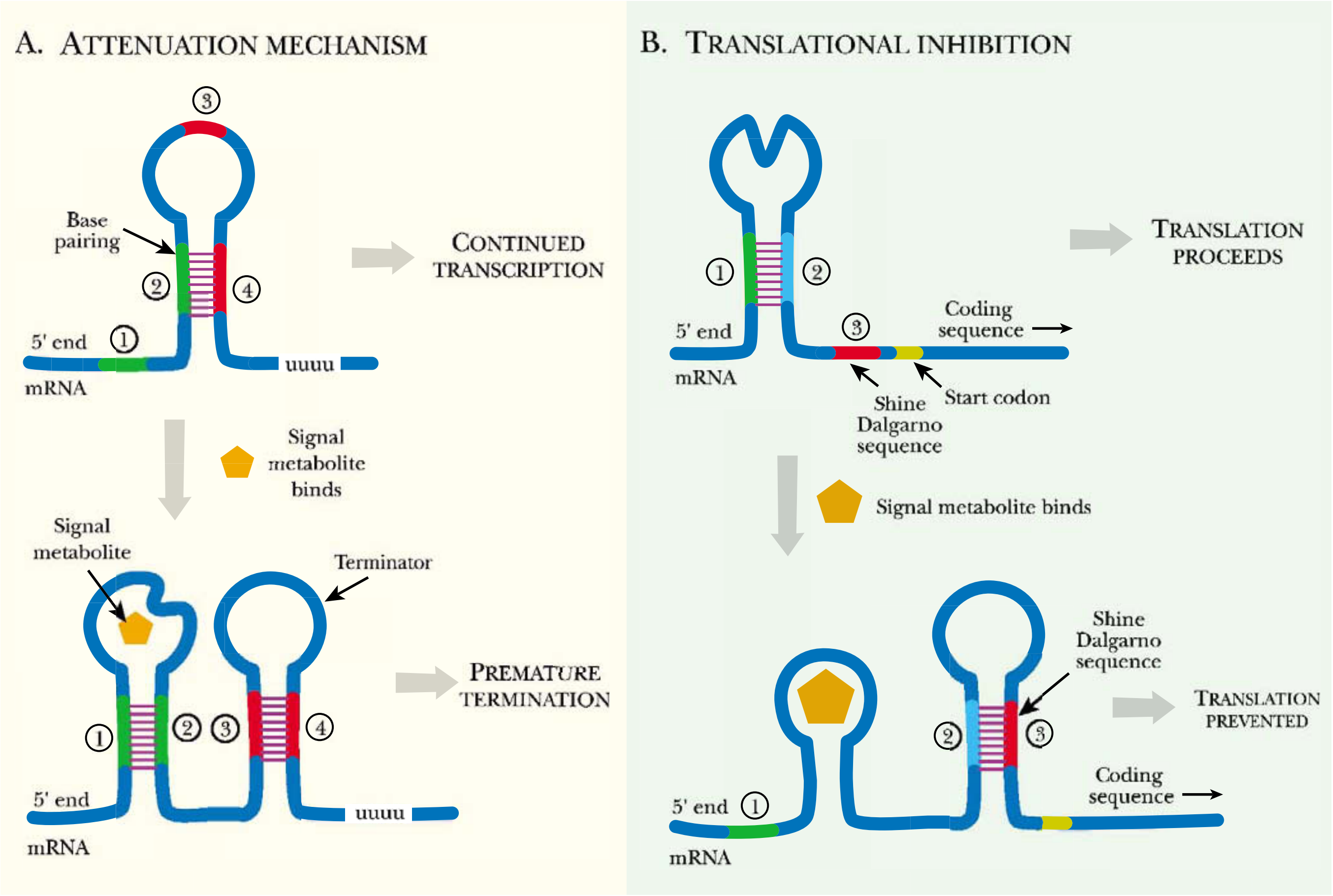


FIGURE 11.18 ***Riboswitch Mechanisms***

Riboswitches alternate between two alternative stem and loop structures depending on the presence or absence of the signal metabolite. (A) In the attenuation mechanism, the presence of the signal metabolite results in formation of the terminator structure and transcription is aborted. (B) In the translational inhibition mechanism, the presence of the metabolite results in sequestration of the Shine-Dalgarno sequence, which prevents translation of the mRNA.

However, sequence analysis has revealed that the genomes of certain fungi and plants contain equivalent sequences, implying that they probably have riboswitches too.

Biosynthetic pathways that make metabolites such as amino acids and vitamins are generally induced when the metabolite is in short supply but are shut down when there is a plentiful supply of the metabolite. The genes for such pathways are often controlled by repressors or by attenuation and are repressed in response to high concentrations of the metabolite in question. In these cases, the metabolite is bound by a regulatory protein as already described, such as the ArgR repressor of *E. coli* which binds arginine, or the tryptophan attenuation protein (TRAP) of *Bacillus* that binds tryptophan.

In riboswitches, the metabolite is directly bound by an RNA sequence at the 5′-end of the messenger RNA. For example, the thiamine riboswitch of *E. coli* contains a sequence that binds the vitamin/cofactor thiamine pyrophosphate with great specificity and is known as the THI box. Similarly the RFN box of the riboflavin riboswitch in *Bacillus subtilis* binds flavin mononucleotide. When these vitamins are in short supply the biosynthetic genes are turned on without the intervention of any regulatory protein. Conversely, when the vitamin is present at high levels the genes are turned off. Riboswitches are presently known for several vitamins, a few amino acids (methionine and lysine) and the purine bases adenine and guanine.

Binding of the metabolite to its RNA box changes the conformation of the whole riboswitch domain. Riboswitches exist in two alternative conformations that have different stem and loop structures. This in turn controls gene expression by one of two related mechanisms, premature termination of transcription (i.e. attenuation) or translational inhibition:

> In the attenuation mechanism (Fig 11.18A), the riboswitch controls whether or not the mRNA for the biosynthetic genes will be prematurely terminated. Here the riboswitch sequesters the terminator sequences in the absence of the signal metabolite and transcription continues. When the metabolite binds, the riboswitch changes to a conformation that allows the formation of a terminator stem and loop, which causes premature termination of the mRNA. Consequently, the genes are not expressed.
>
> In the translational inhibition mechanism (Fig 11.18B), the riboswitch controls whether or not the mRNA will be translated. When the signal metabolite is absent the Shine-Dalgarno sequence is free to bind to ribosomal RNA and translation proceeds. When the signal metabolite binds, the riboswitch sequesters the Shine-Dalgarno sequence and translation is prevented.

Riboswitches can respond to temperature as well as to metabolite concentrations.

Riboswitches can also respond to physical conditions as opposed to small molecules. The RNA thermosoensor is a specialized kind of riboswitch that responds to temperature and controls mRNA translation by sequestration of the Shine-Dalgarno sequence. The principle is the same as above, but formation of the alternative stem and loop structures depends directly on temperature. At high temperature one of the stems is unstable and the riboswitch flips to its high temperature form. The *rpoH* gene of *E. coli* is involved in the heat shock response, as described in Chapter 9. In addition to the regulation described there, translation of *rpoH* mRNA is prevented by an RNA thermosoensor at low temperature but allowed as the temperature increases. Other genes whose translation is controlled by RNA thermosoensors include certain activator genes involved in the virulence of pathogenic bacteria such as *Yersinia* and *Listeria*. Outside the host, these genes are switched off by translational inhibition. Inside the hot-blooded mammalian host the temperature is warmer and the activator proteins are expressed. The activator proteins then proceed to activate several other genes involved in bacterial virulence.

CHAPTER TWELVE

Processing of RNA

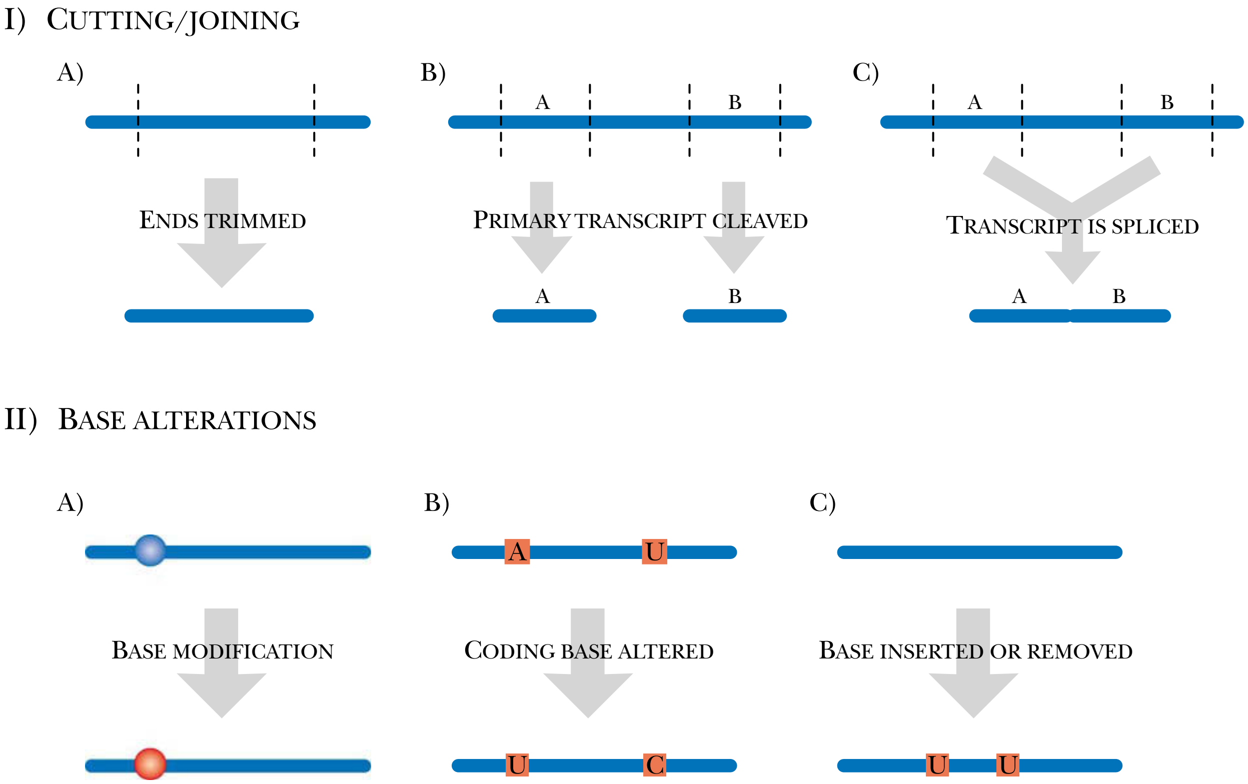


FIGURE 12.01 ***Types of RNA Processing***

RNA processing can be divided into cutting/joining or base alteration strategies.

RNA is Processed in Several Ways

Many RNA molecules are modified in a variety of ways after being synthesized.

RNA is made by RNA polymerase, using a DNA template, in the process known as transcription (see Ch. 6). Sometimes the RNA molecule is ready to function immediately after it has been transcribed (e.g. most bacterial mRNAs). However, in many cases, the RNA needs further processing before it is functional. In these cases, the original RNA molecule, before any further processing occurs, is known as the **primary transcript**. For specific classes of RNA, the precursor (i.e., primary transcript) may be referred to as pre-mRNA, pre-tRNA etc. The term, hnRNA (heterogeneous nuclear RNA) was also used previously, before the relationship of precursor RNA to the final processed RNA product was understood.

Three major types of processing, base modification, cleavage and **splicing** (Fig. 12.01) apply to all classes of RNA. In addition, eukaryotic mRNA undergoes capping and tailing as well as splicing. Like tRNA, which contains modified bases that are made after transcription (Chapter 8), rRNA, especially from higher organisms, contains modified bases. Certain RNA molecules are made as longer precursors that are trimmed to the correct length. In other, related cases, several RNA molecules are included in the same primary transcript, which is then cleaved into several parts. This applies to the rRNA molecules of both prokaryotes and eukaryotes. Splicing involves the removal of long segments from an RNA molecule by cleavage and rejoining of the ends. This is characteristic of eukaryotic mRNA due to the presence of non-coding introns in many eukaryotic genes (see Ch. 4). The primary

primary transcript The original RNA molecule obtained by transcription from a DNA template, before any processing or modification has occurred
splicing Removal of intervening sequences and re-joining the ends of a molecule; usually refers to removal of introns from RNA

transcript still contains the introns, which must be removed at the RNA level by splicing.

In the simpler cases, mRNA processing relies on typical enzymes consisting of proteins. However, as shown below, more complex RNA processing involves other RNA molecules. These RNAs are involved both in sequence recognition and in the actual chemical reactions of cutting and splicing. In fact, certain introns are self-splicing, that is they cut themselves out in a reaction that does not require any protein components (see below). Such RNA enzymes are known as **ribozymes**. As already mentioned in Chapter 8, the transpeptidation reaction in protein synthesis is catalyzed by ribosomal RNA, not by ribosomal proteins. The involvement of RNA in such fundamental processes as protein synthesis and RNA processing has led to the idea that ribozymes were more common in early life. Indeed the **"RNA world"** hypothesis suggests that the original enzymes were all RNA and that protein only assumed this role later in evolution. The RNA world scenario is discussed in more detail in Chapter 20, "Molecular Evolution".

The processing of RNA sometimes involves other RNA molecules, either as guides or as actual enzymes—ribozymes.

Coding and Non-Coding RNA

In bacterial cells, RNA makes up about 20% of the organic material. In eukaryotes, it only accounts for about 3–4%. Although most genes are transcribed to give the mRNA that encodes proteins, this mRNA is only a small fraction of the total RNA. RNA may be divided into coding RNA (i.e. mRNA) and **non-coding RNA**, which includes tRNA, rRNA and a variety of other RNA molecules that function directly as RNA and are not translated into protein.

Coding RNA (i.e. mRNA) is used only to carry information whereas non-coding RNA is not translated but performs a variety of active roles as RNA.

Although there are many different molecules of mRNA, each is only present in relatively few copies. In *E. coli* there are an average of 3–4 copies of about 400 different mRNAs. In contrast, there are many copies of rRNA and tRNA. For example, *E. coli* contains 10–20 thousand ribosomes each associated with one copy of each rRNA. Ribosomal RNA thus accounts for about 80% of the total RNA and tRNA for 14–15%. The mRNA only makes up 4–5% by weight of the RNA.

Ribosomal RNA and transfer RNA are found in all living cells. The other types of non-coding RNA vary from organism to organism. Bacteria contain several small regulatory RNAs (see Ch. 9) as well as tmRNA (transfer and messenger RNA in a single unit) that rescues ribosomes trapped by defective messages (see Ch. 8). In eukaryotes we find **small nuclear RNA (snRNA)**, **small nucleolar RNA (snoRNA)** and **small cytoplasmic RNA (scRNA)** molecules. The snRNA and snoRNA (sometimes called **U-RNA** as they are rich in U) are involved in processing other RNA molecules in the eukaryotic nucleus (see below). The scRNA is a miscellaneous group that comprises molecules with various functions. An increasing number of small regulatory RNA molecules are being found in eukaryotes and, to a lesser extent, in prokaryotes. In eukaryotes the two major classes are siRNA (short interfering RNA), involved in RNA interference, and miRNA (microRNA), short RNA molecules involved in regulating gene expression (see Ch. 11).

Many non-coding RNA molecules are found inside the eukaryotic nucleus.

non-coding RNA RNA molecule that functions without being translated into protein; includes tRNA, rRNA, snRNA, snoRNA, scRNA, tmRNA and some regulatory RNA molecules
ribozyme An RNA molecule that shows enzymatic activity
RNA world Theory that early life depended largely or entirely on RNA for both enzyme activity and for carrying genetic information and that DNA and protein emerged later in evolution
small cytoplasmic RNA (scRNA) Small RNA molecules of varied function found in the cytoplasm of eukaryotic cells
small nuclear RNA (snRNA) Small RNA molecules that are involved in RNA splicing in the nucleus of eukaryotic cells
small nucleolar RNA (snoRNA) Small RNA molecules that are involved in ribosomal RNA base modification in the nucleolus of eukaryotic cells
U-RNA Uracil-rich small RNA (includes snRNA and snoRNA)

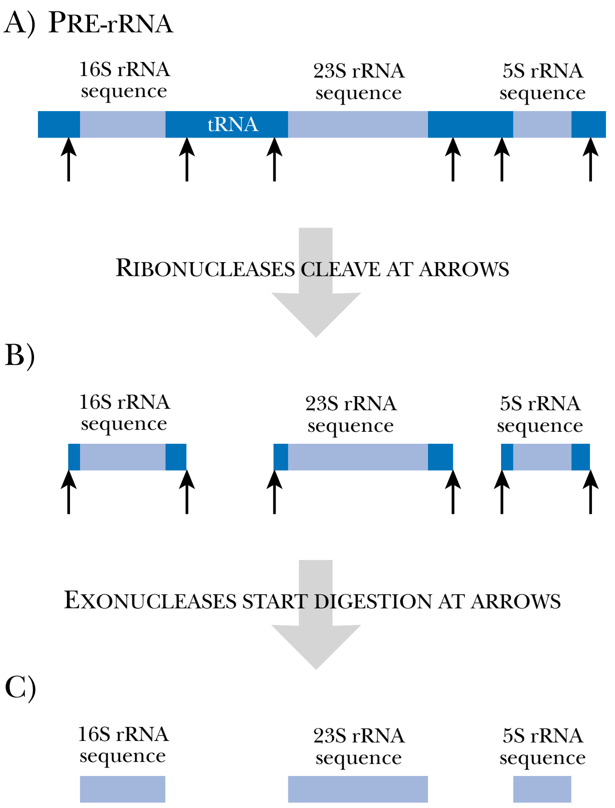


FIGURE 12.02 ***Cleavage of rRNAs From Their Precursor In Prokaryotes***

The pre-rRNA contains sequences for all three rRNA molecules as well as one or two tRNA molecules. Initial processing involves ribonucleases that cut the primary transcript at the sites shown by arrows. The ends must then be further trimmed. (Only the processing of the rRNA molecules is shown in full here; the tRNA is also trimmed after release.)

Processing of Ribosomal and Transfer RNA

Ribosomal RNAs are transcribed together as one long transcript that must be cut apart and trimmed at the ends.

The three rRNA molecules of bacteria are transcribed together to give a single pre-rRNA. This contains 16S rRNA, 23S rRNA and 5S rRNA joined by linker regions (Fig. 12.02). In bacteria, this pre-rRNA transcript also includes some tRNAs. In most bacteria there are several copies of the rRNA genes, seven in *E. coli*, for example.

The mature rRNAs are made by cleavage of the precursor by ribonucleases. This occurs in two stages (Fig. 12.02). First, internal cuts are made, separating the three rRNAs. Ribonucleases III, P and F recognize sites where the pre-rRNA is folded into double-stranded regions held together by base pairing. After this cleavage, the ends are trimmed by several exonucleases.

In eukaryotes, there are four ribosomal RNAs. The 5S rRNA is made separately and does not need processing. The other three (18S rRNA, 28S rRNA and 5.8S rRNA) are made as a single pre-rRNA and processed much as in bacteria.

Transfer RNA precursors must be processed to give functional tRNA molecules.

Transfer RNAs are transcribed as longer precursors that also need processing (Fig. 12.03). Some tRNAs are made singly, others are transcribed together and in bacteria, some are included in the pre-rRNA transcript. The 5′-end of bacterial tRNA is trimmed by **ribonuclease P**. This enzyme is of note because it is a ribozyme. Ribonuclease P consists of both an RNA molecule and a protein, but the catalytic activity is due to the RNA. The protein component merely modulates the activity of the RNA.

Eukaryotic Messenger RNA Contains a Cap and a Tail

In eukaryotic cells, transcription of genes to give messenger RNA is much more complex than in prokaryotes. First, eukaryotic genes are inside the nucleus, not free in the cytoplasm. Second, most eukaryotic genes are interrupted by segments of non-

ribonuclease P A ribonuclease involved in processing tRNA in bacteria that consists of an RNA ribozyme plus an accessory protein

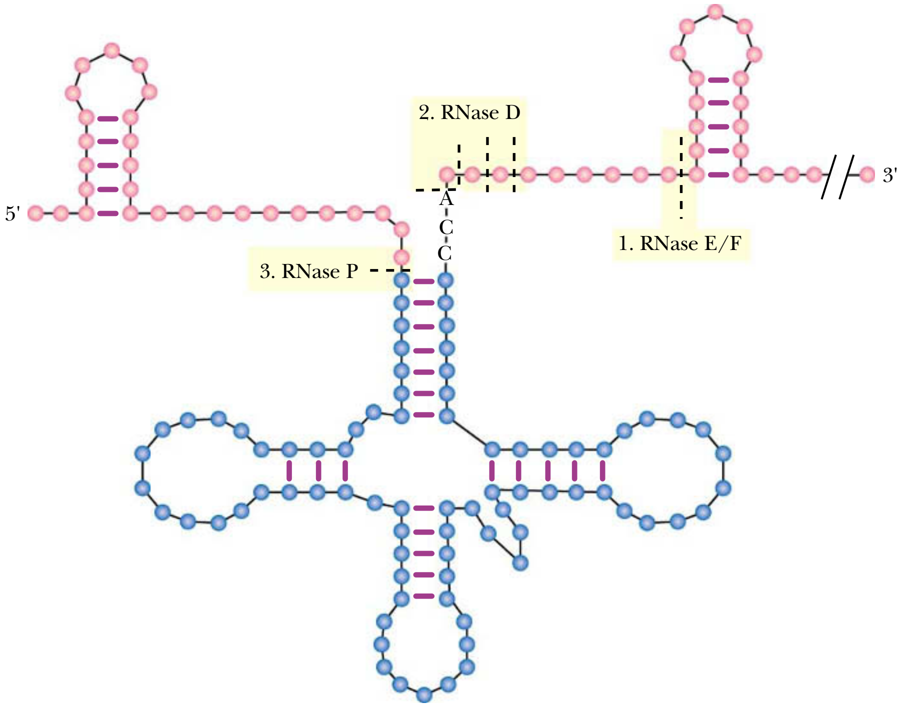


FIGURE 12.03 ***Processing of Transfer RNA***

Nucleotides shown in red are removed. First (1) ribonuclease E or F cleaves the precursor RNA near the 3′ end. Second (2), ribonuclease D chews off bases from the new 3′ end leaving the CCA at the end of the acceptor stem. Third (3), ribonuclease P cleaves the 5′ end precisely.

In eukaryotes, the primary transcript is converted to mRNA in three steps:
1) adding a cap at the front
2) adding a tail to the end
3) removing the introns

coding DNA, the introns. As discussed previously (see Ch. 4) the DNA sequence of a eukaryotic gene consists of regions which code for part of the final protein, the exons, alternating with regions of non-coding DNA, the introns.

The RNA molecule resulting from transcription is known as the primary transcript. It is not yet genuine mRNA because it, too, has exons alternating with introns. If the primary transcript were translated it would result in a huge, dysfunctional protein containing many extra stretches of random amino acids due to the intron regions. The primary transcript is trapped inside the nucleus until the introns are removed. This process is known as splicing and involves cutting out the introns and joining the ends of the exons to generate an RNA molecule which has only the exons; i.e., it contains an uninterrupted coding sequence (Fig. 12.04).

In order to be recognized as a bona fide messenger RNA molecule, and allowed to exit the nucleus, two other modifications must be made. These are the addition of a cap structure to the front and a tail to the rear of the RNA molecule. In fact, these are added before splicing out the introns.

Capping is the First Step in Maturation of mRNA

Before leaving the nucleus, RNA molecules destined to become messenger RNA have a **cap** added to their 5′-ends and a tail added to their 3′-ends. This occurs inside the nucleus and before splicing. Shortly after transcription starts, the 5′-end of the growing RNA molecule is capped by the addition of a guanosine triphosphate (GTP) residue (Fig. 12.05). This is added in a backward orientation relative to the rest of the bases in the RNA. After addition of the GTP, the guanine base has a methyl group attached at the 7 position. This structure is known as a "cap0" structure. Lower eukaryotes only proceed this far.

The cap on a eukaryotic mRNA consists of GTP in reverse orientation.

Further methyl groups may be added to the ribose sugars of the first one or two nucleosides of the original mRNA in some higher eukaryotes (Fig. 12.05). This gives respectively the "cap1" and "cap2" structures. If the first base of the original mRNA

cap Structure at the 5′-end of eukaryotic mRNA consisting of a methylated guanosine attached in reverse orientation

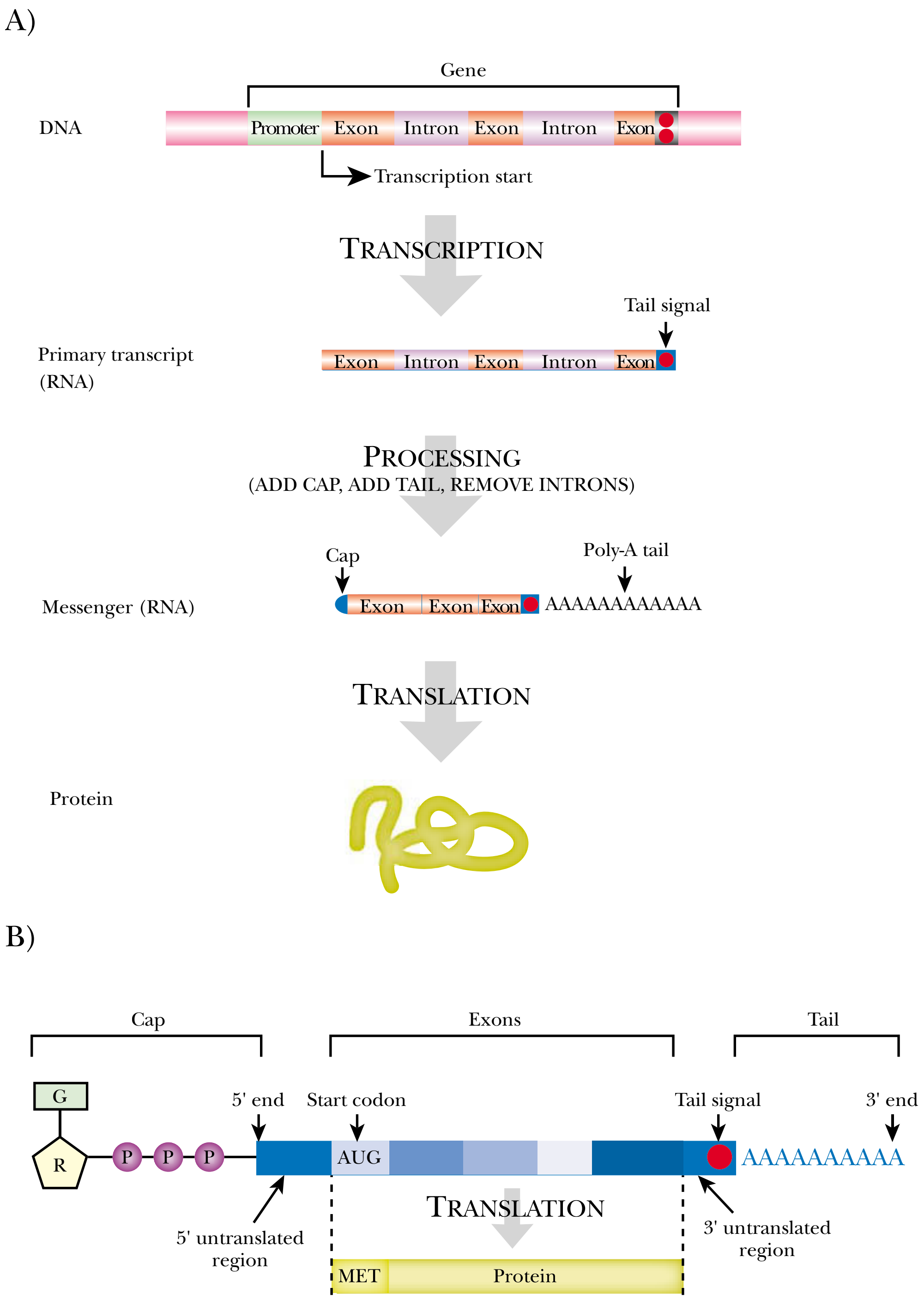


FIGURE 12.04 ***Splicing Out of the Introns and Maturation of Eukaryotic mRNA—An Overview***

The DNA, containing exons and introns, is transcribed to the RNA primary transcript, which contains the RNA version of both introns and exons plus a tail signal. Processing the RNA involves removing the introns, adding a cap and a poly-A tail. Translation of the fused exons forms the protein. B) The cap consists of guanosine triphosphate (attached in the reverse direction) preceding a small untranslated region; the tail consists of a poly-A tract and is preceded by a brief 3′ untranslated segment. When translated, only the information from the exons is used to build the protein.

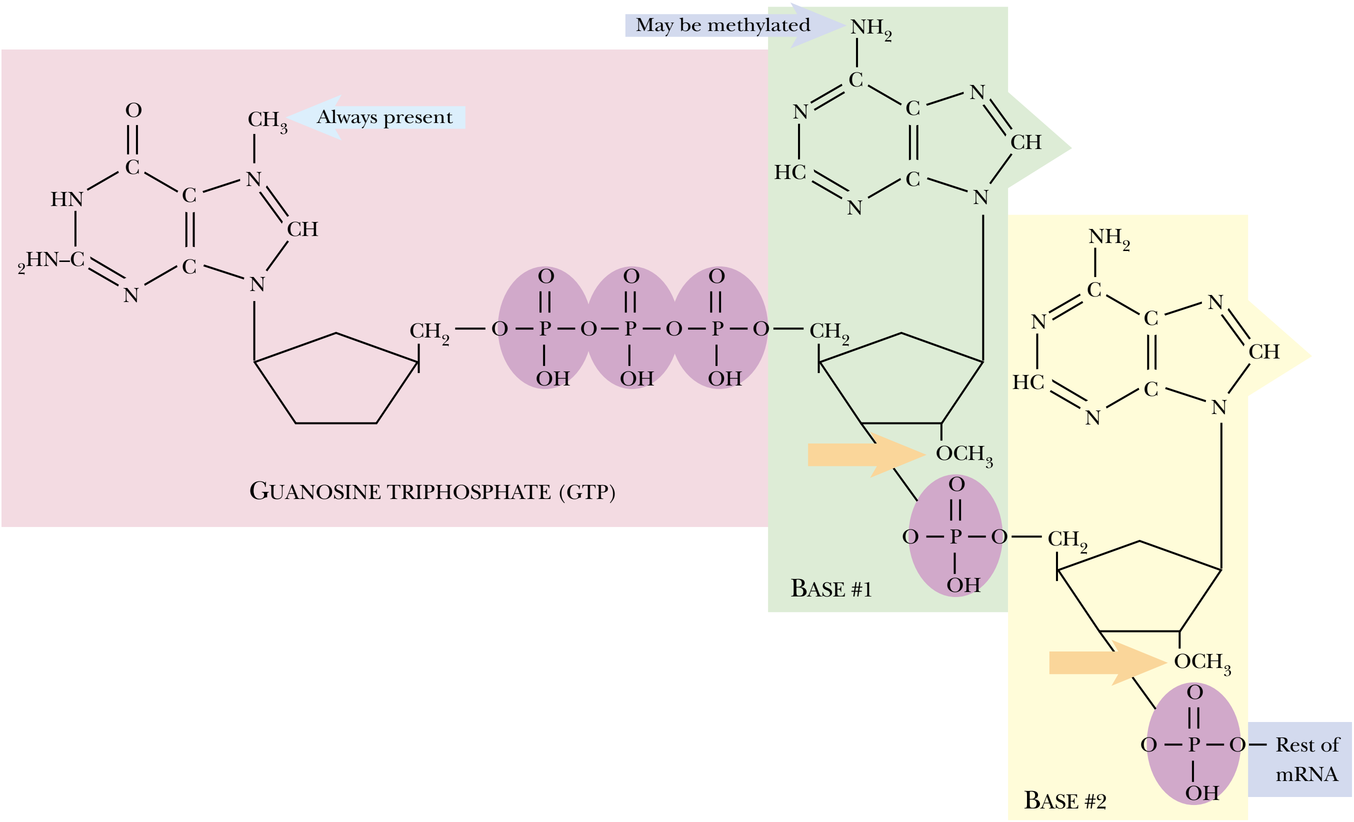


FIGURE 12.05 ***Capping of Eukaryotic mRNA***

A capping enzyme is responsible for adding a guanosine triphosphate to the 5′-end of mRNA. The guanine of the cap is methylated at the 7 position. Possible additions of other methyl groups to the ribose of nucleotides #1 and #2 are indicated.

is adenine this is sometimes methylated at the N^6 position. Whether or not a particular mRNA always has exactly the same cap structure is uncertain.

A Poly(A) Tail is Added to Eukaryotic mRNA

The tail of eukaryotic mRNA is a long string of A's at the 3′ end.

After being capped, the growing RNA has a **poly(A) tail** added (Fig. 12.04). Transcripts destined to become mRNA have a tail recognition sequence—AAUAAA—close to the 3′-end. The RNA polymerase that is making the RNA molecule bypasses this point. However, a specific endonuclease recognizes this sequence and cuts the growing RNA molecule 10 to 30 bases downstream, just after a 5′-CA-3′ dinucleotide. Beyond the cutting site is a GU-rich tract that is also involved in recognition, although it is lost after cutting.

The AAUAAA and GU-rich sites both bind proteins. Cleavage and polyadenylation specificity factor (CPSF) binds to the AAUAAA sequence and cleavage stimulation factor (CST) binds to the GU-rich tract. These two proteins provide a platform for assembly of cleavage factor and **poly(A) polymerase** as well as the **poly(A)-binding protein (PABP)**. Once this **polyadenylation complex** is assembled, the RNA is cut by the cleavage factor (an endonuclease) and a poly(A) tail is added by the poly(A) polymerase. The tail consists of 100 to 200 adenine residues (Fig. 12.06).

poly(A)-binding protein (PABP) Protein that binds to mRNA via its poly(A) tail
polyadenylation complex Protein complex that adds the poly(A) tail to eukaryotic mRNA
poly(A) polymerase Enzyme that adds the poly(A) tail to the end of mRNA
poly(A) tail A stretch of multiple adenosine residues found at the 3′-end of mRNA

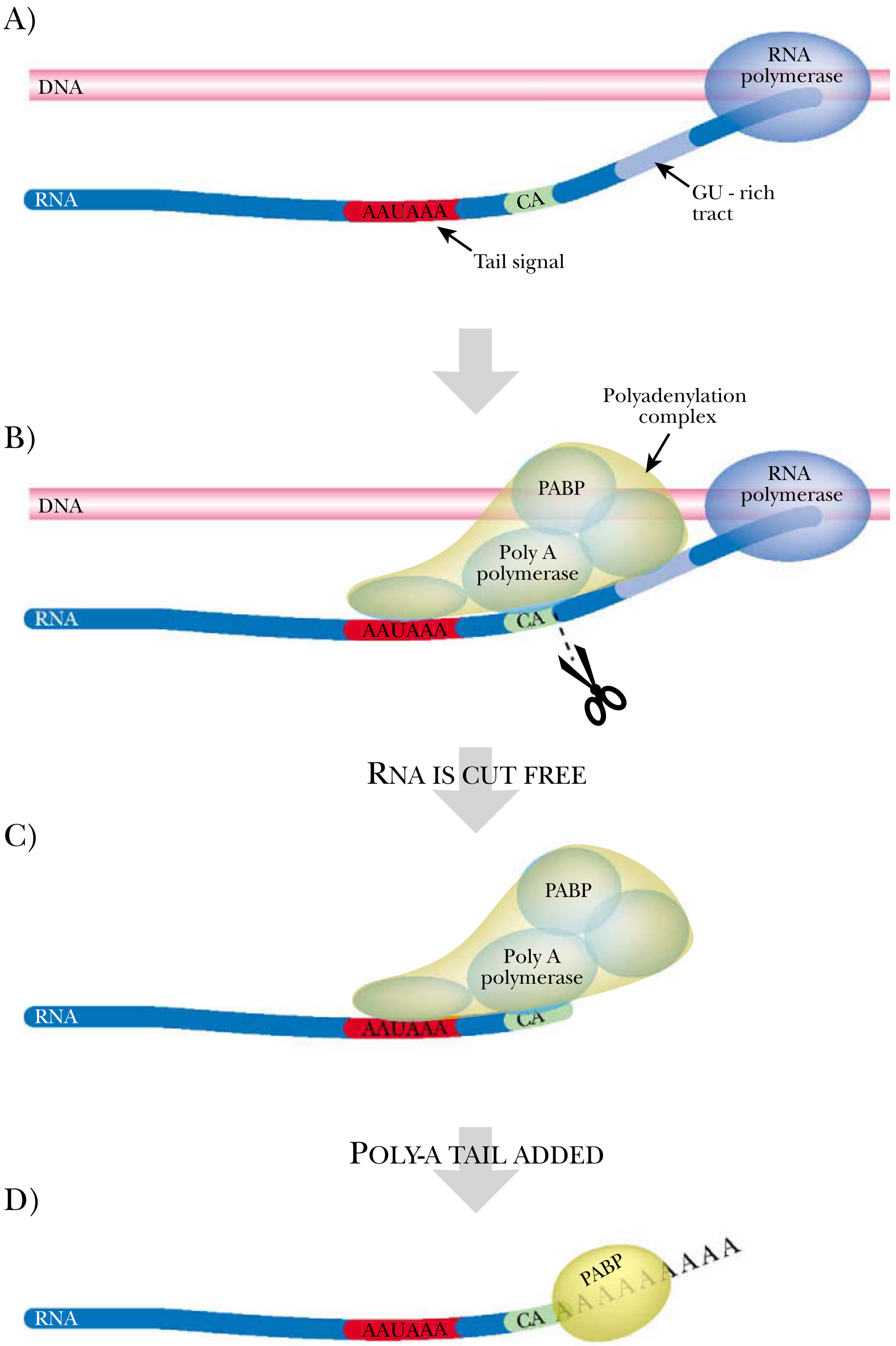


FIGURE 12.06 ***Addition of Poly(A) Tail to Eukaryotic mRNA***

A) During transcription RNA polymerase continues on beyond the end of the coding sequence until the RNA is cut free. Three important sequences beyond the end of the coding sequence are involved in cutting and tail addition: the tail signal (AAUAAA), a CA dinucleotide a few bases downstream, and a GU-rich tract. B) The polyadenylation complex consists of several proteins that bind to these sequences, including poly(A) polymerase itself, cleavage factor and poly(A)-binding protein (PABP). C) The growing RNA is cut just beyond the CA site by cleavage factor. D) Poly(A) polymerase adds the poly(A) tail to the free 3′-end of the mRNA. The completed poly(A) tail is bound by PABP.

TABLE 12.01 Classes of Intron

Class of Intron	Location of Genes
GT-AG (or GU-AG) introns	eukaryotic nucleus (common)
AT-AC (or AU-AC) introns	eukaryotic nucleus (rare)
Group I introns	organelles, prokaryotes (rare) rRNA in lower eukaryotes
Group II introns	organelles (of plants and fungi), some prokaryotes
Group III introns	organelles
Twintrons	organelles
Pre-tRNA introns	tRNA of eukaryotic nucleus
Archeal introns	archaebacterial tRNA and rRNA

The poly(A)-binding protein (PABP) stays associated with the mRNA and binds to the poly(A) tail. It is suspected that the PABP may bind both ends of the mRNA as it also appears to protect the cap from being cut off (see below). Whether the presence of the poly(A) tail and PABP make the mRNA more stable is doubtful; some stable mRNAs have very short tails. *The poly(A) tail is required for translation.* Certain mRNAs in early embryos are stored without a poly(A) tail and cannot be translated. When they are needed for translation, the poly(A) tail is added. Thus, mRNA may be stored.

Some mRNAs of bacteria are also polyadenylated. However, the role of the poly(A) tail is quite different in prokaryotes. In fact, the poly(A) tail triggers degradation of prokaryotic mRNA. The bacterial poly(A) polymerase is associated with the ribosomes and the tails are relatively short (10–40 bases). Addition of a poly(A) tail to mRNA in chloroplasts also promotes its degradation. This is another indication of the prokaryotic ancestry of these organelles (see Ch. 20).

Introns are Removed from RNA by Splicing

Introns are removed and the exons forming the coding sequence are joined together by the spliceosome.

After capping and tailing, the next step in processing pre-mRNA is the splicing out of the introns. Splicing must be accurate to within a single base since a mistake would throw the whole coding sequence out of register and totally scramble the protein resulting from translation of the mRNA. The overall result of this cutting and pasting has been depicted in Figure 12.01. There are several classes of introns (Table 12.01). The most frequent class of intron in eukaryotic nuclear genes is the GT-AG (or GU-AG in RNA code) group of introns. We will therefore discuss these first, before surveying the other variants.

snRNA molecules recognize the ends of the intron as well as the future branch site.

The splicing machinery is known as the **spliceosome** and consists of several proteins and some specialized, small RNA molecules found only in the nucleus (Fig. 12.08). Each small nuclear RNA (snRNA) plus its protein partners forms a **small nuclear ribonucleoprotein (snRNP)** or **"snurp"**. There are five snRNPs—numbered from U1 to U6 with U3 missing! (U3 is actually a snoRNA found in the nucleolus—see below.)

The snRNAs of the snurps recognize three sites on the pre-mRNA. These are the **5′ and 3′ splice sites** and the **branch site**. The vast majority of introns start with GU and end with AG. Recognition is due to base pairing between the snRNA and the

3′ splice site Recognition site for splicing at the downstream or 3′-end of the intron
5′ splice site Recognition site for splicing at the upstream or 5′-end of the intron
branch site Site in the middle of an intron where branching occurs during splicing
small nuclear ribonucleoprotein (snRNP) Complex of snRNA plus protein
snurp snRNP or small nuclear ribonucleoprotein
spliceosome Complex of proteins and small nuclear RNA molecules that removes introns during the processing of messenger RNA

Eukaryotic mRNA may be isolated by taking advantage of its poly(A) tail. Artificial strands of **oligo(U)** or **oligo(dT)** will base pair with poly(A) tracts. Generally, the oligo(U) or oligo(dT) is immobilized on a column and the mixture containing mRNA is poured through the column. The mRNA is trapped by binding of its poly(A) tail (Fig. 12.07). Other molecules, in particular non-coding RNA, pass through.

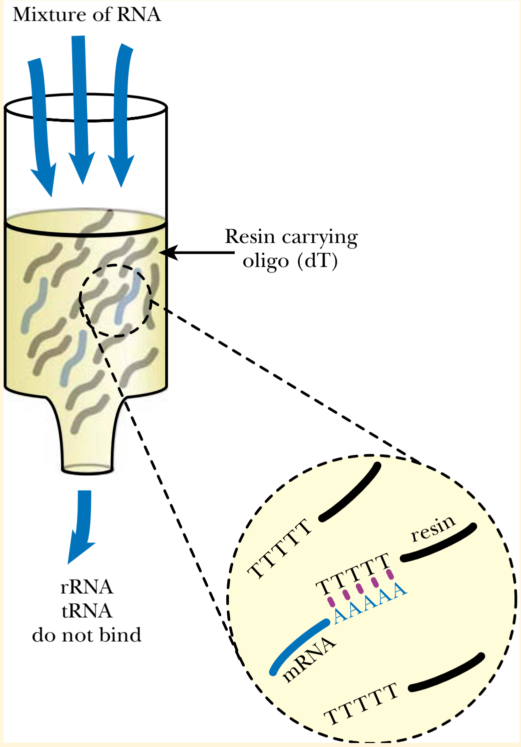


FIGURE 12.07 ***Binding of Poly(A) Tail Allows Isolation of Eukaryotic mRNA***

Oligo (dT) attached to a resin binds the mRNA molecules by base pairing with their poly(A) tails. Ribosomal RNA and transfer RNA molecules are not bound and exit the column. The mRNA is then eluted in purified form.

pre-mRNA. The protein part of the snurp supervises the cutting and joining reactions. This is shown in detail in Figure 12.09 for the **U1** snurp which recognizes the 5′ splice site. In the middle of the intron is a special adenine residue used as a branch site during splicing. The consensus recognition sequences for the 5′ splice site, 3′ splice site and branch site are as follows (residues in bold are most highly conserved as they are involved in the splicing mechanism):

The spliceosome is made of several snRNA molecules plus accompanying proteins.

5′ splice site:	5′–AG ⇓ **GU**AAGU-3′	
3′ splice site:	5′–YYYYYYNC**AG** ⇓ –3′	(Y = any pyrimidine)
branch site:	5′–UACUA**A**C-3′	

The snurps assemble onto the pre-mRNA, so forming the spliceosome (Fig. 12.10). U1 recognizes the 5′ splice site, **U2** binds the branch site (Fig. 12.09B) and a protein

oligo(dT) DNA strand consisting only of thymidine
oligo(U) RNA strand consisting only of uridine
U1 Snurp (snRNP) that recognizes the upstream splice site
U2 Snurp (snRNP) that recognizes the branch site

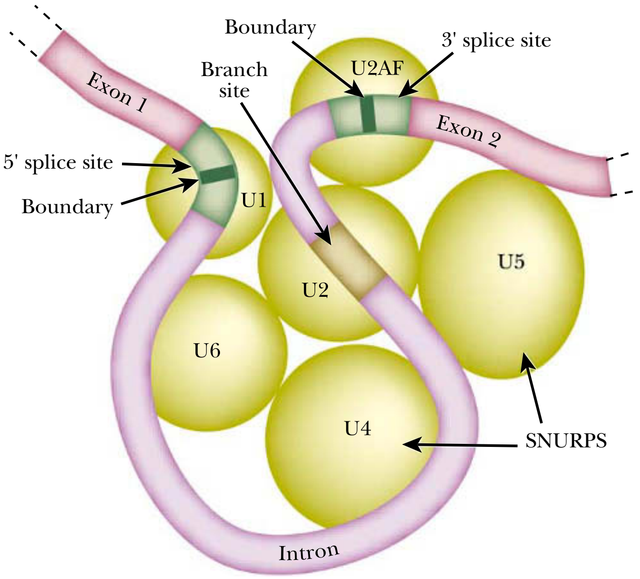
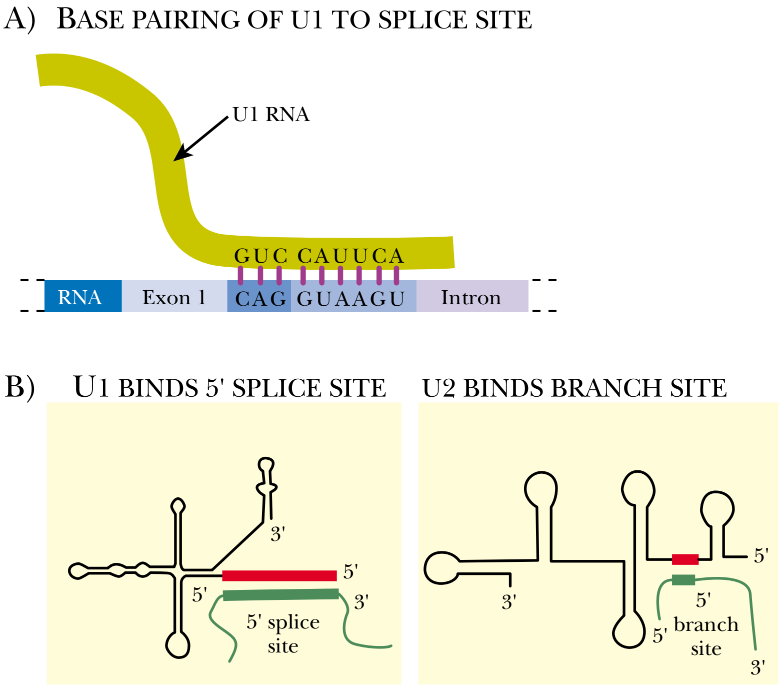


FIGURE 12.08 ***Spliceosome Recognizes Intron/Exon Boundaries***

The spliceosome consists of several ribonucleoproteins (U1 to U6), also known as "snurps", which are involved in splicing. These assemble at the splice sites at the intron/exon boundaries.


A) BASE PAIRING OF U1 TO SPLICE SITE

B) U1 BINDS 5' SPLICE SITE U2 BINDS BRANCH SITE

FIGURE 12.09 ***Recognition of 5′ Splice Site by U1 Snurp***

A) The 5′ splice site at the beginning of the intron is detected by base pairing with the RNA of the U1 snurp. B) The overall binding of U1 at the 5′ splice site and of U2 at the branch site are shown.

called **U2AF (U2 accessory factor)** binds the 3′ splice site. U4/U6 binds to U2 and U5 then arrives, binding first to the downstream exon and migrating to the intron/exon boundary. This makes the intron DNA loop appear as shown in Figure 12.10. The complex then rearranges itself. In particular U6 displaces U1 from the 5′ splice site and U1 and U4 are lost from the complex.

Finally, splicing proceeds in two steps (Fig. 12.11). First, the intron and exon are cut apart at the 5′ splice site and the free 5′-end of the intron loops around and is joined to the adenine at the branch site. Second, the free 3′-end of the upstream exon displaces the intron from the 3′ splice site and the two exons are joined together. The intron is released as a branched **lariat structure** that is later degraded.

The intron is removed and forms a loop with a branch. The exons are joined together.

lariat structure Branched, lariat-shaped segment of RNA generated by splicing out an intron
U2AF (U2 accessory factor) Protein involved in splicing of introns that recognizes the down stream splice site

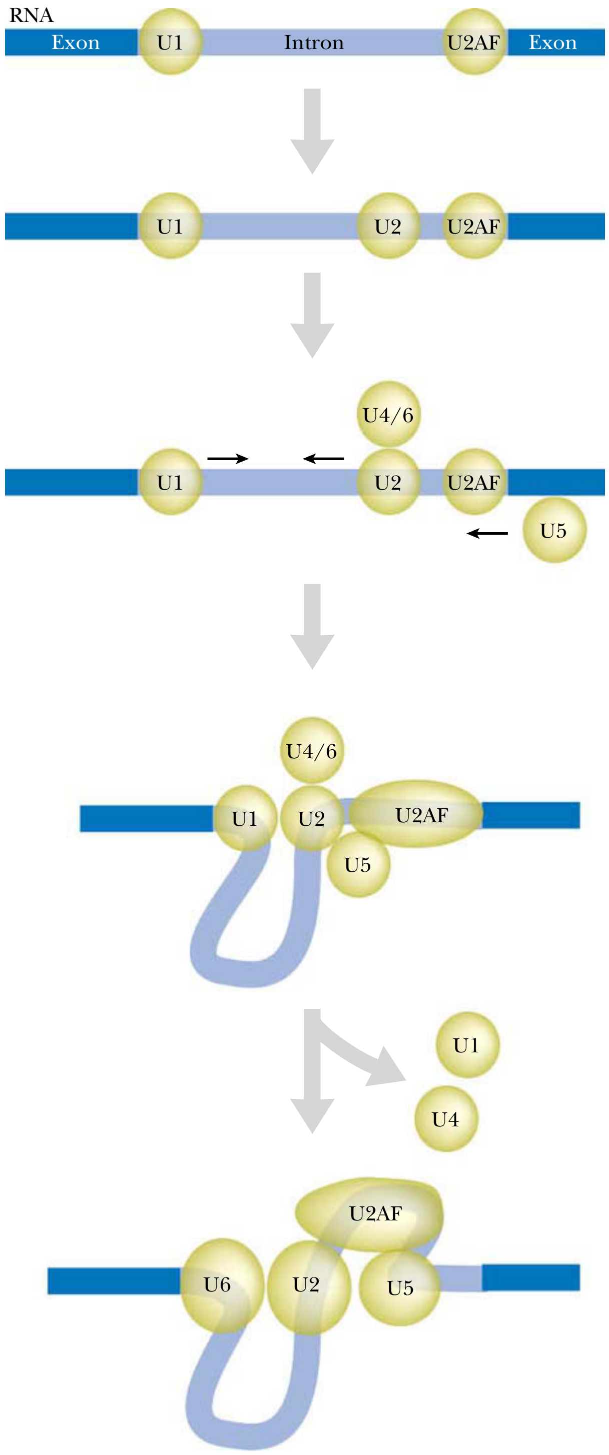


FIGURE 12.10 ***Stages in Spliceosome Assembly***

A) U1 binds the 5′ splice site and U2AF binds the 3′ splice site. B) U2 binds the branch site. C) U4 and U6 bind to U2 and U5 binds to the downstream exon. D) A loop forms by association of U1 and U2. E) U6 displaces U1 from the spliceosome and U4 departs as well.

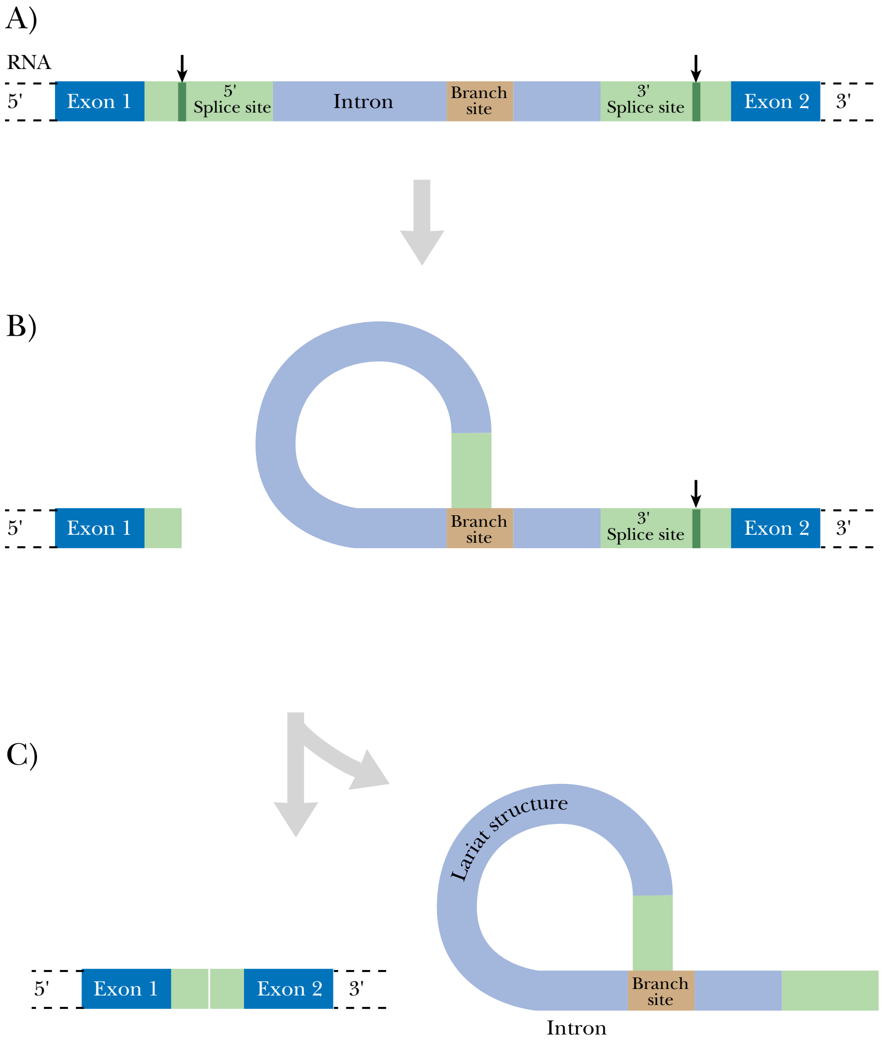


FIGURE 12.11 ***The Splicing Reactions***

A) Two exons and the intervening intron with their 5′ and 3′ splice sites and branch site are shown prior to splicing. B) The 5′ splice site is cut first and the free end of the intron loops back to bind the branch site. The 3′ splice site is severed next and the ends of the two exons are joined. The intron is released as a lariat structure.

Different Classes of Intron Show Different Splicing Mechanisms

Certain introns act as ribozymes and splice themselves out.

There are several classes of introns (Table 12.01). The GT-AG (or GU-AG in RNA code) introns described above are by far the most frequent in eukaryotic nuclear genes. The AT-AC (or AU-AC) introns are extremely similar to the GT-AG introns except for their different intron boundary sequences. They are processed in an almost identical manner, by a different, but closely related, set of splicing factors.

Group I introns are **self-splicing**. The RNA itself provides the catalytic activity and thus acts as an RNA enzyme or ribozyme. No proteins are required for splicing. The folding of the RNA into a series of base-paired stem and loop structures is needed for ribozyme activity. The 3-dimensional structure is folded so as to bring the two splice sites together and to strain the bonds that will be broken. The reaction pathway starts with the guanosine of any of GMP, GDP or GTP attacking the 5′ splice site (Fig. 12.12) and cutting the exon and intron apart. Note that the guanosine nucleotide is free in solution and is not part of the RNA. The free exon-3′-OH then reacts with the downstream splice site. Group I introns include those in the rRNA of lower eukaryotes, such as the single-celled, ciliated freshwater protozoan, *Tetrahymena*. However, most are found in genes of mitochondria and chloroplasts. Occasional cases occur in bacteria and bacteriophages.

self-splicing Splicing out of an intron by the ribozyme activity of the RNA molecule itself without the requirement for a separate protein enzyme

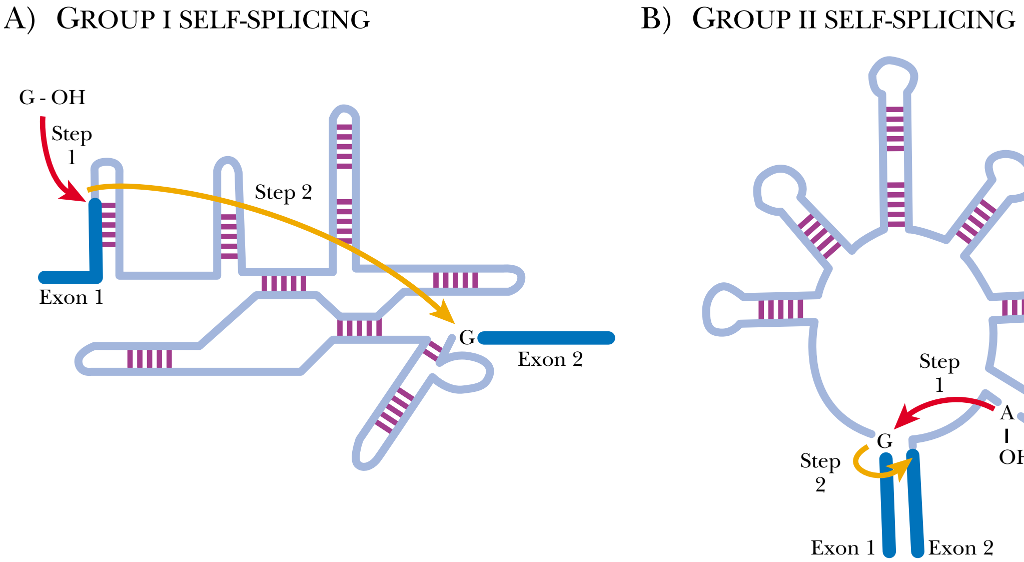


FIGURE 12.12 ***Group I and Group II Intron Self-Splicing Reactions***

In both Group I and Group II self-splicing introns, the intron folds up so as to bring the ends of the two exons together (not shown). For clarity, here we have indicated only the first level of intron folding by base pairing. A) In Group I introns, the 5′ splice site is attacked by a soluble guanosine nucleotide which cuts the exon and intron apart. Next, the free 3′ OH group of the exon reacts with the 3′ splice site and promotes the joining of the two exons (which are actually held close together). B) The self-splicing reaction in Group II introns is similar to Group I, except that an internal adenosine initiates splicing.

Group II introns are found in the organelles of fungi and plants and occasional examples occur in prokaryotes. Group III introns are found in organelles. Both classes are also self-splicing. However, the reaction is started by attack of an internal adenosine (not a free nucleotide as in Group I introns) (Fig. 12.12). This results in a lariat structure being formed, as in the typical nuclear pre-mRNA introns described above. These three types of intron may thus have a common evolutionary origin. Group III introns are similar to Group II introns, but are smaller and have a somewhat different 3-dimensional structure.

Twintrons are complex arrangements in which one intron is embedded within another. They consist of two or more Group I, Group II or Group III introns. Since introns are embedded within other introns, they must be spliced out in the correct order, innermost first, rather like dealing with parentheses in algebra.

Introns are found in Archaea but are removed by simple ribonucleases without needing a splicesome.

Archeal introns are found in tRNA and rRNA and are similar in some respects to eukaryotic pre-tRNA introns. No complex splicing occurs; no snurps are needed and no ribozymes are involved. The tRNA and rRNA precursors fold up into their normal 3D structures with the intron forming a loop. This loop can be cut out by a ribonuclease and the ends joined by an RNA ligase. Their stable 3-dimensional structures hold the two halves of the tRNA and rRNA molecules together during cleavage and ligation and there is no need for extra factors such as snRNPs for recognition or processing.

Alternative Splicing Produces Multiple Forms of RNA

Although any particular splice junction must be made with total precision, eukaryotic cells can sometimes choose to use different splice sites within the same gene. Generally, **alternative splicing** is used by different cell types within the same animal. This

alternative splicing Alternative ways to make two or more different final mRNA molecules by using different segments from the same original gene

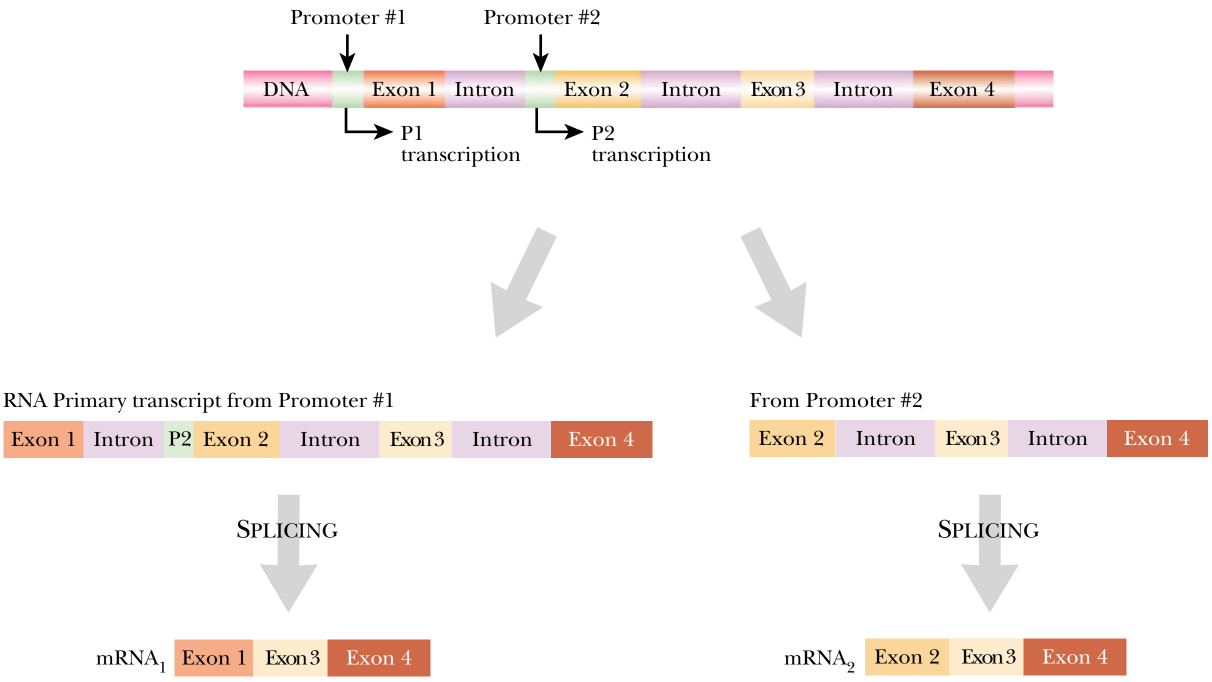


FIGURE 12.13 ***Alternative Promoter Selection***

The DNA contains two promoters that can produce two primary transcripts. If promoter #1 is used, then the segment containing promoter #2 and exon #2 is spliced out. If promoter #2 is used, then exon #1 is not even part of the primary transcript and is therefore not in the mRNA.

allows a single original DNA sequence to be used to make several different proteins that have distinct but overlapping functions. At first glance it might seem that alternative splicing provides, at least in theory, a way for each gene to encode multiple proteins, hence increasing the total number of different proteins available to an organism. However, selection of which alternative splice site to use in a particular cell or tissue must itself be controlled, and this often requires several additional proteins.

There are three main and one rare type of alternative splicing:

Different proteins can be made from the same gene by alternative splicing during formation of the mRNA.

Alternative Promoter Selection occurs when two alternative promoters are utilized. The choice of which promoter to use depends on cell-type specific transcription factors. As figure 12.13 shows, two alternative transcripts result in two different messenger RNAs.

Alternative Tail Site Selection is operative when alternative sites for adding the poly(A) tail are possible. The choice between them again depends on cell type. In this case, cleavage at the earlier poly(A) site results in loss of the distal exon (Fig. 12.14). If the later poly(A) site is chosen, then the earlier poly(A) site and the exon just in front of it are spliced out. This mechanism is used to produce antibodies which recognize the same invading, foreign molecule but which have different rear ends. One type of antibody is secreted into the blood, whereas the other type remains attached to the cell surface.

Alternative Splicing by Exon Cassette Selection involves a genuine choice between actual splicing sites. Depending on the choice made, a particular exon may or may not appear in the final product as shown in Figure 12.15. Here the primary transcripts are actually the same. They are drawn differently in Figure 12.15 to illustrate the splicing plans. Some cell-type specific factor which recognizes the different possible splice sites must come into play here, but the details are still obscure.

exon cassette selection Type of alternative splicing that makes different mRNA molecules by choosing different selections of exons from the primary transcript

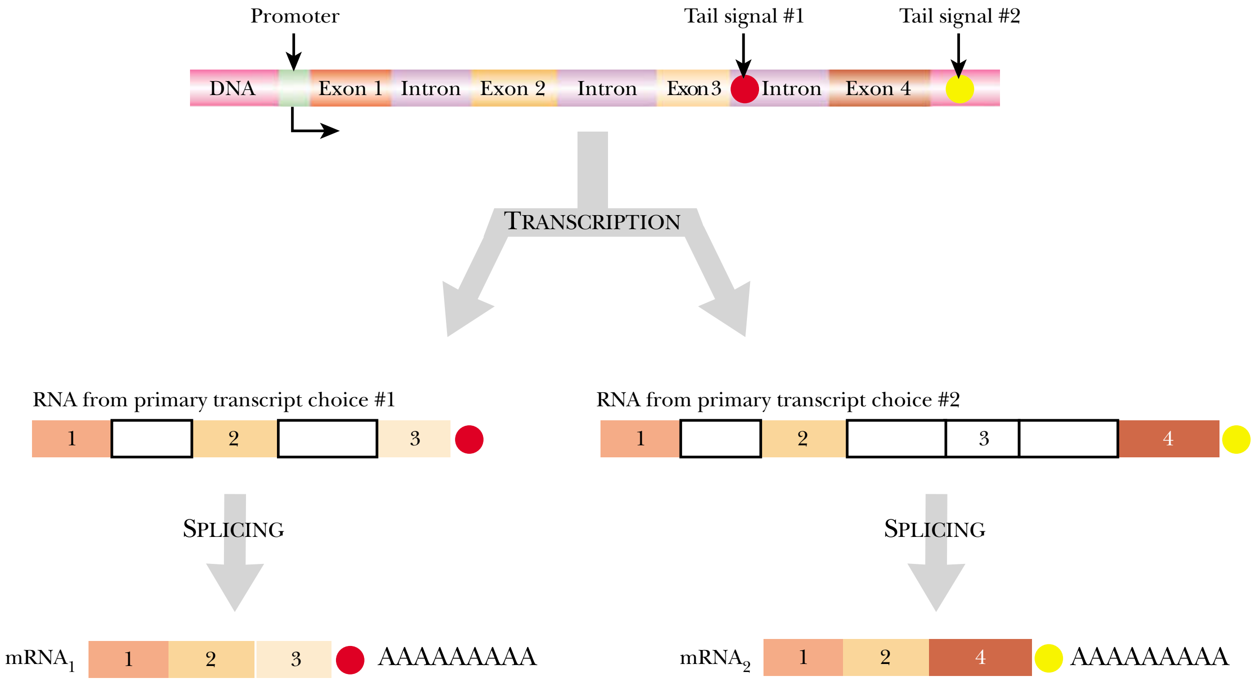


FIGURE 12.14 ***Alternative Tail Site Selection***

In the DNA shown there are two tail signals. Use of the first tail signal results in an mRNA including exons #1, #2 and #3. Use of the downstream tail signal results in an mRNA containing exons #1, #2 and #4.

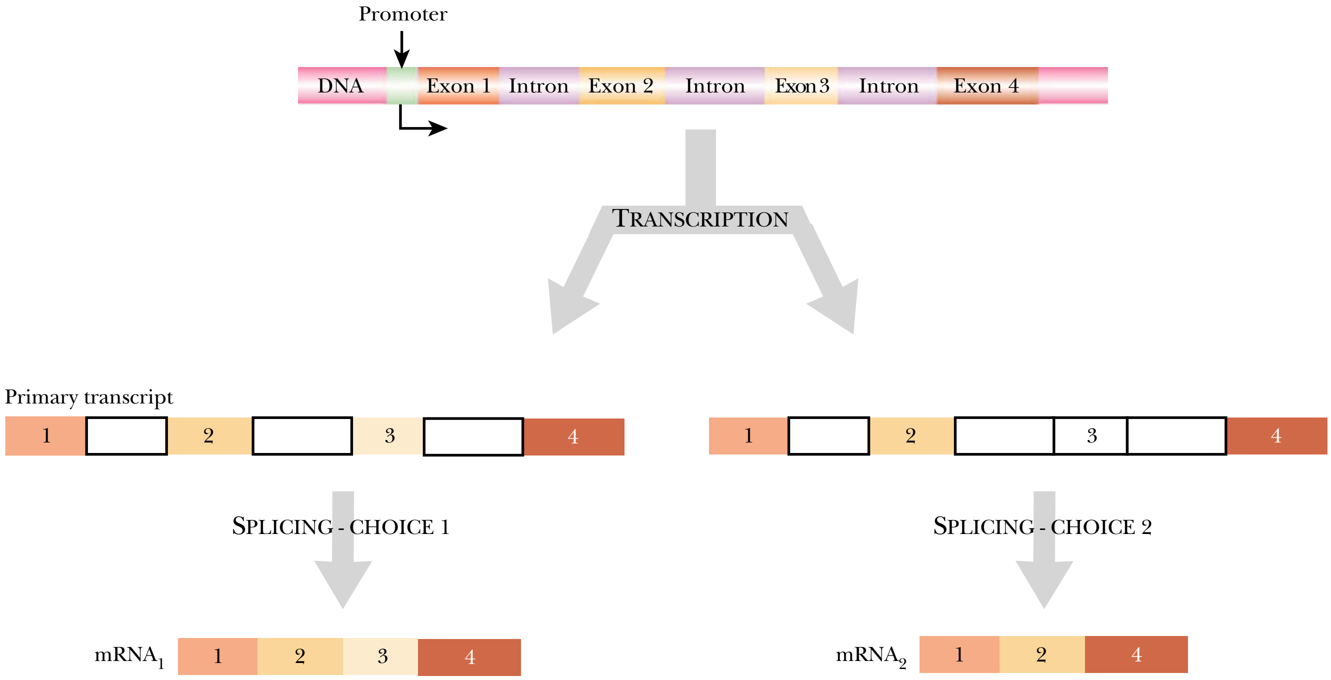


FIGURE 12.15 ***Alternative Splicing by Exon Cassette Selection***

In this example, the DNA contains four exons with intervening introns. Transcription yields a single primary transcript that is spliced in two alternative ways. The same primary transcript is drawn twice in different ways to illustrate the two splicing plans. In one case all four exons are found in the final mRNA whereas in the other case exon #3 plus the two surrounding introns is removed as a unit.

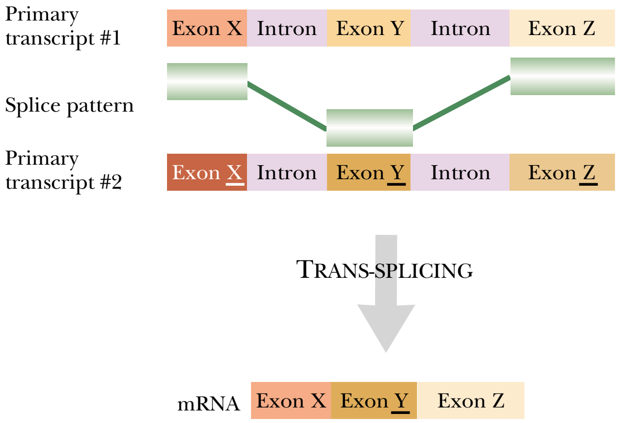


FIGURE 12.16 ***Trans-Splicing of Trypanosome mRNA***

Two primary transcript RNA molecules are depicted. The splicing shown combines two of the exons from the first RNA with one from the second RNA to give the final messenger RNA.

Exon cassette selection occurs in the gene for the skeletal muscle protein troponin T. In the rat this gene has 18 exons. Of these, 11 are always used. Five (exons 4 through 8) may be used in any combination (including none used) and the final two (exons 17 and 18) are mutually exclusive, and one or the other must be chosen. This gives a theoretical mind-boggling 64 possible final mRNAs. The result is that muscle tissue contains multiple forms of this structural protein. The details of troponin splicing vary substantially among different vertebrates. Other muscle proteins quite often show similar multiple forms.

Trans-Splicing, although rare, splices together segments from two different primary transcripts. **Trypanosomes** are parasitic single-celled eukaryotes that cause sleeping sickness and other tropical diseases. They evade immune surveillance by constantly changing the proteins on their cell surfaces by the genetic trick of shuffling gene parts (see Ch. 19). In addition they indulge in the **trans-splicing** of many genes (Fig. 12.16). On the other hand, trypanosomes do not appear to have introns and so do not have normal splicing! Although it has not (yet!) been found in vertebrates, trans-splicing of segments from one RNA molecule into another also occurs in nematodes and in the chloroplasts of plant cells.

Inteins and Protein Splicing

Intervening sequences that splice themselves out are occasionally found in proteins.

Occasional intervening sequences are found that are spliced out at the protein level. Such protein splicing is rare, which is why it was only noticed relatively recently. **Inteins** and **exteins** are the protein analogs of the introns and exons found in the DNA and RNA. In other words, inteins are intervening sequences in proteins that are present when the protein is first made, but are later spliced out. The final protein is made of the exteins that are now joined together (Fig. 12.17). Inteins have been found in yeasts, algae, bacteria and archaebacteria.

Intein splicing involves no accessory enzymes. The intein segment catalyzes its own release as a free polypeptide. Certain specific amino acids must be present at the extein/intein boundaries for the splicing reaction to work. The splicing occurs in two steps, with a branched intermediate. Serine (or cysteine) must be the first amino acid of the downstream extein, as its hydroxyl group (or sulfhydryl if cysteine is used) is needed to carry the upstream extein during the branched stage (Fig. 12.18).

Usually there is just a single intein per protein, but one example is known where two inteins are inserted into the same host protein. More bizarre is the case of the

extein A segment of a protein that remains after the splicing out of any inteins
intein An intervening sequence in a protein—a segment of a protein that can splice itself out
trans-splicing Splicing of a segment from one RNA molecule into another distinct RNA molecule
trypanosome Type of single-celled eukaryotic microorganism that lives as a parasite in higher animals and causes diseases such as sleeping sickness

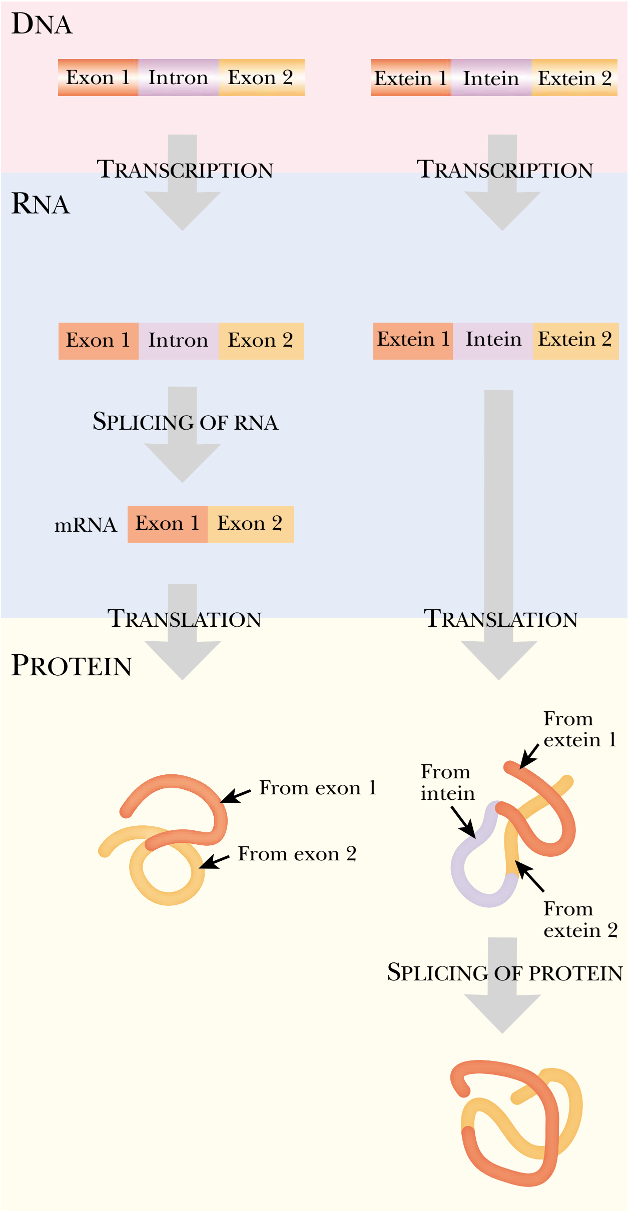


FIGURE 12.17 ***Inteins and Exteins in Proteins***

On the left is the standard scheme by which introns are eliminated during RNA splicing. The intron is eliminated at the level of RNA and is never translated into protein. On the right is the scheme for removal of intervening sequences at the protein level. Regions remaining in the final protein are called exteins and those destined to be lost are called inteins. The major difference from RNA splicing is that in protein splicing the inteins are cut out after the protein is made.

dnaE gene of *Synechocystis* (a blue-green bacterium). This gene is split in two and each half has part of the DNA sequence for an intein attached. These two half-genes are transcribed and translated separately to give two proteins (Fig. 12.19). These fold up together and even though the intein is separated into two segments it still manages to cut itself out. The two halves of the DnaE protein are joined together as the intein splices itself out. The intein is released as two fragments.

If the DNA segment that codes for the intein were deleted, the protein would be made in one step, with no intein and no need for protein splicing. And the intein itself would be extinct. However, the spliced-out intein polypeptide is not just a waste product; it is a site-specific deoxyribonuclease (DNase). Its function is to protect the

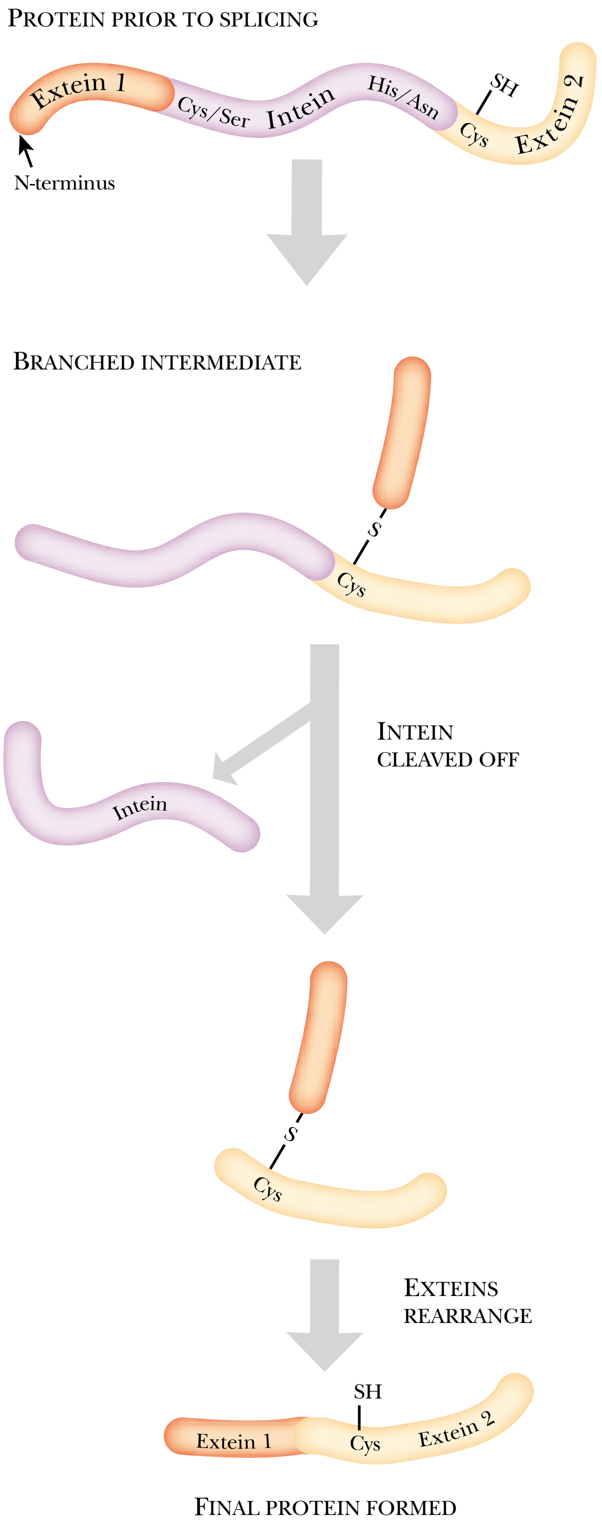


FIGURE 12.18 ***Branched Intermediate during Intein Splicing***

The intervening intein segment splices itself out in two stages. The intein has a Cys or Ser at the boundary with extein 1 and a basic amino acid at its boundary with extein 2. The downstream extein (#2) has a Cys residue at the splice junction. Extein 1 is cut loose and attached to the sulfur side chain of the cysteine at the splice junction. This forms a temporary branched intermediate. Next the intein is cut off and discarded and the two exteins are joined to form the final protein.

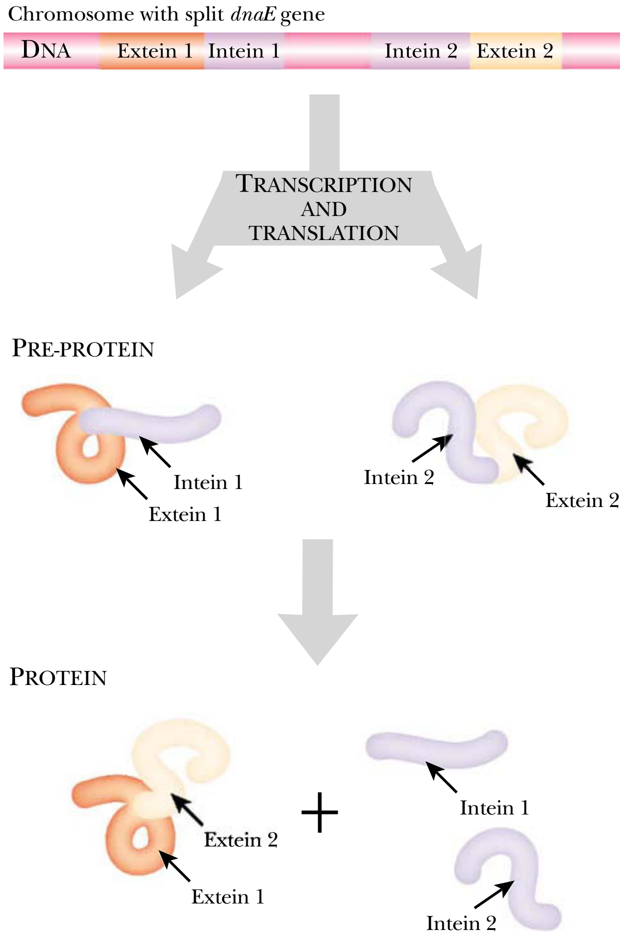


FIGURE 12.19 ***Intein Splicing Reconstitutes the DnaE Protein***

The DNA coding for the DnaE protein of *Synechocystis* is transcribed and translated into two separate proteins, each containing an intein and an extein. The exteins of the two proteins are spliced together by the inteins. During splicing both inteins are lost.

existence of the intein. If a mutation occurs that deletes the intein DNA sequence from the middle of the host gene, the previously formed intein DNase cuts the cell's DNA at this point—a potentially lethal move. Thus, any cell with a single copy of DNA that deletes the useless intein DNA will be killed by the intein protein. Only cells that keep the intein survive. Inteins appear unnecessary for cell survival and intein-encoding DNA may be therefore be regarded as a form of selfish DNA. The origin of inteins is obscure.

Cells that delete the intein DNA are killed by the intein protein.

In eukaryotes there are two copies of each gene. If one copy loses the intein DNA sequence, it will be cut into two fragments by the intein DNase. Yeast and many other eukaryotes can mend double-stranded breaks in their DNA by a special form of recombination. The second (undamaged) copy of the gene where the break occurred lends one strand of its DNA to mend the break. After the single stranded regions are filled in, the result is a repaired copy of the gene that is identical to the undamaged copy. Now both copies again have the intein DNA sequence inserted. This type of repair process is known as **gene conversion** (Fig. 12.20).

Although most inteins have DNase activity, some shorter inteins exist that do not. Possibly they are defective and have lost the original sequences that encoded the nuclease. Inteins are not alone in their attempts to kill cells that delete their encoding DNA sequence. Certain introns use the same tactic. In this case, the DNA of the intron is not just meaningless, but encodes a DNase that cuts in two any copies of the host gene that have lost the intron. Certain plasmids also have mechanisms to kill any cell that loses the plasmid. They use a different and more complicated approach, as plasmids are not inserts in host DNA and so there is no insertion site to recognize. In all such cases, the idea is that only host cells that carry the selfish DNA will survive.

gene conversion Recombination and repair of DNA during meiosis that leads to replacement of one allele by another. This may result in a non-Mendelian ratio among the progeny of a genetic cross

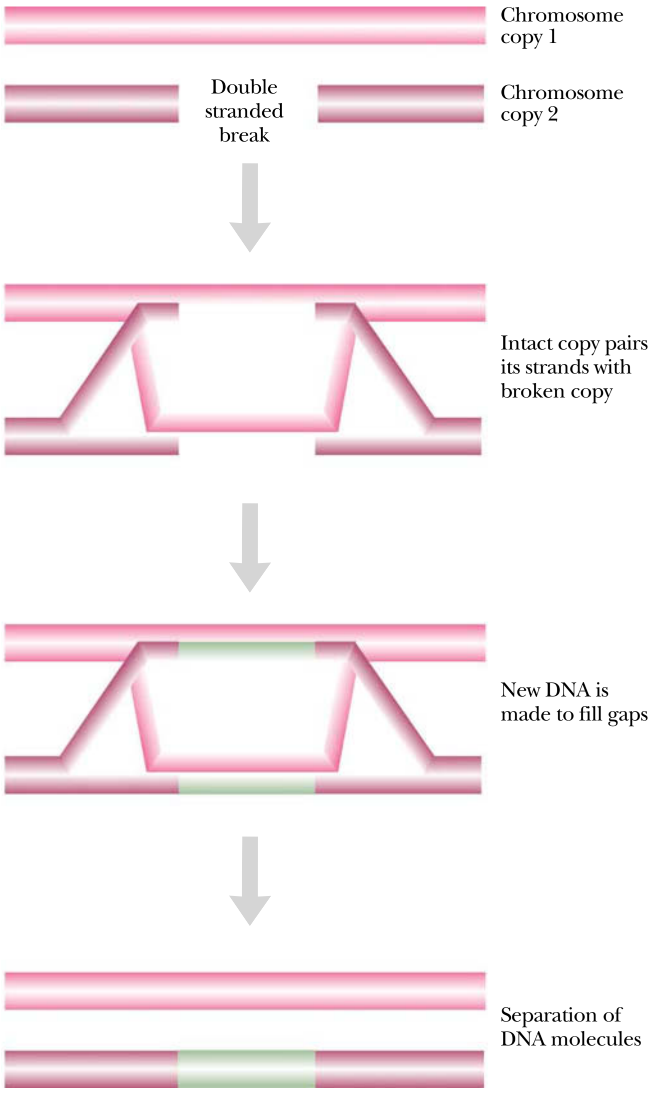

Chromosome copy 1
Chromosome copy 2
Double stranded break
Intact copy pairs its strands with broken copy
New DNA is made to fill gaps
Separation of DNA molecules

FIGURE 12.20 ***Gene Conversion Repairs Broken Chromosomes in Eukaryotes***

A double-stranded break in one member of a pair of chromosomes can be repaired. Repair makes use of the intact chromosome and involves base pairing of the fragmented DNA with the intact DNA. After recognition, the base pairs in the gap are filled in and, when complete, the chromosomes separate.

Base Modification of rRNA Requires Guide RNA

Modified bases are frequently found in tRNA and rRNA.

RNA molecules often contain modified bases. These are made by chemical modification of pre-existing bases. This is especially true of tRNA, which contains a relatively high percentage of many different modified bases (Ch. 8). However, ribosomal RNA has several modified bases and even occasional mRNA molecules may have a modified base or two.

Short RNA guide molecules are used to locate the sites for base modification in eukaryotic rRNA.

In the case of tRNA, individual enzymes are sufficient for the various base modifications. These recognize particular bases in specific regions of various tRNA molecules and modify them. This scenario also applies to bacterial rRNA, which is modified in only a handful of locations. In the case of eukaryotic rRNA, modification occurs at multiple sites and requires small RNA molecules in addition to the modification enzymes. These small RNA molecules are needed to locate the correct sites for modification, which they do by base pairing over a short region with the rRNA. Synthesis and processing of rRNA in eukaryotes occurs in the **nucleolus** and so the guide RNAs are known as small nucleolar RNAs (snoRNAs).

nucleolus Region of the nucleus where ribosomal RNA is made and processed

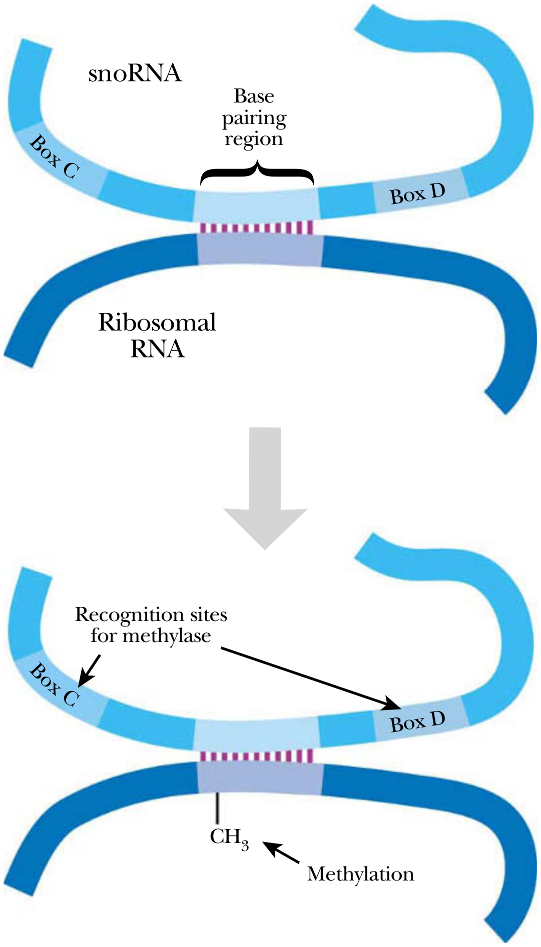


FIGURE 12.21 ***Recognition of Modification Sites on rRNA by snoRNA***

The site on the rRNA to be modified is identified by base pairing to specific sequence on the snoRNA. After snoRNA/rRNA pairing, the methylase binds to the *BoxC* and *BoxD* sequences and then methylates one of the bases of the rRNA.

Methylated bases and pseudouridine are the most common modifications in eukaryotic rRNA.

Nucleotides in eukaryotic rRNA are modified by methylation of the 2′-OH group of the ribose or by converting uridine to **pseudouridine**. Although the number of different types of modification is limited, the number of sites is very large. Thus human pre-rRNA is methylated at 106 positions and pseudo-uridylated at 95. The base sequences around the modification sites are rarely related and so there is no consensus sequence for a modifying enzyme to use. Instead there is a different snoRNA for each modification site. Each snoRNA is 70–100 nucleotides long and has a unique sequence that recognizes the modification site on the rRNA. In addition, all snoRNAs that recognize potential methylation sites share a sequence motif that is recognized by the methylase that modifies the bases (Fig. 12.21). Similarly, the family of snoRNAs that recognize pseudo-uridylation sites have a motif that binds the pseudo-uridylation enzyme.

Interactions between rRNA and snoRNA often involve G-U base pairing. This non-standard base pair is stable in **dsRNA** and also occurs in the base pairing between the guide RNA and mRNA in RNA editing (see below).

Many guide RNA molecules (snoRNA) are encoded inside introns.

Since there are many base modifications there are several hundred different snoRNAs per cell in eukaryotes. Only a few snoRNAs are transcribed from standard genes. Most snoRNAs are encoded in the introns of other genes. The snoRNA is made by cutting up the intron after it has been spliced out of the mRNA (Fig. 12.22). Many of the genes whose introns contain snoRNA sequences encode ribosomal components, e.g. U16 snoRNA is encoded by part of intron 3 from the gene for L1 ribosomal protein.

dsRNA Double-stranded RNA
pseudouridine An isomer of uridine that is introduced into some RNA molecules by post-transcriptional modification

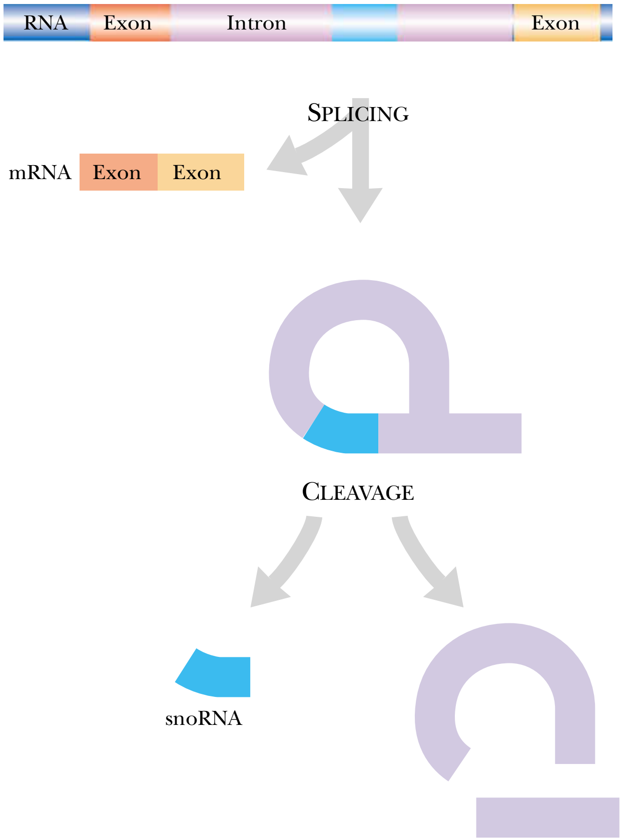


FIGURE 12.22 ***Generation of snoRNA from Intron***

After splicing the mRNA, the intron assumes a lariat structure. The snoRNA is spliced from the intron, leaving smaller fragments of RNA.

RNA Editing Involves Altering the Base Sequence

RNA editing involves changing the actual base sequence of messenger RNA.

Perhaps the most bizarre modification that messenger RNA may undergo is the alteration of its base sequence—known as **RNA editing**. Altering bases within the coding sequence of mRNA usually changes the final protein product that will be obtained. Perhaps not surprisingly RNA editing is very rare in most organisms. In mammals RNA editing is restricted to base substitution and may consist of C-to-U or A-to-I (inosine) changes and relatively few specific cases of mRNA editing are known. In plants both C to U and U to C editing occur quite frequently. More radical editing of mRNA occurs in certain protozoa where bases are inserted and deleted.

An example of C-to-U editing occurs in the human gene for apolipoprotein B, which encodes a protein of 4,563 amino acids, one of the longest polypeptide chains on record.

The full-length protein, apolipoprotein B100, is made in liver cells and secreted into the bloodstream. ApoB100 is required for the assembly of very low density lipoprotein (VLDL) and low density lipoprotein (LDL) particles which carry lipids, including cholesterol, around the body. A short version with only 2,153 amino acids, apolipoprotein B48, is made by intestinal cells. It is secreted into the intestine where it plays a role in the intestinal absorption of dietary fats. The shorter apoB48 lacks the region of apoB100 (between residues 3,129 to 3,532) that is bound by the LDL receptor. Consequently, dietary fat carried by apoB48 is largely absorbed by the liver whereas VLDL or LDL particles with apoB100 can deliver cholesterol to peripheral tissues whose cells possess the LDL receptor.

RNA editing Changing the coding sequence of an RNA molecule after transcription by altering, adding or removing bases

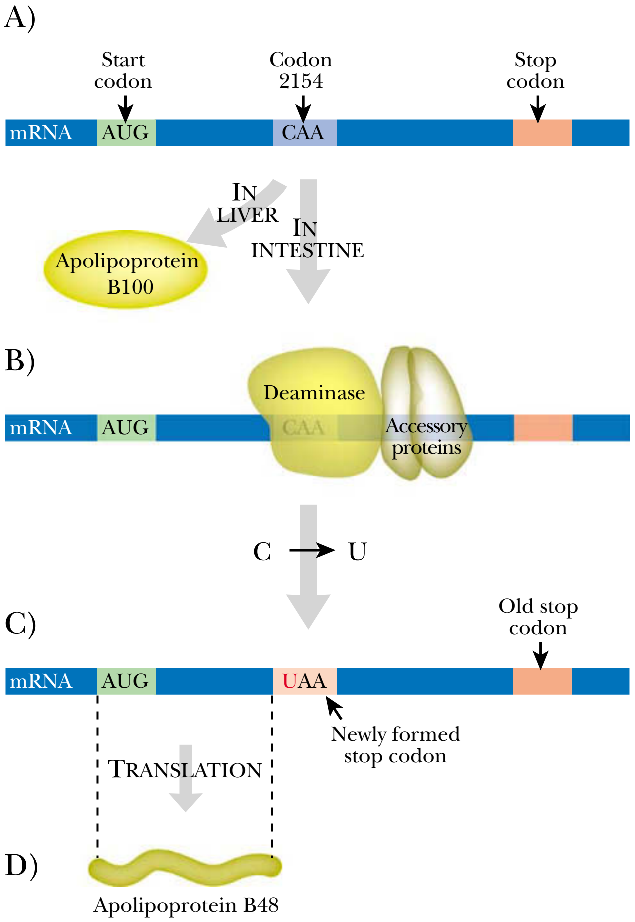


FIGURE 12.23 *Editing of Apolipoprotein B mRNA*

A) The apolipoprotein B gene normally makes apolipoprotein B100 in the liver. B) In the intestine the mRNA is altered by base editing to make apolipoprotein B48. A deaminase binds to a CAA codon in association with accessory proteins. C) The cytosine in the CAA codon is converted to uracil giving UAA. D) The UAA serves as a new stop codon that halts translation in the middle of the coding region. This results in a truncated protein, apolipoprotein B48.

The short version, apoB48, is encoded by the same gene as apoB100 and is made by editing the mRNA. The CAA codon (for glutamine) at position 2,154 is changed to a UAA stop codon by an enzyme that deaminates this specific cytosine, so converting it to uracil. Several accessory proteins are needed to make sure that the **deaminase** binds only at the correct site (Fig. 12.23).

Although only a single gene is needed to produce two versions of apolipoprotein B, editing needs several extra proteins to recognize the editing site and convert the C to U. Having two apolipoprotein B genes of different lengths would surely be more economical and there is no known reason why the more complicated procedure of mRNA editing is used.

A-to-I editing of mRNA also occurs in mammals. In this case adenosine is converted by deamination into inosine by dsRNA adenosine deaminase. Recognition is due to formation of double-stranded regions by base pairing between the modification sites and sequences from neighboring introns. Thus, the intron sequence affects the final coding sequence of the mature mRNA. Consequently, this editing must occur before removal of the introns. The inosine generated acts as guanosine during translation and therefore A-to-I editing may change the sequence of the resulting protein if it occurs within a coding region. This happens in the mRNAs for both the glutamate and serotonin receptor proteins of the mammalian nervous system. Defects in editing result in severe neurological symptoms. A-to-I editing also occurs in the non-coding regions of a significant number of other genes. The effects of these changes are for the most part still uncertain.

Several proteins of the mammalian nervous system undergo A to I editing in the coding sequence of their mRNAs.

deaminase An enzyme that removes an amino group

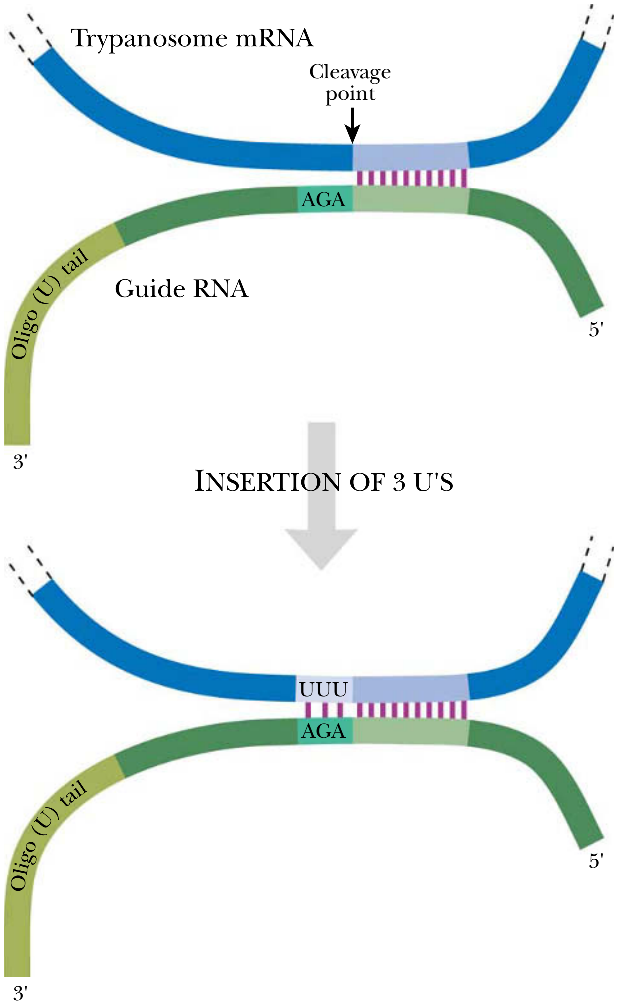


FIGURE 12.24 ***Editing of Trypanosome mRNA***

The guide RNA base pairs along a specific region of the trypanosome mRNA. The extra adenine (A) of the AGA sequence in the guide RNA is used as a template to insert uracil (U) into the mRNA. (Note: some slightly distorted duplex RNA structures allow guanine, as in the AGA sequence shown, to pair with uracil.)

C to U and U to C editing of mRNA are also found in the mitochondria and chloroplasts of most major groups of plants. Typically from three or four to twenty bases are changed in those transcripts that are edited in plant organelles. In most cases, such editing results in changes in the amino acid sequence of the encoded protein that are necessary for full activity. However, silent editing is occasionally observed. The ultimate in pointlessness appears to be the case of the tobacco chloroplast *atpA* gene where a CUC codon is edited to CUU in the mRNA. Both codons encode serine. Conceivably, such silent editing might be an adjustment to the differential availability of tRNAs with different anticodons, however there is no evidence for this.

Trypanosomes frequently edit their mRNA by inserting or removing bases.

Trypanosomes practice RNA editing relatively often. Moreover, they do not merely modify bases chemically but actually insert or remove them. Some of the primary transcripts of trypanosomes, especially those from the mitochondrial genes, are altered by insertion or removal of multiple uridine nucleotides, one at a time, before the final mRNA is generated (Fig. 12.24). In these cases, the coding sequences found on the DNA have incorrect reading frames. If the trypanosome did not edit its mRNA, the result would be defective, frame-shifted proteins made from out of phase coding sequences.

Multiple insertions of U in trypanosome mRNAs occur at positions specified by short **guide RNAs**. These are complementary to short stretches of the mRNA but have an extra A. The U residues are inserted into the mRNA opposite the extra A on the guide RNA.

guide RNA Small RNA used to locate sequences on a longer mRNA during RNA editing

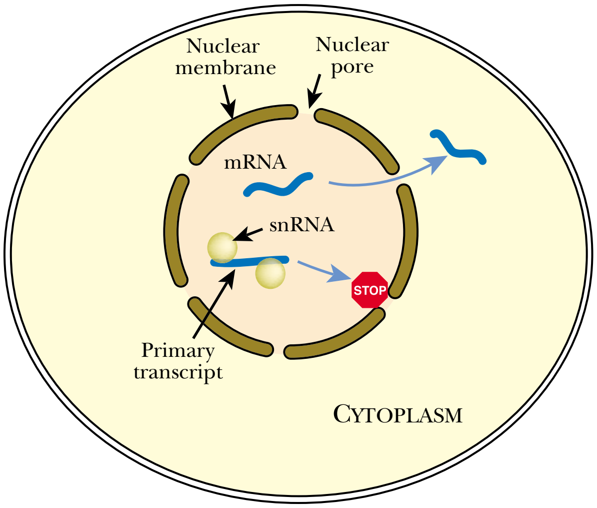


FIGURE 12.25 ***Transport of Eukaryotic mRNA out of the Nucleus***

Processed mRNA is free to leave the nucleus whereas primary transcript still attached to snRNA is prevented from leaving.

Transport of RNA out of the Nucleus

The nucleus is surrounded by a double membrane. Each nucleus has many pores that allow molecules in or out in a carefully controlled manner. Each **nuclear pore** is surrounded by a cluster of proteins that control entry and exit, but the details of precisely which molecules are allowed in or out are still murky. We do know that once messenger RNA has received its cap and tail and had its introns spliced out, it is free to exit the nucleus. Binding of the spliceosome to the RNA prevents it from leaving until splicing is finished (Fig. 12.25).

Transport out of the nucleus of large molecules like RNA and proteins requires energy. This is obtained by the hydrolysis of GTP. Protein factors known as exportins and importins control the exit and entry of specific classes of molecules through the nuclear pores. For example exportin-t is specific for export of tRNA.

Degradation of mRNA

Messenger RNA is degraded after a relatively short lifetime.

Messenger RNA molecules are relatively short-lived and in bacteria the half-life is generally only a couple of minutes. Messenger RNA that is not bound to ribosomes is especially vulnerable to degradation. Bacteria contain multiple **ribonucleases**. These are involved both in processing of tRNA and rRNA and the degradation of mRNA. These ribonucleases can often substitute for one another at least to some extent and so mutants that have lost only one ribonuclease are usually still viable. Bacterial mRNA is degraded in two stages (Fig. 12.26). First, an **endonuclease**, usually ribonuclease E, cleaves regions that are unprotected by ribosomes. Next, **exonucleases** that move in a 3′ to 5′ direction degrade the fragments. Note that *overall* degradation moves in a 5′ to 3′ direction, due to the endonuclease following the ribosomes.

Eukaryotic mRNA must have the tail and cap removed before degradation can proceed.

Degradation of mRNA follows a different route in eukaryotes such as yeast. First the poly (A) tail and then the cap must be removed before actual nuclease digestion. Once the poly (A) tail is shortened to 10–20 bases the poly (A)-binding protein (PABP) is released. Only after the PABP has gone can the cap be removed. Once the cap has also gone, an exonuclease, Xrn1, degrades the mRNA in the 5′ to 3′ direction (Fig. 12.27).

endonuclease A nuclease that cuts a nucleic acid in the middle
exonuclease A nuclease that cuts a nucleic acid at the end
nuclear pore Pore in nuclear membrane that allows proteins and RNA into and out of the nucleus
ribonuclease A nuclease that cuts RNA

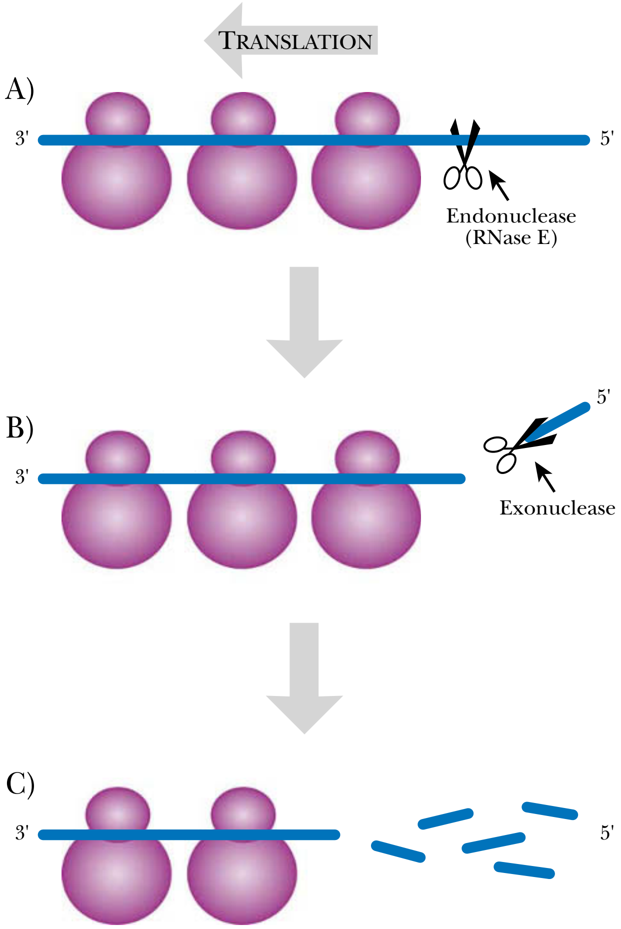


FIGURE 12.26
Degradation of Prokaryotic mRNA

A) The mRNA being translated has a 5′ end unprotected by ribosomes. An endonuclease (RNase E) cuts near the 5′ end. B) The released fragment is then cut by exonuclease, starting at its 3′ end. C) The mRNA is shortened and the severed segment is degraded.

The stability of eukaryotic mRNA depends on the presence or absence of destabilizing sequences. Short-lived mRNA often contains an AU-rich sequence of about 50 bases (known as an ARE) in its 3′-UTR. The consensus for the ARE is multiple repeats of the 5 base sequence AUUUA (hence ARE = AUUUA repeat element). An ARE-binding protein recognizes the ARE and promotes deadenylation and degradation (Fig. 12.28).

Nonsense Mediated Decay of mRNA

Special mechanisms detect and destroy defective mRNA molecules.

Eukaryotic cells possess a special RNA surveillance mechanism that destroys mRNA molecules that contain premature stop codons. Such defective mRNA molecules may result from expression of genes with nonsense mutations, in which a codon that codes for an amino acid is mutated into a stop codon, also referred to as a nonsense codon. Consequently, the mechanism for destroying these mRNAs is known as nonsense-mediated decay or NMD.

NMD plays a protective role in eukaryotic organisms that are heterozygotes with one functional allele and one allele carrying a nonsense mutation. Full expression of the nonsense allele would result in production of a truncated protein. Sometimes this is merely a waste of resources. However, many polypeptides form multi-subunit complexes (either with themselves or with other polypeptides). In this case, aberrant forms of the protein may still bind to the complex and interfere with normal function. Hence, some truncated proteins are actively dangerous. Degradation of mRNA carrying the nonsense allele prevents the synthesis of the aberrant truncated proteins and so protects heterozygotes from possible deleterious effects.

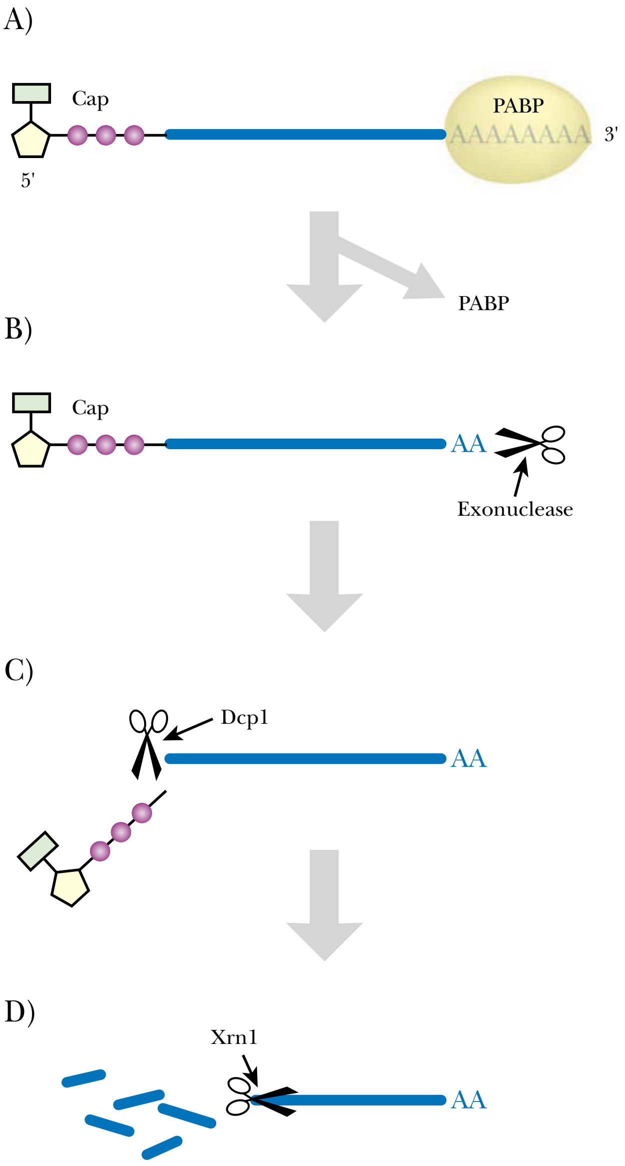


FIGURE 12.27
Degradation of Eukaryotic mRNA

A) The mRNA is shown with the poly (A) tail bound to poly (A)-binding protein (PABP). B) An exonuclease sequentially removes the poly (A) tail. C) A decapping protein (Dcp1) removes the cap. D) Degradation of the mRNA proceeds in the 5′ to 3′ direction due to exonuclease Xrn1.

Despite its name, nonsense-mediated decay probably evolved to deal with defective mRNA that results from errors in the expression of normal genes rather than from inherited mutations. In particular, errors during the complex RNA splicing process that removes introns can lead to defective mRNA molecules. It is notable that nonsense-mediated decay is only found in eukaryotes but not in prokaryotes where splicing is largely absent.

Defective mRNA with premature stop codons may result from several events:

I. Expression of a mutant gene with an internal nonsense mutation.

II. Errors during expression of normal genes that create nonsense mutations.

  a. Errors during transcription that insert incorrect bases resulting in a premature stop codon.

  b. Errors during splicing that alter the reading frame and so create in frame stop codons.

  c. Errors during splicing that result in the retention of all or part of an intron whose sequence includes in frame stop codons.

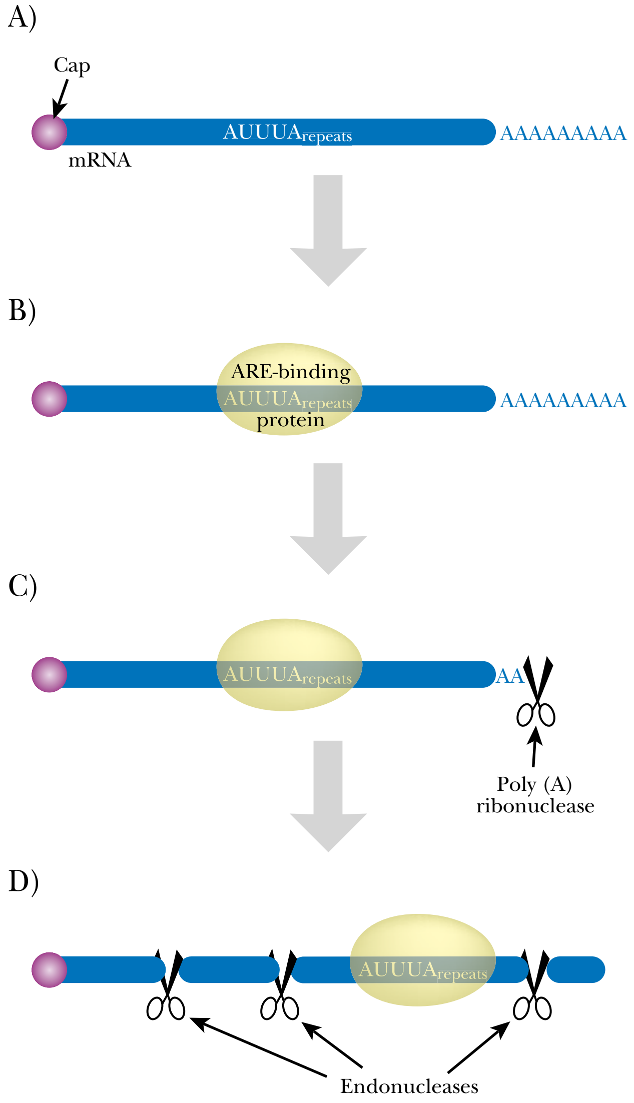


FIGURE 12.28 ***ARE Sequence Facilitates Digestion of Eukaryotic mRNA***

A) The structure of undegraded eukaryotic mRNA shows the AUUUA repeat sequence, the cap and the poly (A) tail. B) ARE-binding protein recognizes a repeated AUUUA sequence. C) Poly (A) ribonuclease degrades the poly (A) tail. D) Endonucleases sever the mRNA at multiple sites.

d. Errors during RNA editing (presumably, although not directly demonstrated).

Nonsense-mediated decay of mRNA is triggered by the presence of premature stop codons.

Nonsense-mediated decay is triggered whenever there is a stop codon more than 50–55 nucleotides upstream of the final exon-exon junction created during splicing. This requires that the location of the exon-exon splice junctions should somehow be marked on the mature mRNA. In animal cells exon-exon junctions are labeled when the primary transcript is converted to the mature mRNA during the splicing process. A complex of proteins, known as the exon junction complex (EJC) is bound to the mRNA about 20–24 nucleotides upstream of each exon-exon junction (Fig. 12.29).

Two of the three Upf proteins then bind to the exon junction complex and some of the original members of the EJC are lost. Upf3 binds first, while the mRNA is still inside the nucleus. Upf2 binds after the mRNA has exited the nucleus. During the first round of translation, the ribosome displaces the EJCs as it moves along the mRNA. If there is a premature stop codon, the ribosome finishes translating before all of the EJC complexes have been bumped off the mRNA. In this case, the termination complex, which includes the release factor plus Upf1, interacts with the remaining EJC (Fig. 12.29). This apparently involves the binding of Upf1 by the other two Upf proteins. This in turn triggers destruction of the mRNA molecule.

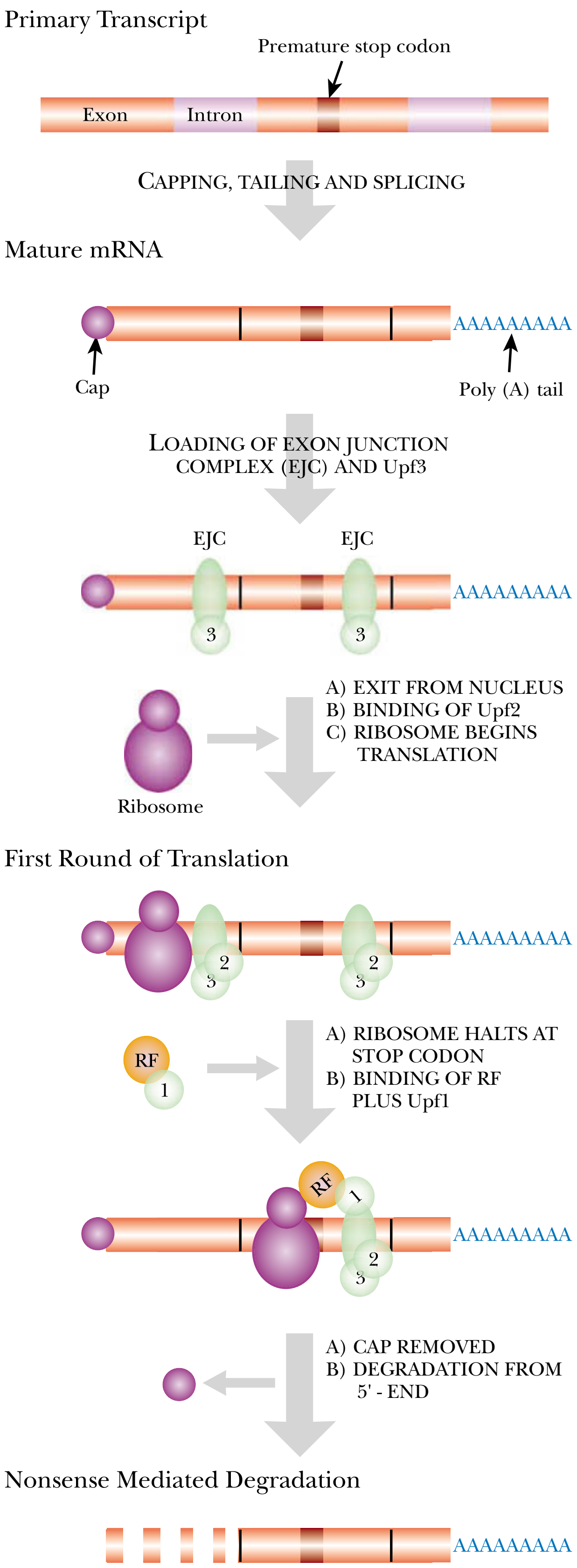

Primary Transcript
Premature stop codon
Exon
Intron
CAPPING, TAILING AND SPLICING
Mature mRNA
AAAAAAAAA
Cap
Poly (A) tail
LOADING OF EXON JUNCTION COMPLEX (EJC) AND Upf3
EJC
EJC
3
3
AAAAAAAAA
A) EXIT FROM NUCLEUS
B) BINDING OF Upf2
C) RIBOSOME BEGINS TRANSLATION
Ribosome
First Round of Translation
2
3
2
3
AAAAAAAAA
RF
1
A) RIBOSOME HALTS AT STOP CODON
B) BINDING OF RF PLUS Upf1
RF
1
2
3
AAAAAAAAA
A) CAP REMOVED
B) DEGRADATION FROM 5' - END
Nonsense Mediated Degradation
AAAAAAAAA

FIGURE 12.29 ***Nonsense-Mediated Decay of Eukaryotic mRNA***

The primary transcript is capped and tailed and the introns are spliced out to give mature mRNA. In this case the mRNA contains a premature termination codon. During splicing the mRNA is loaded with exon junction complexes (EJC) upstream of each exon-exon junction. Upf3 protein binds to the EJC. The mRNA leaves the nucleus and Upf2 protein binds to the EJC. The ribosome loads onto the mRNA and the first round of translation occurs. If translation does not remove all of the EJCs, then nonsense-mediated decay is triggered by binding of release factor (RF) plus Upf1 to the remaining EJC. Numbered circles represent Upf1, Upf2 and Upf3.

In nonsense-mediated decay, the first step is removal of the cap from the mRNA. (This contrasts with normal mRNA degradation where the poly (A) tail must be removed before the cap—see above). Next, the mRNA is degraded from the exposed 5′-end.

In yeast, less than 5% of genes contain introns. Therefore most yeast genes do not require splicing of the primary transcript to generate the mRNA. Consequently there are no exon-exon junctions to serve as markers for nonsense-mediated decay. Instead, most yeast mRNA contains downstream sequence elements (DSE). The DSE sequences are rather ill-defined but are AU rich. The DSE sequences serve as binding sites for the proteins involved as markers in nonsense-mediated decay.

Nonsense-mediated decay in yeast also differs from animals in another respect. In both mammals and in the roundworm, *Caenorhabditis elegans*, NMD is regulated by phosphorylation of protein Upf1. This does not occur in yeast. Phosphorylation probably occurs during the translation termination process and is necessary for NMD to proceed. Addition and removal of phosphate from Upf1 requires several other proteins. These are absent in yeast.

Yeast mutants with knockouts in the *Upf* genes grow nearly normally on many media, but show impairment of mitochondrial function. *C. elegans* with impaired NMD is viable. However, the reproductive system develops abnormally and fertility is greatly reduced. In mammals, defects in the *Upf* genes appear to be lethal.

CHAPTER **THIRTEEN**

Mutations

Mutations Alter the DNA Sequence

If nature were to follow all the rules laid out in the preceding chapters, there would never be any variation in DNA from one generation to the next. However, nature is not perfect and mistakes happen. Errors may occur in any of the processes of molecular biology. An error in a cell's genetic material is known as a **mutation**. As might be expected, many mutations are detrimental. However, detrimental mutations tend to be overestimated because they are more noticeable. Very often the negative effect is minimal, and in fact, the majority of mutations have little or no significant effect on the survival of an organism—they are essentially neutral. Furthermore, occasional mutations may turn out to be beneficial to the survival and reproduction of the organism. The accumulation of such beneficial mutations allows the organism to evolve in response to changing environmental conditions (see Ch. 20).

Mutations are heritable alterations in the genetic material of any organism or gene creature.

At the molecular level, mutations are alterations in the DNA molecules of which the genes are made. Consequently, when a DNA molecule replicates, any changes due to mutation of the original DNA base sequence will be duplicated and passed on to the next generation of cells. In single-celled organisms, mutations are passed on from one generation to the next when the organism divides. Among multi-cellular organisms, the situation is more complicated. Mutations are inherited by the next generation of organisms only if they occur in the cells of the germ line and are passed on during sexual reproduction. Mutations that occur in somatic cells will only be passed on to the descendents of those cells. Such mutant cell lines will be restricted to the original multi-cellular organism where the mutation occurred. Somatic mutations that result in unregulated cell growth are responsible for the emergence of cancers. Other somatic mutations merely result in particular cell-lines or organs being genetically different from the rest of the body.

Since the DNA is used as a template in transcription to make an RNA copy, a mutation in the DNA sequence within a cell will be passed on to the mRNA molecule. Finally, the mRNA is translated to yield protein. An altered RNA sequence may be translated into an altered and possibly defective protein. [Errors may also occur during transcription and translation resulting in occasional defective RNA or protein molecules. These are not regarded as mutations as they are not passed on to the descendents of the cell where they occurred.]

Most mutations cause little observable harm. A few unusually severe mutations are responsible for inherited disease.

When humans carry a mutation in their reproductive cells which leads to an observable defect in their children, the resulting condition is referred to as **inherited disease**. In fact, all humans are mutants many times over, with a substantial number of errors in their genes. However, there are many different types of mutation and most have only minor effects; in fact, many appear to cause no noticeable defect at all. Relatively few mutations cause such large changes that they attract attention. Moreover, higher organisms have two copies of each gene. This means that if one copy is damaged

inherited disease Disease due to a genetic defect that is passed on from one generation to the next
mutation An alteration in the DNA (or RNA) that comprises the genetic information

by mutation, there is a back-up copy which can be used to produce the correct protein. This often suppresses the potential defect, unless the mutation is dominant. It has been estimated that a typical human carries enough harmful mutations to total approximately eight lethal equivalents per genome. Put another way, if humans were haploid, with only a single copy of each gene, the average person would be dead eight times over. Due to mutations accumulated over the centuries, all humans are genetically different from their ancestors.

The Major Types of Mutation

Many different types of mutation occur. Some affect a single base, others affect large segments of DNA.

A single mutation is a single event and a multiple mutation is the result of several events. A single mutational event, however large or complex its effect, is regarded as a single mutation. A mutation that involves only a single base is known as a **point mutation**. A **null mutation** totally inactivates a gene; the expression "null mutation" is a genotypic term. Complete absence of a gene product may or may not cause a detectable phenotype. A **tight mutation** is one whose phenotype is clear-cut. The complete loss of a particular enzyme may result in no product in a particular biochemical pathway. For example, the complete inability of a bacterium to grow if provided with a certain sugar is an example of a tight mutation. A **leaky mutation** is one where partial activity remains. For example, 10 percent residual enzyme activity might allow a bacterium to still grow, albeit very slowly.

The sequence of a DNA molecule may be altered in many different ways. Such mutations have a variety of outcomes that depend in part on the nature of the change and in part on the role of the DNA sequence that was altered. The major types of sequence alteration are as follows, and will be discussed separately below:

Base substitution: one base is replaced by another base.

Insertion: one or more bases are inserted into the DNA sequence.

Deletion: one or more bases are deleted from the DNA sequence.

Inversion: a segment of DNA is inverted, but remains at the same overall location.

Duplication: a segment of DNA is duplicated; the second copy usually remains at the same location as the original.

Translocation: a segment of DNA is transferred from its original location to another position either on the same DNA molecule or on a different DNA molecule.

Mutations may affect genes that encode proteins, genes encoding non-coding RNA, regulatory sequences and recognition sites.

Much of the discussion below considers what happens when mutations occur within genes that encode proteins. However it is important to realize that mutations may also occur within those genes whose products are tRNA, rRNA, or other nontranslated RNA molecules. Alterations in these molecules may have drastic effects on ribosome function, splicing or other vital processes. Furthermore, mutations may also fall within promoter sequences or other regulatory sites on the DNA that do not actually encode any gene product. Nonetheless, such regulatory sites are important for gene expression and altering them may have major effects.

base substitution Mutation in which one base is replaced by another
deletion Mutation in which one or more bases is lost from the DNA sequence
duplication Mutation in which a segment of DNA is duplicated
insertion Mutation in which one or more extra bases are inserted into the DNA sequence
inversion Mutation in which a segment of DNA has its orientation reversed, but remains at the same location
leaky mutation Mutation where partial activity remains
null mutation Mutation that totally inactivates a gene
point mutation Mutation that affects a single base pair
tight mutation Mutation whose phenotype is clear-cut due to the complete loss of function of a particular gene product
translocation Mutation in which a segment of DNA is transferred from its original location to another site on the same or a different DNA molecule

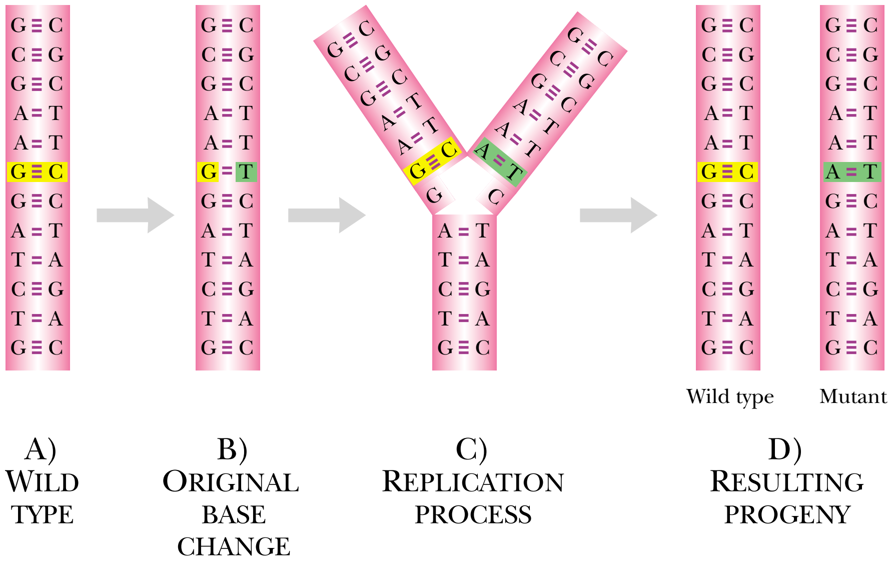


FIGURE 13.01 ***Segregation of Base Alterations in DNA***

A) A mutation has occurred causing a C to be replaced by a T. B) During replication, one DNA molecule (in yellow) matches the original base and the other strand (green) matches the mutated base. C) The strands segregate into the progeny, giving one wild type and one mutant.

Base Substitution Mutations

If one base is replaced by another, a base substitution mutation has occurred. These may be subdivided into **transitions** and **transversions**. In a transition a pyrimidine is replaced by another pyrimidine (i.e., T is replaced by C or vice versa) or a purine is replaced by another purine (i.e., A is replaced by G or vice versa). A transversion occurs when one base is replaced by another of a different type; for example, a pyrimidine is replaced by a purine or vice versa.

DNA molecules are double stranded. If a mutation occurs and a single base is replaced with another, the DNA molecule will temporarily contain a pair of mismatched bases (Fig. 13.01). When the DNA molecule replicates, complementary bases will be incorporated into the new strands opposite the bases making up the mismatch. The result is one wild-type daughter molecule and one mutant DNA molecule.

When mutations are induced by experimental treatment, it is necessary to allow the cells time to divide after treatment before imposing any selection. This allows the original DNA strands to separate and the cell to make new DNA molecules that are either fully wild-type or fully mutant. This process is sometimes referred to as **segregation**, as the originally mutated cell segregates the mutation and the wild-type into separate daughter cells upon cell division.

Missense Mutations May Have Major or Minor Effects

When a change in the base sequence alters a codon so that one amino acid in a protein is replaced with a different amino acid, this is called a **missense mutation**. Overall, this is the most frequent outcome of changing a single base. The severity of a missense mutation depends on the location and the nature of the amino acid that was substituted.

Mutant versions of genes are numbered as described in Ch. 1. Thus *genX123* refers to the 123rd mutation isolated in the gene, *genX*. A mutation which results in a codon for one amino acid being replaced by another may be written *genX123* (Arg185Leu),

missense mutation Mutation in which a single codon is altered so that one amino acid in a protein is replaced with a different amino acid
segregation Replication of a hybrid DNA molecule (whose two strands differ in sequence) to give two separate DNA molecules, each with a different sequence
transition Mutation in which a pyrimidine is replaced by another pyrimidine or a purine is replaced by another purine
transversion Mutation in which a pyrimidine is replaced by a purine or vice versa

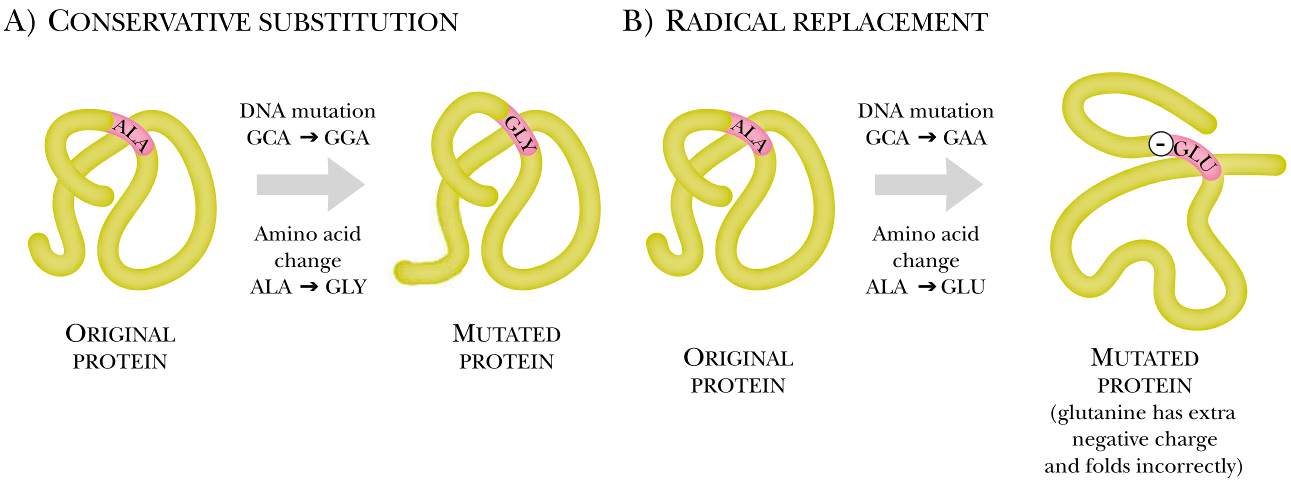


FIGURE 13.02 ***Conservative Substitution and Radical Replacement***

A) A mutation resulting in DNA change from GCA to GGA will result in the conservative substitution of an alanine for a glycine. Since both amino acids have similar properties, it is likely that the mutant protein will fold similarly to the wild type. B) A mutation resulting in the substitution of a glutamate for an alanine is a radical replacement as the glutamate has an extra negative charge that will probably cause the protein to fold quite differently from the wild type.

or in one-letter code (R185L). This indicates that arginine at position 185 has been replaced by leucine.

Proteins must assume their correct three-dimensional structure in order to function properly. Moreover, most proteins, especially enzymes, contain an active site whose role is critical. This region contains relatively few of the many amino acids that make up a typical protein. Sequence comparison of the same protein from different organisms usually shows that only the amino acids in a few positions are invariant or nearly so. These highly conserved amino acid residues generally include those in the active site(s) plus others that are critical for correct folding of the protein. Although the protein must fold up correctly, the precise identity of the amino acids at many positions may vary substantially without causing major changes in overall structure. Thus, mutations that alter active site residues will usually have major effects. Mutations that alter residues important for structure will also have a major impact. However, mutations affecting less vital parts of the protein will often have minor effects and substantial activity may remain.

Replacing an amino acid with a chemically similar one often has little effect on a protein.

The chemical properties of the original amino acid and the one replacing it in the mutant are also important. Suppose the codon UCU, which codes for serine, is changed to ACU, which codes for threonine. Both serine and threonine are small, hydrophilic amino acids with hydroxyl groups. Replacing one amino acid with another that has similar chemical and physical properties is known as a **conservative substitution**. Swapping serine for threonine in the less critical regions of a protein will probably not alter its structure radically and the protein may still work, at least partially. In rare instances, the protein may actually work better. On the other hand, if the exchange is made in a critical region of the protein, such as the active site, even a conservative substitution may completely destroy activity. Nonetheless, since the critical regions of most proteins occupy only a small proportion of the total sequence, most conservative substitutions will be relatively mild and usually non-lethal (Fig. 13.02A).

Replacing an amino acid with one that has very different properties often causes significant damage to the protein.

Replacing one amino acid with another that has different chemical and physical properties is known as a **radical replacement** (Fig. 13.02B). Suppose the codon GUA, which codes for valine, is changed to GAA; this then yields a glutamic acid. This

conservative substitution Replacement of an amino acid with another that has similar chemical and physical properties
radical replacement Replacement of an amino acid with another that has different chemical and physical properties

replaces a bulky hydrophobic residue with a smaller, hydrophilic residue that carries a strong negative charge. Under most circumstances, replacing valine with glutamic acid will seriously cripple or totally incapacitate most proteins. If the residue in question is on the surface of the protein, it is sometimes possible to get away with a radical replacement, provided that the change does not affect a critical binding site or alter the solubility of the protein too drastically.

Mutant proteins may sometimes be defective only under certain conditions, such as high temperature.

An interesting and sometimes useful type of missense mutation is the **temperature sensitive (ts) mutation**. As its name indicates, the mutant protein folds properly at low temperature (the "permissive" temperature) but is unstable at higher temperatures and unfolds. Consequently, the protein is inactive at the higher or "restrictive" temperature. If a protein is essential, a missense mutation will often be lethal to the cell. However, a temperature sensitive mutant can be grown and used for genetic experiments at the lower permissive temperature, where it remains alive. To analyze the damage caused by the mutation, the temperature is then shifted upward to the restrictive temperature at which the protein is inactivated and the organism may eventually die. An example in the fruit fly, *Drosophila*, is the *para(ts)* mutation. This affects a protein that forms sodium channels necessary for transmitting nerve impulses. At high temperatures the mutant protein is inactive and the flies are paralyzed. At lower temperatures, they are capable of normal flight (Fig. 13.03).

Mutations whose effects vary depending on a variety of environmental conditions are well known.

Naturally occurring temperature sensitive mutations have given rise to the patterns of fur coloration in some animals. Many light colored animals have black tips to their paws, tails, ears and noses. This is due to a temperature-sensitive mutation in the enzyme that synthesizes melanin, the black skin pigment of mammals. In these cases, the mutant enzyme is inactive at normal mammalian body temperature, but active at the lower temperatures found at the extremities. Consequently, melanin is made only in the cooler outlying parts of the body (Fig. 13.04).

Mutations whose effects vary depending on the environment are known as **conditional mutations**. Cold-sensitive mutations do occur but are much rarer than high or normal temperature-sensitive mutations. Multi-subunit proteins are often held together by hydrophobic patches on the surfaces of the subunits (see Ch. 7). The hydrophobic interaction is weaker at lower temperatures. It is therefore possible to get altered proteins, whose hydrophobic bonding is weaker, that fail to assemble at low temperatures but are normal at higher temperatures. For example, microtubule proteins are temperature dependent. Microtubules are cylinders made from the helical assembly of the monomer tubulin. In *Saccharomyces cerevisiae*, residues whose mutation caused cold sensitivity were concentrated at the interfaces between adjacent alpha-tubulin subunits. Mutations that respond to the osmotic pressure or ionic strength of the medium are also known.

Nonsense Mutations Cause Premature Polypeptide Chain Termination

Not all codons encode amino acids. Three (UAA, UAG and UGA) are stop codons that signal the end of a polypeptide chain. A **nonsense mutation** occurs when the codon for an amino acid is mutated to give a stop codon. Suppose that the codon UCG for serine is changed by replacing the middle base, C, with A. This gives the stop codon UAG. When the ribosome translates the mRNA, it comes to the mutant codon that used to be serine. But this is now a stop codon, so the ribosome stops and the rest of the protein does not get made. Release factor recognizes the premature stop codon and releases the partly-made polypeptide. Hence, nonsense mutations are sometimes called **chain termination mutations**. Usually, the shortened polypeptide

chain termination mutation Same as nonsense mutation
conditional mutation Mutation whose phenotypic effects depend on environmental conditions such as temperature or pH
nonsense mutation Mutation due to changing the codon for an amino acid to a stop codon
temperature-sensitive (ts) mutation Mutation whose phenotypic effects depend on temperature

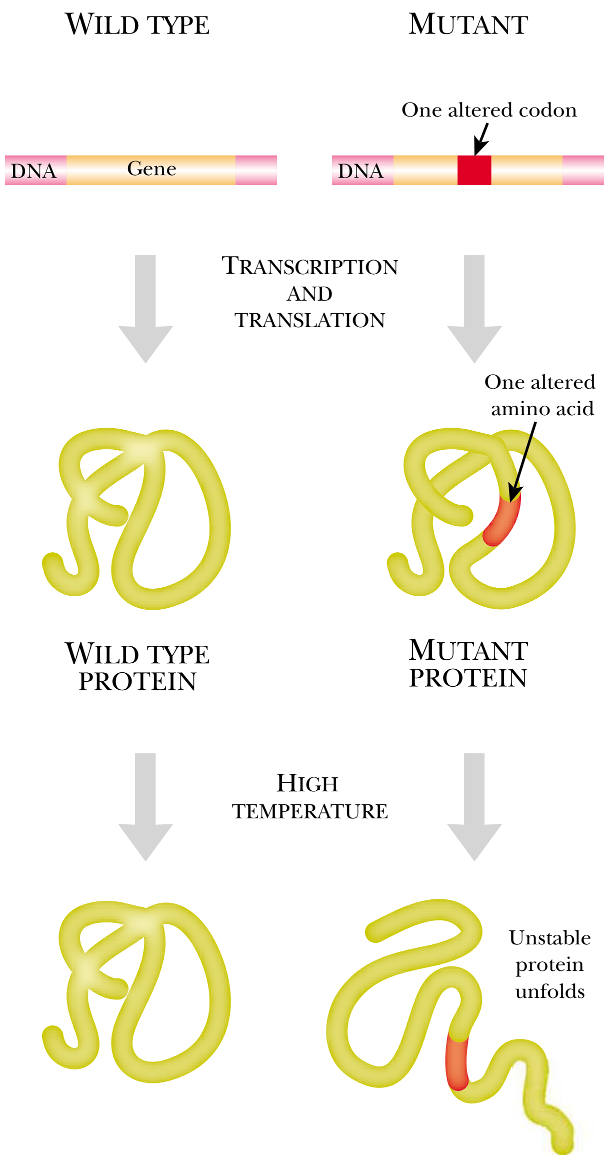


FIGURE 13.03
Temperature Sensitive Mutation

The wild-type gene encodes a protein that folds similarly at high and low temperature. The mutant protein folds normally at low temperature but unfolds at high temperature, and consequently, no longer works properly.

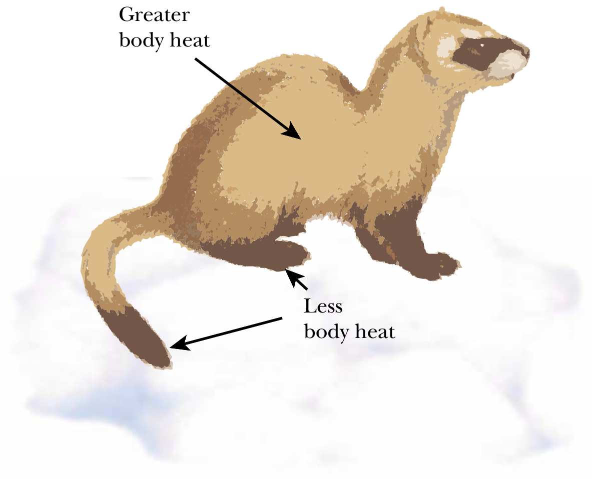


FIGURE 13.04
Temperature-Sensitive Fur Coloration

Some animals change colors in their extremities when body temperature is lowered in response to seasonal temperature change. The mutant gene for melanin is active at lower temperatures.

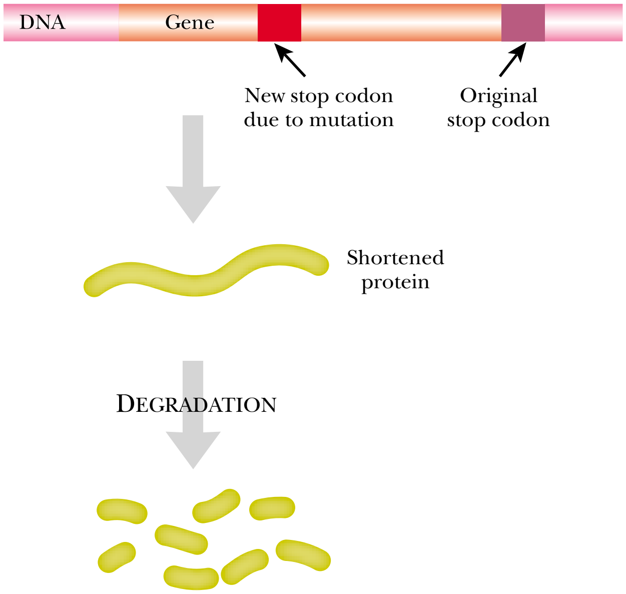


FIGURE 13.05 ***Nonsense Mutation***

A mutation in the DNA has produced a new stop codon that causes premature termination of the protein. The shortened protein remains unfolded and is usually detected by the cell and degraded.

chain cannot fold properly (Fig. 13.05). Such misfolded proteins are detected and degraded by the cell (see Ch. 8). The result, in practice, is normally the total absence of this particular protein. Nonsense mutations are often lethal if they affect important proteins.

Deletion Mutations Result in Shortened or Absent Proteins

Mutations that remove one or more bases are known as deletions and those that add extra bases are known as insertions. Clearly, the effect of a deletion (or insertion) depends greatly on how many bases are removed (or inserted). In particular, we should distinguish between point mutations where one (or a very few) bases are affected, and gross deletions and insertions that affect long segments of DNA. Point deletions and insertions may have major effects due to disruption of the reading frame—see below. Here we will consider the effects of larger deletions.

Deletions are indicated by the symbol Δ or by DE. Thus Δ(*argF-lacZ*) or DE(*argF-lacZ*) indicates a deletion of the region (of the *E. coli* chromosome in this case) from the *argF* to the *lacZ* gene. Obviously, deletion of the DNA sequence for a whole gene means that no mRNA and no protein will be made (Fig. 13.06). If the protein is essential, then the deletion will be lethal. Large deletions may remove part of a gene, an entire gene or several genes. Deletions may also remove part or all of the regulatory region for a gene. Depending on the precise region removed, gene expression may be decreased or increased. For example, a deletion that removes the binding site for a repressor may result in a large increase in activity of the gene in question. Thus loss of DNA may result in elevated activity. Again, deletions may remove largely functionless DNA, such as the non-coding sequences between genes or the introns found within genes. In this case, the effects may be small or marginal.

Deletions may remove critical segments of DNA or largely functionless regions of DNA.

Deletion mutations are surprisingly frequent. About 5 percent of spontaneous mutations in bacteria such as *E. coli* are deletions. Although bacteria lack introns and the intergenic spacer regions are very short, it is still possible to generate non-lethal deletions of considerable size. The main reason is that many genes are needed only under certain limited environmental conditions. Thus deletions of the entire *lac* operon in *E. coli* prevent the organism from using lactose as a source of carbon, yet have no other deleterious effects.

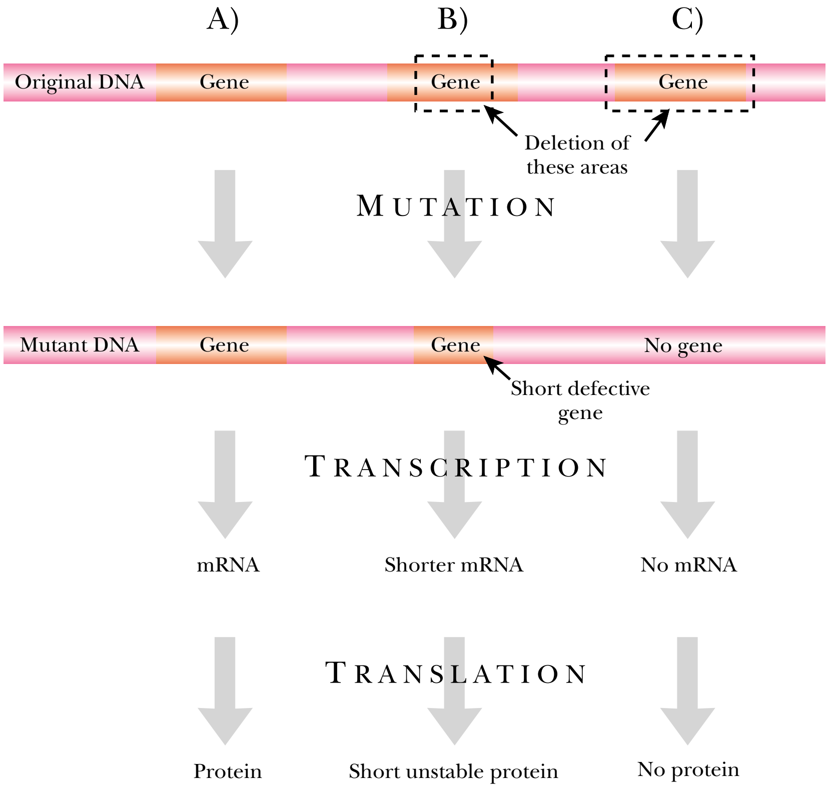


FIGURE 13.06 ***Effects of Deletion Mutations***

A) The wild-type gene produces a normal mRNA and a normal protein. B) A large deletion causes a shorter mRNA and a short unstable protein. C) Deletion of an entire gene results in no mRNA and no protein.

Insertion Mutations Commonly Disrupt Existing Genes

Genes may also be inactivated by insertions of DNA. If a foreign segment of DNA is inserted into the coding region, then the gene is said to be disrupted. Usually the gene will be completely inactivated; however, the precise result will vary depending on the length and sequence of the inserted DNA (Fig. 13.07) as well as on its precise location. Thus, if an insertion occurs very close to the 3′-end of a gene, most of the coding sequence will remain intact and sometimes a more or less functional protein may still be made. The cause of insertion mutations may be divided into two distinct categories. Some of these mutations are the result of **mobile genetic elements**, usually thousands of bases long, inserting themselves into a gene. Other insertion mutations, usually only one or a few bases long, are caused by mutagenic chemicals or by mistakes made by DNA polymerase. These short insertions are discussed below under "Frameshift Mutations."

Most naturally occurring insertions are due to mobile elements, including transposons, retroposons and certain viruses.

Under natural conditions, most insertion mutations are due to mobile genetic elements inserting themselves into the DNA. Such elements include insertion sequences, **transposons** and **retroposons** (see Ch. 15) and certain viruses that may integrate their DNA into the host chromosome (see Ch. 17). Insertions are indicated by the symbol :: between the target gene and the inserted element. Thus *lacZ*::Tn10 indicates the insertion of the transposon Tn10 into the *lacZ* gene. Insertion of such large genetic elements disrupts the target gene and completely disrupts its proper function. The presence of transposons greatly increases the frequency of various other DNA rearrangements, such as deletions and inversions. The precise mechanisms are uncertain but are probably due to abortive transposition attempts (see Ch. 15).

Large insertions within operons often prevent transcription of genes downstream from the insertion site.

Transposable elements and viruses usually contain multiple transcriptional terminators. Consequently, RNA polymerase cannot transcribe through them. Therefore

mobile genetic element A discrete segment of DNA that is able to change its location within larger DNA molecules by transposition or integration and excision
retrotransposon or retroposon A transposable element that uses reverse transcriptase to convert the RNA form of its genome to a DNA copy
transposon Same as transposable element, although the term is usually restricted to DNA-based elements that do not use reverse transcriptase

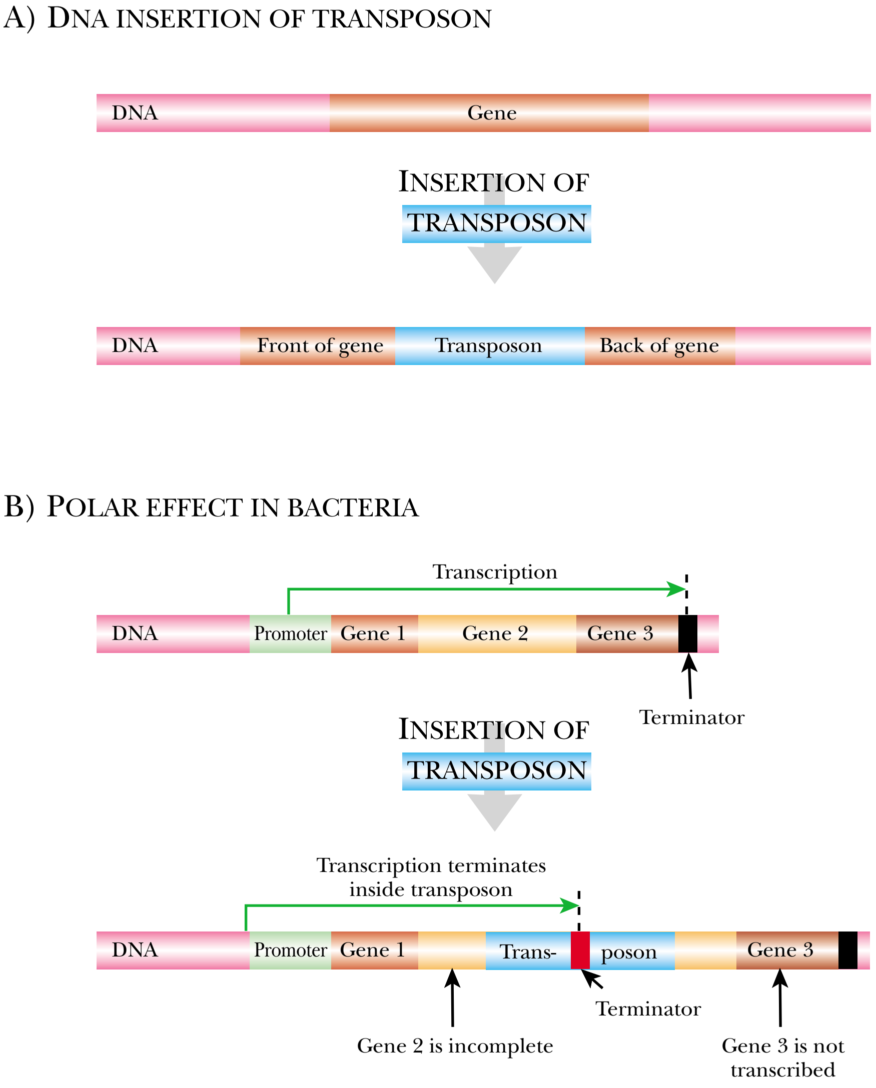


FIGURE 13.07 ***Effects of Insertion Mutations***

A) Insertion of a transposon into the middle of a gene interrupts the coding sequence. B) Insertion of a transposon into the second gene of a bacterial operon with three genes. Gene 1 is the only gene correctly transcribed since the transposon disrupts gene 2 and causes premature termination. Gene 3 will not be transcribed, although its coding sequence is still intact.

their presence blocks transcription of any other downstream genes that share the same promoter as the gene that suffered the insertion event. This effect is referred to as **polarity**. Since bacterial genes are often found clustered in operons and are co-transcribed onto the same mRNA (see Ch. 6), they are much more likely than eukaryotic genes to show polarity effects due to insertions.

Occasionally, insertions may activate genes. If an insertion occurs in the recognition site for a repressor, binding of the repressor will be prevented and activation of the gene may result. In addition, a few transposons are known to have promoters close to their ends, facing outwards (Fig. 13.08). Insertion of these may activate a previously silent gene. Examples are known of "cryptic" genes that have potentially functional gene products that cannot be expressed due to defective promoters. Thus the *bgl* operon of *E. coli* is inactive in the wild type and only expressed in mutants when a transposon carrying an outward-facing promoter is inserted just upstream of the operon and reactivates it.

polarity When the insertion of a segment of DNA affects the expression of downstream genes, usually by preventing their transcription

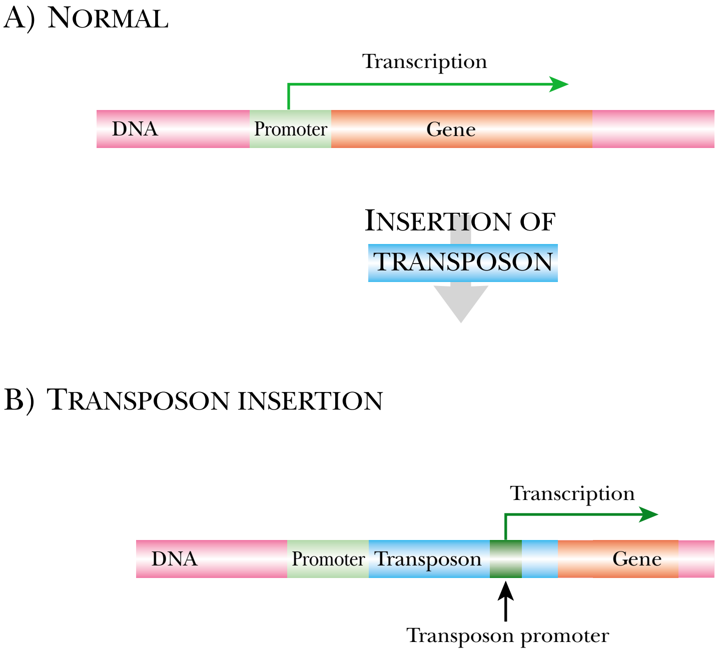


FIGURE 13.08 ***Unusual Activating Effects of Insertion Mutations***

A) The gene shown is under the control of its own promoter. B) A transposon is inserted between the normal promoter and the structural gene. The gene is now expressed under control of a promoter carried by the transposon.

Frameshift Mutations Sometimes Produce Abnormal Proteins

A mutation that alters the reading frame usually disrupts protein function completely.

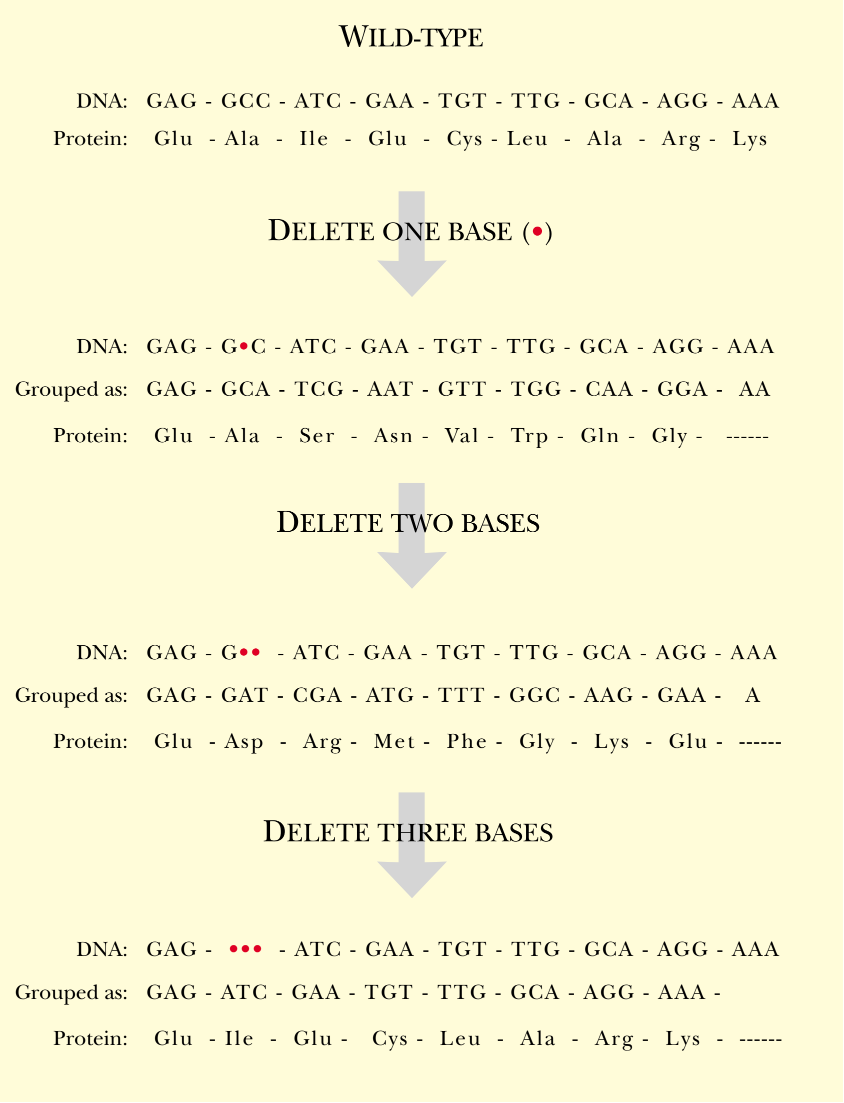
Bases are read as codons, that is in groups of three, when translated into amino acids during protein synthesis (Ch. 8). The introduction or removal of one or two bases can have drastic effects since the alteration changes the reading frame of the afflicted gene. If a single base of a coding sequence is inserted or removed, the reading frame for all codons following the insertion or deletion (Fig. 13.09) will be changed. The result will be a completely garbled protein sequence. Such **frameshift mutations** usually completely destroy the function of a protein, unless they occur extremely close to the far end. Insertion or deletion of two bases also changes the reading frame and alters protein function.

However, insertion or deletion of three bases adds or removes a whole codon and the reading frame is retained. Apart from the single amino acid that is gained or lost, the rest of the protein is unchanged. If the deleted (or inserted) amino acid is in a relatively less vital region of the protein, a functional protein may be made. Adding or deleting more than three bases will give a similar result as long as the number is a multiple of three. In other words, a whole number of codons must be added or subtracted to avoid the consequences of changing the reading frame.

DNA Rearrangements Include Inversions, Translocations, and Duplications

An inversion is just what its name implies, an inverted segment of DNA (Fig. 13.10A). Reading a stretch of DNA backwards results in drastic changes. Inversions within genes are usually highly detrimental. On the other hand, inversions do not always disrupt genes. If the endpoints of an inversion are in intergenic DNA, then inversion of a DNA segment carrying one or more intact genes, together with their promoters, may have only mild effects. In this case the orientation of the gene(s) will be reversed relative to the rest of the chromosome and it will be transcribed in the opposite direction.

frameshift Mutation in which the reading frame of a structural gene is altered by insertion or deletion of one or a few bases



WILD-TYPE

DNA: GAG - GCC - ATC - GAA - TGT - TTG - GCA - AGG - AAA
Protein: Glu - Ala - Ile - Glu - Cys - Leu - Ala - Arg - Lys

DELETE ONE BASE (•)

DNA: GAG - G•C - ATC - GAA - TGT - TTG - GCA - AGG - AAA
Grouped as: GAG - GCA - TCG - AAT - GTT - TGG - CAA - GGA - AA
Protein: Glu - Ala - Ser - Asn - Val - Trp - Gln - Gly - ------

DELETE TWO BASES

DNA: GAG - G•• - ATC - GAA - TGT - TTG - GCA - AGG - AAA
Grouped as: GAG - GAT - CGA - ATG - TTT - GGC - AAG - GAA - A
Protein: Glu - Asp - Arg - Met - Phe - Gly - Lys - Glu - ------

DELETE THREE BASES

DNA: GAG - ••• - ATC - GAA - TGT - TTG - GCA - AGG - AAA
Grouped as: GAG - ATC - GAA - TGT - TTG - GCA - AGG - AAA -
Protein: Glu - Ile - Glu - Cys - Leu - Ala - Arg - Lys - ------

FIGURE 13.09 ***Frameshift Mutations***

The effect of deleting one, two or three bases on the reading frame and the protein produced is illustrated. Note that by deleting three consecutive bases of a reading frame, the amino acid sequence stays the same although there is one amino acid missing.

A translocation is the removal of a section of DNA from its original position and its insertion in another location, either on the same chromosome or on a completely different chromosome (Fig. 13.10B). If an intact gene is merely moved from one place to another, it may still work and little damage may result. On the other hand, if, for example, half of one gene is moved and inserted into the middle of another gene, the results will be doubly chaotic.

A duplication occurs when a segment of DNA is duplicated and both copies are retained. In most cases, the duplicate is located just following the original copy (Fig. 13.10C); in other words, there is a **tandem duplication**. Duplication within a gene will seriously disrupt the gene product, whereas duplication of a large segment of DNA may generate extra copies of genes. Such duplication followed by sequence divergence is thought to be a major source of new genes over the course of evolution (see Ch. 20).

tandem duplication Mutation in which a segment of DNA is duplicated and the second copy remains next to the first

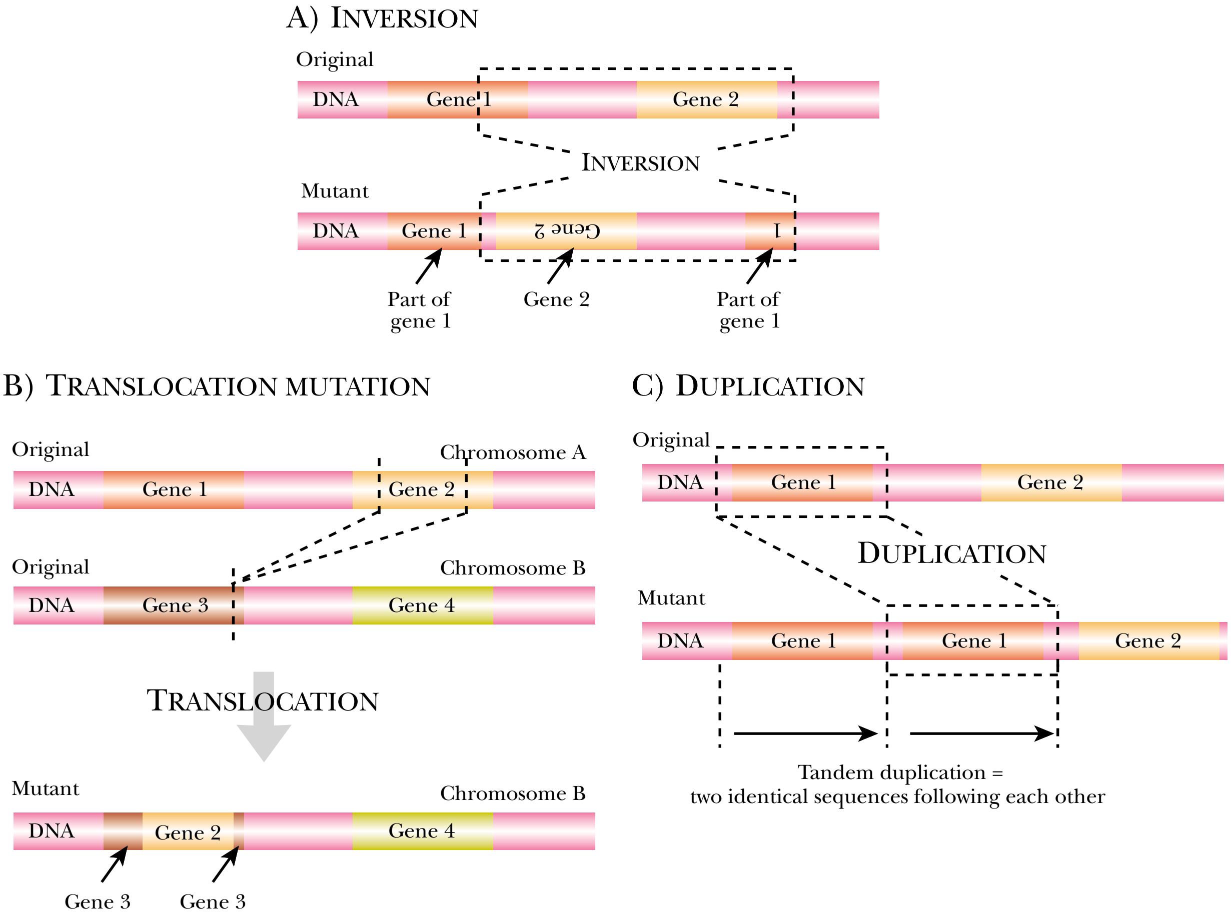


FIGURE 13.10 ***Inversions, Translocations and Duplications***

A) The inversion shown encompasses gene 2 and part of gene 1. The inverted regions are indicated by the backward spelling. B) Two chromosomes are shown with four genes. Part of gene 2 is moved from its location on chromosome A to chromosome B to a position that splits gene 3. C) Gene 1, along with some non-coding DNA, is duplicated in a tandem duplication.

Phase Variation Is Due to Reversible DNA Alterations

A few genes are switched on or off by inverting the regulatory sequences in front of them.

Phase variation is due to the reversible inversion of a DNA segment. When the DNA segment faces forward, a promoter close to its end can transcribe downstream genes (Fig. 13.11). If the segment is inverted, the promoter now faces backwards and the genes cannot be transcribed. Interconversion of the two alternative forms occurs at high frequency. It is catalyzed by a specific enzyme, an **invertase**, that recognizes specific sequences at the two ends of the invertible segment. Thus the system alternates between two phases, one in which the gene(s) is switched off, and the other in which it is turned on. Usually just the regulatory sequences are inverted and only rarely are structural genes actually disrupted by phase variation.

The flip-flop rate for phase variation is about 1 in 10,000 bacterial cells per generation, about a hundred times more frequent than most point mutations in the same organism. Since an inversion is undeniably a change in DNA sequence, which is by definition a mutation, phase variation may be viewed both as a form of genetic regulation and as an aberrant type of mutation.

The best known examples of phase variation occur in bacteria such as *Salmonella* and *E. coli*, where the switch is between expressing two alternative surface proteins.

invertase (Strictly, DNA invertase) An enzyme that recognizes specific sequences at the two ends of an invertible segment and inverts the DNA between them

phase variation Reversible inversion of a segment of DNA leading to differences in gene expression

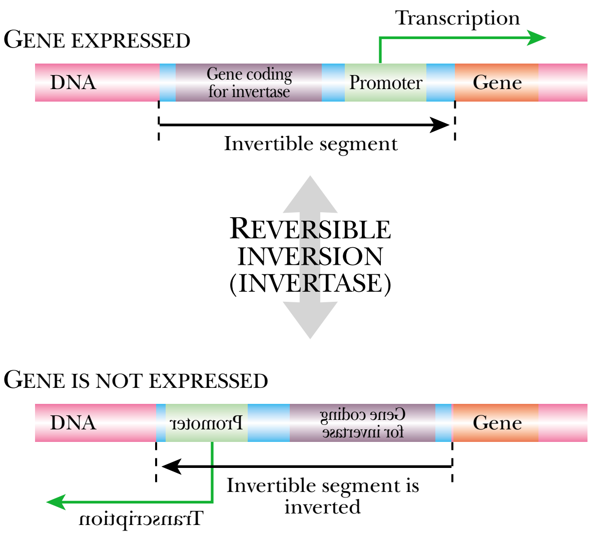

GENE EXPRESSED
Transcription
DNA
Gene coding for invertase
Promoter
Gene
Invertible segment
REVERSIBLE INVERSION (INVERTASE)
GENE IS NOT EXPRESSED
DNA
Promoter
Gene coding for invertase
Gene
Invertible segment is inverted
Transcription

FIGURE 13.11 ***Inversion Causes Phase Variation***

Some segments of DNA can vary in orientation. In the form at the top the DNA segment is oriented so that the promoter is in the correct position to transcribe the gene. In the inverted form (below) the gene is not expressed because the promoter faces the wrong direction.

In *Salmonella*, the flagella can be made from two alternative proteins, H1 and H2. The genes for H2 protein and the rh1 repressor lie downstream of the promoter on the invertible segment (Fig. 13.12). When these are expressed, rh1 represses the gene for H1 protein, which is located elsewhere on the bacterial chromosome. When the segment inverts, due to Hin invertase, the H2 and rh1 genes are no longer expressed. Lack of the rh1 repressor allows instead expression of the H1 gene. Thus either H1 or H2 is expressed, but never both at once. These flagellar proteins (known historically as H-antigens) are strong antigens and found on the surface of bacteria. Periodically switching at random between different flagellar proteins helps bacteria avoid detection by the mammalian immune system.

Pathogenic strains of *E. coli* often bind to intestinal cells by thin helical filaments of protein known as pili or fimbriae. These are also strong surface antigens and are often subject to phase variation by a similar mechanism as described above. Appropriately enough, certain viruses that infect enteric bacteria also use phase variation to change the tail fiber proteins that recognize bacterial surface proteins during infection. This phase variation results in different host ranges for the alternative types of virus particle. For example, bacteriophage Mu flip-flops between recognizing *E. coli* and the closely related *Citrobacter*.

Silent Mutations Do Not Alter the Phenotype

Many mutations have no effect on the phenotype—they are "silent".

A **silent mutation** is an alteration in the DNA sequence that has no effect on the operation of the cell and is therefore not so much silent as invisible from the outside. In other words, silent mutations do not alter the phenotype. One obvious kind of silent mutation is a base change occurring in the non-coding DNA between genes. Therefore, no genes are damaged and no proteins are altered. Higher organisms possess intervening sequences, the introns, within many of their genes. Since introns are cut out and discarded when the messenger RNA is made, most alterations to the sequence of an intron will not affect the final protein.

Not all base changes in an intron are harmless. Changes in the few critical bases at the splice recognition sites will result in failure to splice out the intron or in aberrant splicing. This will give a severely damaged protein product when the misspliced mRNA is translated. Further, many of the small nucleolar RNAs are derived from introns in other genes (Ch. 12). In addition, occasional cases are known where an intron

silent mutation An alteration in the DNA sequence that has no effect on the phenotype

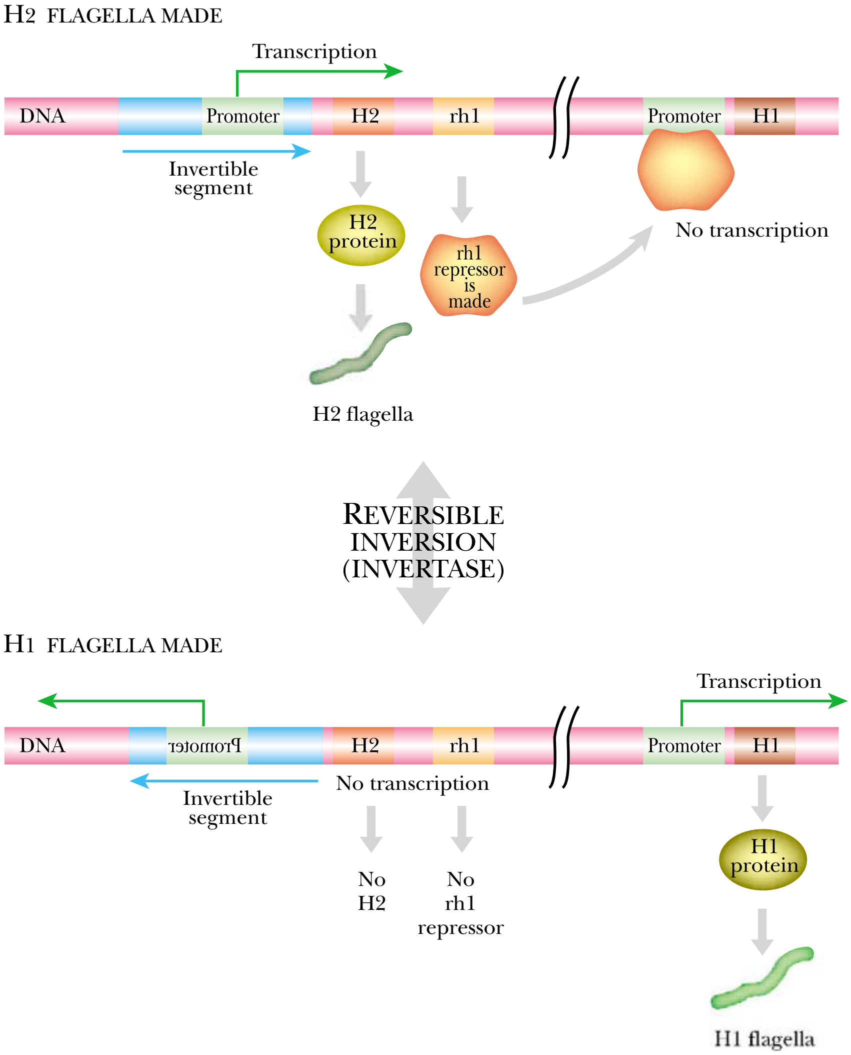


FIGURE 13.12 *Phase Variation of Flagella in Salmonella*

The top diagram illustrates the state where flagellar protein H2 is made. The invertible segment has its promotor aligned to express H2 and the rh1 repressor, which then represses the gene for H1. Therefore in this orientation only H2 is made. In the bottom diagram the invertible segment is reversed. Neither H2 nor the repressor for H1 are made. Thus H1 is free to be expressed. The two forms of the gene are interconvertable.

is needed to guide base modifications to a neighboring exon (see Ch. 12). Nevertheless, most base changes within most introns are silent mutations.

Since many amino acids have several codons, changing the third base of a codon often leaves the amino acid unchanged.

The third main type of silent mutation occurs within the coding region of a gene and does get passed on to the messenger RNA. Remember that each codon, or group of three bases, is translated into a single amino acid in the final protein product. However, because there are 64 different codons, most of the 20 possible amino acids have more than one codon (see Codon Table, Fig. 8.02). So a base change that converts the original codon into another codon that codes for the same amino acid will have no effect on the final structure of the protein.

For example, the amino acid alanine has four codons: GCU, GCC, GCA and GCG. (Note that the sequences are discussed in RNA language; these are the codons as found on mRNA.) Since they all have GC as the first two bases, any codon of the form GCN (N = any base) will give alanine. A mutation in an original GCC sequence changing the last C to an A or a G or a U results in change in the sequence of the codon, but there is no change in the amino acid produced (alanine) in the resulting protein. Many other amino acids (such as valine, threonine and glycine) also have sets of four codons in which the last base does not matter. This pattern is referred to as **third base redundancy**. Very often, altering the third base of a codon has no effect on the protein that will be made. In other words, about a third of single base substitutions will be silent,

third base redundancy Situation where a set of four codons all code for the same amino acid and thus the identity of the third codon base makes no difference during translation even if they occur within the protein coding region of a gene. [Note: the term "codon degeneracy" refers to the fact that a single amino acid may be encoded by multiple codons. In most, but not all cases, this is due to third base redundancy. However, both arginine and serine have six codons each; these inevitably differ among themselves by more than just the third base.]

Third base mutations that do not alter protein identity can sometimes have effects due to differential codon usage and tRNA bias. Different codons for the same amino acid often vary in their frequency of use and the corresponding tRNAs are often present in levels that are related to the frequency of codon usage. Consequently, if changing the third base converts a frequently used codon to a rarely used codon for the same amino acid, translation may be slowed due to shortage of the appropriate tRNA. (Conversely, changing a rare to a common codon may speed up translation.) This effect is usually only significant for proteins that are made at such high levels that their synthesis uses a significant proportion of the available tRNA in the cell.

Chemical Mutagens Damage DNA

DNA may be damaged by a variety of chemicals and by radiation.

Mutations that are caused by agents that damage the DNA are known as **induced mutations**. Agents that mutate DNA are called **mutagens** and are of three main types: mutagenic chemicals, radiation and heat. Even if there are no dangerous chemicals or radiation around, mutations still occur, though less frequently. These are **spontaneous mutations**. Some of these are due to errors in DNA replication. The enzymes of DNA replication are not perfect and sometimes make mistakes. In addition, DNA undergoes certain spontaneous chemical reactions (alterations) at a low but detectable rate and this rate goes up with increasing temperature.

The most common mutagens are toxic chemicals that react with DNA and alter the chemical structure of the bases. For example, EMS (ethyl methane sulfonate) is widely used by molecular biologists to mutagenize growing cells. It adds an ethyl group to bases in DNA and so changes their shape and their base-pairing properties. Nitrite converts amino groups to hydroxyl groups and so converts the base cytosine to uracil (Fig. 13.13). Nitrite is used experimentally to mutate purified DNA, such as a cloned gene carried on a plasmid, while the plasmid is in the test tube. The mutagenized DNA is then transferred back into a cell to identify the mutations that were generated. During DNA replication, the DNA polymerase misidentifies these altered bases and puts in the wrong bases in the new complementary strand of DNA it is making (Fig. 13.13).

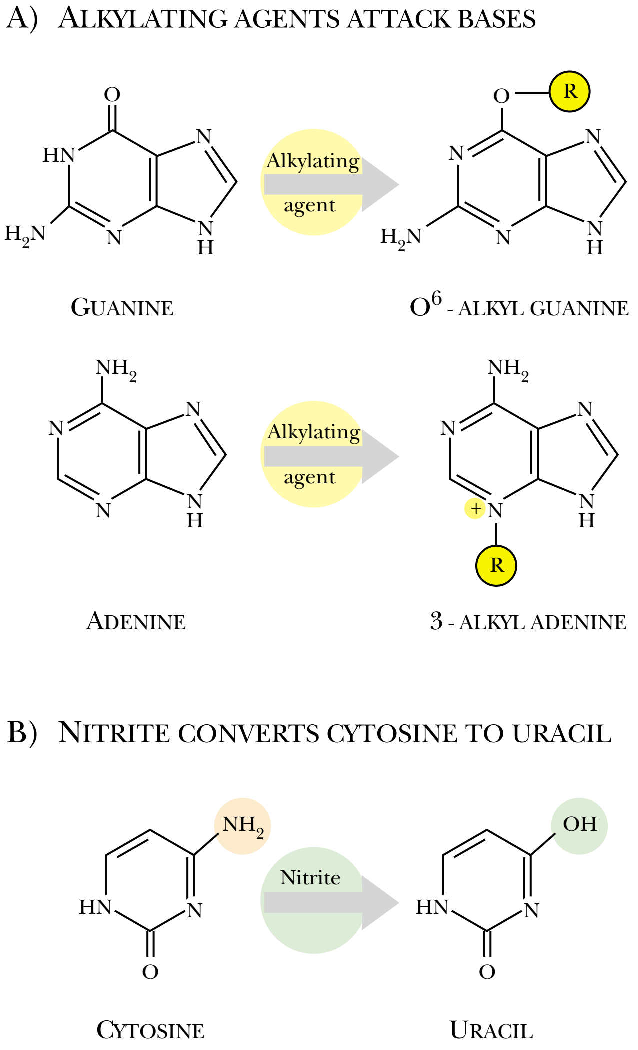
Base analogs are mistaken by the cell for the natural nucleic acid bases.

Base analogs are chemical mutagens that mimic the bases found in natural DNA. For example, bromouracil resembles thymine in shape. It is converted by the cell to the DNA precursor, bromouridine triphosphate, which DNA polymerase inserts where thymine should go. Unfortunately, bromouracil can flip-flop between two alternative shapes (Fig. 13.14). In its alternate form, bromouracil resembles cytosine and pairs with guanine. If bromouracil is in its misleading form when DNA polymerase arrives, a G will be put into the new strand opposite the bromouracil instead of A.

Intercalating agents result in the insertion of an extra base pair during DNA replication.

Some mutagens imitate the structure of a base pair rather than a single base. For example, **acridine orange** has three rings and is about the size and shape of a base pair. Acridine orange is not chemically incorporated into the DNA. Instead, it squeezes in between the base pairs in the DNA (Fig. 13.15), a process referred to as **intercalation**. During DNA replication, the DNA polymerase mistakes the intercalating agent for a base pair and puts in an extra base when making the new strand. As discussed above, insertion of an extra base will change the reading frame of the protein encoded by a

acridine orange A mutagenic agent that acts by intercalation
base analog Chemical mutagen that mimics a DNA base
induced mutation Mutation caused by external agents such as mutagenic chemicals or radiation
intercalation Insertion of a flat chemical molecule between the bases of DNA, often leading to mutagenesis
mutagen Any agent, including chemicals and radiation, that can cause mutations
spontaneous mutation Mutation that occurs "naturally" without the help of mutagenic chemicals or radiation

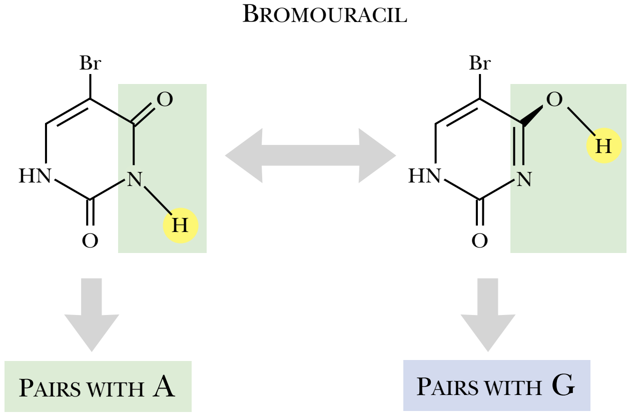


A) ALKYLATING AGENTS ATTACK BASES

B) NITRITE CONVERTS CYTOSINE TO URACIL

FIGURE 13.13 ***Base Alteration by Chemical Mutagens***

A) Alkylating agents alter the structure of bases by adding alkyl groups. B) Nitrite will convert cytosine to uracil (which pairs with adenine).



BROMOURACIL

FIGURE 13.14 ***Bromouracil Acts as a Base Analog***

Bromouracil has two alternative forms, one of which (left) looks like thymine and pairs with adenine; the other (right) looks like cytosine and pairs with guanine.

gene. Since this will completely destroy the function of the protein, intercalating agents are highly hazardous mutagens.

A **teratogen** is an agent that causes abnormal development of the embryo, which results in gross structural defects. Teratogens may or may not cause mutations. The most famous example is thalidomide, which resulted in the birth of malformed

teratogen An agent that causes abnormal embryo development leading to gross structural defects or monstrosities

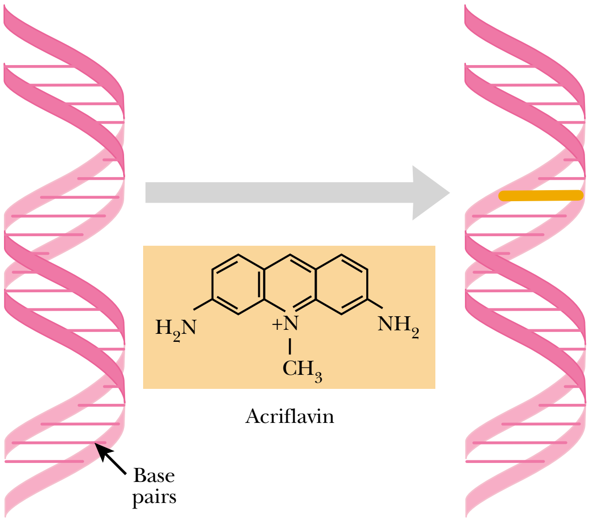


FIGURE 13.15
Intercalating Agents

An intercalating agent, such as acriflavin, can insert itself between base pairs and mimic a whole extra base pair. When replication occurs, the intercalating agent causes an extra base pair to be inserted into the new DNA. Commercial acriflavin is actually a mixture of the structure shown plus the derivative without the N-methyl group.

children with missing limbs. Thalidomide interferes with the development of embryos as opposed to causing mutations. Although the mechanism responsible for the malformations remains uncertain, it is known that thalidomide prevents blood vessels from forming (i.e. it is anti-angiogenic), which may partly explain the drug's ability to cause birth defects.

Radiation Causes Mutations

High energy radiation damages DNA.

Some types of radiation cause mutations. High frequency electromagnetic radiation, ultraviolet radiation (UV light), X-rays and gamma rays (γ-rays), directly damage DNA. X-rays and γ-rays are **ionizing radiation**; that is, they react with water and other molecules to generate ions and free radicals, notably hydroxyl radicals. Ionizing radiation is responsible for about 70 percent of the radiation damage to DNA. The other 30 percent of the radiation damage is due to direct interaction of X-rays and γ-rays with DNA itself. In the early days of molecular biology, X-rays were often used to generate mutations in the laboratory. X-rays tend to produce multiple mutations and often yield rearrangements of the DNA, such as deletions, inversions and translocations.

Ultraviolet radiation promotes formation of thymine dimers.

Ultraviolet radiation is electromagnetic radiation with wavelengths from 100 to 400 nm. It is nonionizing and acts directly on the DNA. The bases of DNA show an absorption peak at around 254 nm and UV close to this wavelength is absorbed very efficiently by DNA. In particular, UV causes two neighboring pyrimidine bases to cross-react with each other to give dimers. Thymine dimers are especially frequent (Fig. 13.16). Although DNA polymerase can proceed by skipping over thymine dimers, this leaves a single-stranded region that needs repairing. The repair process in turn causes the insertion of incorrect bases in the newly synthesized strand (see Ch. 14 for details on error-prone repair). This therefore results in mutation.

Ultraviolet radiation is emitted by the sun. Most of it is absorbed by the ozone layer in the upper atmosphere, so it does not reach the surface of the earth. Damage to the ozone layer by the chlorinated hydrocarbons used in aerosol sprays and refrigerants has allowed more UV radiation to reach the surface of this planet, especially in certain areas. This has probably contributed to the increased frequency of skin cancer noted in recent years.

In addition to electromagnetic radiation, there are other forms of radiation, such as the α-particles and β-particles emitted by radioactive materials along with

ionizing radiation Radiation that ionizes molecules that it strikes

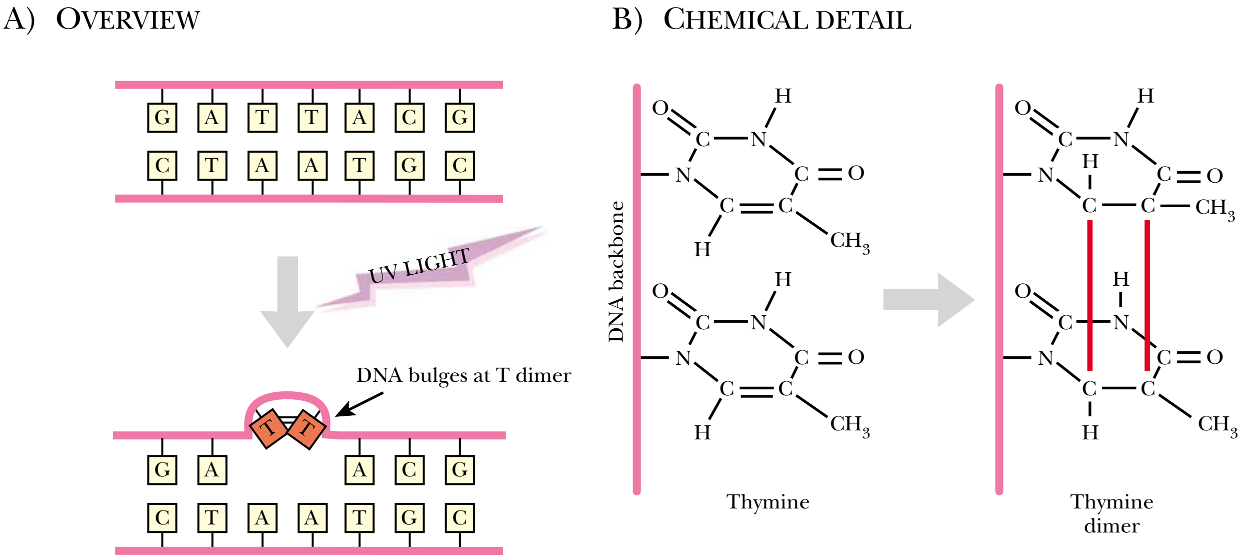


FIGURE 13.16 ***Thymine Dimers***

A) Ultraviolet light (UV) sometimes results in the formation of a thymine dimer (red). B) The detailed chemical structure of the thymine dimer is shown.

γ-rays. Most α-particles are too weak even to penetrate skin but β-particles may cause significant damage to DNA and other biological molecules. However, α-emitters can be mutagenic if they have entered the body, for example by being breathed in or swallowed.

Spontaneous Mutations Can Be Caused by DNA Polymerase Errors

DNA polymerase makes spontaneous mistakes that result in mutations.

The enzymes that replicate DNA during cell division are not perfect. They make errors at a rate that is low, but nonetheless significant over a long period of time. As discussed in Ch. 5, DNA polymerases carry out **proofreading** and check recently inserted nucleotides for mistakes before moving on. In some cases, the proofreading ability is part of the polymerase itself. In other cases, it is due to an accessory protein such as the DnaQ protein associated with *E. coli* DNA polymerase III. Cells carrying mutations that abolish or damage these proofreading abilities show much higher rates of spontaneous mutation. Genes that give rise to altered mutation rates when they themselves are mutated are known as **mutator genes**. Hence, *E. coli dnaQ* mutants were originally named *mutD* (for mutator D).

The error rate for DNA replication in *Escherichia coli* is approximately one base in 10 million. About 20 times as many errors occur in the lagging strand as in the leading strand. This probably results from DNA polymerase I having a less effective proofreading capability than DNA polymerase III. The lagging strand is made discontinuously (see Ch. 5) and the gaps are filled in by DNA polymerase I, whereas the leading strand is all made by PolIII.

DNA polymerase may slip when replicating short sequence repeats.

In addition to putting in an occasional wrong base, DNA polymerase may very rarely omit bases or insert extra bases. This is due to strand slippage. If a run of several identical bases occurs, the template strand and newly synthesized strand of DNA may

mutator gene Gene whose mutation alters the mutation frequency of the organism, usually because it codes for a protein involved in DNA synthesis or repair

proofreading Process that checks whether the correct nucleotide has been inserted into new DNA. Usually refers to DNA polymerase checking whether it has inserted the correct base

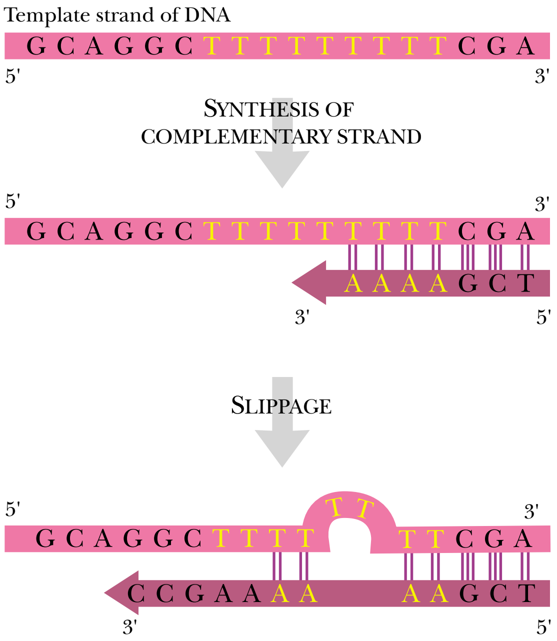


FIGURE 13.17 ***Strand Slippage Creates Small Insertions or Deletions***

The template strand of DNA shown contains numerous thymines (T) in a row (shown in yellow). When replication occurs thymine pairs with adenine (A). However, a long tract of identical bases may cause confusion and some thymines may slip and pair out of register. The extra T residues of the template strand do not pair and form a bulge. In the case shown a small deletion of two bases has occurred.

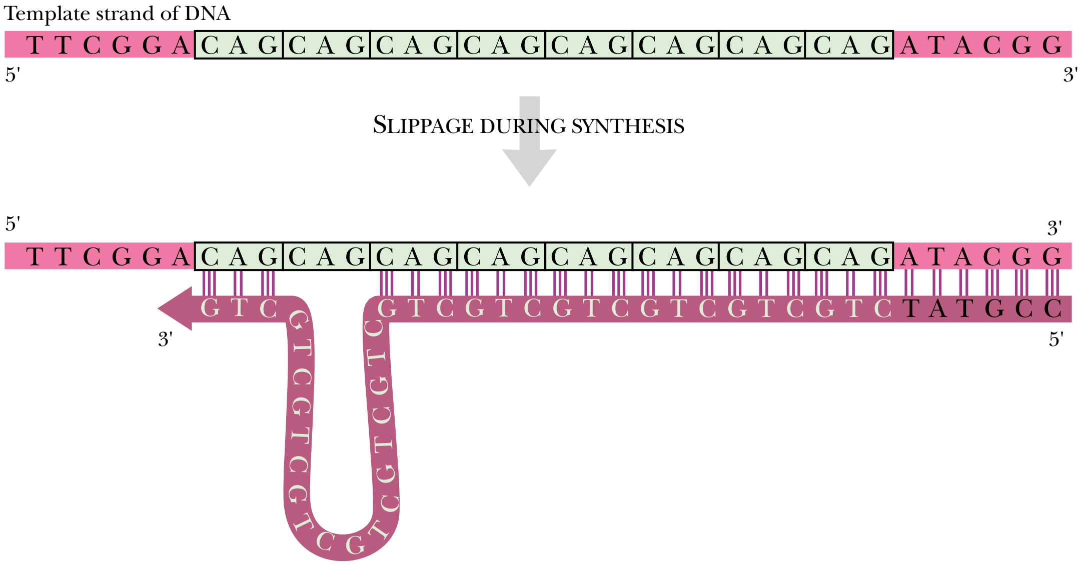

Template strand of DNA
T T C G G A C A G C A G C A G C A G C A G C A G C A G C A G A T A C G G
5'
3'
SLIPPAGE DURING SYNTHESIS
5'
3'
T T C G G A C A G C A G C A G C A G C A G C A G C A G C A G A T A C G G
G T C
G T C G T C G T C G T C G T C G T C T A T G C C
G T C G T C G T C G T C G T C
3'
5'

FIGURE 13.18 ***Strand Slippage of Trinucleotide Repeat***

Multiple trinucleotide repeats, such as CAG, may cause strand slippage during DNA replication. In the case illustrated, looping out has occurred in the newly synthesized strand of DNA. The result will be an insertion of six trinucleotide repeats.

become misaligned (Fig. 13.17). Depending on which strand slips, a base may be inserted or omitted during replication.

Slippage may also occur in regions of DNA where there are multiple repeats of a short sequence, perhaps two or three bases (Fig. 13.18). In this case, a whole repeat unit of several bases will be added or deleted. Well known cases occur in the human trinucleotide repeat expansion diseases, such as fragile X syndrome and Huntington's disease. Here copies of a three-base repeat are added or lost due to slippage.

Mutations Can Result from Mispairing and Recombination

Recombination may occurs between closely related sequences of DNA, such as two alleles of the same gene. Many DNA rearrangements, including deletions, inversions, translocations and duplications may result from mistaken pairing of similar sequences followed by recombination. The mechanism of recombination is dealt with in Ch. 14; here, the overall result of mispairing will be considered. If the similar sequences are in the same orientation, mispairing followed by crossing over will generate a duplication on one molecule of DNA and a corresponding deletion on the other (Fig. 13.19).

If two copies of a sequence are on the same DNA molecule but face each other (i.e., are in opposite orientations), mispairing followed by crossing over will generate an inversion (Fig. 13.20). For example, the chromosome of *E. coli* contains seven copies of the genes for ribosomal RNA. Strains of *E. coli* are known in which the whole segment of the bacterial chromosome between two of these rRNA operons has been inverted. Such strains grow slightly slower but nonetheless, are viable.

Spontaneous Mutation Can Be the Result of Tautomerization

However sophisticated DNA polymerase may be, there are chemical limits on the accuracy of DNA replication. Even if DNA polymerase inserts the correct base, errors may still occur. This is due to the **tautomerization** of the bases that constitute DNA. Each base may exist as two possible alternative structures that interconvert. Such structural isomers that exist in dynamic equilibrium are known as tautomers. In each case, one isomer is much more stable and the vast majority of the base is found in this form. However, the less stable alternative tautomer will appear very rarely. If this happens just as the replication fork is passing, the rare tautomer may cause incorrect base pairing.

Thymine has keto and enol tautomers (Fig. 13.21). The common, keto-form pairs with adenine, but the rare enol-tautomer base pairs with guanine. Guanine also has keto and enol tautomers. In this case the rare enol-guanine base pairs with thymine rather than cytosine. Similarly, adenine equilibrates between common amino and rare imino tautomers. The rare imino-adenine base pairs with cytosine instead of thymine. Cytosine alone does not have the potential to introduce mismatches. Although it does have amino and imino tautomers, both pair with adenine. As the temperature increases, the probability that a base is in the incorrect tautomeric state also increases and so, therefore does the mutation frequency.

Spontaneous Mutation Can Be Caused by Inherent Chemical Instability

Although DNA is relatively stable, some of its components do show a low level of spontaneous chemical reaction. Several bases undergo slow but measurable loss of their amino group; i.e., **deamination**. Adenine, guanine and cytosine may all spontaneously deaminate, but by far the most frequent is the deamination of cytosine to give uracil (Fig. 13.22). In addition, the modified base, 5-methyl-cytosine, is especially prone to deamination, so giving "methyl-uracil;" in other words, thymine. The result, in both cases, is the replacement of C by T. Deamination of A (to hypoxanthine) and G (to xanthine) occurs at only 2 to 3 percent of the rate for cytosine. Both hypoxanthine and xanthine usually (but not exclusively) base pair with C, so mutations may be introduced in some cases.

deamination Loss of an amino group
tautomerization Alternation of a molecule, in particular a base of a nucleic acid, between two different isomeric structures

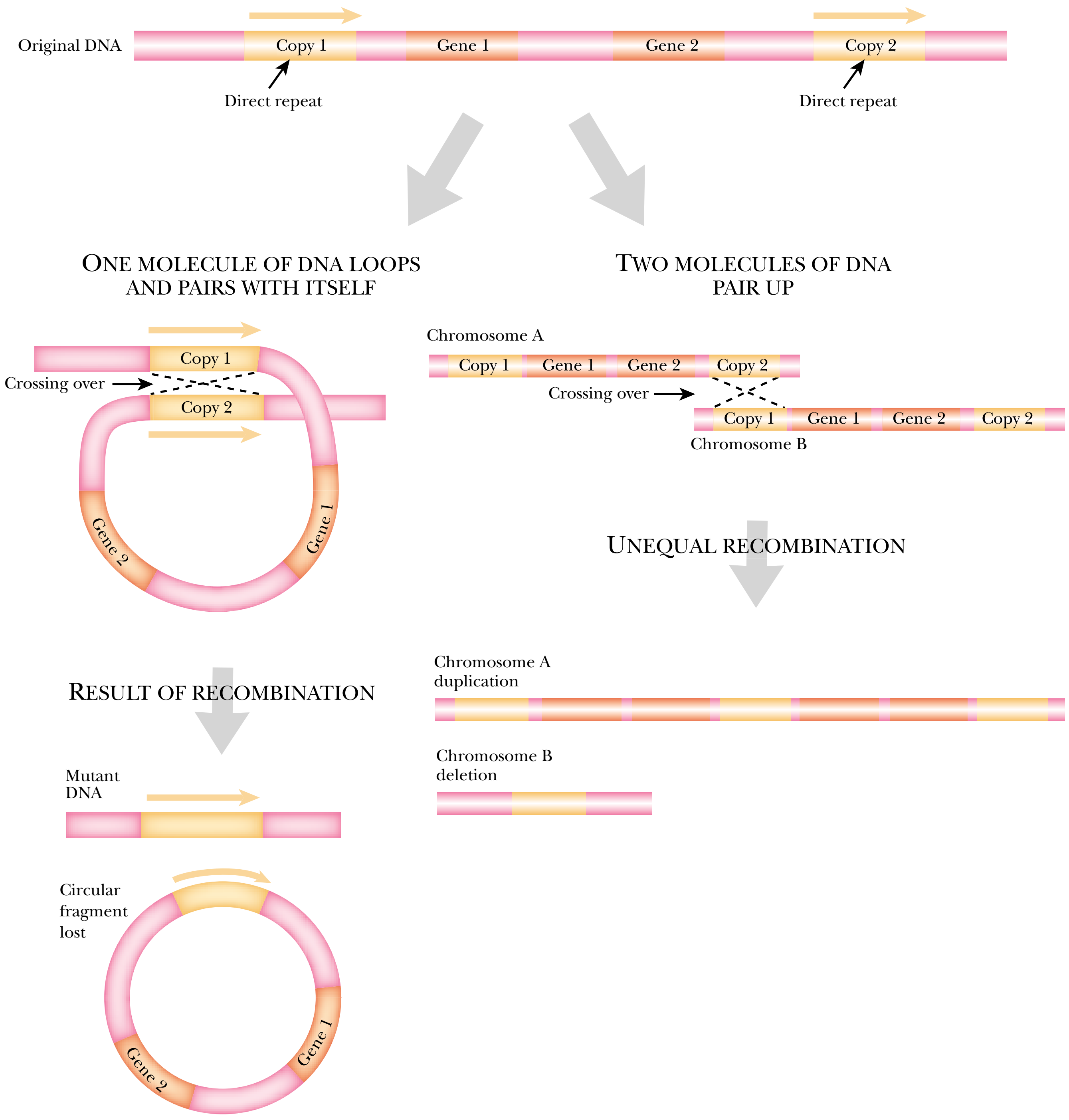


FIGURE 13.19 ***Mispairing of Direct Repeats Generates Deletions and Duplications***

Direct repeats in a DNA molecule may undergo two fates. On the left, the two repeats in a single DNA molecule pair up and recombine. This yields two products; the original DNA molecule suffers a deletion of DNA between the two repeats and a separate circular molecule of DNA is released. On the right, a repeat on chromosome A pairs with another repeat on chromosome B. The result after recombination is a duplication on chromosome A and a deletion of the corresponding region from chromosome B.

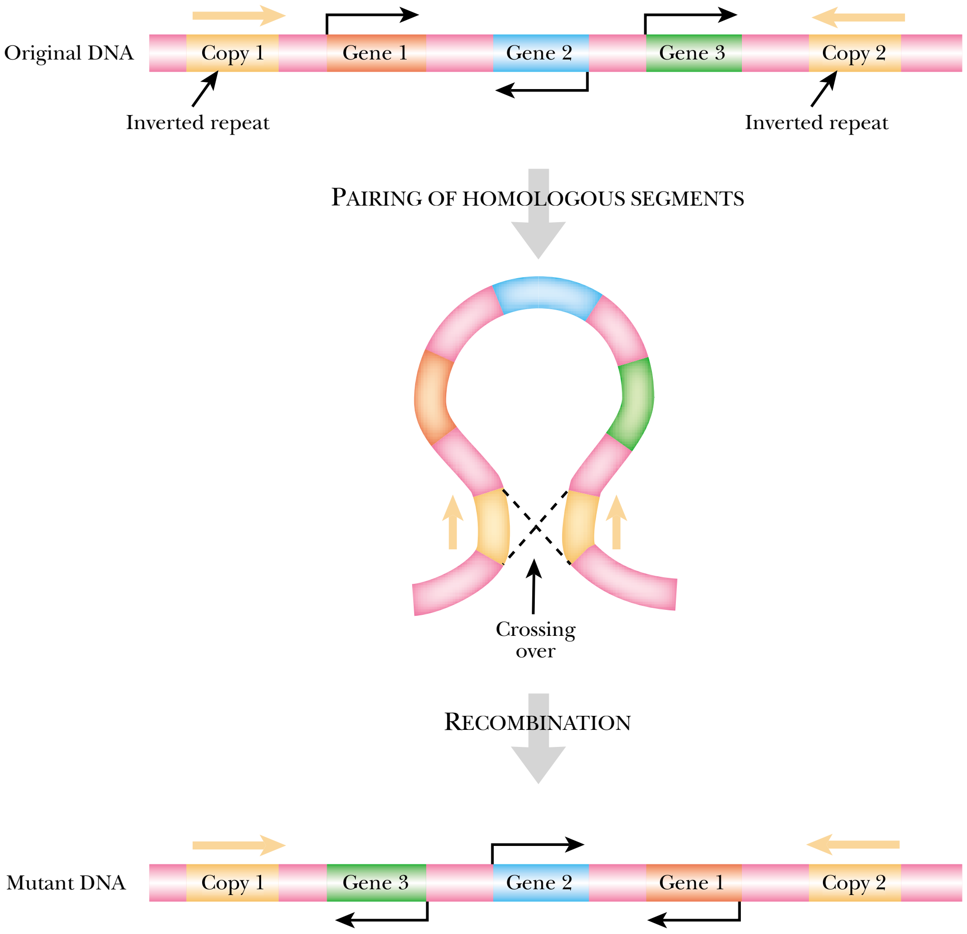


FIGURE 13.20 ***Inversion of DNA by Mispairing of Inverted Repeats***

The DNA molecule shown has two copies of a sequence that are inverted relative to each other. Three intervening genes (Genes 1, 2 and 3) with their directions of transcription (arrows) are also shown. The duplicate sequences may pair up, forming a stem and loop, and undergo recombination. The result is an inversion of the region between the duplicate sequences. This reverses the direction of transcription of the three enclosed genes with respect to the DNA molecule.

Oxidative damage to DNA is also significant. Hydroxyl and superoxide radicals derived from molecular oxygen will attack several bases. The most common target is guanine, which is oxidized to 8-hydroxy-guanine, which pairs preferentially with A. Hence a G · C base pair may be mutated into a T · A pair.

Non-enzymatic methylation of bases occurs at a low frequency. The methyl donor, S-adenosyl-methionine, is normally used by enzymes that attach methyl groups to their substrates. However, it is sufficiently reactive to attack several bases at a low rate spontaneously. The major problem is the formation of 3-methyl-adenine, which tends to block DNA elongation.

Occasionally, the bonds linking the bases of DNA to deoxyribose may spontaneously hydrolyze. This occurs more often with purines than with pyrimidines, generating empty, apurinic sites. Such missing bases tend to block DNA replication and are also an invitation to DNA polymerase to insert an incorrect base.

Mutations Occur More Frequently at Hot Spots

If the same gene is mutated thousands of times, are the mutations all different and are they distributed at random throughout the DNA sequence of that gene? Many of them

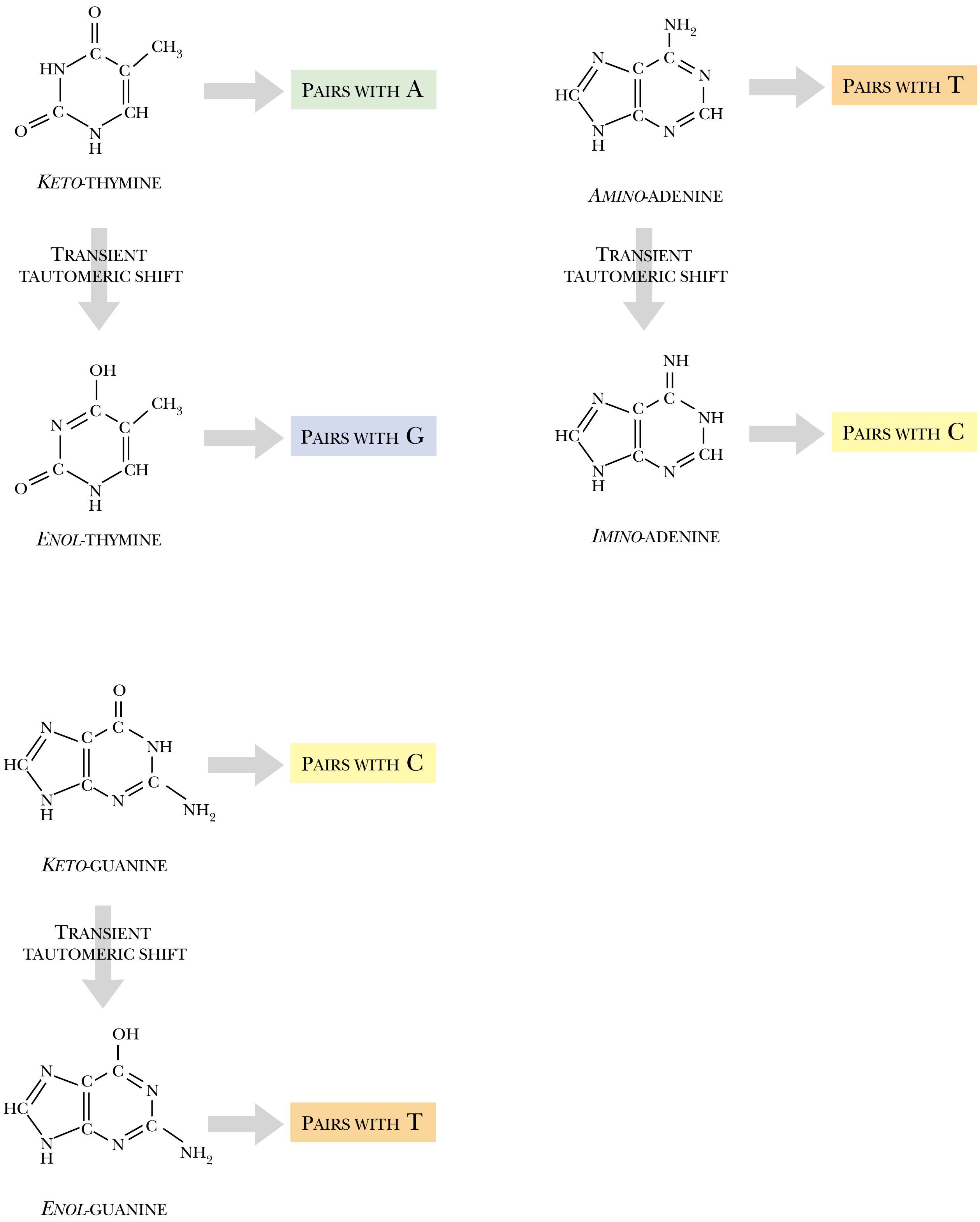


FIGURE 13.21 ***Base Tautomerization May Cause Mismatches***

The tautomers of thymine, adenine and guanine are shown. In each instance the lower, short-lived tautomer pairs with the inappropriate base.

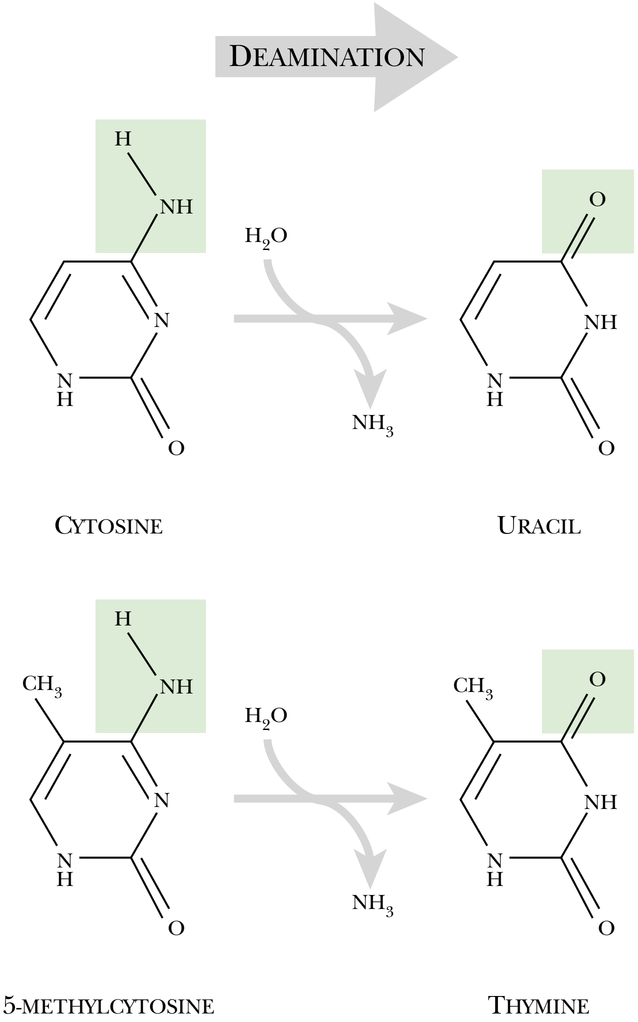


FIGURE 13.22 ***Deamination of Cytosine and 5-Methyl-cytosine***

The deamination of cytosine yields uracil and the deamination of the methylated form of cytosine yields thymine—both inappropriate bases.

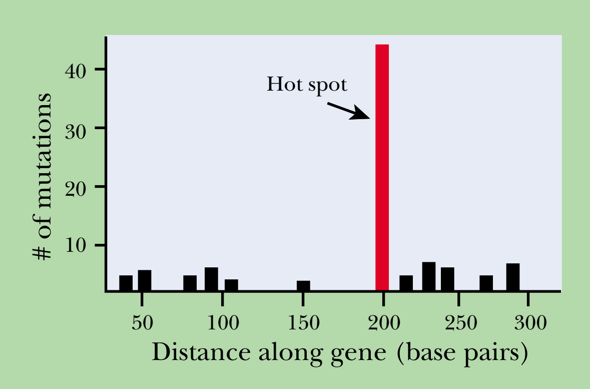


FIGURE 13.23 ***Hot Spots in Distribution of Mutations***

The frequency of mutations along a gene is graphed, showing that one particular location at around 200 base pairs along the gene receives far more mutational events than other regions.

are, but here and there in the DNA sequence are locations where mutations happen many times more often than average (Fig. 13.23). All the mutations occurring at such a site will usually be identical. These sites are called **hot spots**.

Most hot spots are due to the presence of occasional methyl-cytosine bases in the DNA (see Ch. 4 for DNA methylation). These are made from cytosine after DNA synthesis and they pair correctly with guanine, just like normal cytosine. However, every

hot spots Site in DNA or RNA where mutations are unusually frequent

TABLE 13.01 Mutation Rates in DNA Genomes

		Mutation Rate per Generation		
Organism	Genome Size (kilobases)	Per kb	Per genome (uncorrected)	Per effective genome
Bacteriophage M13	6.4	7.2×10^{-4}	0.005	0.005
Bacteriophage Lambda	49	7.7×10^{-5}	0.004	0.004
Escherichia coli	4,600	5.4×10^{-7}	0.003	0.003
Saccharomyces cerevisiae	12,000	2.2×10^{-7}	0.003	0.003
Caenorhabditis elegans	80,000	2.3×10^{-7}	0.018	0.004
Drosophila	170,000	3.4×10^{-7}	0.058	0.005
Human	3,200,000	5.0×10^{-8}	0.160	0.004

now and then methyl-cytosine spontaneously deaminates to give thymine (=methyl-uracil). This pairs with adenine, not with guanine, and so when the DNA is replicated next, an error results.

Hot spots also occur for deletions, insertions and other major DNA rearrangements. Depending on the mechanism of mutation, certain sequence motifs will favor particular genetic events. As noted above, many rearrangements are due to illegitimate recombination between two nearby regions of DNA with similar sequences.

How Often Do Mutations Occur?

The rates of mutation per generation for several well known organisms are shown in Table 13.01. Defining the rate of mutation is not as obvious as it might appear. Should mutations be expressed as the rate of alterations to the DNA sequence or should they be expressed as the rate of alterations in functional genes? In viruses and bacteria, where most DNA encodes genes, it makes little difference which method of expression is utilized. However, in higher organisms, most DNA is non-coding and alterations in the non-coding DNA rarely have significant effects on the viability of the organism. The concept of "**effective genome**" (i.e., the coding portion of the genome) allows for this and can be used when considering mutation rates. In Table 13.01, the mutation rates cited for the multi-cellular organisms per "effective genome" differ from the raw values. For the single-celled organisms, these two rates are, of course, nearly identical.

All organisms have similar mutation rates, if the effective genome size is considered.

Overall, the more DNA per genome, the lower the mutation rate per kilobase of DNA. Higher organisms have both more accurate DNA replication and more sophisticated DNA repair systems in order to cope with their extra DNA. However, because they have such vast amounts of DNA, the mutation rates per genome are still much higher for the more advanced organisms. In contrast, mutation rates per effective genome are very similar for all organisms, suggesting that this is the level at which evolutionary constraints act. Too high a mutation rate will cause too much damage; too low a rate will fail to provide enough new mutations to drive evolution.

Another ambiguity resides within the term "generation." For single-celled organisms, there is no problem in defining generation, but, again, higher organisms are problematic. Does the term generation refer to an individual cell or of the multicellular organism as a whole? For example, there are many cell generations that transpire between a fertilized egg and the next generation of reproductive cells (i.e., the egg or sperm cells). The values in Table 13.01 refer to each cellular generation.

effective genome The portion of the genome that consists of useful genetic information and ignores the intervening and non-coding DNA. Only applicable to eukaryotic organisms

In the roundworm *Caenorhabditis elegans*, there are only about 10 cell divisions between the zygote and the next generation of germ cells, so the mutation rate per generation of whole animals is 0.04 (about ten times the rate per cell generation). For animals with more cell divisions between generations, this number is much higher—in flies it is 0.13 and for humans 1.6. In flies and humans, the number of cell divisions leading to sperm is significantly greater than the number leading to egg cells. Consequently, sperm carry many more new mutations than eggs and the male parent contributes a greater portion of mutations to the offspring than the mother.

RNA genomes have extremely high mutation rates.

Unlike DNA polymerases, RNA polymerases lack the ability to proofread. Consequently, RNA-based genomes have much higher rates of spontaneous mutation. The mutation rates per genome per generation range from 1 to 5 for small RNA viruses—a rate approximately 1000-fold higher than for the DNA-based cells in Table 13.01. This rate is so high that a significant fraction of the virus particles are defective and it has been estimated that a three-fold higher mutation rate would cause total lack of viability. RNA is only used as the genetic material by certain viruses with relatively small genomes, such as influenza or AIDS. These viruses evade immune surveillance by mutating very rapidly (see Ch. 17). In practice, RNA is not used as the genetic material by any living cell, or even by larger viruses such as bacteriophage T4 or smallpox. Presumably the much higher error rate characteristic of RNA would cause severe problems in organisms where many gene products interact.

Reversions Are Genetic Alterations That Change the Phenotype Back to Wild-type

A reversion that precisely restores the original DNA sequence is highly unlikely.

Obviously, it is possible for DNA that already carries one mutation to be mutated again. There is a small chance that the second mutation will reverse the effect of the first. This process is called **reversion** and refers to the observable outward characteristics of an organism. Reversion is thus a term used to describe a phenotype.

What is the chance a single preselected base will mutate to revert itself to the wild-type? The likelihood that precisely the one base out of millions that were previously mutated will be the very one to mutate again is extremely low. Those rarities where the original base sequence is exactly restored are known as **true revertants**. More often, revertants actually contain a second base change that cancels out the effect of the first one. These are therefore known as **second-site revertants** and the second mutation is known as a **suppressor mutation**.

Most revertants have changes in their DNA that cancel out the effects of the previous mutation.

Not surprisingly, mutations that involve more than a single base change are much less likely to revert. Since part of the original DNA sequence has been completely lost, deletions are completely nonrevertible, at least as far as restoring the original DNA sequence is concerned. Precise reversal of insertions, inversions and translocations is also extremely rare, though not theoretically impossible. Nonetheless, in such cases, reversion is almost always due to compensatory changes in another gene(s).

A second-site reversion can occur if the original mutation was a frameshift mutation caused by the deletion or insertion of a single base. The frameshift mutation alters the reading frame and garbles the protein sequence, as shown in Fig. 13.24. But suppose an extra base is inserted a little way farther along the sequence. This second-site insertion will restore the original reading frame. Although the DNA sequence is not identical to its original state, because of base redundancy the protein has been exactly restored. Similarly, an insertion mutation can be corrected by a second-site deletion.

A less obvious but more frequent case of reversion occurs where the original mutation was a base change. Again, the key to successful reversion is to restore activity to

reversion Alteration of DNA that reverses the effects of a prior mutation
second-site revertant Revertant in which the change in the DNA, which suppresses the effect of the mutation, is at a different site to the original mutation
suppressor mutation A mutation that restores function to a defective gene by suppressing the effect of a previous mutation
true revertant Revertant in which the original base sequence is exactly restored

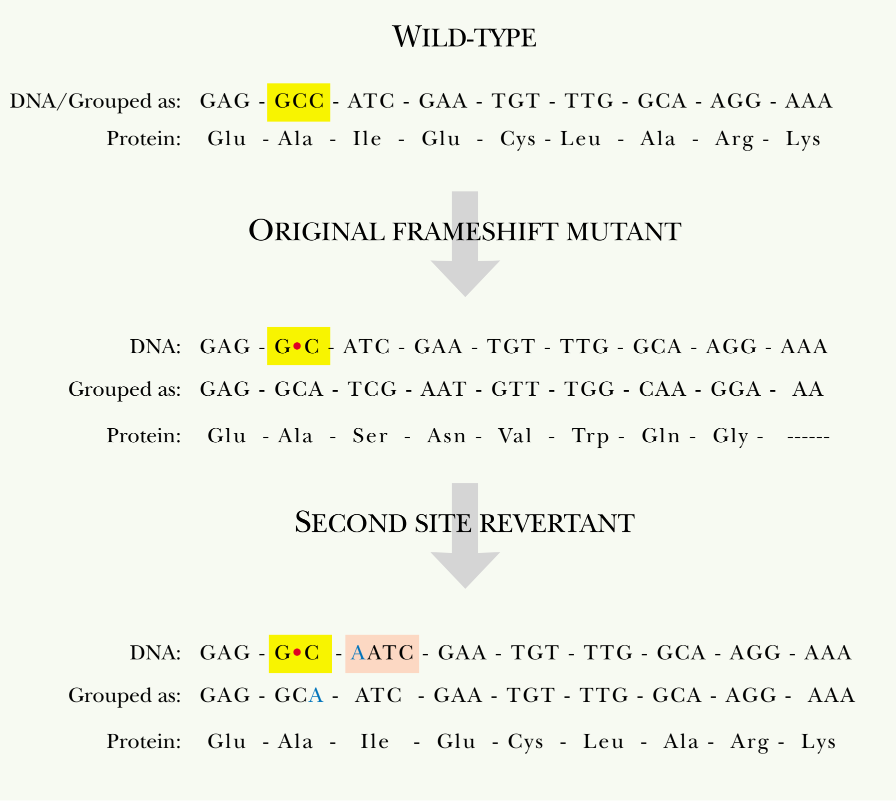

WILD-TYPE
DNA/Grouped as: GAG - GCC - ATC - GAA - TGT - TTG - GCA - AGG - AAA
Protein: Glu - Ala - Ile - Glu - Cys - Leu - Ala - Arg - Lys
ORIGINAL FRAMESHIFT MUTANT
DNA: GAG - G•C - ATC - GAA - TGT - TTG - GCA - AGG - AAA
Grouped as: GAG - GCA - TCG - AAT - GTT - TGG - CAA - GGA - AA
Protein: Glu - Ala - Ser - Asn - Val - Trp - Gln - Gly - ------
SECOND SITE REVERTANT
DNA: GAG - G•C - AATC - GAA - TGT - TTG - GCA - AGG - AAA
Grouped as: GAG - GCA - ATC - GAA - TGT - TTG - GCA - AGG - AAA
Protein: Glu - Ala - Ile - Glu - Cys - Leu - Ala - Arg - Lys

FIGURE 13.24 ***Second-site Reversion of Frameshift Mutation***

The DNA sequence and the encoded amino acids are shown for wild-type, an original frameshift mutant and a second site revertant. In the original mutant a single base deletion alters the reading frame. The second site revertant has an extra base inserted, which reverses the original frameshift. Although the DNA sequence is not identical to the wild type, the amino acid sequence of the protein has been restored.

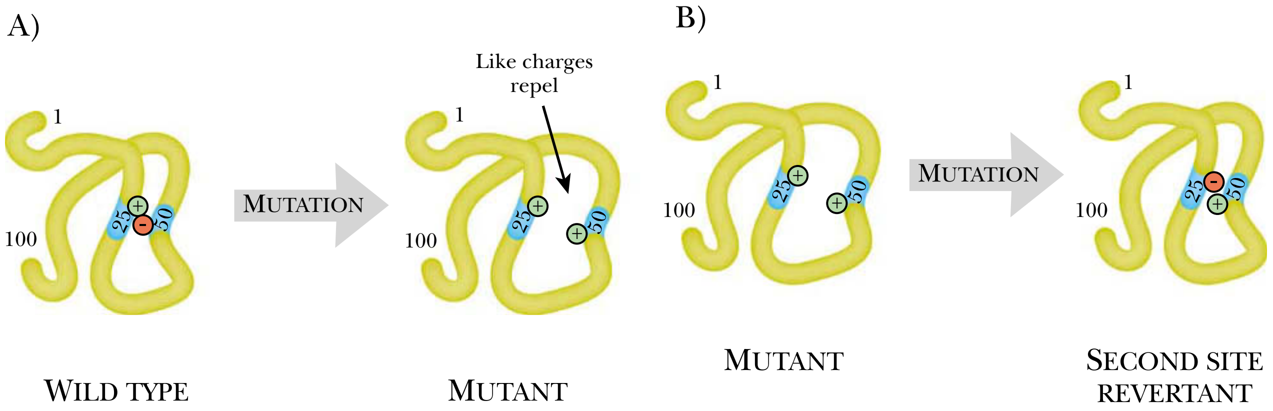


FIGURE 13.25 ***Second-site Reversion of Base Change Mutation***

A) The original mutation alters amino acid #50 from negatively to positively charged. This causes a change in conformation due to charge repulsion. B) A second mutation alters amino acid #25 from positively to negatively charged. This restores the attraction between position #25 and #50 and the protein reverts to its original conformation.

the protein. The precise restoration of the original DNA sequence is less important. Consider a protein whose correct three-dimensional structure depends on the attraction between a positively charged amino acid at, say, position 25 and a negatively charged one at position 50. Suppose the original mutation changes codon 50 from GAA for glutamic acid (negatively charged) to AAA, which encodes lysine, a positively charged amino acid (Fig. 13.25). The folding of the protein is now disrupted due to charge repulsion. A true revertant could be made by replacing AAA with GAA. However, the attraction between residues 25 and 50 can be restored by mutating codon

The stop codons were originally identified by mutations in bacteriophage T4. The first one identified was UAG, the amber codon, which received its name in a curiously convoluted manner. The laboratory of Seymour Benzer at Caltech was looking for a mutation that would allow a certain kind of bacteriophage mutant to grow. Benzer said that whoever identified the mutation, would have it named after him. The mutation was eventually isolated by a student named Harris Bernstein. Since "Bernstein" is German for "amber" UAG was named the amber codon. The second stop codon to be found (UAA) was called "ochre" to keep the color theme. The third stop codon, (UGA) is less common and so the use of "opal" or less often "umber" is less frequent and not fully settled.

25 to give a negatively charged amino acid. This yields a negative charge at position 25 and a positive charge at position 50. Because the attraction between these two regions has been restored, the protein may fold properly again (Fig. 13.25). Will such a revertant protein work correctly? Sometimes yes, sometimes no; it depends on a variety of other factors, such as whether folding is completely restored and whether the alterations damage the active site.

Reversion Can Occur by Compensatory Changes in Other Genes

During selection of revertants of a particular gene based on phenotypic differences, a mixture of new mutations will be found. Experimentally this is most easily done using cultures of bacteria or yeast, but the principle applies to higher organisms, too. Occasional true revertants will regain the original DNA sequence. Various possible second site revertants will occur that restore at least some activity to the protein. These may lie within the gene that was originally mutated or in other genes. Restoring activity to a protein by a second mutation within the same gene is sometimes known as **intragenic suppression**. This is to contrast with **extragenic suppression**, where the effects of a mutation are suppressed by a compensatory mutation in a second, quite separate gene.

Sometimes reversion is due to compensatory changes in a completely different gene.

Extragenic suppression is also called intergenic suppression, to indicate that the two genes involved interact in some way. A whole variety of possible mechanisms exists for such effects. Since an alteration in one gene is making up for a defect in another, extragenic suppression rarely restores function completely.

Two examples of extragenic expression are presented; one due to metabolic compensation and the other due to altered tRNA (see below). Many organisms have multiple genes that code for closely related proteins. For example, the fumarate reductase and the succinate dehydrogenase of bacteria such as *E. coli* both catalyze the same reaction, the interconversion of fumarate with succinate. The fumarate reductase (FRD) is used to make succinate from fumarate during anaerobic growth, whereas the succinate dehydrogenase (SDH) functions in the Krebs Cycle to convert succinate to fumarate during aerobic growth. Although there are slight differences in the DNA and protein sequence, the major difference is in the mode of regulation; the *frd* genes are normally expressed only anaerobically and the *sdh* genes only in the presence of oxygen. Consequently, a mutation in *frd* may be suppressed by a regulatory mutation that allows expression of succinate dehydrogenase anaerobically. Conversely, a mutation in *sdh* may be suppressed by a regulatory mutation that turns on fumarate reductase in the presence of oxygen.

extragenic suppression Reversion of a mutation by a second change that is within another distinct gene
intragenic suppression Reversion of a mutation by a second change at a different site but within the same gene

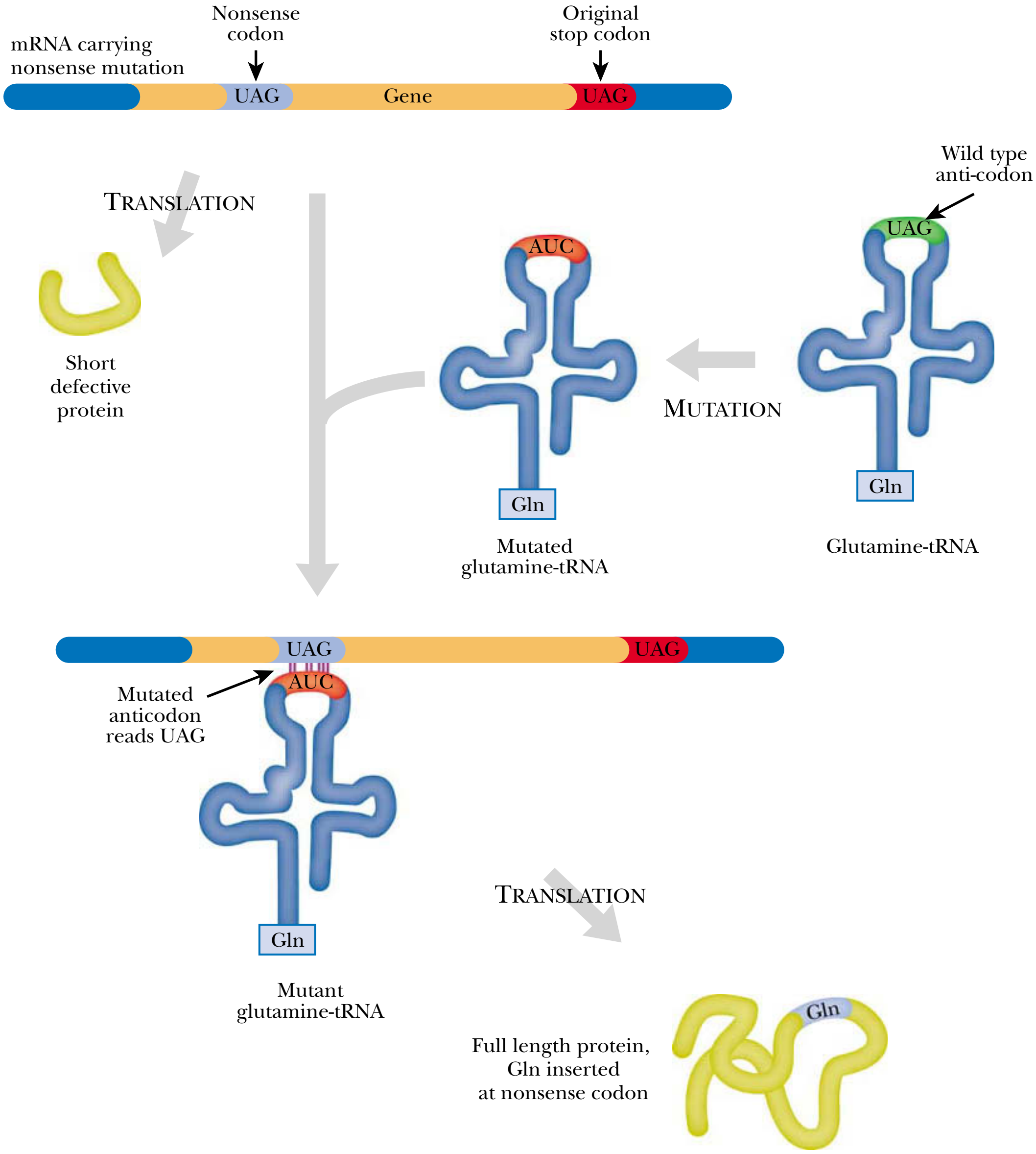


FIGURE 13.26 ***Mechanism of Nonsense Suppression***

A gene containing a nonsense codon suffers a premature stop during translation and a short defective protein is made. However, a tRNA whose anticodon is mutated (from GUC to AUC) can recognize the stop codon and insert an amino acid (glutamine in this case). A full length protein will be made that has only one amino acid different from the original wild type.

Altered Decoding by Transfer RNA May Cause Suppression

Mutant tRNA molecules are known that read stop codons and insert amino acids. This suppresses nonsense mutations.

Nonsense mutations can be suppressed by alterations in tRNA. As discussed above, a nonsense mutation occurs when a codon that should code for an amino acid is changed to a stop codon. This results in a truncated and usually nonfunctional protein. Such a defect may be suppressed, at least partially, by changing the anticodon sequence of a tRNA molecule so that it recognizes the stop codon instead. Consider the stop codon UAG. Altering the anticodon of tRNA^{Gln} from GUC (reading CAG for Gln) to AUC will make it recognize UAG instead. Such an altered tRNA will insert glutamine wherever it finds a UAG stop codon (Fig. 13.26).

Such altered tRNA molecules are known as **suppressor tRNAs**. The UAG stop codon is known as *amber* and the UAA stop codon as *ochre*. The UGA stop codon has no universally accepted name, but is sometimes called *opal* or *umber*. Amber suppressors are mutant tRNAs that read UAG instead of their original codon. Ochre suppressor tRNAs read both UAA and UAG due to wobble. Opal suppressors are rare.

Suppressor tRNA mutations can only occur if a cell has more than one tRNA that reads a particular codon. One may be mutated while the other must carry out the original function; otherwise, the loss of the original tRNA would be lethal. In practice, cells often have multiple tRNA genes and so suppressor mutations are reasonably common, at least in microorganisms. Bacterial suppressor mutations have been found in tRNAs for glutamine, leucine, serine, tyrosine and tryptophan. The amino acid inserted by the suppressor tRNA may be identical to the original amino acid whose codon mutated to give the stop codon. In this case, the protein made will be fully restored. Alternatively, a different amino acid may be inserted and a partially active protein may be produced.

Remember that stop codons are normally recognized by release factor, and have no cognate tRNAs. Since suppressor tRNA competes with release factor, suppression is never complete and typically ranges from 10 to 40 percent. This may provide enough of the suppressed protein for the cells to survive. However, the suppressor tRNA will also suppress other stop codons in the same cell and so generate longer (and incorrect) versions of many proteins whose genes were never mutated. Not surprisingly, cells with suppressor mutations grow more slowly. Only bacteria and lower eukaryotes (e.g., yeasts, roundworms) can tolerate suppressor mutations. In both insects and mammals, suppressor mutations are lethal.

Frameshift suppressor tRNAs are also occasionally found among bacteria. These mutant tRNA molecules have an enlarged anticodon loop and a four-base anticodon. This enables them to insert a single amino acid by reading four bases in the mRNA. They can suppress the effects of a frameshift mutation that was caused by the insertion of a single extra base. Frameshift suppressor tRNAs with five-base anticodons have been made artificially, but have not been isolated naturally.

Mutagenic Chemicals Can Be Detected by Reversion

Testing chemicals for tumor formation in animals is expensive and time-consuming. However, since cancer is due to DNA alterations, most carcinogens are in fact also mutagens. Consequently, chemicals suspected of being carcinogenic are routinely screened for possible mutagenic effects by testing against bacteria. The **Ames test** makes use of multiple strains of the bacterium *Salmonella typhimurium* carrying well characterized mutations in the genes for histidine synthesis. It is used routinely by industry and government agencies to screen food colorings and preservatives, cosmetics such as hair dyes, and many other industrial chemicals for possible mutagenic effects.

Screening for reversion using well characterized mutations allows the detection of mutagenic chemicals.

Mutants of *Salmonella typhimurium* carrying mutations in the *his* genes can no longer make the amino acid histidine and cannot grow unless given histidine. When large numbers of these mutant bacteria are placed on growth medium lacking histidine, just a handful of colonies appear. These are revertants, and since reversions are merely mutations back to the original state, the frequency of reversion is also increased by mutagenic agents. To test a suspect chemical, samples of *Salmonella his* mutants are mixed with the agent and then plated onto minimal medium with just a trace of histidine. The amount of histidine added is growth-limiting. Therefore the bacteria can only divide a few times and run out of histidine before making visible colonies. If the added

Ames test Test for mutagenic activity that makes use of bacteria
suppressor tRNA A mutant tRNA that recognizes a stop codon and can insert an amino acid when it reads a stop codon on the mRNA

chemical does induce mutation then His+ revertants will be formed during these few cell divisions, then each resulting revertant can grow into a visible colony. Different types of original mutations, for example, base changes or frameshift mutations, are used to screen for different classes of mutagenic agents.

The *Salmonella typhimurium* strains used in practice for mutagen testing have several alterations that make them more sensitive to mutagens. Firstly, they carry mutations that make the bacterial outer membrane more permeable to large and/or hydrophobic molecules. Secondly a variety of alterations have been made to inactivate bacterial DNA repair mechanisms (see Ch. 14). For example, the *uvrB* gene may be deleted to eliminate excision repair of DNA.

Certain chemicals (pro-mutagens) are only mutagenic after metabolic conversion to active derivatives. In animals this is usually due to liver enzymes such as cytochrome P450 that are intended to detoxify harmful chemicals by oxidation. When testing for pro-mutagens, an extract containing such rat liver enzymes is mixed with the bacteria in the Ames test. Recently, genes for some variants of human cytochrome P450 have been cloned and successfully expressed in the *Salmonella* strains used for mutagen testing. The resulting bacteria synthesize the liver enzymes internally and are much more sensitive in their response to pro-mutagens.

Experimental Isolation of Mutations

The emergence of molecular biology has relied greatly on the analysis of mutations in a variety of organisms. Obviously, such analysis requires a supply of mutants to work with. These may be obtained by a wide range of approaches. Mutations may be spontaneous or artificially induced. Artificial mutagenesis may be carried out using living organisms (*in vivo* mutagenesis) or performed using isolated DNA (*in vitro* mutagenesis). In addition, some means is needed for identifying mutants, whether they occur spontaneously or are made deliberately.

Since the frequency of spontaneous mutations is low, it is only possible to rely on this source of mutations when a population of millions of organisms can be surveyed in a reasonable time. Consequently, this approach is largely restricted to bacteria and single-celled eukaryotes, such as yeast. For a gene of average size (~1000 bp) the mutation rate is approximately 0.5 per million per generation in bacteria such as *E. coli* (see Table 13.01). Thus, a typical culture of several 1,000 million bacteria per ml that has resulted from several generations of growth may contain half a dozen spontaneous mutants per million cells (or several thousand mutants per ml of culture) affecting any particular gene of interest (assuming such mutations are not lethal).

From a practical viewpoint, the problem is not so much causing mutations as finding and isolating the ones desired.

The problem then becomes how to isolate these mutants. It is clearly impractical to examine millions of microorganisms individually. Therefore the isolation of spontaneous mutants relies on some form of direct selection. Usually, samples of the culture are spread on the surface of solid medium designed to allow only the desired mutants to grow. For example, mutations in DNA gyrase make bacterial cells resistant to quinolone antibiotics, such as nalidixic acid. Therefore medium containing nalidixic acid kills the vast majority of bacteria and can be used to isolate gyrase mutants. Sometimes bacterial cultures may be enriched for the required mutation by growth for several generations in liquid selective medium before transferring to solid selective medium for the final isolation. This increases the proportion of the required mutants in the population.

A large number of direct selections have been used to isolate bacterial mutants. The major categories of selection and the kinds of genes affected are as follows:

A. Resistance to antibiotics. Alterations in genes whose products are targets of the antibiotic or that are involved in entry of the antibiotic into the cell. For example, streptomycin resistance selects alterations in ribosomal protein S12, rifampicin resistance selects alterations in RNA polymerase, nalidixic acid resistance selects mutations in DNA gyrase.

B. Resistance to analogs of metabolites. Alterations in genes whose products are involved in synthesis, degradation or transport of the metabolite. For example, chlorate (an analog of nitrate) selects mutants defective in nitrate reductase, chloroethanol selects mutants defective in alcohol dehydrogenase, various selenium compounds select mutants with altered sulfur metabolism.

C. Resistance to bacteriophage. Alterations in genes encoding bacteriophage receptor or components needed for entry of viral DNA. For example, resistance to bacteriophage lambda selects for loss of LamB protein on cell surface, resistance to bacteriophage T1 selects for loss of TonB protein needed to energize viral DNA entry.

D. Growth in the absence of certain metabolic supplements. Usually used to select revertants from mutants defective in synthesis of amino acids, nucleotides, vitamins etc.

E. Growth on certain substances as carbon source. Usually used to select revertants from mutants defective in metabolism of sugars, organic acids by known pathways. Growth on novel compounds is sometimes selected. For example, selection for growth of *E. coli* on propanediol selects for aerobic expression of a pathway normally only expressed anaerobically during the fermentation of deoxysugars.

In Vivo versus *In Vitro* Mutagenesis

The frequency of mutation can be greatly increased by a variety of chemical agents or certain types of radiation, as already described in the earlier part of this chapter. These agents are often used deliberately on living cells (*in vivo* mutagenesis). In particular, sublethal concentrations of mutagenic chemicals may be added directly to growing cultures of bacteria or single celled eukaryotes. It is also possible to treat higher organisms, though the approach is limited by the necessity of screening large numbers of large organisms. Nonetheless, relatively small multi-cellular organisms, such as flies and roundworms, have been successfully mutagenised by this approach.

Alternatively, purified DNA may be treated with the mutagenic agent (*in vitro* mutagenesis—see below). In this case, the DNA must be transformed back into the bacteria before screening for mutations is performed.

Mutations in certain genes (mutator genes) involved in DNA synthesis and repair can themselves increase the frequency of mutation. The presence of such defects may be used to generate mutations in bacteria and some single celled eukaryotes. However, such defects tend to be lethal in more complex organisms.

Insertion and deletion mutations may also be made *in vivo* by using transposable elements. Genetic elements such as insertion elements, transposons or bacteriophage Mu insert semi-randomly into host DNA. If the insertion point lies within a host cell gene, the result is an insertion mutation that inactivates the gene in question. Certain transposable elements excise themselves from DNA inaccurately, leaving behind deletions at the point of prior insertion. The behavior of transposable elements is described in more detail in Ch 15, Mobile DNA.

If the frequency of mutation has been increased to where several bacteria per thousand carry mutations in the gene of interest, it is possible to identify mutants by screening (as opposed to selection). A culture of bacteria that has been mutagenized is diluted and samples are spread onto the surface of solid media so as to give single colonies (between 50 and a few hundred colonies per plate are typically obtained). The colonies are then screened by a variety of techniques such as indicator media or replica plating. The term "phenotypic screening" refers to the analysis of the mutants by their phenotype, rather than by examination of the genes.

Bacterial mutants are often screened for their responses on a variety of specialized culture media.

Indicator media are growth media that change color in response to metabolic reactions. The simplest show alterations in pH due inclusion of a pH indicator in the medium. These are typically used to monitors sugar breakdown by bacteria as this generates acid and so alters the pH of the growth medium. Redox indicators, such as

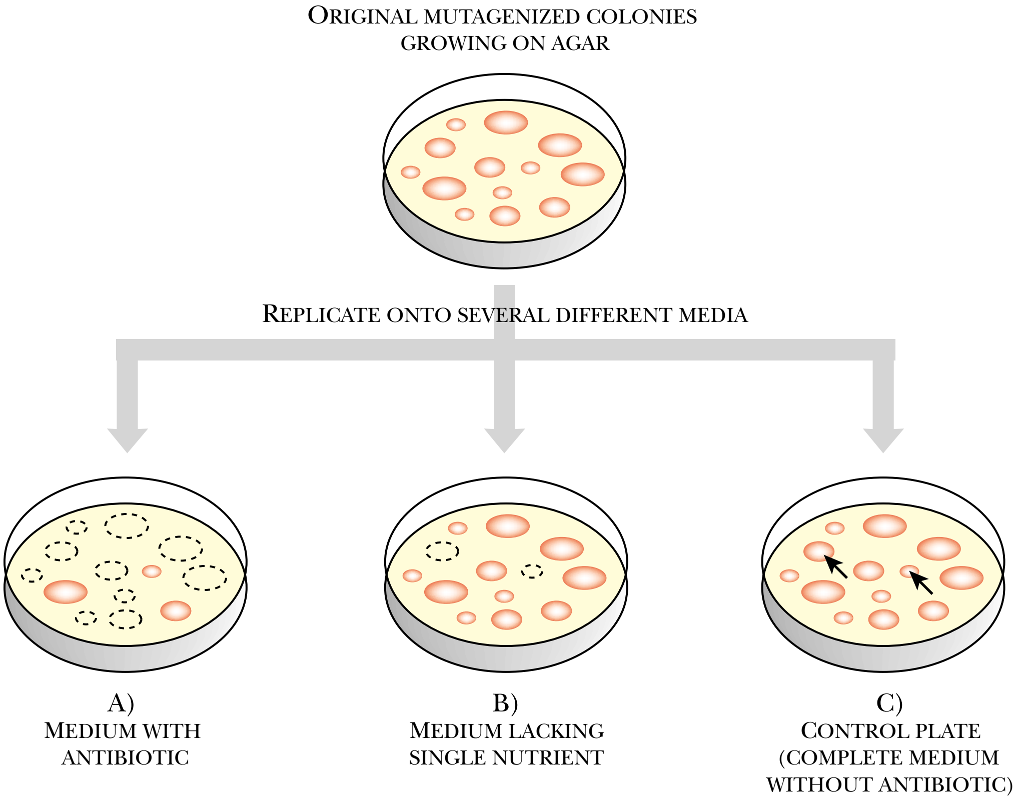


FIGURE 13.27 ***Replica Plating***

After treatment with mutagens, bacterial colonies are grown on normal medium in a Petri dish. Replicas are made using pads of velvet or filter paper to transfer samples from each original colony simultaneously to fresh media. Growth on the different media is compared. Colonies present only on a medium with a positive selection, such as an antibiotic, may be retrieved directly (A). However, colonies missing from a particular test medium, such as medium lacking a particular nutrient (B) must be retrieved from the control plate (C). Since these mutants will look identical to the other colonies, they are identified by their position, as indicated by the arrows.

tetrazolium dyes, respond to oxidation reduction reactions. They may be used to monitor the ability of bacteria to oxidize certain growth substrates.

More specific indicators, such as substrates for individual enzymes, may also be used. For example, X-gal is a substrate for β-galactosidase that releases a blue dye when split by this enzyme (see Ch. 7). Bacterial colonies expressing significant levels of β-galactosidase turn blue in the presence of X-gal. One of the most cheerful indicators is the use of selenium salts. Bacteria that can reduce selenate or selenite accumulate granules of elemental selenium that are bright red. In this case several steps of a specific pathway, rather than just one enzyme, are involved. Note that for an indicator system to work well, the colored product must be insoluble, otherwise it will diffuse through the agar and the color will no longer be localized to the colony that performed the reaction.

Replica plating allows screening for mutants that fail to grow under the test conditions.

Replica plating is another widely used form of phenotypic screening. It is particularly useful when searching for mutants that have lost the ability to grow under certain conditions. The same mutagenized bacterial colonies are tested for growth on a variety of media and colonies that fail to grow on the medium of interest are kept. For example, media with different carbon sources, growth supplements or growth inhibitors may be used. Since it is not possible to subculture a colony that failed to grow on a test medium, the mutagenized colonies are first grown on normal (i.e. non-selective) medium. Filter paper or velvet pads are then pressed onto the colonies and pick up bacteria from each original colony. Bacteria corresponding to each colony are then transferred to assorted test media by pressing the filter paper or velvet pads onto the surface of the fresh medium (Fig. 13.27). This technique preserves the arrangement of the colonies on the agar and allows colonies missing on the test medium to be retrieved from the original master plates.

Site-Directed Mutagenesis

DNA may be manipulated in the test tube and the altered DNA construct may then be inserted into the target organism. The simplest form of such *in vitro* mutagenesis is

TABLE 13.02 Techniques used for *In Vitro* Mutagenesis

Chemical mutagenesis of cloned DNA
The gene to be mutagenized is cloned onto a suitable vector, usually a plasmid. DNA carrying the target gene is extracted and purified and treated with a chemical mutagen *in vitro*. The altered DNA is then transformed back into the original organism and screening is carried out to identify organisms that received a mutant version of the gene.

Gene disruption by restriction and ligation (See Ch. 22, Recombinant DNA)
A DNA cassette, often carrying a gene for resistance to some antibiotic to allow selection, is inserted into the target gene by using restriction enzymes and DNA ligase. This approach is often used if convenient restriction sites are available. If not, then PCR-based introduction of extra DNA is a good alternative.

In vitro DNA synthesis (See Ch. 24, Genomics)
Single stranded DNA is sometimes generated for sequencing by using M13 vectors. In vitro DNA synthesis may be performed using such ssDNA as template using T7 polymerase and a supply of nucleoside triphosphates. DNA polymerization may be initiated using artificially synthesized primers whose sequence has been altered by a few bases. This will generate a mutagenized product that incorporates these changes. This technique has largely been replaced by PCR based methods.

PCR based techniques (See Ch. 23, PCR)

a) Introduction of Specific Base Changes
Using PCR primers whose sequence has been altered will generate a PCR product that incorporates these changes.

b) Localized random mutagenesis
Manganese ions cause errors in PCR reactions. Hence random mutations may be introduced into the segment of DNA being amplified.

c) Generation of Insertion or Deletion by PCR
Using PCR primers that include sequences homologous to the target location allows replacement of a region of chromosome with a segment of DNA generated by PCR.

Transgenic technology
Transgenic technology creates genetically modified organisms. It may therefore be regarded as a form of mutagenesis. Extra DNA sequences may be introduced from other organisms, by a variety of techniques.

to treat purified DNA with mutagenic chemicals or radiation. However, a variety of more sophisticated techniques have been used to deliberately construct mutations utilizing genetic engineering technology. These techniques are usually known as **directed mutagenesis** (or, sometimes, **site-directed mutagenesis**, when the site of mutation is carefully controlled). Some techniques introduce changes in one or a few bases, whereas others involve more drastic alterations. Some in vitro techniques generate semi-random base changes whereas others are extremely specific. Assorted methods have been used, in particular PCR (see Ch. 23) is now widely used and has replaced many of the older approaches. These applications are discussed together with the appropriate techniques later in this book and are summarized here for reference in Table 13.02.

DNA alterations are often constructed by a variety of genetic engineering techniques.

directed mutagenesis Deliberate alteration of the DNA sequence of a gene by any of a variety of artificial techniques
site-directed mutagenesis Deliberate alteration of a specific DNA sequence by any artificial technique

CHAPTER FOURTEEN

Recombination and Repair

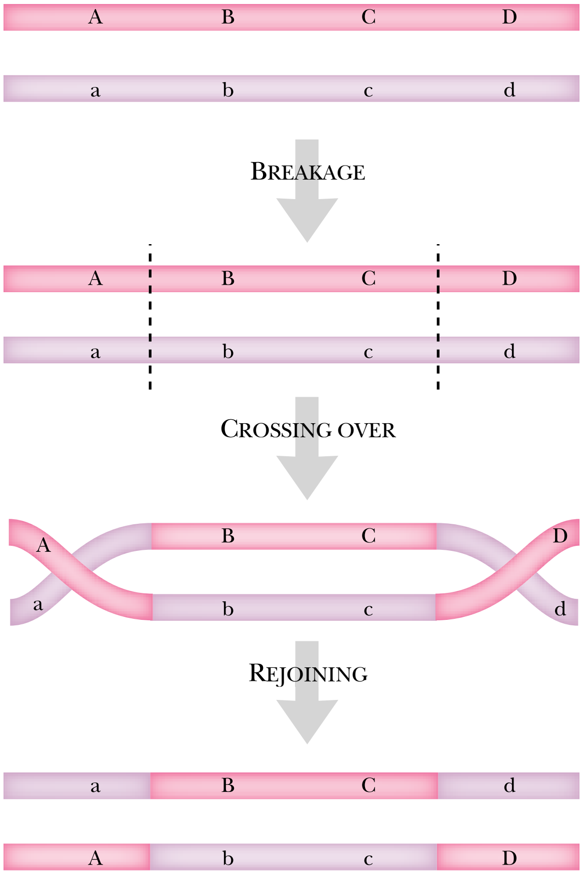


FIGURE 14.01 ***Two Crossovers Result in Recombination***

Two paired double-stranded DNA molecules align related regions (A with a; B with b, etc.). Breakage occurs in each paired DNA molecule followed by crossing over and rejoining of the ends. This exchanges part of one DNA strand with another.

Overview of Recombination

Different DNA molecules may swap segments, usually of related sequence.

The exchange of genetic information between chromosomes occurs in a variety of organisms and under a variety of circumstances. At the molecular level, this involves exchanging segments of DNA molecules by a mechanism known as **recombination**. During sexual reproduction in eukaryotes, the process of meiosis allows the exchange of segments of DNA between homologous chromosomes. This generates greater genetic diversity among the offspring, which in turn allows much greater opportunity for evolutionary selection. Although prokaryotes do not practice sexual reproduction in the same manner as higher organisms, nonetheless they have several mechanisms to promote genetic exchange. Among bacteria, fragments of DNA may be recombined into the chromosome after entering the cell as a result of transformation, transduction or conjugation (see Ch. 18 for a description of DNA transfer mechanisms among prokaryotes). Even virus genomes may undergo recombination under certain circumstances.

Two crossovers between DNA molecules are needed for recombination to occur.

In all cases of recombination, two DNA molecules are broken and rejoined to each other forming a **crossover** (Fig. 14.01). A single crossover usually forms short-lived hybrid DNA molecules. If two crossovers occur, the segment of DNA between them will be transferred from one DNA molecule to the other. This is recombination. [In fact, a single crossover can promote recombination by exchanging the ends of a pair of linear chromosomes. However, a single crossover cannot cause recombination between two circular molecules of DNA.]

crossover Structure formed when the strands of two DNA molecules are broken and joined to each other
recombination Exchange of genetic information between chromosomes or other molecules of DNA

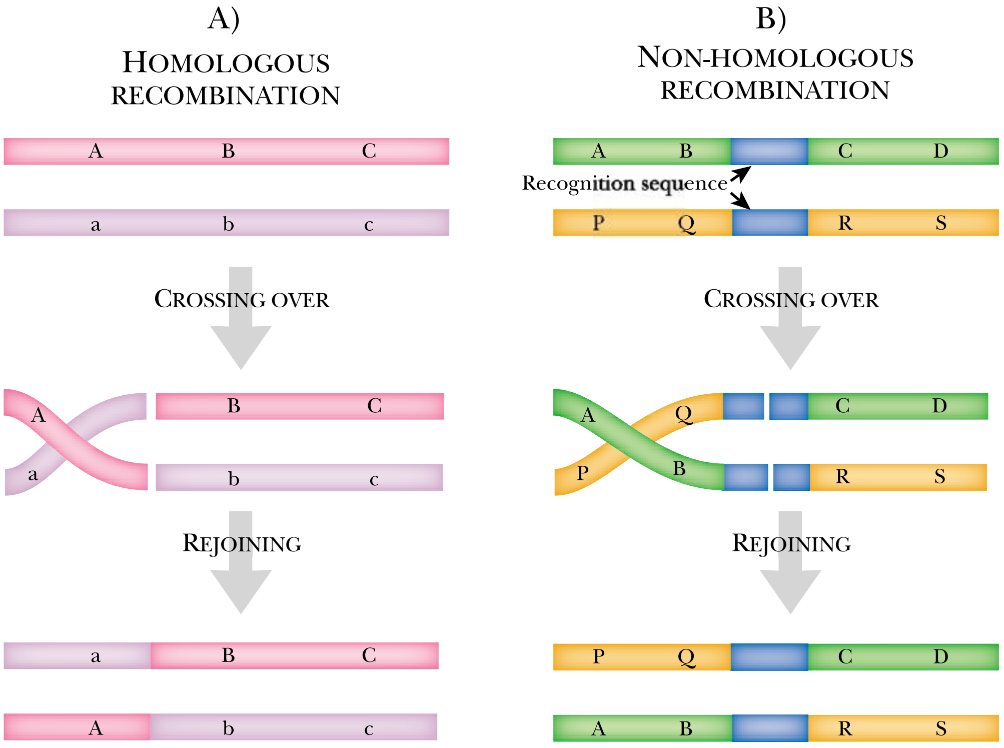


FIGURE 14.02
Homologous versus Non-homologous Recombination

A) In homologous recombination, two DNA molecules have similar sequences such that the pink (top) strand can align with the purple (bottom) strand. If a double-stranded break occurs within the aligned regions, a crossover event will exchange regions of the DNA. B) In non-homologous recombination, related protein recognition sequences lie within two unrelated regions of DNA. Proteins bind to the recognition sequences and carry out recombination. The proteins direct double-stranded breakage and crossing over. Genetic exchange can thus occur between two unrelated DNA molecules. [This event could also theoretically be classified as a translocation.]

Recombination may be divided into **homologous** and **non-homologous recombination**. For homologous recombination to occur, the DNA sequences at the crossover region must be sufficiently similar to base pair. In practice, homologous recombination normally occurs between two copies of the same chromosome (as in meiosis) or between two copies of closely related DNA. Crossovers due to base homology may occur in DNA as short as 20–30 bases, however, 50 to 100 bases is needed for reasonable crossover frequencies. Non-homologous recombination is much rarer and involves specific proteins that recognize particular sequences and supervise the formation of crossovers between them (Fig. 14.02). In both cases, the molecular details come mostly from bacteria, especially *E. coli,* and the details in higher organisms remain much more vague.

Recombination between unrelated DNA sequences can occur due to the involvement of specific recognition proteins.

Molecular Basis of Homologous Recombination

During homologous recombination two double-stranded DNA molecules recognize each other and form a crossover. This involves breaking one strand of each DNA duplex, exchanging strands and rejoining the ends (Fig. 14.03). This results in the formation of a **Holliday junction**, named after Robin Holliday who proposed this model in 1964. The Holliday junction contains two **heteroduplex** regions where single strands from the two separate DNA molecules have paired up. [A heteroduplex is any region of double-stranded nucleic acid, DNA or RNA, where the two strands come from two different original molecules.]

The Holliday junction can twist around and rearrange itself. The two interconvertible forms shown in Fig. 14.04 do not require any change in bonding or base pairing, they are simply alternative conformations. The important issue is that two

heteroduplex A DNA double helix composed of single strands from two different DNA molecules
Holliday junction DNA structure formed during recombination and found at the crossover point where the two molecules of DNA are joined
homologous recombination Recombination between two lengths of DNA that are identical, or nearly so, in sequence
non-homologous recombination Recombination between two lengths of DNA that are largely unrelated. It involves specific proteins, that recognize particular sequences and form crossovers between them. Same as site-specific recombination

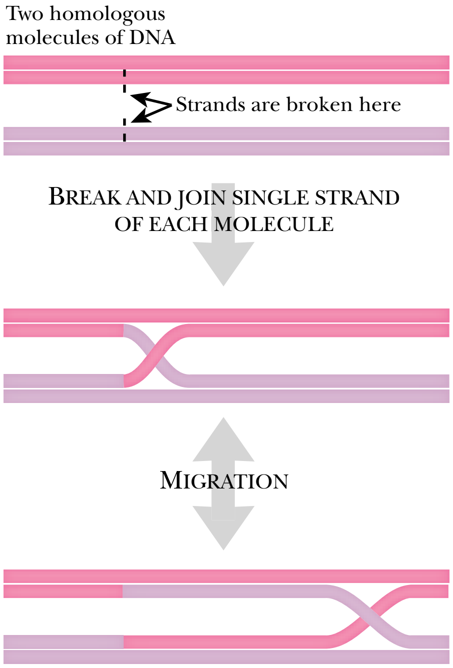

Two homologous molecules of DNA
Strands are broken here
BREAK AND JOIN SINGLE STRAND OF EACH MOLECULE
MIGRATION

FIGURE 14.03 ***Formation of a Crossover***

Two homologous molecules of DNA align in regions of similar sequence. A single-stranded break occurs in the backbone of each molecule. The two ends switch with each other creating a crossover of single-stranded DNA. This crossover can rearrange itself via the intermediate chi form.

genuinely different products can be formed from the breakdown or **resolution** of the Holliday junction. Which product is obtained depends on which conformation the junction is in when it is resolved. One possible result is the regeneration of the two original DNA molecules. In fact they are not absolutely the same as before, and are sometimes known as "**patch recombinants**" as a short patch of heteroduplex remains in each molecule. The alternative is the formation of two hybrid DNA molecules by crossing-over. Resolution of the Holliday junction to give two separate DNA molecules requires an enzyme, known as a **resolvase**. In *E. coli* the RuvC and RecG protein both act as resolvases and can substitute for each other. Resolvases cut and rejoin the second (previously unbroken) strands at the junction, generating the complete, double-stranded crossover.

Another interesting property of the Holliday junction is that it can migrate along the DNA (Fig. 14.05), a process called "branch migration". This involves breaking and re-forming hydrogen bonds. This should, in theory, require no overall energy input, as an equal number of bonds are broken as re-formed. However, in practice, spontaneous migration is extremely slow and energy-dependent enzymes are needed to speed up the process. In *E. coli*, the RuvA protein binds to the junction and RuvB drives migration.

Single-Strand Invasion and Chi Sites

The mechanism of crossover formation involves a temporary triple helix and specific recognition sequences, the chi sites.

A major question is how the homologous sequences of DNA actually find and recognize each other. Remember that we are not talking about two single strands of DNA base-pairing but about the merger of two separate double helixes of DNA in which the bases are turned inwards. The approach seems to differ somewhat between bacteria and higher organisms.

In bacteria, the major proposed mechanism is single-strand invasion. This requires a single-stranded region to form in one of the DNA double helixes. This single strand

patch recombinant DNA double helix with a short patch of heteroduplex due to transient formation of a crossover
resolution Cleavage of the junction where two DNA molecules are fused together so releasing two separate DNA molecules. Refers to the breakdown both of crossovers formed during recombination and of cointegrates formed by transposition
resolvase Enzyme that carries out resolution of DNA

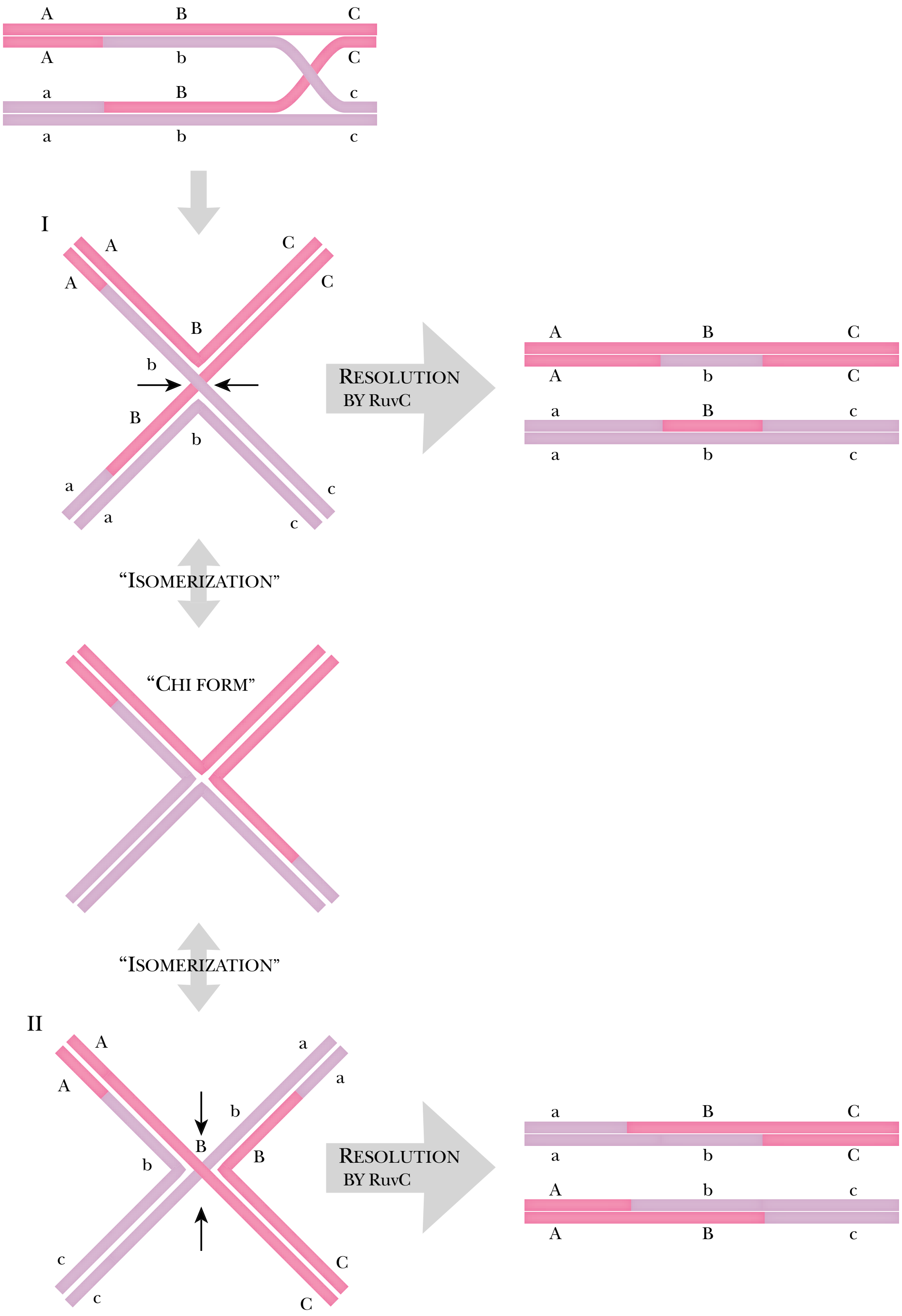


FIGURE 14.04 ***Rearrangement and Resolution of a Holliday Junction***

The Holliday junction can isomerize or change into two alternate conformations. Resolution of form I exchanges gene "B" from the pink molecule with gene "b" from the purple molecule resulting in a patch recombinant. The other strands of the two DNA molecules are unaffected, thus only a small region of heteroduplex DNA results. When form I isomerizes, an intermediate "chi form" appears first, then the crossover reforms. Notice that form I and II have identical base pairing. The difference between forms I and II is the crossover arrangement. In form I, the broken and rejoined strands crossover whereas in form II, the unbroken strands crossover. Resolution by RuvC hybridizes both DNA strands of the double helix rather than just one.

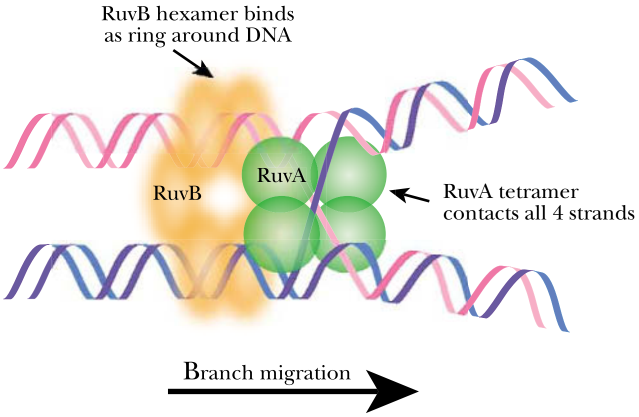

RuvB hexamer binds as ring around DNA
RuvB
RuvA
RuvA tetramer contacts all 4 strands
Branch migration

FIGURE 14.05 ***Migration of a Holliday Junction***

A complex of four RuvA proteins and six RuvB is able to break and reform hydrogen bonds between base pairs thus allowing the crossover to migrate along the DNA helix.

then intrudes into the second DNA double helix to give a triple-stranded helix. The two stages in this process depend on the RecBCD and RecA proteins, respectively. Originally it was thought that crossovers could form between any two homologous sequences, but specific sequences called **chi sites** are also needed. However, since chi sequences (5′-GCTGGTGG-3′) are very common, crossovers do form more or less at random in any sufficient length of bacterial DNA. The chi site was named because of the resemblance of a crossover to the Greek letter chi, χ.

During single-stranded invasion, RecBCD binds to DNA at double-strand breaks. It then moves along the DNA unwinding the double helix until it reaches a chi site (Fig. 14.06). Here the RecD endonuclease clips one of the strands to the 3′-side of the chi sequence and then dissociates from the DNA. The RecBC helicase continues unwinding the DNA generating a single strand.

The **RecA protein** binds to the single strand with the free 3′-end and inserts it into another DNA double helix to give a temporary triple helix (Fig. 14.07). RecA stabilizes the single-stranded region of DNA. This strand invasion causes displacement of one of the strands of the second double-helix. This displaced strand will eventually pair with the remaining single strand of the DNA that was originally unwound by RecBCD. The resulting crossover is resolved as described in Fig. 14.04, above.

The question remains—how did the double-stranded break appear in the first place? Bacteria avoid generating double-stranded breaks in their chromosomes. Furthermore, bacteria are haploid, therefore, do not contain pairs of homologous chromosomes that recombine during sexual reproduction. In practice, recombination in bacteria occurs between the resident bacterial chromosome and shorter fragments of incoming DNA. These fragments enter the bacterial cell by a variety of processes (see Ch. 18 for details). They may be taken up as free DNA from the outside medium (transformation), carried inside a virus particle (transduction), or received from another cell during mating (conjugation). In most cases the incoming DNA will consist of relatively short linear fragments that provide the ends necessary for recognition by RecBCD. These mechanisms provide genetic diversity to haploid, non-sexual bacteria, allowing them to adapt to changing environments.

Site-Specific Recombination

Recombination can occur between two molecules of DNA that have little sequence similarity. This is known as non-homologous or **site-specific recombination**. Instead of aligning regions of DNA by sequence homology, the DNA contains a short

chi sites Specific sequences on the DNA of eukaryotes where crossovers form
RecA protein Protein involved in recombination and repair of DNA in *E. coli* that binds single-stranded DNA
site-specific recombination Recombination between two lengths of DNA that are largely unrelated. It involves specific proteins that recognize particular sequences and form crossovers between them. Same as non-homologous recombination

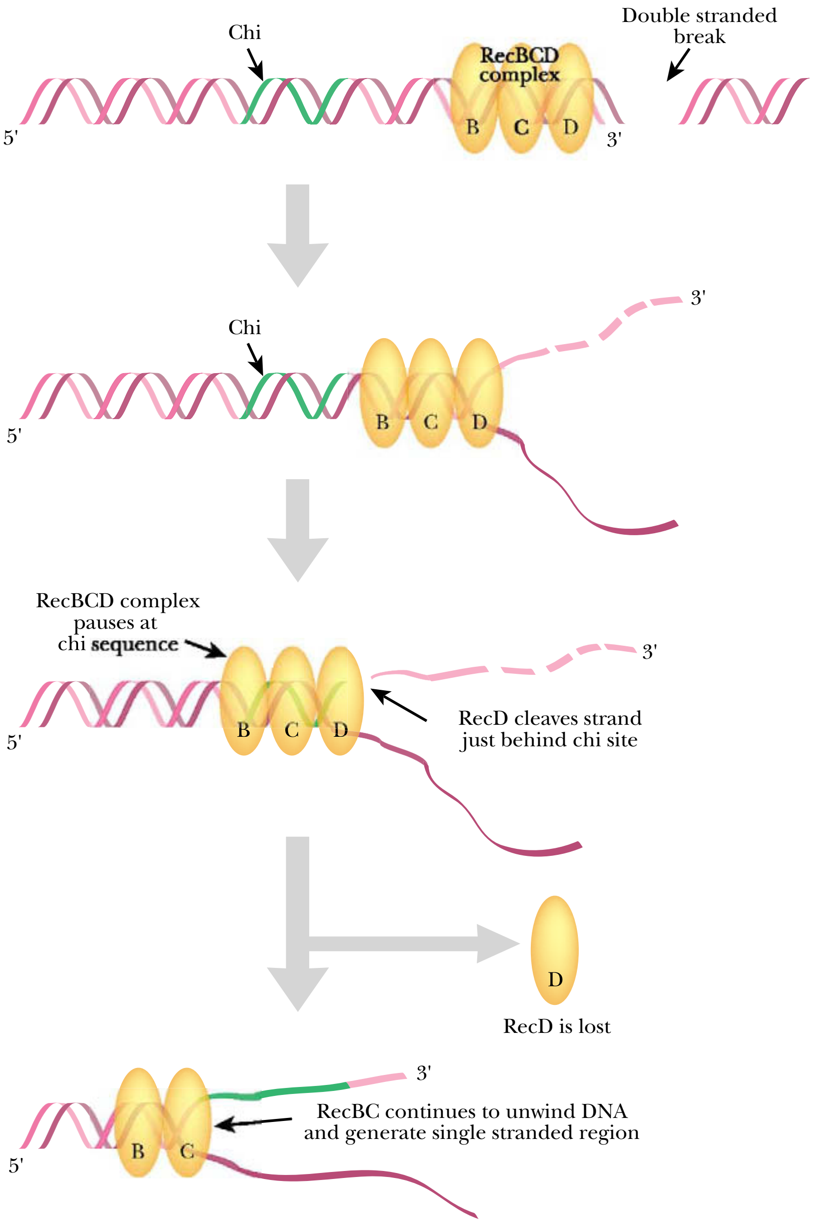

Chi
Double stranded break
RecBCD complex
5'
3'
B
C
D
Chi
3'
5'
RecBCD complex pauses at chi sequence
3'
RecD cleaves strand just behind chi site
5'
D
RecD is lost
3'
RecBC continues to unwind DNA and generate single stranded region
5'

FIGURE 14.06 ***RecBCD Recognizes Chi Sites***

The complex of RecB, RecC, and RecD proteins recognizes the ends of a double-stranded break, and travels along the DNA until reaching the closest chi site. RecD cleaves the backbone of one strand and dissociates from the complex. RecBC continue to unwind the DNA beyond the chi site creating a stretch of single-stranded DNA.

Integration of lambda into the chromosome of *E. coli* depends on recognition of the attachment site by a specific integrase protein.

recognition sequence for a specific protein. The protein then initiates the recombination event.

The classic case is the integration of the DNA of bacteriophage lambda (λ) into the chromosome of *Escherichia coli*. Each of these contains a λ **attachment site (*att*λ)**. These are designated *attBOB'* (on the bacterial chromosome) and *attPOP'* (on the lambda genome). The central core of 15 bases (designated O) of these is identical, but the outermost regions (B and P) differ in size and sequence between host and phage. The core region of both attachment sites is recognized by lambda **integrase** or **Int protein**. This makes a staggered double-strand cut in each core sequence. The ends are joined, and the result is that the circle of lambda DNA is inserted into the bacterial chromosome (Fig. 14.08).

***att*λ** λ attachment site, recognition site on DNA used during integration of lambda DNA into *E. coli* chromosome
integrase Enzyme that inserts a segment of dsDNA into another DNA molecule at a specific recognition sequence. In particular, lambda integrase inserts lambda DNA into the chromosome of *E. coli*
Int protein Same as integrase

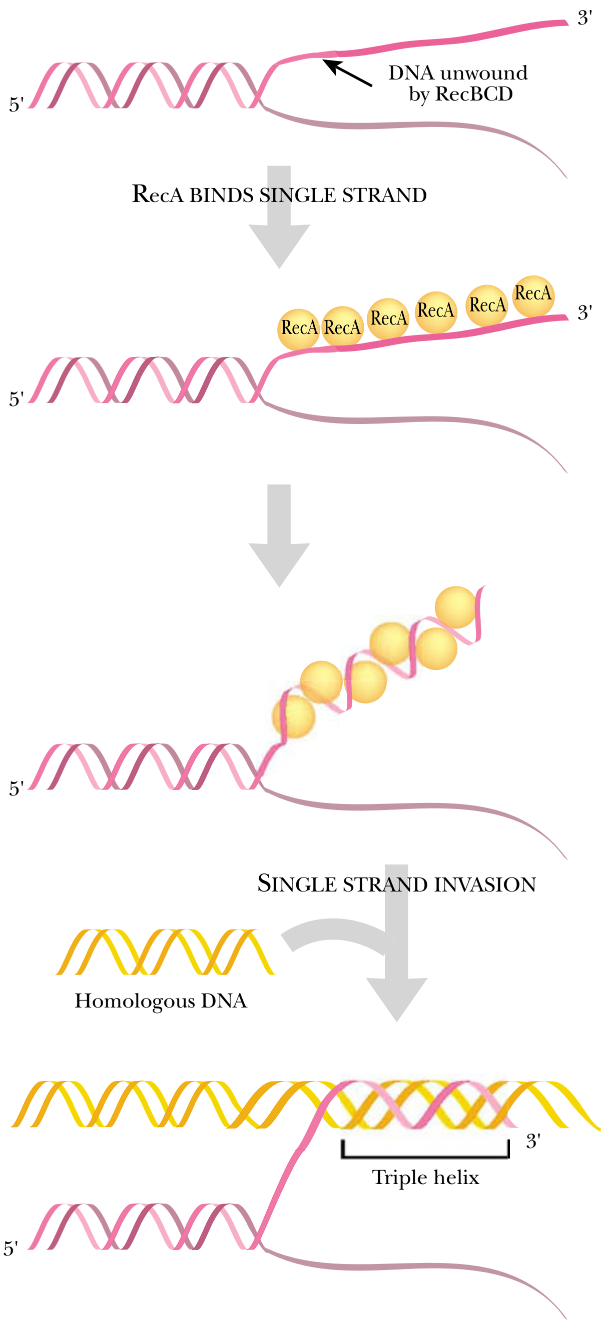


FIGURE 14.07 *RecA Promotes Strand Invasion*

RecA binding stabilizes the unwound single-stranded DNA from Fig. 14.06. The stabilized strand is able to invade the homologous double-stranded DNA forming a triple helix.

In fact, the strands are cut and joined one at a time. The first round of cutting and joining gives a Holliday junction and the second round resolves it leading to **integration**. The arrangement at the crossover is shown in Fig. 14.09. Note that a single crossover event is sufficient for integration of a circular molecule of DNA into another molecule. [Although lambda DNA is linear inside the virus particle, it circularizes upon entering the bacterial cell and before integration (see Ch. 17 for life cycle of lambda).] After integration, lambda DNA is flanked by two hybrid ***att*λ** sites, *attBOP'* and *attPOB'*. Int protein cannot carry out recombination between these hybrid sites and cannot therefore reverse the integration event. Excision of lambda requires **Xis protein**

integration Insertion of a segment of dsDNA into another DNA molecule at a specific recognition sequence
lambda attachment site (*att*λ) Recognition site on DNA used during integration of lambda DNA into *E. coli* chromosome
Xis protein Enzyme that reverses DNA integration by removing a segment of dsDNA and resealing the gap leaving behind an intact recognition sequence. Same as excisionase. Not to be confused with Xist RNA involved in X chromosome silencing

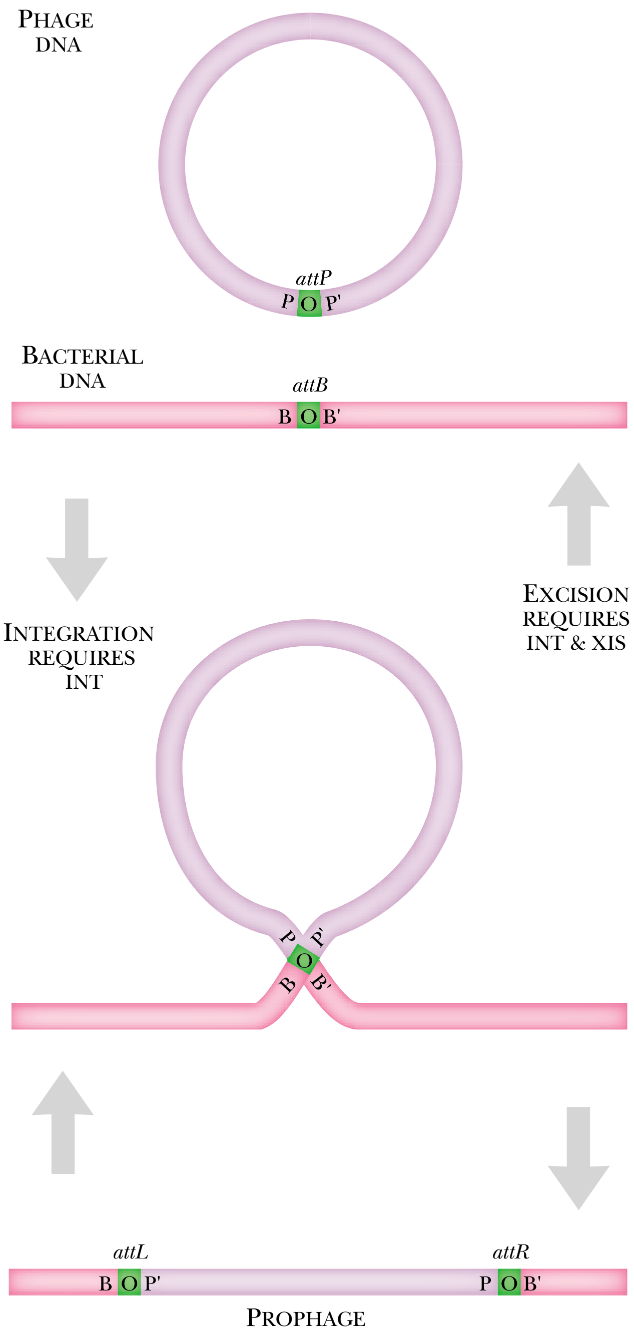


FIGURE 14.08 ***Integration of Lambda DNA—Overview***

Bacterial DNA and λ phage DNA are aligned at the "O" region of the attachment sites. Int protein induces two double-stranded breaks that are resolved, giving a crossover. Since the two recognition sites are altered in their flanking regions, λ cannot be excised by Int alone but needs another protein, known as excisionase or Xis in addition.

("excisionase") in addition to Int. The control of Xis and Int activity determines whether or not λ stays latent in the bacterial chromosome or emerges and replicates. This decision plays a large part in controlling the life cycle of the virus.

Recombination in Higher Organisms

Recombination in eukaryotes occurs mostly during the early stages of meiosis. For crossing over to occur between the pairs of homologous chromosomes, double-stranded breaks must be introduced into them—a hazardous procedure. Double-

excisionase Enzyme that reverses DNA integration by removing a segment of dsDNA and resealing the gap. In particular, lambda excisionase removes integrated lambda DNA

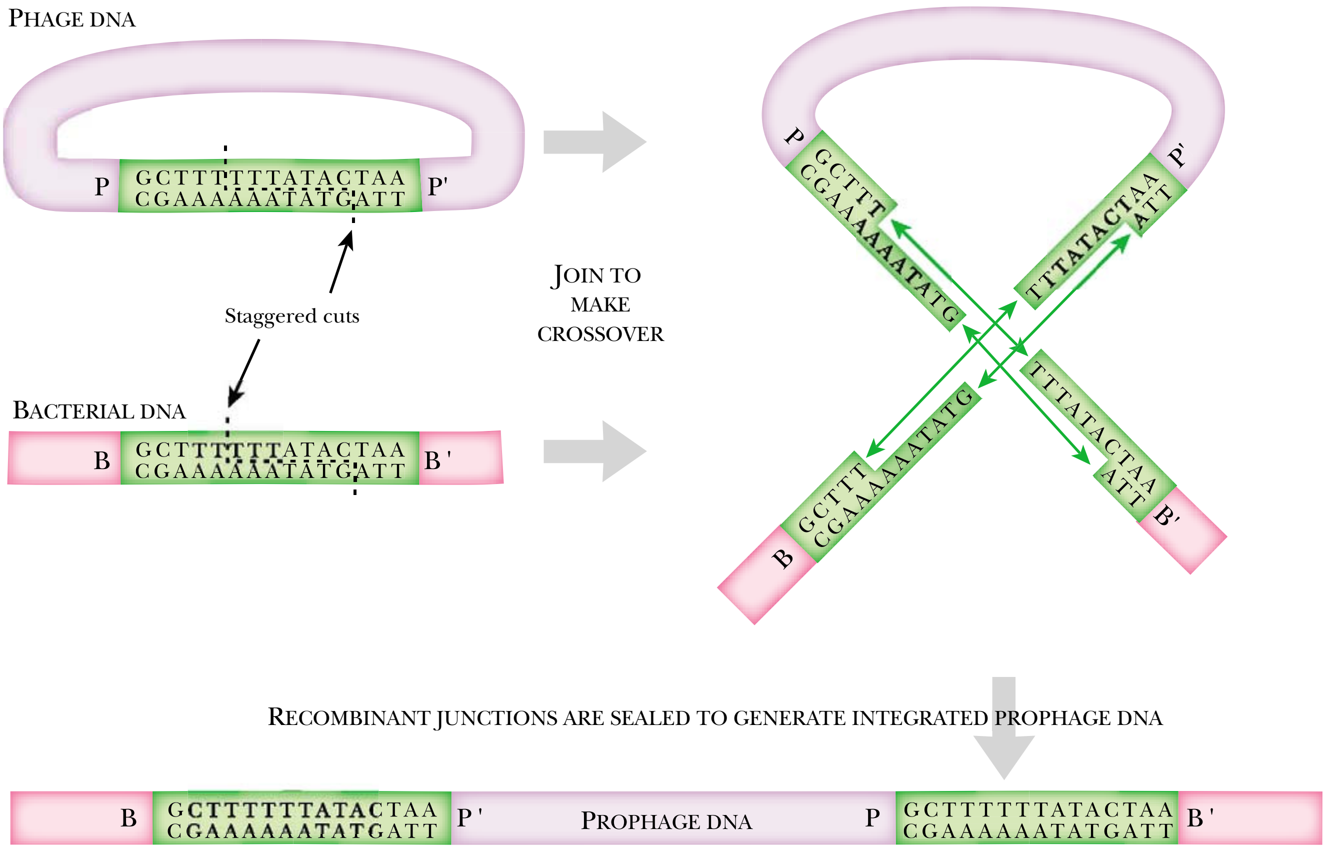


FIGURE 14.09 ***Integration of Lambda DNA—Detail of Crossover***

The "O" region (green) is cut so that an overhang is generated. Rejoining the cut ends of λ with the bacterial chromosome allows λ to integrate its DNA into the host cell.

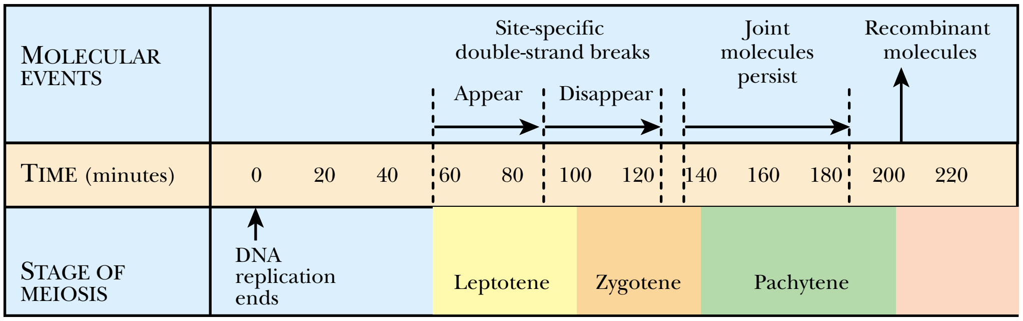


FIGURE 14.10 ***Timeline of Eukaryotic Recombination in Yeast***

Site-specific double-stranded breaks in DNA appear 60–90 minutes after DNA replication. The breaks disappear as hybrid molecules are made during the zygotene phase of meiosis. Resolution of the hybrids then occurs during pachytene. Recombinant molecules appear approximately 120 minutes after the appearance of double-stranded breaks. Therefore, eukaryotic recombination occurs in a span of approximately 2 hours.

stranded breaks appear in eukaryotic chromosomes during the first stage of meiosis, known as leptotene and the paired chromosomes are joined together during the next stage (zygotene) to form the hybrid junction structures needed for recombination (Fig. 14.10). It is assumed that these resemble the Holliday junction of bacteria but the details are obscure. Resolution of the crossovers then occurs during the third stage of meiosis (pachytene). Finally, the crossovers dissociate, releasing recombinant chromosomes in the final stage of meiosis (diplotene).

During meiosis in eukaryotic cells, frequent recombination occurs between pairs of homologous chromosomes.

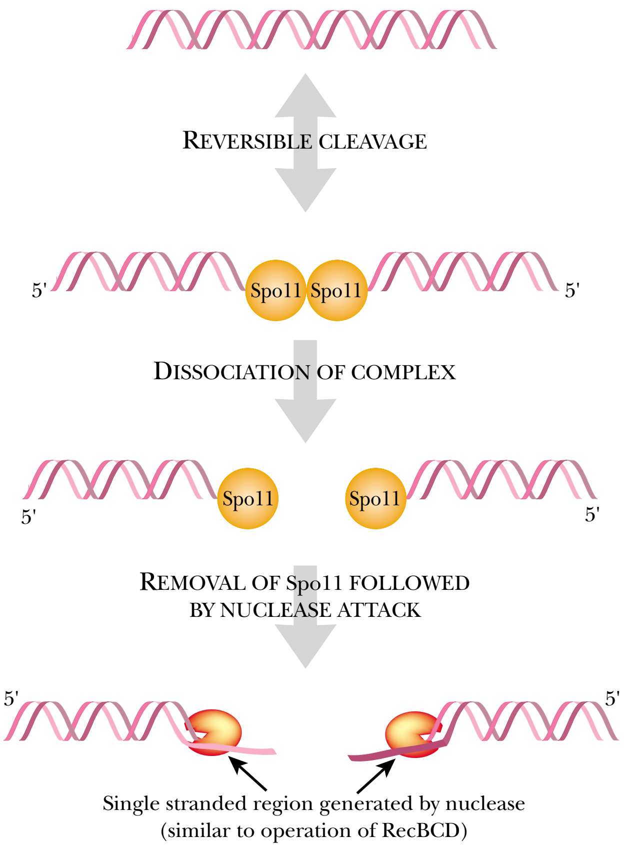

REVERSIBLE CLEAVAGE
5'
Spo11
Spo11
5'
DISSOCIATION OF COMPLEX
Spo11
Spo11
5'
5'
REMOVAL OF Spo11 FOLLOWED BY NUCLEASE ATTACK
5'
5'
Single stranded region generated by nuclease (similar to operation of RecBCD)

FIGURE 14.11 ***Spo11 Promotes Double Strand Breaks***

Spo11 binds and cleaves double-stranded DNA in yeast. After cleaving the DNA, Spo11 is replaced by a nuclease that generates the single-stranded region. The two DNA fragments are held together by other proteins that are not shown here.

A complex of a dozen or more proteins, many poorly characterized, is needed to generate the double-stranded breaks. In yeast, the Spo11 protein is probably responsible for making the double-stranded breaks (Fig. 14.11). Other proteins then displace it, but keep hold of the loose ends of the DNA. Single strands with free 3′-ends are then generated. As in bacteria, this single strand is then used to invade another DNA double helix. In yeast the **Rad proteins** oversee this process. In particular, the Rad51 protein of yeast is homologous to RecA of bacteria and binds to free single strands of DNA forming helical, protein-covered filaments (see Fig. 14.07, above).

Single-strand invasion requires several other Rad proteins, in addition to Rad51. Defects in the *RAD* genes of yeast cause radiation sensitivity as many of them are also involved in repair of radiation damage to DNA (see below)—hence their name. Recombination during mitosis depends on the Rad system alone. However, recombination during meiosis is 100-fold more frequent and needs additional factors. The Dmc1 protein is needed for crossing over between homologous chromosomes. Dmc1 forms rings that surround either single or double-stranded DNA and assist single strands to invade a homologous double-stranded DNA.

Overview of DNA Repair

DNA may be damaged by a variety of agents including chemical mutagens and radiation or by spontaneous reactions of the DNA itself as discussed in the previous chapter. Damage to DNA may cause breaks in the chromosomes or may block replication and so result in the death of the cell. Some types of DNA damage may cause mutations, but other types do not. A mutation may occur if the damage is not repaired

Rad proteins Group of proteins involved in recombination and repair of DNA damage in yeast and animal cells. Rad51 corresponds to the prokaryotic RecA protein

TABLE 14.01 DNA Repair Systems of *Escherichia coli*

Repair System	Genes	Mechanism or Function
Mismatch repair	*dam* *mutSHL*	DNA adenine methylase base mismatch recognition and excision
Nucleotide excision "cut and patch" repair	*uvrABCD*	finds and excises incorrect nucleotides
Guanine oxidation repair	*mutMYT*	removal of oxidized guanine derivatives
Alkylated base repair	*ada* *alkA*	alkyl removal and transcriptional activator alkylpurine removal (glycosylase)
Uracil removal	*ung*	uracil-N-glycosylase removes uracil from DNA
Base excision	*xthA, nfo*	AP endonucleases
Very short patch repair	*dcm* *vsr*	DNA cytosine methylase endonuclease cutting on 5′ side of T in TG mismatch
Photoreactivation	*phr*	photolyase
Recombination repair	*recA* *recBCD* *recFOR*	single strand binding double strand break repair recombination functions
SOS Repair system "error prone repair"	*recA, lexA* *umuDC* *dinB*	regulation of SOS system DNA polymerase V DNA polymerase IV

or if the replication apparatus proceeds through the damaged zone and incorporates incorrect bases.

Even if the DNA has been damaged, all is not lost. Most cells contain a variety of damage control systems and some of these can repair damaged DNA. There are several different DNA repair systems designed to deal with different problems. Some repair systems act in a general manner by looking for overall distortions of DNA structure whereas others focus on specific chemical defects. In some cases, the DNA repair process itself may lead to mutations.

All organisms contain a variety of systems that repair damage to the DNA.

The details of DNA repair are derived mostly from investigation of the bacterium *Escherichia coli*. However, most organisms repair their DNA and possess a variety of repair systems that are probably similar in many respects to those of *E. coli*. Defects in certain DNA repair systems of humans lead to a higher rate of mutation in both somatic and germ line cells. This results in a higher incidence both of heritable defects and of cancer. Table 14.01 lists the known DNA repair systems of *E. coli* together with some of the genes involved. [The synthetic enzymes involved in replacing damaged DNA are not included in the Table if they are also used in normal DNA replication.] Selected repair systems will be discussed in more detail below.

DNA Mismatch Repair System

Some repair systems monitor the DNA double helix for structural defects, rather than looking for any specific chemical error. In *E. coli* there are two of these, the **mismatch repair system** and the **excision repair system** (see below). Both detect structural distortions of the DNA double helix but the mismatch system is more sensitive than the excision repair system. A variety of alterations result in base pairs that don't actually pair properly. If two opposite bases do not match (e.g. G·A) and therefore do not hydrogen bond correctly, a slight bulge will form in the DNA helix. Mismatches of

excision repair system Also known as "cut and patch" repair. A DNA repair system that recognizes bulges in the DNA double helix, removes the damaged strand and replaces it

mismatch repair system DNA repair system that recognizes mispaired bases and cuts out part of the DNA strand containing the wrong base

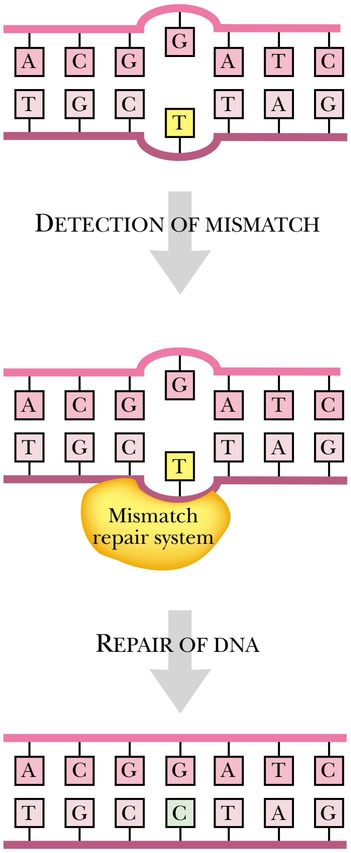


FIGURE 14.12 ***Principle of Mismatch Repair***

Base pairs with incorrect hydrogen bonding cause distortions in the double helix. The mismatch repair system identifies and corrects these distortions.

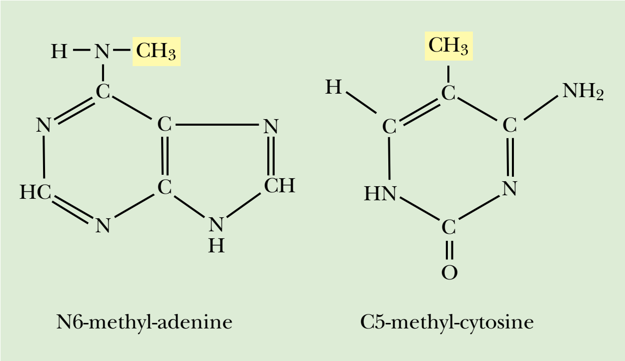


FIGURE 14.13 ***Methylated Bases—Chemical Structure***

Sites for methylation of adenine and cytosine are highlighted.

genuine DNA bases, the presence of base analogs, chemically altered bases and frameshifts all cause distortions that alert the mismatch repair system.

The mismatch repair system cuts out part of the DNA strand containing the wrong base. The gap is then filled in by DNA polymerase III ("PolIII") which hopefully inserts the correct bases to give correctly matched base pairs (Fig. 14.12). The involvement of PolIII is unusual as most repair systems use DNA Polymerase I to replace short damaged regions of DNA.

But how does the cell know which of the two mispaired bases was the wrong one? Since most mismatches arise during DNA replication, the repair systems need to know which strand came from the mother cell and which was the recently synthesized (and error-carrying) daughter strand. In *E. coli*, the chromosome is methylated by two systems that serve to distinguish the new and old strands of DNA. **DNA adenine methylase** (**Dam**) (product of the *dam* gene) converts adenine in the sequence GATC to 6-methyladenine. **DNA cytosine methylase** (**Dcm**) converts cytosine in the sequences CCAGG and CCTGG to 5-methylcytosine. Note that all these recognition sequences are palindromic so that the DNA will be methylated equally on both strands. These methylated bases do not perturb base pairing, as 6-methyladenine and 5-methylcytosine form correct base pairs with T and G respectively (Fig. 14.13). Recall that thymine, which is actually 5-methyluracil, and uracil both base pair identically with adenine.

Immediately after DNA replication, the old strands will be methylated but the new strands of DNA will not. The Dam and Dcm enzymes take just a couple of minutes to methylate the new strands. Until this is done, there is a brief period in which the DNA is said to be **hemi-methylated** (Fig. 14.14). During this period, a variety of repair systems check the DNA, looking for mismatches where the wrong base was inserted. The difference in methylation allows them to tell which is the old, correct, strand and which is the newly made strand. The delay in fully methylating new DNA is also involved in controlling the initiation of new rounds of DNA replication in bacteria, in a manner not fully understood (see Ch. 5). Different groups of bacteria use different recognition sequences, but the principle of distinguishing old and new strands by methylation remains the same.

The major mismatch repair system of *E. coli*, the MutSHL system, uses the methylation of GATC sequences by DNA adenine methylase to monitor which strand is newly made (Fig. 14.15). Many of the genes involved in DNA repair in *E. coli* are named *mut* for "mutator" since defects in these genes result in a higher mutation rate. The mismatch repair system consists of three genes, *mutS, mutH* and *mutL*. The MutS protein recognizes the distortion caused by mismatched base pairs. The MutH protein finds the nearest GATC site and nicks the non-methylated strand. The MutL protein

Mismatch repair corrects mispaired bases.

In bacteria, the parental strand of DNA is identified by methylation.

Newly made DNA is non-methylated and is checked for errors before methylation occurs.

DNA adenine methylase (Dam) A bacterial enzyme that methylates adenine in the sequence GATC
DNA cytosine methylase (Dcm) A bacterial enzyme that methylates cytosine in the sequences CCAGG and CCTGG
hemi-methylated Methylated on only one strand

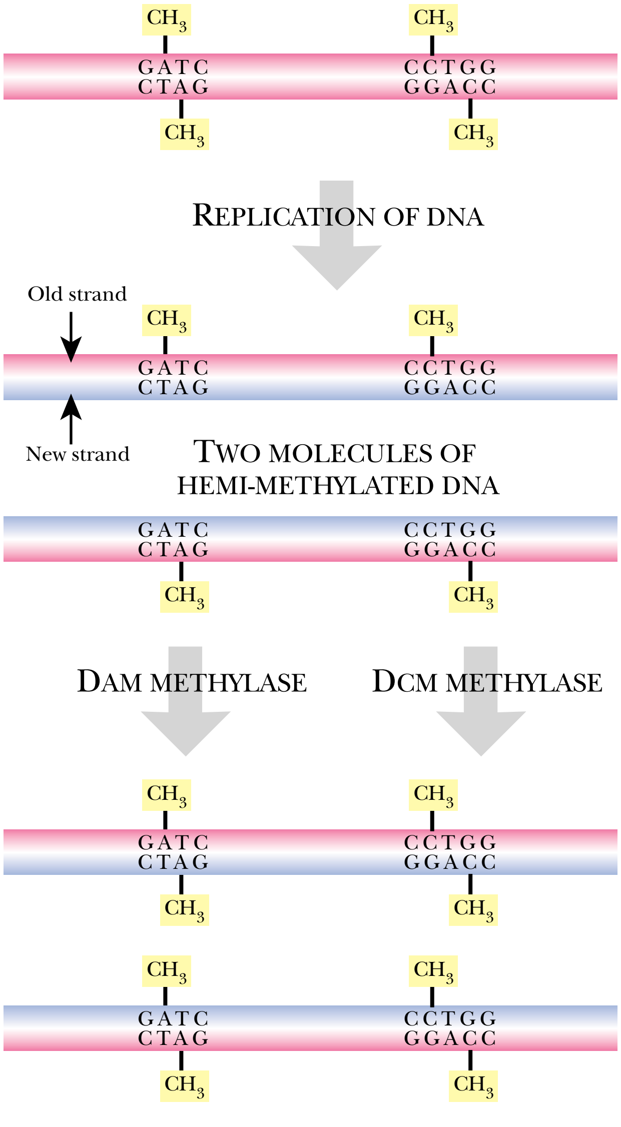


FIGURE 14.14 ***Hemi-methylated DNA: Old Strands Versus New***

When DNA is replicated, the old strand is methylated but the new strand is originally non-methylated and thus the DNA double helix, as a whole, is "hemi-methylated." Dam methylase and Dcm methylase subsequently methylate the new strand at their specific recognition sequences as described in the text.

apparently holds the complex together. The nearest GATC site may be some distance from the mismatch and so it is thought that the DNA is pulled through the MutSHL complex forming a loop until a GATC site is reached (Fig. 14.15). PolIII then attaches and repairs the gap created by the MutSHL system.

General Excision Repair System

Cut and patch repair recognizes distortions of the DNA double helix.

The most widely distributed system for dealing with damaged DNA is excision repair, often referred to as "cut and patch" repair. This system recognizes bulges in the DNA double helix, but is not as sensitive as the mismatch system. The excision repair system does not detect mismatches, base analogs or certain methylated bases that cause only slight distortions. It does repair most damage due to UV radiation such as thymine dimers and other cross-linked products. Defects in the genes for excision repair result in lowered resistance to UV light and are therefore named *uvr* (UV resistance).

The UvrAB excision repair system is similar to the mismatch system. The UvrAB complex cruises the DNA looking for bulges. When it finds a defect, UvrA departs and is replaced by UvrC protein. UvrB nicks the DNA to the 3′-side of the damage and UvrC nicks to the 5′-side (Fig. 14.16). UvrD is a helicase that unwinds the single-stranded region that has been cut out by UvrBC. Next DNA polymerase I (PolI) fills in the gap with a new strand of DNA. Note that PolI possesses both a polymerase

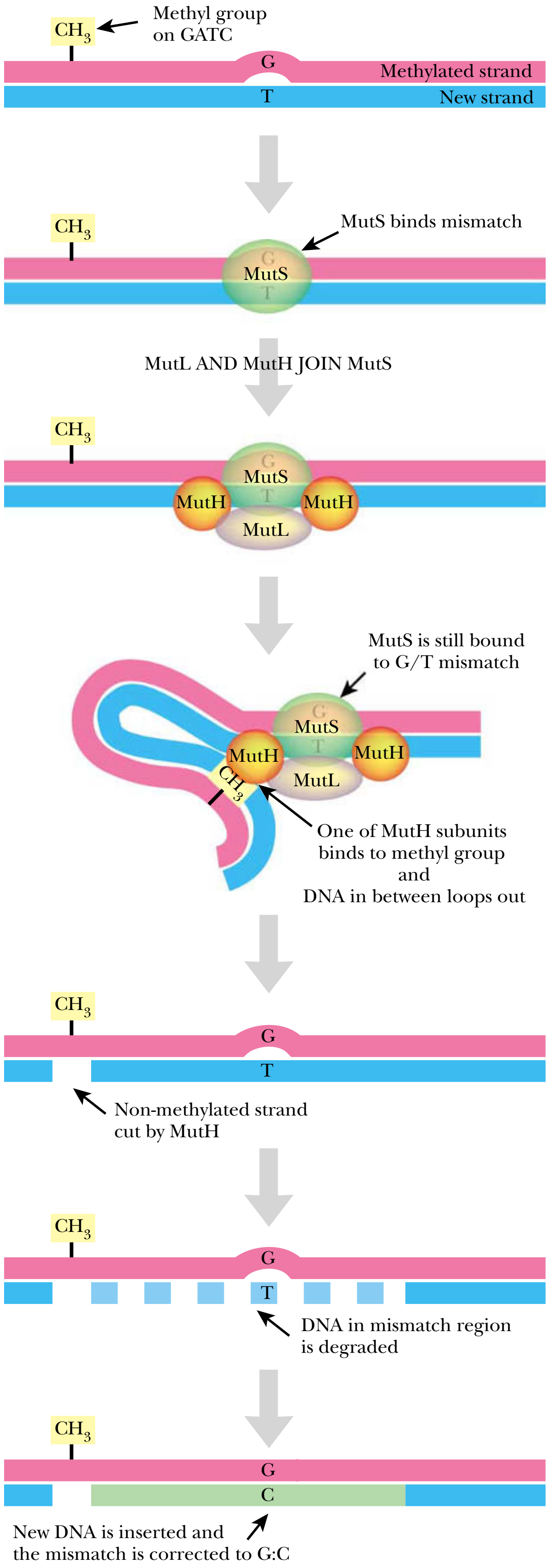

CH_3
Methyl group on GATC
G
T
Methylated strand
New strand
CH_3
MutS binds mismatch
G
MutS
T
MutL AND MutH JOIN MutS
CH_3
G
MutS
T
MutH
MutH
MutL
MutS is still bound to G/T mismatch
G
MutS
T
MutH
MutH
MutL
CH_3
One of MutH subunits binds to methyl group and DNA in between loops out
CH_3
G
T
Non-methylated strand cut by MutH
CH_3
G
T
DNA in mismatch region is degraded
CH_3
G
C
New DNA is inserted and the mismatch is corrected to G:C

FIGURE 14.15 ***The MutSHL Mismatch Repair System***

MutS recognizes a mismatch shortly after DNA replication. MutS recruits MutL and two MutH proteins to the mismatch. MutH locates the nearest GATC of the new strand by identifying the methyl group attached to the "mother" strand. MutH cleaves the non-methylated strand and the DNA between the cut and the mismatch is degraded. The region is replaced and the mismatch is corrected.

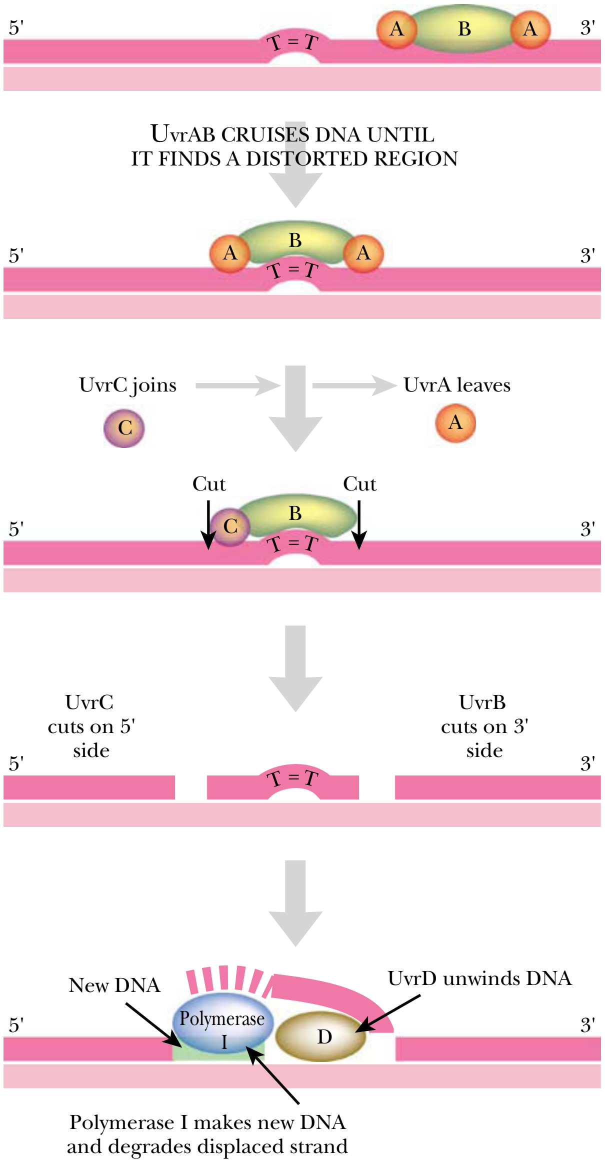


FIGURE 14.16 ***The UvrABC Excision Repair System***

A cross-linked thymine in the top (dark pink) strand is recognized by the UvrAB complex. UvrA protein is replaced with UvrC, which cuts the defective strand just before the thymine dimer (to the 5′ side). UvrB nicks the strand downstream of the thymine dimer (3′ side). UvrD unwinds the damaged strand, which is degraded and replaced by DNA polymerase I.

activity and a 5′-exonuclease activity that allows it to remove the old strand of DNA in front of it. Thus as it moves along synthesizing the new strand, PolI also nibbles away the old strand. Finally, the nicks are closed by DNA ligase.

DNA Repair by Excision of Specific Bases

Bases that do not occur naturally are removed from DNA by a variety of specific repair systems.

In contrast to the general repair systems that recognize distortions in the DNA double helix, there are a variety of repair systems that recognize specific chemical changes in DNA. In particular, methylation and deamination give rise to unnatural bases that are not normally found in DNA (see Ch. 13). In such cases, there is no question which member of a mismatched base pair is wrong. Obviously the non-DNA base should be removed and a variety of enzymes exist that do just this.

Deamination generates hypoxanthine (from adenine), xanthine (from guanine) and uracil from cytosine (see Ch. 13). These three bases are all removed by **DNA glycosylases** that break the bond between the base and the deoxyribose sugar of

DNA glycosylase Enzyme that breaks the bond between a base and the deoxyribose of the DNA backbone

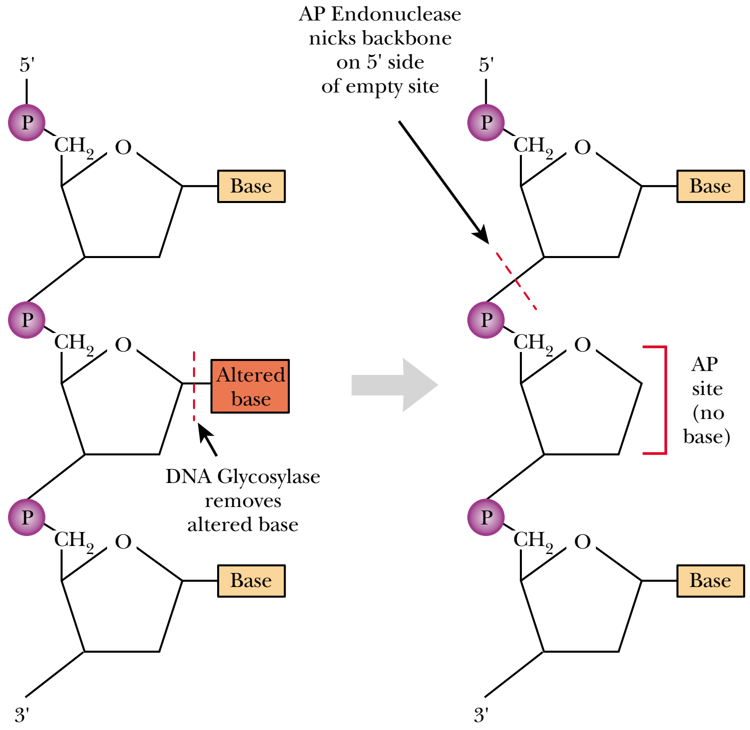


FIGURE 14.17 ***Removal of Unnatural Bases***

An altered base (such as hypoxanthine) is removed by DNA glycosylase leaving an empty, AP site. The AP site is recognized by AP endonuclease, which nicks at the 5′ side of the sugar backbone. This leaves a free 3′–OH on the base pair just upstream (above) the AP site. PolI recognizes the 3′–OH and replaces a stretch of single-stranded DNA that includes the AP site.

the DNA backbone. Specific DNA glycosylases exist for each unnatural base. Thus **uracil-N-glycosylase**, **Ung protein**, removes uracil from DNA. Some methylated derivatives, such as 3-methyl adenine and 3-methyl guanine are removed in the same manner.

Removal of bases leaves an empty space in the DNA known as an **AP-site**. AP stands either for apurinic or apyrimidinic depending on which type of base was removed to create the AP-site (Fig. 14.17). Next an **AP endonuclease** cuts the backbone of the DNA next to the missing base leaving a free 3′-OH group. DNA polymerase I then makes a short new piece of DNA starting with the free 3′-OH group. As PolI moves along it degrades the single strand in front of it by its 5′-exonuclease activity. As usual, the final nick is sealed by DNA ligase.

Some unnatural bases are the result of oxidation. 8-Oxoguanine is especially prevalent and will sometimes mispair with adenine leading to an 8oxoG·A base pair (see Ch. 13) instead of a G·C base pair. A specific DNA glycosylase, the MutM protein, removes 8-oxoguanine from DNA. In addition, the MutY protein removes adenine when (and only when) it is found opposite 8-oxoguanine (Fig. 14.18). In both cases removal generates an AP-site that is processed as described above. Finally, 8-oxoguanine may arise in the nucleotide pool used to synthesize DNA. To repair the defective nucleotides, the MutT protein finds the 8-oxoguanine derivative of GTP and cleaves off two phosphate groups. This prevents the incorporation of pre-formed 8-oxoguanine into DNA.

Specialized DNA Repair Mechanisms

Several more exotic systems exist to deal with special cases. Deamination of 5-methylcytosine (which pairs with G) produces thymine causing a T·G mismatch. Since thymine is a naturally occurring DNA base, the mismatch repair system would only

AP endonuclease Endonuclease that nicks DNA next to an AP-site
AP-site A site in DNA where a base is missing (AP-site = apurinic site or apyrimidinic site depending on the nature of the missing base)
Ung protein Same as uracil-N-glycosylase
uracil-N-glycosylase Enzyme that removes uracil from DNA

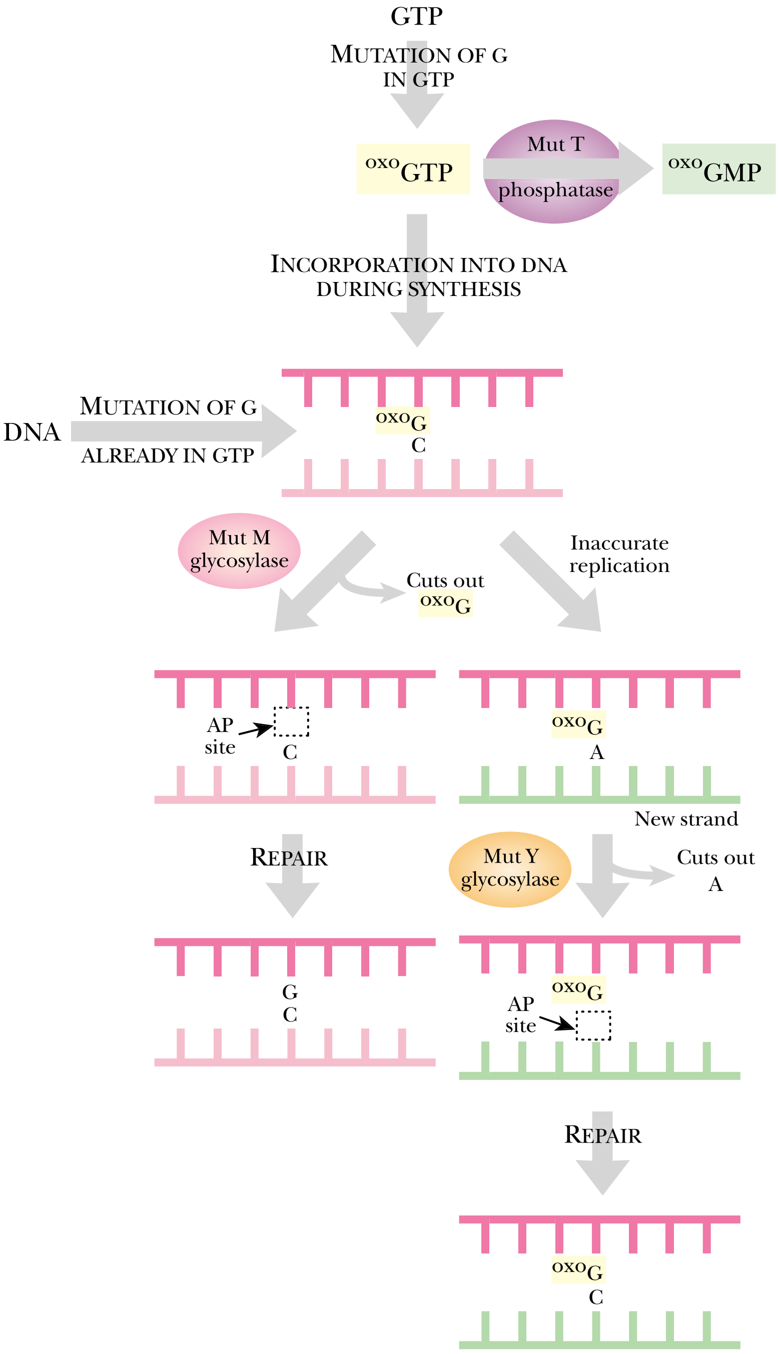


FIGURE 14.18 ***Dealing with Oxidized Guanine***

Three different methods of dealing with 8-oxoG exist. If the 8-oxoG is detected in the nucleotide precursors for DNA synthesis (oxoGTP), MutT phophatase dephosphorylates the precursor. The mono-phosphate (oxoGMP) cannot be incorporated into DNA. If 8-oxoG is already incorporated and is correctly paired with C, it is cut out by MutM glycosylase leaving an AP site. AP endonuclease and PolI then repair the AP site. If the 8-oxoG is incorporated in DNA but is incorrectly mismatched with A, then MutY glycosylase removes the A and this AP site is repaired as above. The resulting 8-oxoG·C base pair can be repaired by the MutM glycosylase.

catch a T·G mismatch if generated as a replication error and found in newly made DNA. However, deamination occurs spontaneously at any time and rarely during replication. Consequently, deamination of 5-methylcytosine often goes un-repaired and the presence of 5-methylcytosine leads to mutational hot-spots as discussed in Chapter 13.

Thymine derived from methyl cytosine is removed by the "very short patch repair" system.

Nonetheless, most of the 5-methylcytosine in *E. coli* is made by Dcm methylase and is found in the sequences CCAGG and CCTGG. Whenever T is found replacing C and therefore paired with G in one of these sequences, it is removed. A specific endonuclease nicks the DNA next to the T of the T·G mismatch. This system is sometimes known as "very short patch repair" and the nicking enzyme as Vsr endonuclease. DNA polymerase I then removes a short length of the strand with the incorrect T and replaces it with a new piece of DNA. Although methylating the cytosine in

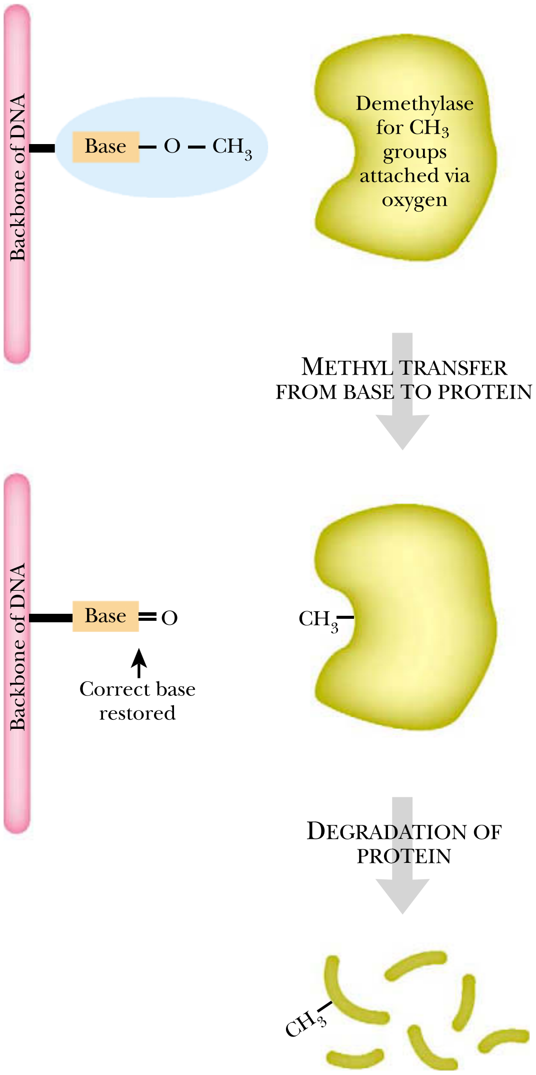


FIGURE 14.19 ***Suicide Demethylase for O-Methyl Bases***

Demethylase protein recognizes CH_3 attached via oxygen (O) to either guanine or thymine. The demethylase transfers the CH_3 to itself then degrades itself. The correct base structure is restored by removal of the methyl group without any other modification.

CCAGG/CCTGG is unique to *E. coli*, other organisms also make 5-methylcytosine and presumably contain analogous repair systems with different sequence specificities.

Although some methylated bases are removed by DNA glycosylases, those with the methyl group attached to oxygen need special treatment. The methyl groups of O^6-methyl guanine or O^4-methyl thymine are removed by suicidal proteins that transfer the methyl group from the base to themselves (Fig. 14.19). This leaves behind the correct base, with no need for synthesizing a new stretch of DNA, but inactivates the protein, which is then degraded. Such single-use proteins are not true enzymes, as they do not catalyze a reaction that occurs multiple times. [Certain restriction enzymes also act in a similar manner, inactivating themselves after one reaction (see Ch. 22).]

Ada protein not only removes methyl groups from both the bases and phosphate backbone of DNA but also acts as a transcriptional activator.

The Ada ("adaptation to alkylation") protein of *E. coli* is especially interesting since it has both C-terminal and N-terminal active sites, both used in removing methyl groups from DNA. When methyl groups from methylated bases are attached near the C-terminus of Ada, it is inactivated. However within the cell, alkylating agents attack the phosphate groups of the DNA backbone as well as individual bases. Ada protein will also remove methyl groups from the phosphate backbone. In this case, the methyl group is transferred to a site near the front (N-terminus) of Ada protein. This turns Ada into a transcriptional activator that increases expression of several genes involved in combating DNA damage by alkylation (Fig. 14.20).

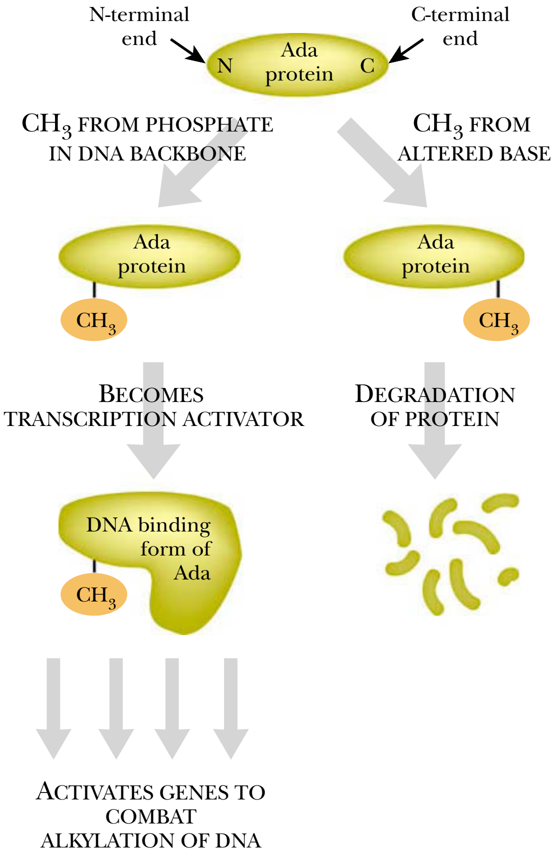


FIGURE 14.20 ***Ada Plays a Dual Role in Removing Alkyl Groups***

Ada protein has two roles in DNA repair. Ada is a classic demethylase that accepts CH_3 from an altered base then degrades itself. Ada can also accept CH_3 from the phosphate backbone of the DNA. The CH_3 is attached to the N-terminal domain of Ada and transforms Ada into a transcriptional activator. This form of Ada turns on genes used to combat DNA alkylation.

Photoreactivation Cleaves Thymine Dimers

Light energy can be used to split apart thymine dimers.

So far we have considered repair systems that are specific for a single altered base. UV light generates pyrimidine dimers, in particular thymine dimers, as well as some other types of damage. Although the Uvr excision repair system will remove pyrimidine dimers due to their distortion of the DNA double helix, a specific repair system also exists. The photoreactivation repair system identifies pyrimidine dimers and cleaves them, thus regenerating the original bases directly (Fig. 14.21). No DNA synthesis is involved. Photo-reactivation only occurs when visible light is present. A single protein is responsible, an enzyme known as photolyase, which absorbs blue light (350–500 nm) and uses the energy to drive the cleavage reaction.

Curiously, photolyase even helps remove pyrimidine dimers in the dark! Under these conditions it binds to the dimer but lacks the energy to remove the crosslink. Nonetheless, binding of photolyase alerts the excision repair system to replace the damaged DNA. Photoreactivation was the first DNA repair system discovered and is present in almost all organisms—except placental mammals, including humans.

Transcriptional Coupling of Repair

If the template strand of DNA is damaged in a gene that is being transcribed the RNA polymerase may grind to a halt. Such damage is repaired preferentially by what is known as **transcription-coupled repair**. If the mutation is found in the non-template strand, the mutation is less likely to be repaired. This implies that repairing the template strand is of higher priority. In bacteria the transcription-repair coupling factor,

transcription-coupled repair Preferential repair of the template strand of DNA that may be transcribed

TRCF, detects a stalled RNA polymerase and directs the UvrAB proteins to the site of the block. It is thought that TRCF detaches the RNA polymerase so it does not hinder the UvrAB repair proteins. The excision repair system then mends the damage as described above.

In eukaryotes, transcription-coupled repair is especially important since coding regions are relatively few and far between. The eukaryotic transcription factor, TFIIH (see Ch. 6)-possesses helicase activity and helps to unwind the DNA when transcription begins. If the DNA is distorted due to chemical damage, TFIIH is thought to recruit the eukaryotic excision repair system. This operates similarly to its bacterial counterpart, except in eukaryotes the damaged single strand is cut out only after the DNA is unwound to produce a bubble (Fig. 14.22). Two nicks are made, at the junctions between the double-stranded and single-stranded DNA.

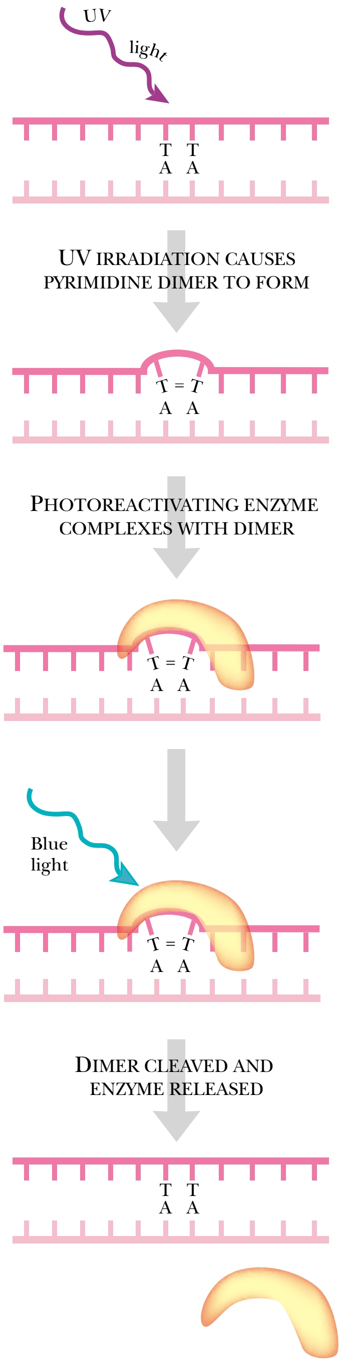

UV light
T T
A A
UV IRRADIATION CAUSES PYRIMIDINE DIMER TO FORM
T = T
A A
PHOTOREACTIVATING ENZYME COMPLEXES WITH DIMER
T = T
A A
Blue light
T = T
A A
DIMER CLEAVED AND ENZYME RELEASED
T T
A A

FIGURE 14.21 ***Photo-Reactivation Cleaves Pyrimidine Dimers***

UV light crosslinks two thymines found next to each other on the DNA molecule. The distortion in the helix is recognized by photolyase. Light activates photolyase to remove the crosslink, restoring the thymines to their original conformation.

Repair by Recombination

Although repairing all damage to DNA is the ideal, in practice cells may have to settle for less. Some cases of DNA damage prevent replication and if left unattended would kill the cell. In such a predicament, avoiding possible future mutation is a secondary issue. For example, thymine dimers cause the replication apparatus to stall. Replication is then re-initiated beyond the blockage. However, this leaves a single-stranded region that was not replicated. If unfilled, this gap would cause a break in the chromosome next time a replication fork passed through the region. Although a gap is left in one strand, the replication enzymes still synthesized a new strand of DNA on the undamaged mother strand. Thus there are two copies of the damaged region, one complete and the other single-stranded with the blockage still in position (Fig. 14.23). To fill this gap, the bacterial RecA protein carries out recombination between the single-stranded region and the complete double-stranded copy. RecA protein binds to the single-strand and then forms a triplex structure with the corresponding double-stranded region. Recombination fills the gap in the defective strand, at the cost of leaving a gap in the other, undamaged version of this region. However, since this gap is in an undamaged strand of DNA, it can be filled by DNA polymerase (Fig. 14.23).

The net result of this process is that replication has occurred. Although the old damaged DNA strand remains un-repaired, the newly made DNA molecule is correct. This process of circumventing a blockage may occur during multiple rounds of replication and eventually only one descendent cell out of thousands will have a defective strand of DNA. Higher organisms contain analogous systems and can even use recombination to repair double-stranded breaks (see below). In the case of damage that is difficult or impossible to repair, such damage limitation systems are extremely useful.

SOS Error Prone Repair in Bacteria

Damage to DNA results in the induction of a variety of genes whose products help to minimize the effects of the damage. Some of these are repair enzymes and others delay cell division until the damage has been repaired. Yet others provide a pathway of last resort—they allow DNA replication to proceed through severely damaged zones, even at the cost of introducing mutations, a process known as **error-prone repair**.

The **SOS system** of *E. coli* was so named because it responds to severe and potentially lethal DNA damage. Many single-stranded regions of DNA are generated by damage, partly due to excision repair and to stalled replication reinitiating beyond the damage. Since single-stranded DNA activates RecA, severe DNA damage activates

error-prone repair Type of DNA repair process that introduces mutations
SOS system An error-prone repair system of bacteria that responds to severe DNA damage

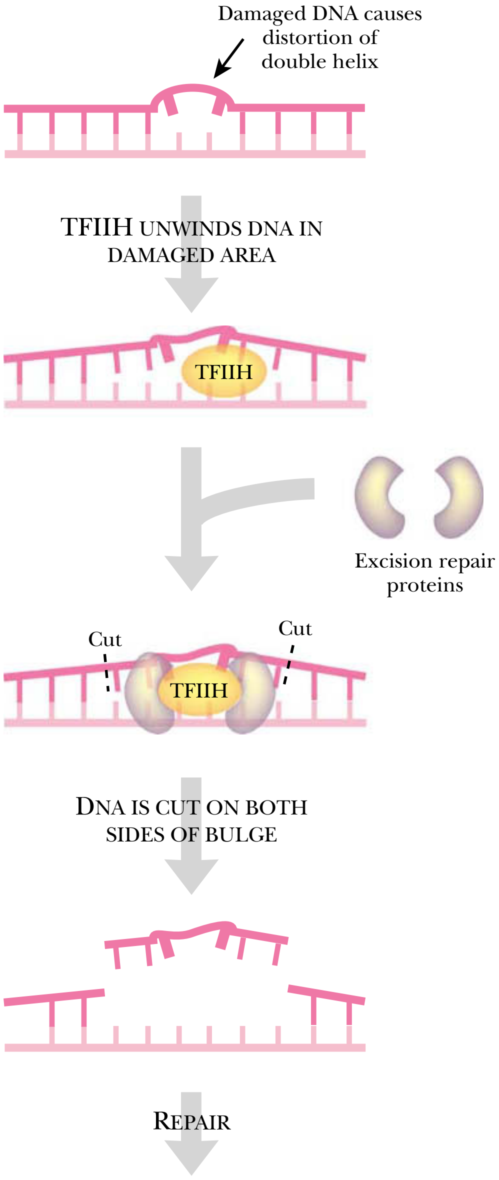

Damaged DNA causes distortion of double helix
TFIIH UNWINDS DNA IN DAMAGED AREA
TFIIH
Excision repair proteins
Cut
Cut
TFIIH
DNA IS CUT ON BOTH SIDES OF BULGE
REPAIR

FIGURE 14.22 ***Eukaryotic Transcription-Coupled Excision Repair***

Distortion of eukaryotic DNA is recognized by TFIIH as it unwinds the DNA during transcription. The excision repair system recognizes TFIIH and nicks the unwound region of DNA, releasing the damaged strand. DNA polymerase repairs the gap and transcription can resume.

The SOS repair system tackles severely damaged DNA, especially regions where the DNA has become single-stranded.

many RecA proteins. Activated RecA induces the SOS system by activating LexA, a transcriptional repressor of the SOS genes, by inducing its self-cleavage. Once cleaved, LexA no longer blocks transcription of SOS genes. The SOS proteins combat the DNA damage (Fig. 14.24).

Among the genes induced by the SOS response two are of special note. These are *umuC* and *umuD* (umu = ultraviolet mutagenesis), which encode **DNA polymerase V**. This polymerase lacks a proofreading subunit so it can replicate past pyrimidine dimers and missing bases (i.e. AP-sites). PolV makes mistakes when passing damaged DNA and, for example, tends to put in GA (rather than the correct AA) opposite a thymine dimer. Normal DNA polymerase (PolIII) cannot replicate past such damage because its proofreading subunit stops it from proceeding until a correct base pair has been inserted. If this is impossible due to damage, then PolIII grinds to a halt and PolV takes over. The PolV subunits, UmuC and UmuD, form a complex containing a dimer of UmuD plus one UmuC protein. However, when first made, UmuD_2C does not act as a polymerase but delays normal DNA replication in order to allow time for repair. Activated RecA then induces UmuD to undergo self-cleavage to UmuD’. When a

Error prone repair fills in dangerous gaps in the DNA at the cost of generating base changes, i.e., point mutations.

DNA polymerase V A repair polymerase in bacteria that can replicate past pyrimidine dimers and AP-sites

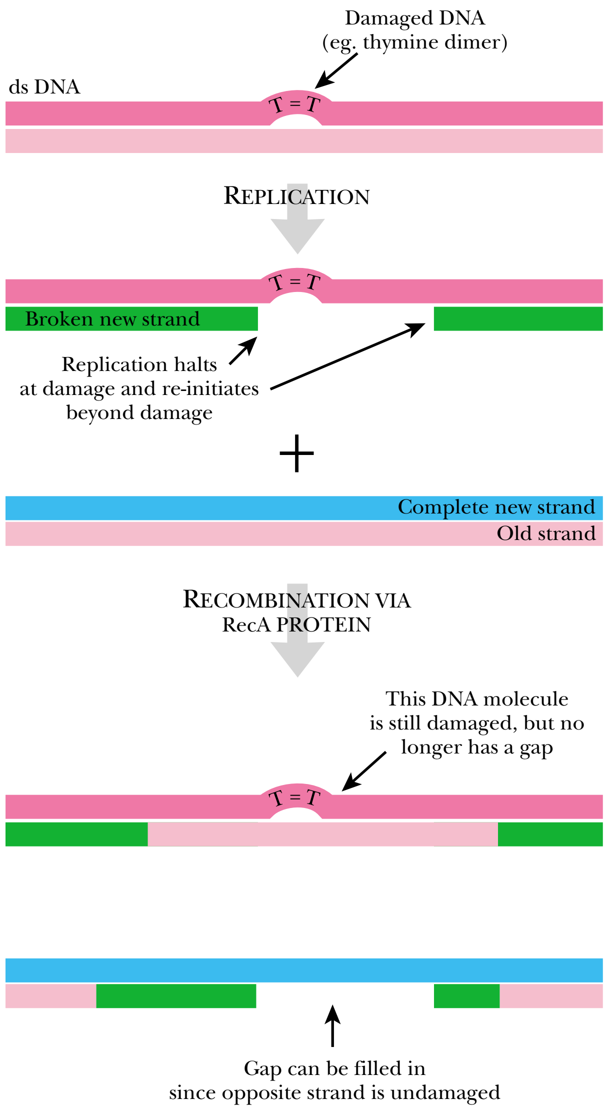


FIGURE 14.23 ***RecA and Recombination Repair***

Since DNA damage stalls the replication enzymes, thymine dimers create large single-stranded gaps (green strand). The other strand (light pink) can be replicated as usual. To repair the gap, RecA directs crossovers on both sides of the single-stranded region. This transfers the gap to another DNA molecule where the correct complementary strand is present. The gap can now be filled in by DNA polymerase. Notice that the thymine dimer is not repaired during this process.

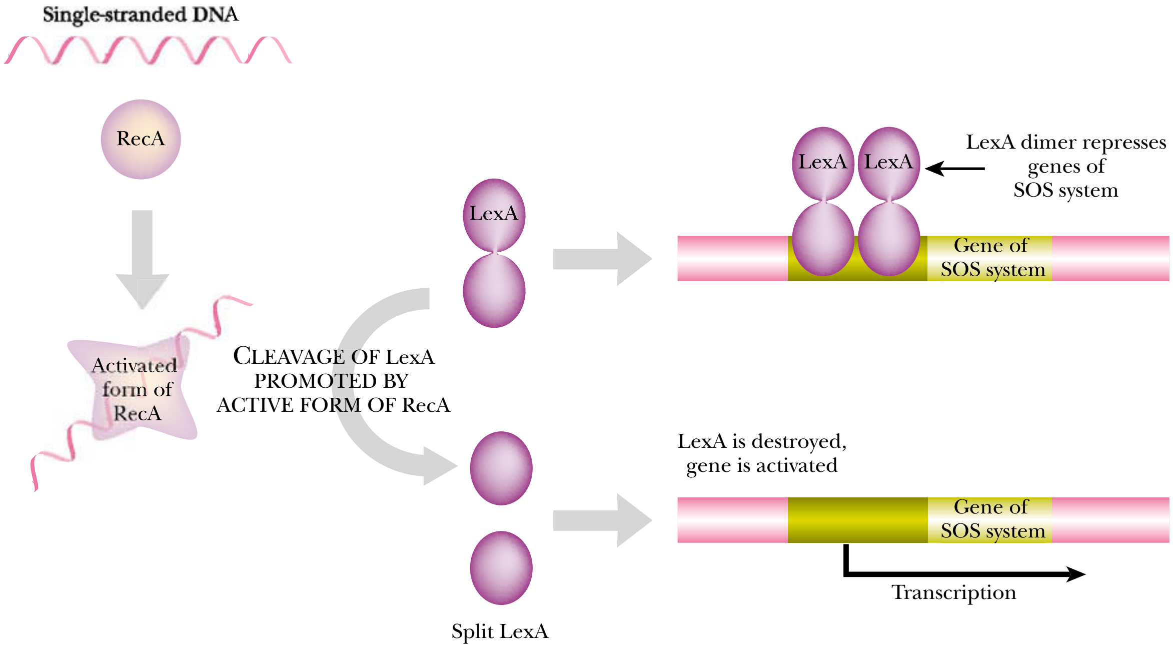


FIGURE 14.24 ***RecA and LexA Control the SOS System***

Binding of RecA to single-stranded DNA activates RecA so that it cleaves LexA protein. Cleaved LexA protein can no longer bind to DNA and is released. The genes of the SOS system are no longer blocked from transcription. The SOS gene products combat DNA damage.

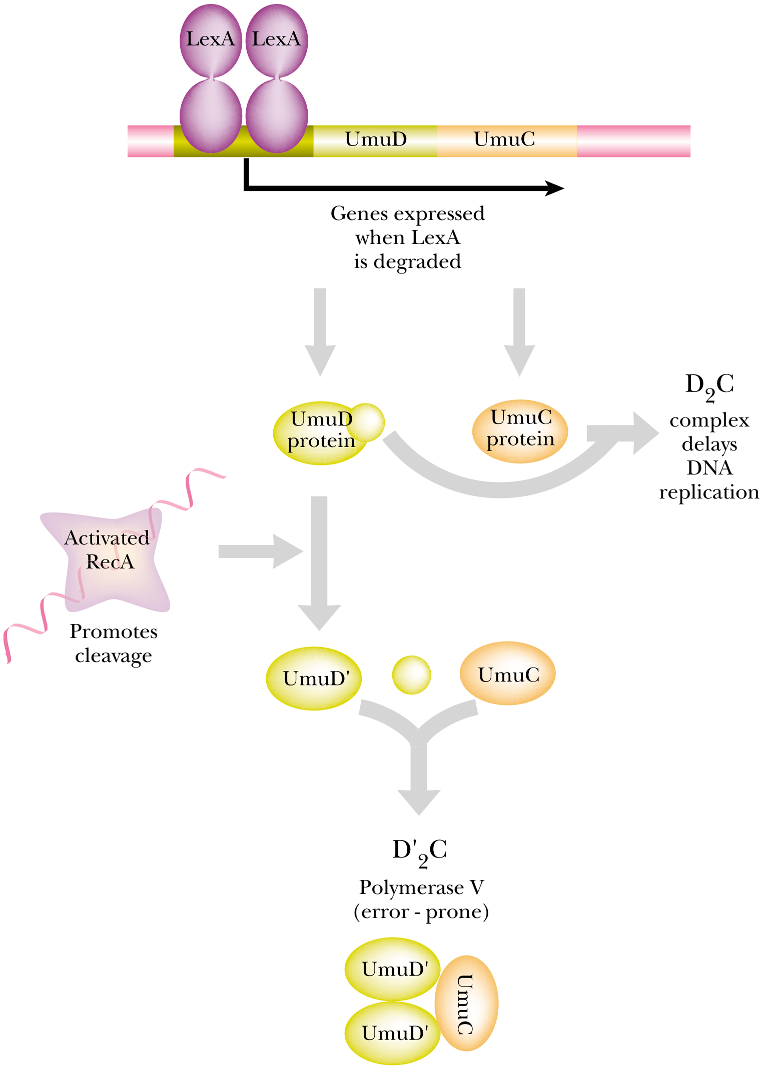

LexA
LexA
UmuD
UmuC
Genes expressed when LexA is degraded
UmuD protein
UmuC protein
D_2C complex delays DNA replication
Activated RecA
Promotes cleavage
UmuD'
UmuC
D'_2C
Polymerase V (error - prone)
UmuD'
UmuD'
UmuC

FIGURE 14.25 ***DNA Polymerase V is Part of the SOS System***

Activated RecA cleaves LexA dimers, allowing expression of UmuC and UmuD. Two UmuD combine with one UmuC protein and this complex slows replication. DNA repair mechanisms then have time to repair some damage. If the damage is extensive, activated RecA (bound to single-stranded DNA) cleaves UmuD to form UmuD′. When two UmuD′ and one UmuC proteins combine, PolV is formed and replicates past any unrepaired damage.

dimer of UmuD’ combines with one UmuC, the error-prone PolV (UmuD’$_2$C) is formed (Fig. 14.25).

Like *E. coli*, yeast, flies and humans all have error-prone DNA polymerases that respond to DNA damage and can replicate past damaged regions. In higher organisms these repair enzymes appear to be more specialized and less error-prone. For example, when human **Polymerase Eta** passes a symmetrical thymine dimer it usually puts in AA. However, although more accurate, it cannot pass other types of pyrimidine dimers like *E. coli* PolV. Although this results in a mutation, it is better than complete failure to replicate the DNA.

Repair in Eukaryotes

Many of the repair systems described for bacteria have counterparts in animals. However, these are usually less well characterized and much of our knowledge comes from finding eukaryotic proteins that are homologous to bacterial DNA repair enzymes. Defects in human DNA repair systems cause assorted health problems. In particular, the higher mutation rates that occur in the absence of DNA repair cause a higher frequency of various forms of cancer.

polymerase eta (η) A repair DNA polymerase in animals that can replicate past thymine dimers

For example, the human *hMSH2* (human MutS homologue 2) gene encodes a protein very similar to the MutS protein of *E. coli*. Defects in this gene greatly increase the likelihood of several types of cancer. Such patients have a relatively high frequency of short deletions and insertions that would normally be corrected by the mismatch repair system (due to MutSLH in *E. coli*). Curiously, when the normal human *hMSH2* gene is cloned and expressed in *E. coli* it increases the bacterial mutation frequency! This is apparently due to interference between hMSH2 and MutS. Patients with a defective *BRCA1* (breast cancer A1) gene are more susceptible to breast and ovarian cancer. The BRCA1 protein is involved in both the mending of double-strand breaks and transcription-coupled excision repair. These processes are defective when the *BRCA1* gene is damaged. The role of mutation and DNA repair in cancer is extremely complex and many details are still obscure.

The recessive hereditary disorder xeroderma pigmentosum is due to a failure of the excision repair system that removes thymine dimers and other bulky base adducts. Defects in any of approximately ten genes involved in excision repair will give xeroderma pigmentosum. The result is hypersensitivity of the skin to sunlight and ultraviolet radiation.

Double-Strand Repair in Eukaryotes

Eukaryotic cells can mend double-stranded breaks in the DNA.

Double-strand breaks may be caused by ionizing radiation and certain chemical mutagens. They may also be left behind when some transposable elements excise themselves and move (see Ch. 15). Both yeast and mammalian cells have a similar system for repairing double-strand breaks by a process known as **non-homologous end joining** (Fig. 14.26). First the two Ku proteins bind, one on either side of the break. DNA-dependent protein kinase (DNA-PK), which is attached to the Ku complex, then activates the XRCC4 protein, which in turn directs DNA ligase IV to repair the break.

Occasionally the non-homologous end joining system joins together lengths of DNA that were never previously attached. This can be viewed as a form of non-homologous recombination and generates new combinations of genetic material, including rarities such as chromosomal rearrangements.

In addition, double-strand breaks may be repaired by homologous recombination. This makes use of the presence of pairs of chromosomes in eukaryotic cells. Here, the corresponding sequence on the sister chromosome is used to repair its damaged partner. This process involves a complex that includes Rad51 and other Rad proteins.

Gene Conversion

Sometimes one allele of a gene is converted into another by mismatch repair.

Generally, crossing over is expected to be symmetrical and different alleles from two different parents will be inherited according to Mendel's laws (see Ch. 1). In other words, when two parents reproduce sexually, different alleles from each parent should appear with equal frequency in the offspring. Occasional exceptions to this occur by a mechanism known as **gene conversion**. The name refers to the fact that one allele is converted to the other. The mechanism involves the mismatch repair system operating upon the intermediate structures generated by recombination.

Under normal circumstances, the two strands of a DNA molecule are complementary in sequence and therefore represent the same genetic information. Consequently, any particular double helical DNA molecule carries a single allele of a gene.

gene conversion Recombination and repair of DNA during meiosis that leads to replacement of one allele by another. This may result in a non-Mendelian ratio among the progeny of a genetic cross

non-homologous end joining DNA repair system found in eukaryotes that mends double-stranded breaks

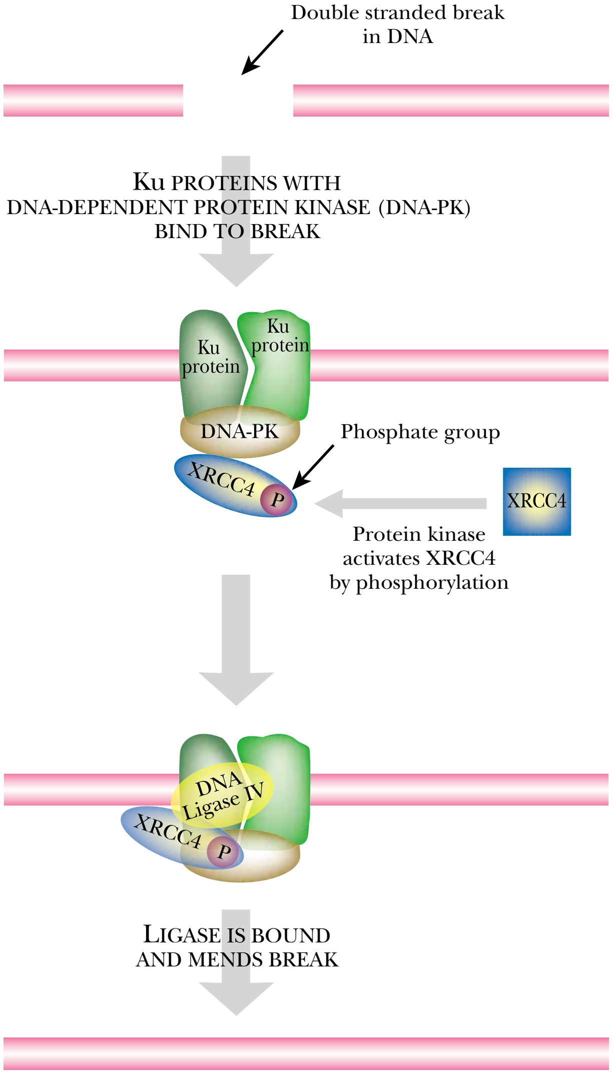

Double stranded break in DNA
Ku PROTEINS WITH DNA-DEPENDENT PROTEIN KINASE (DNA-PK) BIND TO BREAK
Ku protein
Ku protein
DNA-PK
Phosphate group
XRCC4
P
XRCC4
Protein kinase activates XRCC4 by phosphorylation
DNA Ligase IV
XRCC4
P
LIGASE IS BOUND AND MENDS BREAK

FIGURE 14.26 ***Non-Homologous End Joining in Mammals***

Double-stranded breaks are recognized by the Ku proteins, which bind one to each end. The two Ku proteins recruit DNA-PK to the complex. DNA-PK phosphorylates XRCC4 protein, which then recruits DNA ligase IV to join the two broken ends.

However, if the two strands of a DNA molecule do not base pair completely then, strictly speaking, each strand represents a different allele. Such heteroduplex DNA only exists transiently and will soon be corrected by the mismatch repair system; nonetheless, its existence provides the opportunity for gene conversion.

During recombination short heteroduplex regions are created in the DNA next to the crossover point, as discussed above. [This is true whether crossing over generates hybrid DNA molecules or whether the "original" DNA molecules are regenerated from the Holliday junction as shown in Fig. 14.04 above.] Consider a crossover between two DNA molecules carrying different alleles of the same gene. If the crossover occurs within the coding sequence of the gene of interest, then heteroduplexes will be formed in which the two strands represent the two alleles of the gene (Fig. 14.27). Two alternative possibilities now exist. Replication may occur immediately, in which case the four daughter molecules of DNA produced will show a normal Mendelian ratio (a 3 : 1 ratio in the example illustrated in Fig. 14.27). Alternatively, the mismatch repair system may correct the mismatched base pairs in the heteroduplex region before replication. In this case, one strand is altered to match the other which, in effect, converts one allele into the other allele (in the figure R is converted to r) and the ratio of progeny is changed (from 3 : 1 to 4 : 0 in the example in Fig. 14.27).

Gene conversion may be widespread but is difficult to detect under normal conditions.

Such occasional deviations are difficult to detect since gene conversion is equally likely in either direction. Thus the deviations will cancel out and Mendelian ratios will

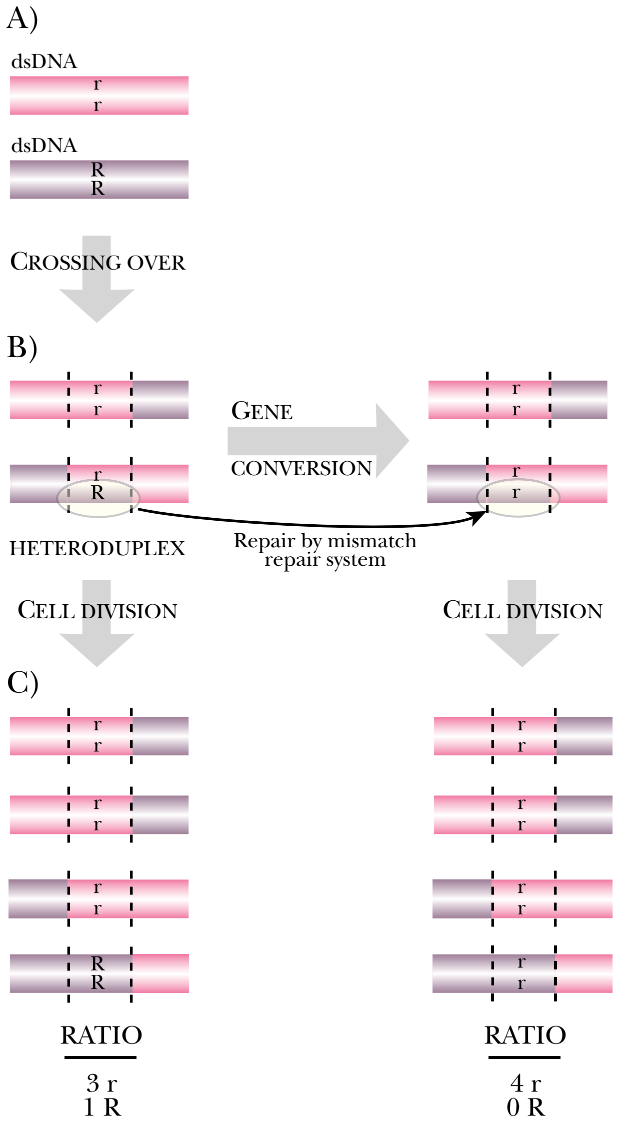


FIGURE 14.27 ***Gene Conversion Following Crossing Over***

Crossing over between two DNA molecules creates a short region of heteroduplex, very close to the Holliday junction (see Fig. 14.04). Sometimes, as shown here, this heteroduplex lies within a gene with two different alleles. A) Two DNA molecules that are about to cross over each carry different alleles (r or R) of the same gene. Crossing over occurs within the gene and a short region of heteroduplex that contains one strand of allele r and the other strand of allele R is generated in one of the daughter DNA molecules. B) Gene conversion occurs when the mismatch repair system corrects the heteroduplex, in this case giving a molecule with the r allele in both strands. C) Gene conversion perturbs the final ratio of the two alleles, as seen following cell division. If no heteroduplex were formed between r and R, a ratio of 2r : 2R would be obtained (not shown). If the heteroduplex is formed, but gene conversion does not occur, a 3r : 1R ratio results, whereas heteroduplex formation followed by gene conversion gives 4r : 0R.

be preserved in a large number of progeny. Gene conversion is thought to occur in all or most organisms, but is only detectable under special circumstances. In practice, it is seen most easily in fungi of the **Ascomycete** group (yeasts, *Neurospora*, etc.) because of their developmental pattern. Sexual reproduction results in the formation of a zygote from the fusion of egg and sperm. Meiosis in these fungi then produces a cluster

ascomycete Type of fungus that produces four (or sometimes eight) spores in a structure known as an ascus

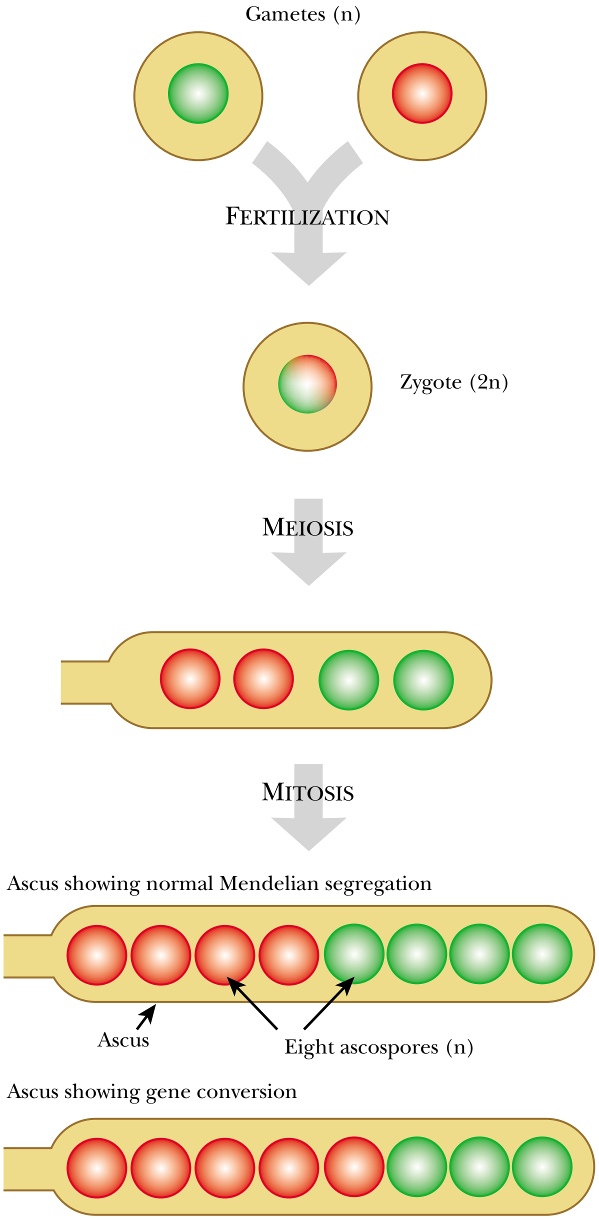


FIGURE 14.28 ***Mendelian Ratios in Ascospore Formation***

In certain fungi a structure known as an ascus keeps together groups of spores ("ascospores") derived from a single zygote. This allows the observation of Mendelian ratios from a single occurrence of meiosis.

of four spores all derived from the same zygote (Fig. 14.28). Sometimes mitosis immediately follows meiosis and results in eight final spores. These four (or eight) "ascospores" all stay together inside a special bag-like structure, the **ascus**. It is thus possible to count the Mendelian ratio separately for each group of four (or eight) offspring derived from the same individual zygote.

ascus Specialized spore forming structure of ascomycete fungus

CHAPTER FIFTEEN

Mobile DNA

Sub-Cellular Genetic Elements as Gene Creatures

In addition to the chromosomes of living cells, a whole variety of **genetic elements** are found in nature. These range in complexity from viruses that possess both a genome plus a protective protein shell to genetic elements such as plasmids and viroids that are merely single molecules of nucleic acid (see Ch. 16 and 17). These sub-cellular elements are all parasitic in the sense that they rely on a host cell to provide energy and raw materials. In some cases, as with many viruses, the host cell may be destroyed. In other cases, as with most plasmids, the host cell is unharmed and may, in fact, benefit by genes carried on the plasmid.

Living cells can be viewed as providing habitats for viruses and other genetic elements.

Since all of these elements possess their own genetic information they may be regarded as life forms of a sort. We will sometimes refer to them as **gene creatures** since they lack their own cells but carry genetic information. From the perspective of a cell, gene creatures are parasites. From the perspective of a gene creature, the cell it inhabits is simply its environment. The latter view may seem strange, but it often helps to understand genetic mechanisms if we view them from the perspective of the genetic element, rather than the cell or organism.

Some of these genetic elements do not even exist as independent genomes but are found as lengths of DNA integrated into other DNA molecules. Some of these have alter egos as viruses or plasmids, whereas others are only ever found as segments of integrated DNA. Some of these integrated DNA segments are able to move around from site to site within host DNA molecules and are known as **transposable elements** or **transposons**. Other stretches of parasitic DNA are stuck permanently where they are and are probably the remains of once mobile gene creatures. They have degenerated into junk DNA. Much of the large human genome is comprised of these types of junk DNA that are no longer active.

Most Mobile DNA Consists of Transposable Elements

Certain segments of mobile DNA possess the ability to move from place to place on the same or different DNA molecules.

Molecules of DNA may be partitioned between dividing cells or they may be transferred from one cell to another. But to a geneticist, the term "**mobile DNA**" has a specialized meaning. Mobile DNA refers to segments of double-stranded DNA that move as discrete units from place to place within other DNA molecules. [Certain mobile DNA elements may exist transiently as separate molecules, but they cannot function independently—e.g. the gene cassettes of integrons, discussed below.] Segments of mobile DNA may move from one site to another on the same larger DNA molecule or from one host DNA molecule to another. Some insert more or less at random whereas others can insert only at specific sequences on the host DNA molecule.

Although the DNA of certain viruses can insert itself into the chromosomes of the host cell, most mobile DNA consists of genetic elements known as transposons or transposable elements. They are also sometimes called "**jumping genes**" because they may hop around from place to place among the chromosomes. The process of jumping from one site to another is called **transposition**. Transposons are not merely dependent on a host cell like plasmids and viruses; they are dependent on a host DNA molecule! Transposons are always inserted into other DNA molecules so they are never free as separate molecules (Fig. 15.01).

gene creature Genetic entitiy that consists primarily of genetic information, sometimes with a protective covering, but without its own machinery to generate energy or replicate macromolecules
genetic element Any molecule or segment of DNA or RNA that carries genetic information and acts as a heritable unit
jumping gene Popular name for a transposable element
mobile DNA Segment of DNA that moves from site to site within or between other molecules of DNA
transposable element A mobile segment of DNA that is always inserted in another, host molecule, of DNA. It has no origin of replication of its own and relies on the host DNA molecule for replication. Includes both DNA-based transposons and retrotransposons
transposition The process by which a transposon moves from one host DNA molecule to another
transposon Same as transposable element, although the term is usually restricted to DNA-based elements that do not use reverse transcriptase

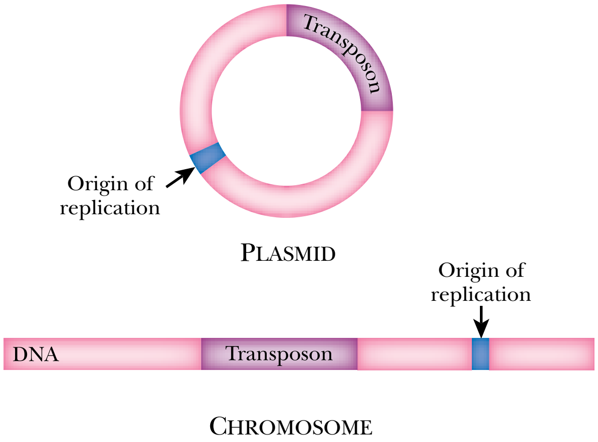


FIGURE 15.01
Transposable Elements are Never Free

Transposable elements are stretches of DNA able to move from one position to another, but are always found within a DNA molecule such as a bacterial plasmid (top) or a eukaryotic chromosome (bottom). Transposons do not contain their own origin of replication, but rely on the host DNA to provide this feature.

As noted in Chapter 5, any molecule of DNA that possesses its own origin of replication is known as a **replicon**. Chromosomes, plasmids and virus genomes are replicons and may be regarded as self-replicating. In contrast, transposons lack a replication origin of their own and are not replicons. They can only be replicated by integrating themselves into other molecules of DNA, such as chromosomes, plasmids or virus DNA. As long as the DNA molecule of which the transposon is part gets replicated, the transposon will also be replicated. If the transposon inserts itself into a DNA molecule with no future, the transposon "dies" with it.

Chromosomes, plasmids and virus genomes are capable of self-replication, transposons are not.

Transposable elements are classified based on their mechanism of movement. The major division is between those that move via an RNA intermediate and need reverse transcriptase to generate this, and those with a DNA-based mechanism. The DNA based transposons are subdivided into two main groups according to whether a new copy is generated during transposition (complex or replicative transposition) or whether the original copy moves, leaving a gap in the DNA in its previous location (conservative or "cut-and-paste" transposition). Table 15.01 summarizes the main groups of transposable elements whose properties will be discussed in more detail below.

The Essential Parts of a Transposon

The enzymes responsible for transposition must recognize repeated sequences at each end of the transposon.

All transposons possess two essential features. First, they have **inverted repeats** at either end. Second, transposons must have at least one gene that encodes the **transposase**, the enzyme needed for movement (Fig. 15.02). The transposase recognizes the inverted repeats and moves the segment of DNA bounded by them from one site to another. The frequency of transposition varies from one transposon to another. Typically, it ranges from 1 in 1,000 to 1 in 10,000 per transposon per cell generation.

In fact, the transposase recognizes two different DNA sequences. First, it recognizes the inverted repeats at the transposon ends and this tells it which piece of DNA must be moved. In addition the transposase must also recognize a specific sequence on the DNA molecule it has chosen as its future home. This is known as the **target sequence** and is usually from three to nine base pairs long. Most often it is an odd number of base pairs, with nine base pairs the most common length. Transposases will often accept a target site with a sequence that is a near match to the preferred target sequence. In some cases, recognition of the target sequence is so lax that it may be difficult to derive a consensus sequence. Due to the short length and low specificity,

inverted repeats Sequences of DNA that are the same but in opposite orientations
replicon A molecule of DNA or RNA that is self-replicating, that is it has its own origin of replication
target sequence Sequence on host DNA molecule into which a transposon inserts itself
transposase Enzyme responsible for moving a transposon

TABLE 15.01 Transposons and Related Elements

Type of Element	Length (approximate)	Terminal Repeats	Mechanism of Mobility	Mechanism to exit cell
DNA-based elements				
Insertion sequence	750–1,500	Inverted	cut-and-paste transposition	No
Simple transposon	1,300–5,000	Inverted	cut-and-paste transposition	No
Composite transposon	2,500–10,000	IS sequences	cut-and-paste transposition	No
Complex transposon	5,000	Inverted	replicative transposition	No
Bacteriophage Mu	37 kb	Inverted	replicative transposition	Virus particle
Conjugative transposon	30 kb–150 kb	None	transfer plus integration	Conjugation
Integron	1,500	None	inserts genes into transposon	No
Retro-elements (possess reverse transcriptase)				
Retrovirus	7,000–10,000	Direct (LTR)	via RNA intermediate	Virus particle
Endogenous retrovirus	7,000–10,000	Direct (LTR)	via RNA intermediate	Defective Virus particle
Retrovirus-like element	7,000–10,000	Direct (LTR)	via RNA intermediate	No
Retrotransposon	6,000	Direct	via RNA intermediate	No
LINE	6,500	None	via RNA intermediate	No
Retron	1,300–2,000	None	unknown	No
Retrointron	2,000–3,000	None	via RNA intermediate	No
Retro-derived elements (or "Retro-transcripts")				
Processed pseudogene	1,000–3,000	None	immobile	No
SINE	300	None	transcription & reintegration	No

Notes:

None of these elements have their own origin of replication, except for conjugative transposons, which have an origin of transfer (but not of vegetative replication).

Length refers to "typical" elements. Since many transposable elements may acquire extra internal DNA by a variety of mechanisms, longer elements carrying a variety of extra genes may be found. In addition, shorter defective elements are frequently found among many classes of transposable element. The lengths given refer to intact functional elements.

The length of Mu DNA inside a virus particle is 39 kb and consists of 37 kb of Mu DNA plus approximately 2,000 bases of chromosomal DNA attached to the Mu ends. Note that when Mu first enters a new host cell the original integration event uses cut-and-paste transposition, not replicative transposition.

All intermediates between fully functional retroviruses and sequences presumed to be derived from defective retroviruses are found in eukaryotic chromosomes, as well as individual LTR elements with no internal sequences. Consequently, the classification and naming of retro-elements varies somewhat among different authors. LINES and related elements are sometimes referred to as non-LTR-retrotransposons to distinguish them from "true" retrotransposons, which have terminal repeats equivalent to the LTRs of retroviruses.

Retrointrons are mobile group I or II introns that encode reverse transcriptase.

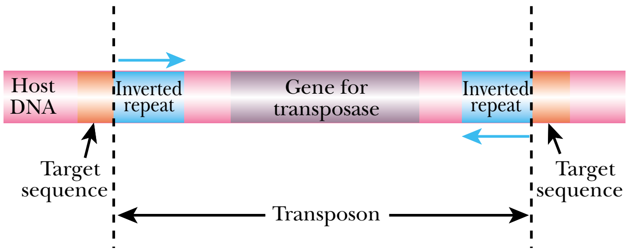


FIGURE 15.02 ***Essential Components of a Transposon***

Essential features of a transposon are two inverted repeats at the ends and a gene for transposase. Other, non-essential, genes may be found between the inverted repeats. During transposition, the transposon duplicates the host DNA target sequence and therefore one copy flanks each side of the transposon.

Although transposons are inserted at target sequences, the specificity is low and transposition appears more or less random.

multiple copies of the target sequence will be found on most DNA molecules of any length and insertion will often be almost random. Whenever a transposon moves, the target sequence is duplicated due to the mechanism of transposition (see below). The result is that two identical copies of the target sequence are found, one on each side of the transposon.

TABLE 15.02 Some Insertion Sequences

Insertion sequence	Total length	Inverted repeat	Target length
IS1	768	23	9
IS2	1327	41	5
IS3	1258	40	3
IS4	1426	18	11–13
IS5	1195	16	4
IS10	1329	22	9
IS50	1531	9	9
IS903	1057	18	9

Many larger transposons carry a variety of genes unrelated to transposition itself. Antibiotic resistance genes, virulence genes, metabolic genes and others may be located within transposons and become mobile within the genome. Some of the first transposons to be analyzed carried genes for antibiotic resistance which protected the bacteria that hosted them from attack by human medicine. The ability to move blocks of genetic material from one DNA molecule to another has made transposons very important in the evolution of plasmids, viruses and the chromosomes of both bacteria and higher organisms.

Insertion Sequences—the Simplest Transposons

The simplest and shortest transposons, known as **insertion sequences**, were first found in bacteria. They are designated IS1, IS2 etc. Typical insertion sequences are 750 to 1,500 base pairs (bp) long with terminal inverted repeats of 10 to 40 bp (Table 15.02). Often the inverted repeats at the ends of insertion sequences are not quite exact repeats. For example, the inverted repeats of IS1 match in 20 out of 23 positions.

The simplest transposable elements contain overlapping genes for a transposase and a regulatory protein located between inverted repeats.

Insertion sequences only encode a single enzyme, the transposase, the enzyme needed for movement. Between the inverted repeats is a region that actually contains two open reading frames, *orfA* and *orfB*. The transposase itself is derived from both open reading frames by a frameshift that occurs during translation (Fig. 15.03). The low frequency of such frameshifting results in the transposase only being expressed at low levels. Unregulated transposition would lead to massive damage to the host chromosome. Consequently, synthesis of transposase needs to be carefully regulated. When no frameshift occurs, the first open reading frame, *orfA*, is expressed. This gene product is a transcriptional regulator that controls the production of transposase and of itself.

Insertion sequences are found in the chromosomes of bacteria and also in the DNA of their plasmids and viruses. For example, several copies each of the insertion sequences IS1, IS2, and IS3 are found in the chromosome of *E. coli*. The F-plasmid has zero copies of IS1, one copy of IS2, and two copies of IS3. When the F-plasmid and chromosome possess identical IS sequences, the plasmid can insert itself into the host chromosome. This, in turn, allows transfer of chromosomal genes by the F-plasmid as explained in Chapter 18.

Insertion sequences contain no genes that provide a convenient phenotype. Originally their presence was recognized because movement of the insertion sequence inactivated genes with a noticeable phenotype. Such insertion mutations usually abolish gene function completely and are often polar to downstream genes of the operon. In addition, they revert only at very low frequencies (see Ch. 13).

insertion sequence A simple transposon consisting only of inverted repeats surrounding a gene encoding transposase

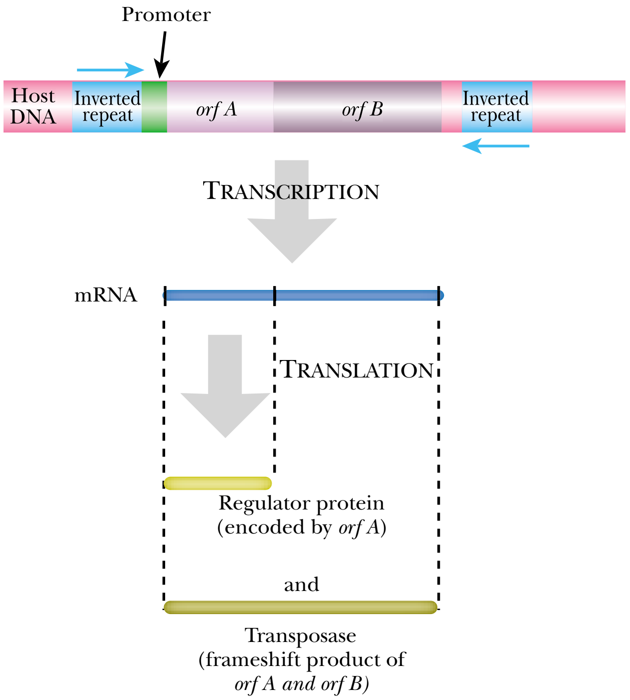

Promoter
Host DNA
Inverted repeat
orf A
orf B
Inverted repeat
TRANSCRIPTION
mRNA
TRANSLATION
Regulator protein (encoded by *orf A*)
and
Transposase (frameshift product of *orf A and orf B*)

FIGURE 15.03 ***Structure of an Insertion Sequence***

Insertion sequences, like all transposons, are flanked by inverted repeats. Two open reading frames within the insertion sequence encode the transposase gene. When a frameshift occurs during translation, transposase is produced and the insertion sequence "jumps" to a new location. If the frameshift does not occur, then only the *orfA* gene product is expressed. This protein regulates transposition by turning off transcription of *orfA* and *orfB* at the promoter.

Movement by Conservative Transposition

The simplest transposons, including the insertion sequences, move by a mechanism known as **conservative or "cut-and-paste" transposition** (Fig. 15.04). This requires three elements: the transposase, the inverted repeats at the ends of the transposon, and a suitable target sequence on another segment of DNA somewhere else in the cell. The transposon may move to another site in the same molecule of DNA or to a separate molecule of DNA. The DNA molecule into which a transposon jumps can be a plasmid, a virus, or a chromosome; any DNA molecule will do, as long as it has a reasonable target sequence.

The transposase starts by binding the inverted repeats at the transposon ends. It then cuts the transposon loose from its original site (Fig. 15.05A). Next the transposase finds a suitable target sequence in the molecule of DNA that will be the transposon's new home. It makes a staggered cut that opens the target sequence to give overhanging ends (Fig. 15.05B). Finally, it inserts the transposon into the gap.

Conservative transposition leaves behind a double-stranded break in the DNA.

This leaves a structure with two short regions of single-stranded DNA. These are recognized by the bacterial host, which synthesizes the second strand. The net result is that the transposon has moved, and the target sequence has been duplicated in the process. This type of transposition process is known as conservative transposition because the DNA of the transposon is not altered during the move.

Where the transposon cut itself out of its original home, it leaves a double stranded break in the DNA. There is a significant likelihood that this damaged DNA molecule will not be repaired and is doomed. Clearly, high frequency transposition would severely damage the host cell chromosomes. Even if the break is sealed, the duplicated target sequence that is left behind may cause a permanent frameshift if within a coding sequence. Consequently, as remarked above, transposition must be tightly regulated.

conservative transposition Same as cut-and-paste transposition
cut-and-paste transposition Type of transposition in which a transposon is completely excised from its original location and moves as a whole unit to another site

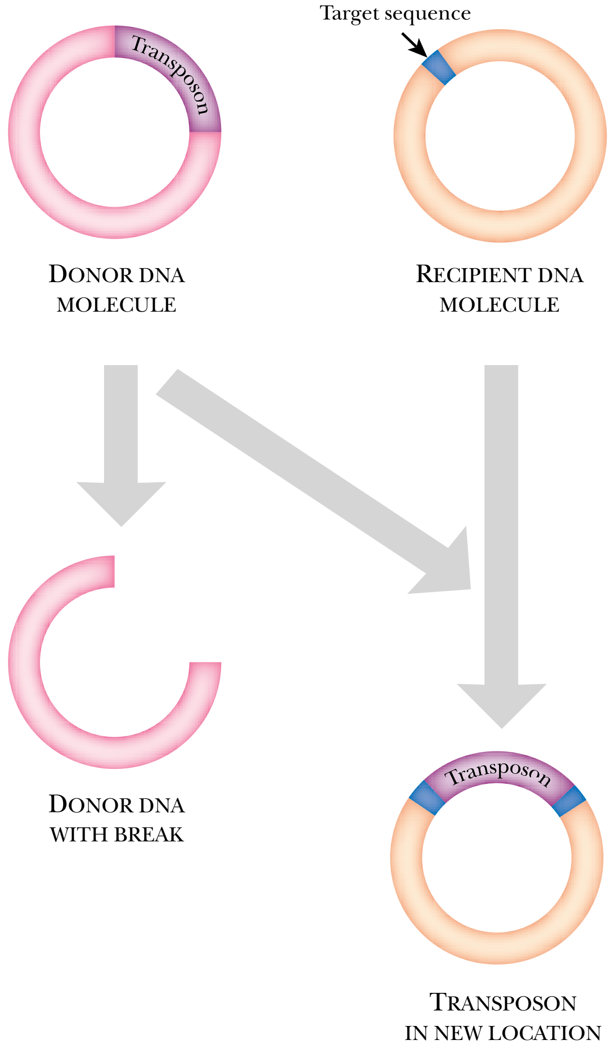


FIGURE 15.04 ***Outline of Conservative Transposition***

The transposon moves from one DNA molecule to another. It inserts into the target sequence on the recipient DNA molecule and leaves behind a double-stranded break in its original location.

Complex Transposons Move by Replicative Transposition

Replicative transposition does not cause damage to the original DNA host molecule.

Replicative transposons need an extra enzyme, resolvase, and a site in the DNA for it to act on.

Transposition does not always leave behind damaged DNA with double stranded breaks. Some transposons are capable of **replicative transposition**, during which the transposon creates a second copy of itself (Fig. 15.06). Consequently, both the original home site and the newly selected target location end up with a copy of the transposon. The original home DNA molecule is not abandoned or damaged. Transposons using this mechanism are known as **complex transposons** because the process is more complex than the simple cut-and-paste mechanism described above.

Complex transposons have a transposase that recognizes their inverted repeats and the host target sequence just like other kinds of transposons. In addition, replicative transposons needs an extra enzyme, **resolvase**, and an extra DNA sequence, the **internal resolution site (IRS)** which is recognized by the resolvase (Fig. 15.07). Although complex transposons are replicated while moving, they are not replicons, as they have no origin of replication. The transposon does not make a new copy

complex transposon A transposon that moves by replicative transposition
internal resolution site (IRS) Site within a complex transposon where resolvase cuts the DNA to release two separate molecules of DNA from the cointegrate during replicative transposition
replicative transposition Type of transposition in which two copies of the transposon are generated, one in the original site and another at a new location
resolvase An enzyme that cuts apart a cointegrate releasing two separate molecule of DNA

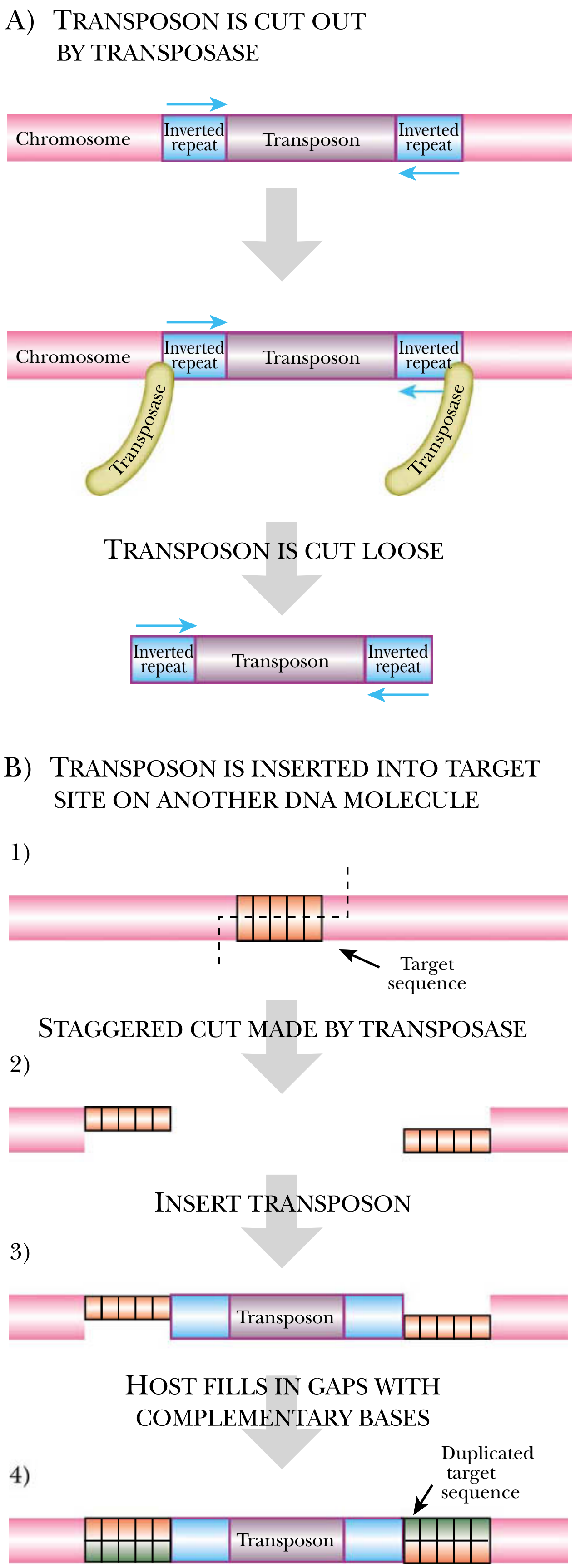


FIGURE 15.05 ***Movement by Conservative Transposition***

A) Transposase, produced by the transposon, recognizes and cuts the inverted repeats, freeing the transposon from the chromosome. The chromosome is left with a double-stranded break that needs to be repaired. B) The transposon bound by transposase identifies a target sequence, and transposase directs a staggered cut into the target DNA. The ssDNA ends of the target sequence are joined to the inverted repeats of the transposon. The resulting single-stranded regions are filled in by the host cell, thus duplicating the target sequence.

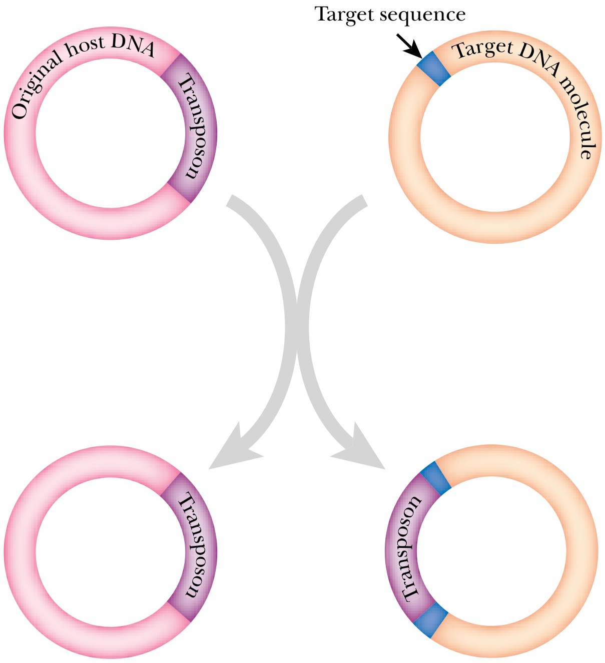


FIGURE 15.06 ***Outline of Replicative Transposition***

The transposon is duplicated as it moves from one DNA molecule to another. It inserts into the target sequence on the recipient DNA molecule and leaves behind a copy of the transposon in the original location.

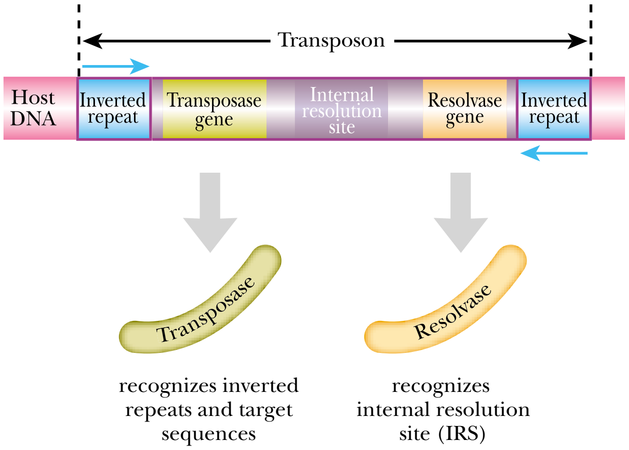


FIGURE 15.07 ***Components of a Complex Transposon***

Complex transposons have a gene for resolvase and an internal resolution site, in addition to the gene for transposase and two flanking inverted repeats.

of itself that is liberated as a free intermediate to find a new home. Instead, complex transposons trick the host cell into duplicating their DNA during the transposition process.

Replicative transposition proceeds as follows. The transposase starts by making single stranded nicks at the ends of both the transposon and the target sequence. Next, it joins the free ends to create a **cointegrate** in which both DNA molecules are linked together via single strands of transposon DNA (Fig. 15.08). The presence of single-stranded DNA alerts the host cell repair systems, which synthesize the complementary strands, thus duplicating the transposon. This leaves a cointegrate of two double-stranded DNA molecules linked by two transposons. Note that each copy of the transposon consists of one old and one new strand of DNA.

The function of resolvase is to resolve the cointegrate and separate the two DNA molecules again. It does this by recognizing the two IRS sequences, in the middle of the two copies of the transposon, and carrying out recombination between them (Fig. 15.08). This generates two free DNA molecules each carrying a single copy of the

cointegrate A temporary structure formed by linking the strands of two molecules of DNA during transposition, recombination or similar processes

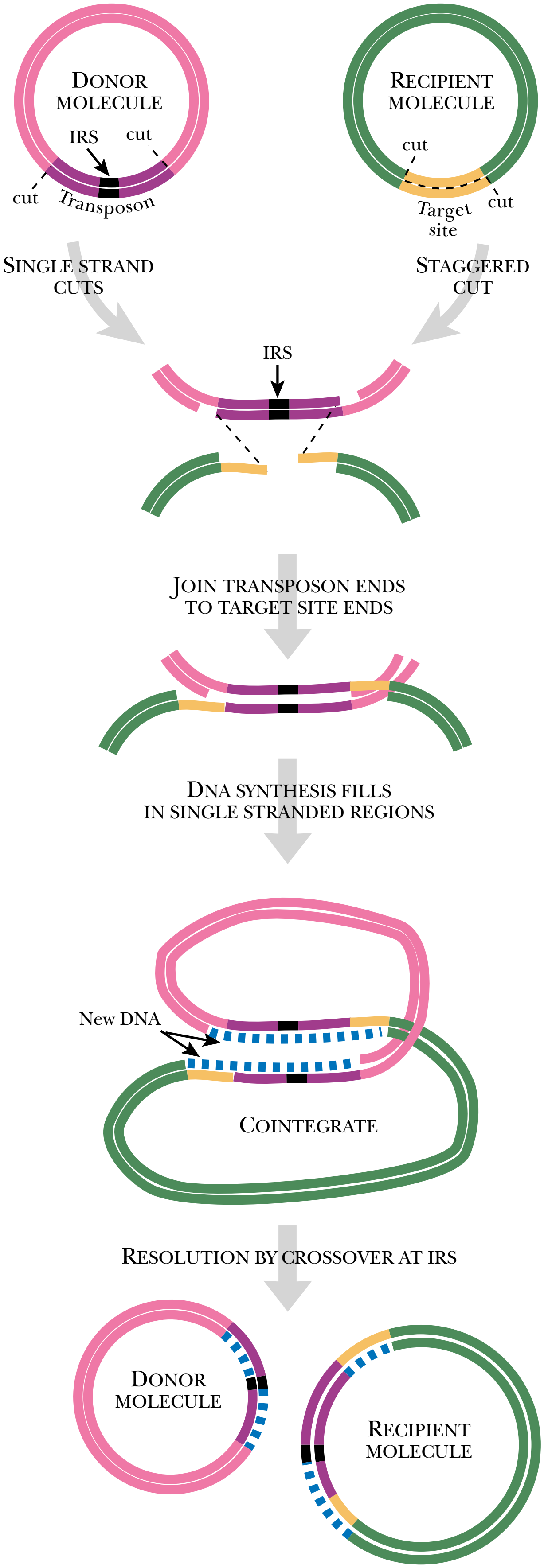


FIGURE 15.08 ***Replicative Transposition Forms a Cointegrate***

Single-stranded cuts are made flanking the transposon in the donor molecule and a staggered cut is made in the target site on the recipient molecule. The ends are joined as shown, which causes the transposon to "split", resulting in two single-stranded copies of the transposon. The single-stranded DNA alerts the host to repair the defect thus making both transposons double-stranded. The recipient DNA is now joined to the donor DNA via duplicated transposons and forms what is called a cointegrate. Resolvase, produced by the transposon, then resolves the cointegrate at the two IRS sequences and releases the donor and recipient molecules. Notice that a copy of the transposon is now located on each molecule of DNA.

transposon. Note that resolution further scrambles the two copies of the transposon as shown in Figure 15.08.

Most of the known complex transposons carry other genes in addition to those involved in transposition and resolution. For example, Tn1 and Tn3 are complex transposons that carry resistance to antibiotics of the penicillin family and are found in both the plasmids and chromosomes of many bacteria. Movement of complex transposons is traced by the expressed phenotype of these genes.

Replicative and Conservative Transposition are Related

Although replicative and conservative transpositions seem quite different, the actual mechanisms of the transposase steps are closely related. In both cases the target sequence is opened by a staggered cut (Fig. 15.09). In both cases, the transposase cuts at the junction between the ends of the transposon and the host DNA. However, in conservative transposition, both strands are cut whereas in replicative transposition only one strand is cut. In either case, the free 3′-ends of the transposon are joined to the 5′-ends of the opened target sequence. This sequence of events moves a conservative transposon to its new position while it creates a cointegrate in the case of the complex transposon.

Several steps in conservative and replicative transposition are very similar at the mechanistic level.

The next step is again very similar. The host cell enzymes fill in the single-stranded regions using the free 3′-ends of the opened target sequence as primers. In conservative transposition the new DNA is merely a handful of nucleotides and this step just duplicates the target sequence. In the case of replicative transposition, the single-stranded regions are longer and this step duplicates the transposon itself.

This similarity is illustrated by the transposon Tn7, which normally operates by the conservative mechanism. Tn7 is unusual in having a transposase consisting of two proteins. TnsA makes single-stranded nicks at the 5′-ends of Tn7, and TnsB carries out the nicking and joining at the 3′-ends of Tn7, therefore, TnsA and TnsB create a double-stranded cut when both are expressed. Mutants of Tn7 exist that have a defective TnsA protein and no longer cut the 5′-strand. However, TnsB continues to cut and rejoin the 3′-strand, forming cointegrates as in replicative transposition. Therefore TnsB resembles the transposase of complex transposons. TnsA protein, which cuts Tn7 free of its original site, has a structure similar to a type II restriction endonuclease (see Ch. 22).

Composite Transposons

A **composite transposon** consists of two inverted repeats from two separate transposons moving together as one unit and carrying the DNA between them (Fig. 15.10). For example, consider a segment of DNA flanked at both ends by two identical insertion sequences. The transposase will move any segment of DNA surrounded by a pair of the inverted repeats that it recognizes. Consequently, when transposition occurs there are several possibilities (Fig. 15.10). First, each of the insertion sequences may move independently. Second, the whole structure between the two outermost inverted repeats may move as a unit, i.e., as a composite transposon.

Transposable elements may be built up in modular fashion from simpler transposons plus entrapped DNA.

Many of the well-known bacterial transposons that carry genes for antibiotic resistance or other useful properties are composite transposons. Three of the best known are Tn5 (kanamycin resistance), Tn9 (chloramphenicol resistance) and Tn10 (tetracycline resistance). Usually the pair of insertion sequences at the ends of the transposon are inverted relative to each other, as in Tn5 and Tn10. The IS elements of Tn5 and Tn10 have not been found alone and are named IS50 and IS10 respectively. Less often

composite transposon A transposon that consists of two insertion sequences surrounding a central block of genes

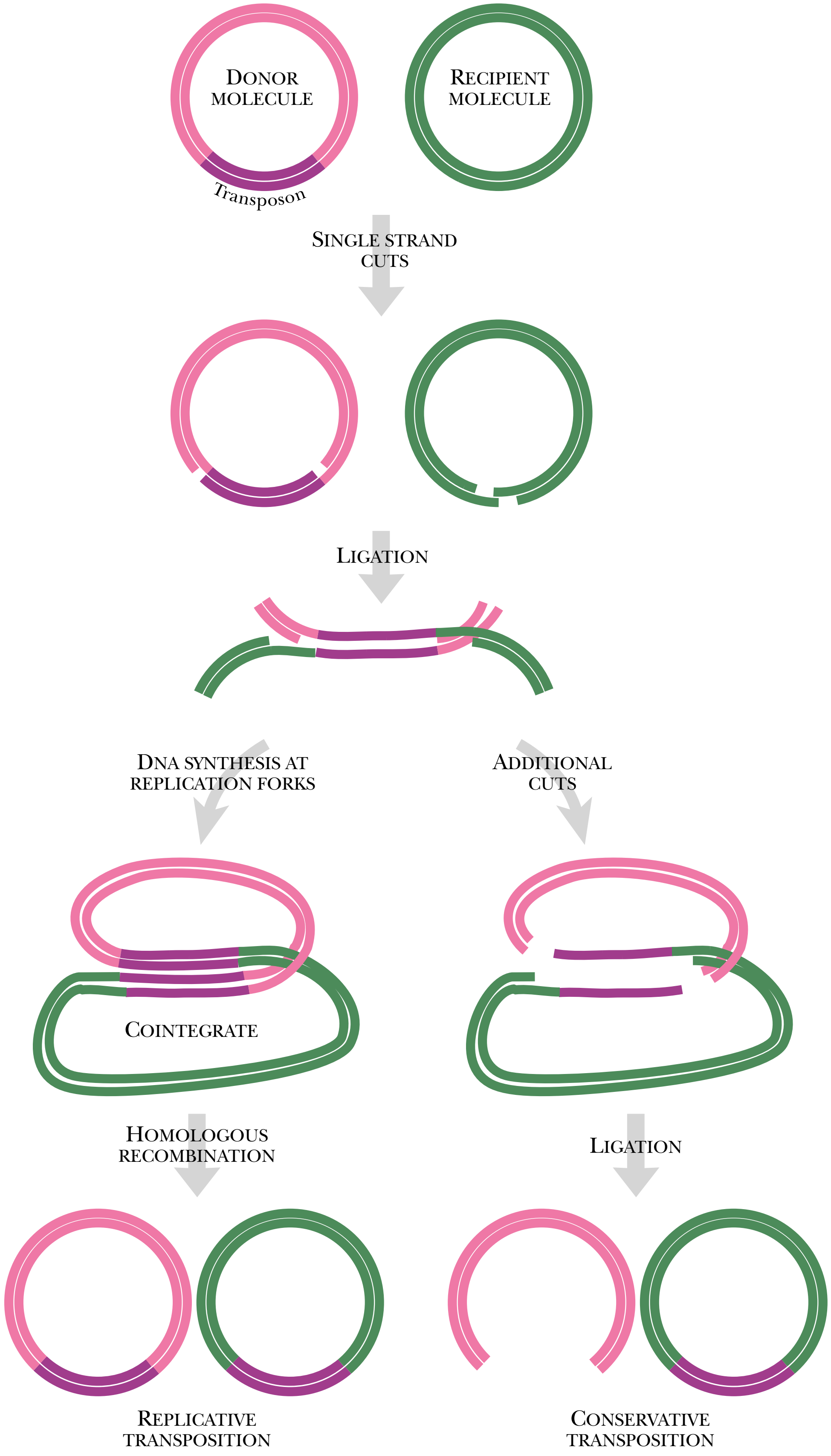


FIGURE 15.09 ***Replicative and Conservative Transposition are Related***

Both conservative and replicative transpositions begin with single-stranded cuts and the joining of ends. In conservative transposition (right side) a second cut completely releases the transposon from the donor molecule after the donor and recipient molecule have joined. In contrast, in replicative transposition, the transposon splits into two single-stranded copies at this stage. Next the host cell fills in the single-stranded gaps, which is just the short target sequence in conservative replication, but the entire transposon in replicative transposition.

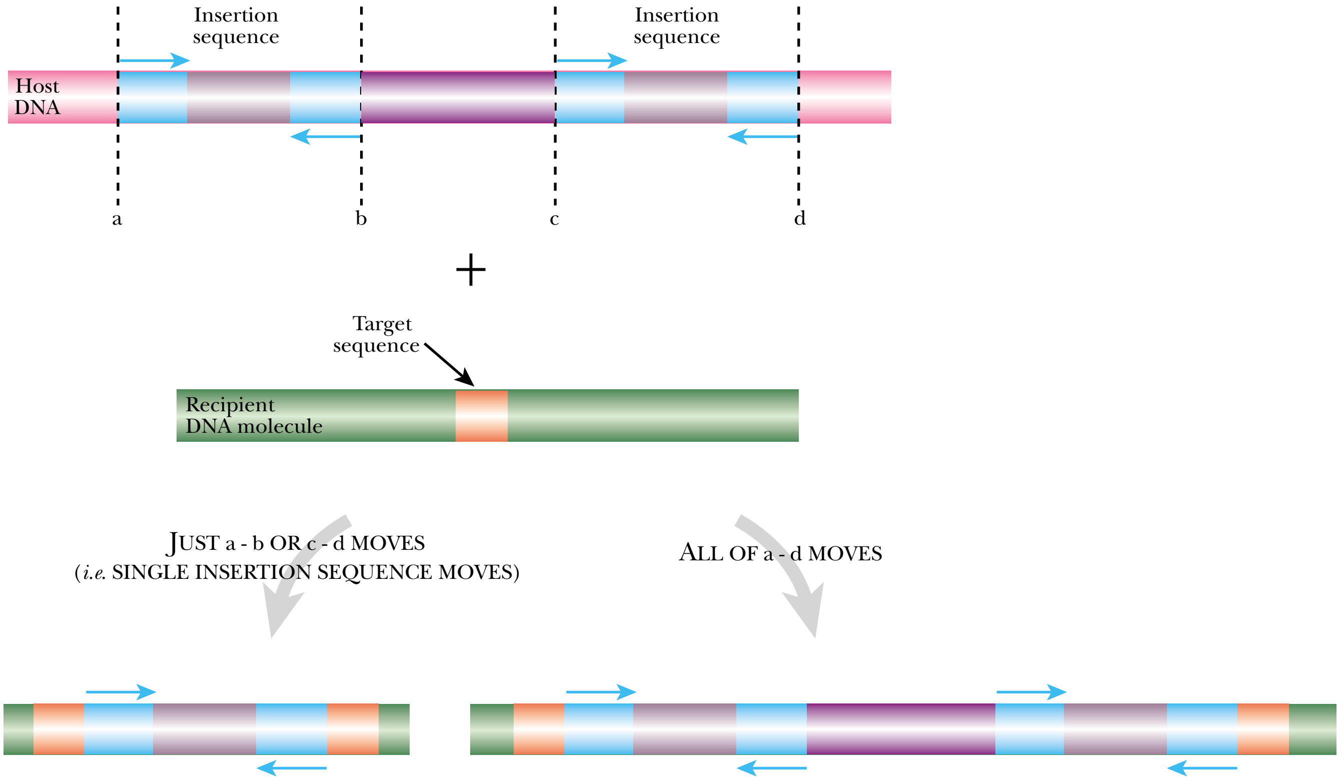


FIGURE 15.10 ***Principle of the Composite Transposon***

Two identical insertion sequences can move as two independent transposons or as one composite transposon. Composite transposition moves both insertion sequences (region a–b and c–d) as well as any intermediate DNA (region b–c) to the new location.

the two IS elements face in the same direction, as in Tn9, which is flanked by direct repeats of IS1.

Once a useful composite transposon has assembled by chance, natural selection will act to keep the parts together. Mutations accumulate that inactivate the innermost pair of inverted repeats, which prevents the insertion sequences from jumping independently. Often, one of the two transposase genes is also lost. The result is that the two ends and the middle are now permanently associated and always move as a unit (Fig. 15.11). In practice, all stages from newly formed to fully fused composite transposons are found in bacteria. For that matter, novel composite transposons can be assembled in the laboratory by genetic manipulation.

Transposition may Rearrange Host DNA

Insertions, deletions and inversions of host cell DNA may result from anomalous or failed attempts at movement by a composite transposon. As remarked above, the transposase will move any segment of DNA surrounded by a pair of correct inverted repeats. Composite transposons have four such terminal repeats. Two of these are "inside ends", relative to the transposon, and two are "outside ends" (Fig. 15.10 above). Any pair that face in opposite directions may be used together. Transposition of a single IS element involves one "inside end" and one "outside end". Normal movement of the whole composite transposon uses the two "outside ends".

Movement of transposons, or their sub-components, may cause rearrangements of the host DNA molecule.

But suppose that the two "inside ends" are used for transposition. The whole of the DNA molecule outside the transposon will be moved. This is easiest to see if we

FIGURE 15.11 ***Evolution of a Composite Transposon***

Since composite transposons no longer require two separate transposase genes and four inverted repeats, mutations accumulate in the non-essential regions. These mutations ensure that the transposon moves as a composite unit. This is especially important if the transposon carries internal genes that enhance the survival of the host cell.

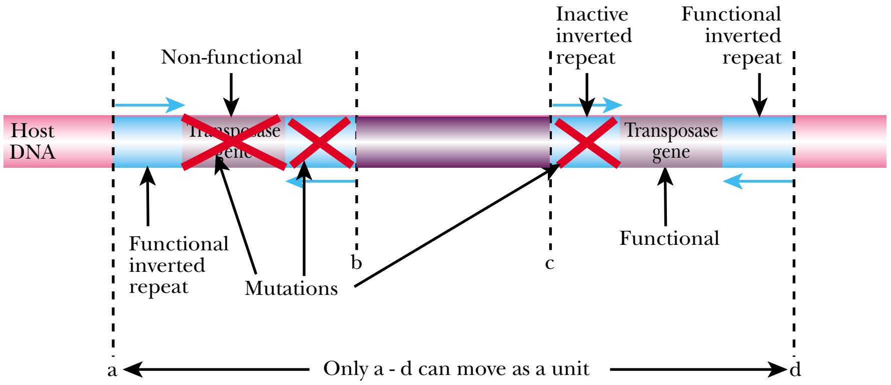
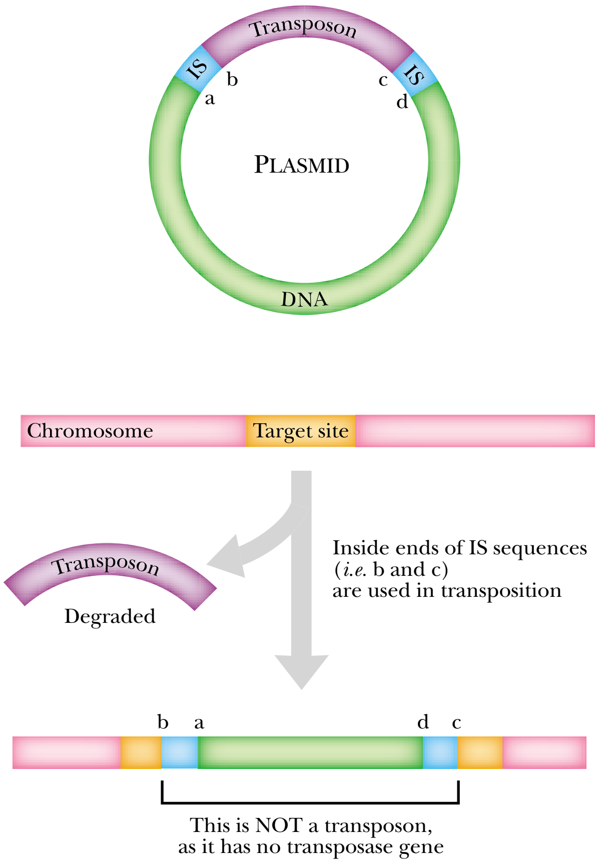


FIGURE 15.12 ***Insertion Created by Using Inside Ends to Transpose***

Plasmid DNA can integrate into a chromosome if the "inside ends" (b and c) of the composite transposon are moved by transposase. Inverted repeats b and c point outwards from the transposon, but the transposase is still capable of moving the DNA between them—i.e., the DNA making up most of the plasmid. If the transposase gene is located inside the transposon (in the purple segment), the DNA that jumped will not be able to move itself again, i.e., it is not itself a genuine transposon.


Transposon
IS
IS
b
a
c
d
PLASMID
DNA
Chromosome
Target site
Transposon
Degraded
Inside ends of IS sequences (*i.e.* b and c) are used in transposition
b
a
d
c
This is NOT a transposon, as it has no transposase gene

consider a small circular DNA molecule, such as a plasmid (Fig. 15.12), that carries a composite transposon. Using the "outside ends" moves the transposon, using the "inside ends" moves the rest of the DNA molecule. [Notice that the "inside ends" and the "outside ends" face in different directions.] The result, in the case illustrated, is the insertion of a segment of plasmid DNA into the host chromosome. Note that the segment that moved may not be able to move again if the gene for the transposase was left behind during this maneuver.

If both of the "inside ends" are used in jumping to another site on the same host DNA molecule the result will be the insertion of the host DNA into itself (Fig. 15.13). Depending on which pairs of ends are re-joined after breakage, the host DNA will suffer either a deletion or an inversion. The inside region of the composite transposon is lost during this process. This process is sometimes known as abortive transposition.

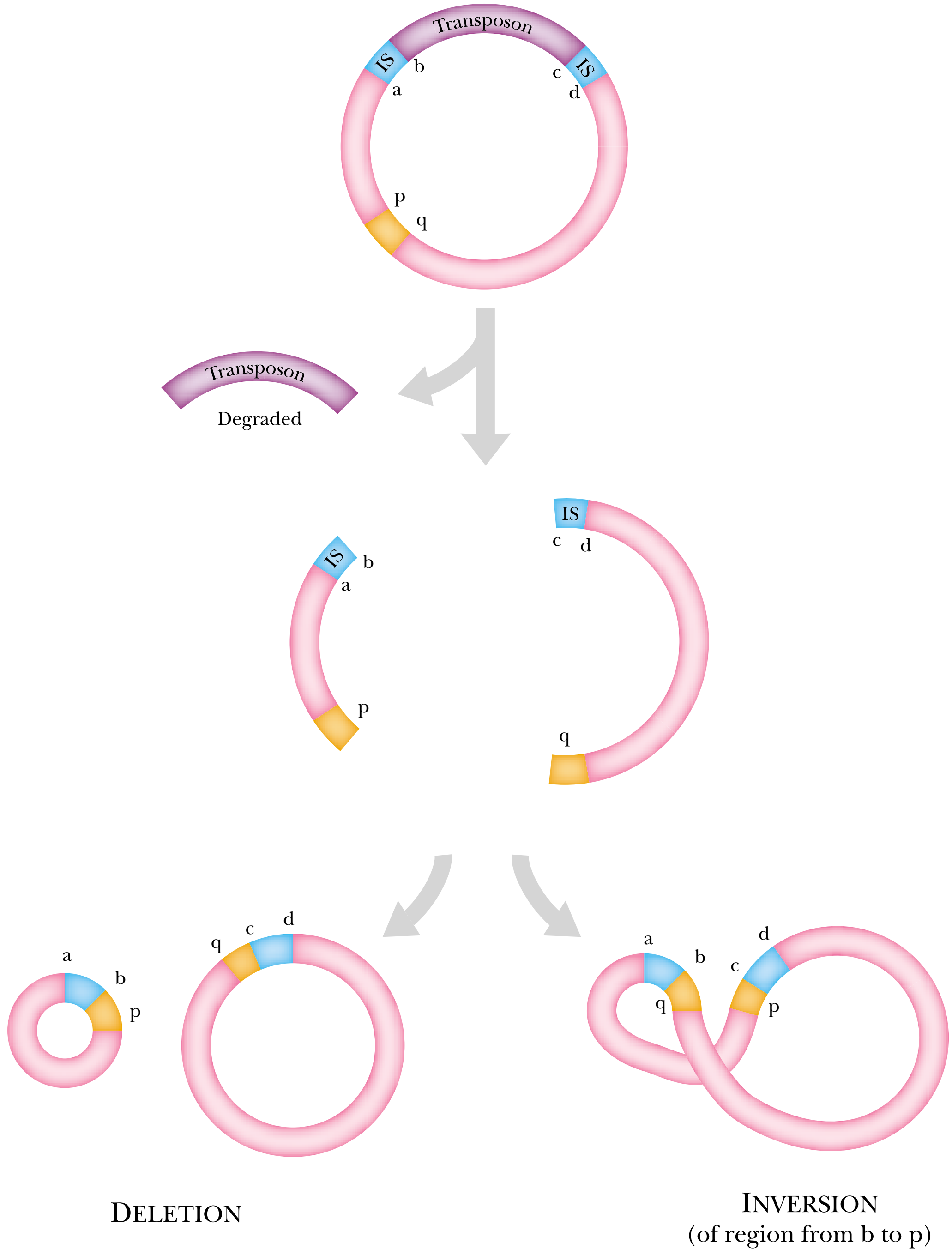


FIGURE 15.13 ***Deletions and Inversions made by Abortive Transposition***

"Inside ends" (b and c) are incorrectly used by transposase to move the host DNA rather than the transposon. If the target sequence (p–q) is found on the same DNA molecule, transposase cuts this site also, creating a double-stranded break. There are two alternative ways of rejoining the ends. Separate molecules may be formed, thus creating a deletion in the host DNA molecule. Alternately, the small piece (p–b) may rejoin the larger fragment such that region p joins IS sequence c–d, and region q joins IS sequence a–b, thus inverting the DNA.

Transposons in Higher Life Forms

Plants frequently contain simple transposons, such as the Ac/Ds elements of corn.

Transposons are scattered throughout the DNA of all forms of life. Although we have used examples from bacteria to illustrate how they work, Barbara McClintock observed the first jumping genes during genetic crosses in maize (corn) plants. She worked in the 1940's before the DNA double helix was even discovered, but nonetheless realized that segments of the plant genetic material must be moving around. When she presented her conclusions in public, in 1951, no one believed her. Later investigations in bacteria revealed the molecular details of transposition, and confirmed what Barbara McClintock had observed in the 1940s. Barbara McClintock received her Nobel Prize in 1983 when the significance of her work was more fully realized.

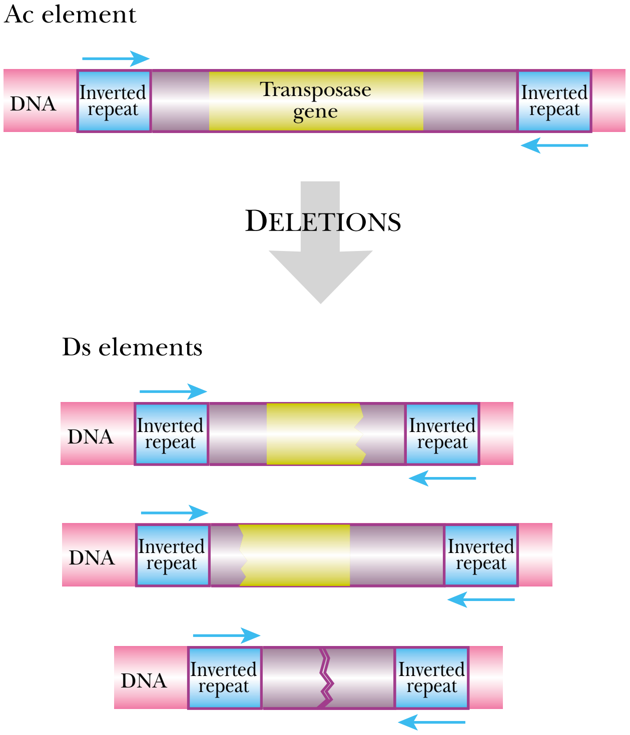


FIGURE 15.14 ***Ac/Ds* *Family of Transposons in Maize***

The simple Ac element of Maize contains two inverted repeats and a functional transposase gene. Ds elements are derived from Ac by deletion of the transposase gene, either completely or partially.

Multiple defective copies of transposons are often found in higher organisms. They rely on an intact copy to help them move.

The Ac/Ds family of transposons in corn is simple and conservative. Family members leave behind double stranded gaps in the DNA when they move. They have inverted terminal repeats of 11 base pairs and insert at an 8 bp target sequence. The **Ac element** is 4,500 bp long and is a fully functional transposon with the ability to move itself. The **Ds elements** vary in size and are defective. They are derived from Ac by deletion of all or part of the transposase gene and they cannot move by themselves. If a cell contains an Ac element anywhere in its DNA, then the transposase enzyme made by Ac can also move any Ds element. Therefore to remain mobile, the Ds elements must keep the inverted repeats, otherwise the Ac transposase will not recognize them (Fig. 15.14). The Ac and Ds elements do not need to be on the same chromosome for transposition to occur.

If a Ds element has been inserted into the gene for purple kernels of corn, the gene is disrupted and the kernels are white. If there is no Ac element in any of the cells of the entire corn kernel the white color is stably inherited. If an Ac element is also present in some of the cells of the corn kernel, it may move the Ds element. The cells in which both elements are present return to the original purple color. All the daughter cells of these will also be purple. Eventually, a patch of purple will appear on the kernel as it grows. If the transposition occurs when the kernel is just beginning to develop, a large patch of purple will appear, and if the tranposition occurs when the kernel is almost fully developed the patch of purple will be very small. This produces a mottled kernel of corn (Fig. 15.15).

Animals and plants frequently contain transposon families with both active and defective members. These may vary greatly in their overall lengths. Not only are there defective members that need help to move, but we also find totally inactive transposons. These have suffered mutations in their terminal repeats which makes them unrecognizable by transposase and therefore immobile.

Conservative transposons of the mariner family are widespread in eukaryotes.

The most widely distributed transposons in higher organisms are those of the Tc1/mariner family. The first members to be found were **Tc1**, from the nematode

Ac element Intact and active version of a transposon found in maize
Ds elements Defective version of a transposon found in maize; cannot move alone but needs the Ac element to provide transposase
Tc1 element Transposon *Caenorhabditis* 1. A transposon of the mariner family found in the nematode *Caenorhabditis*

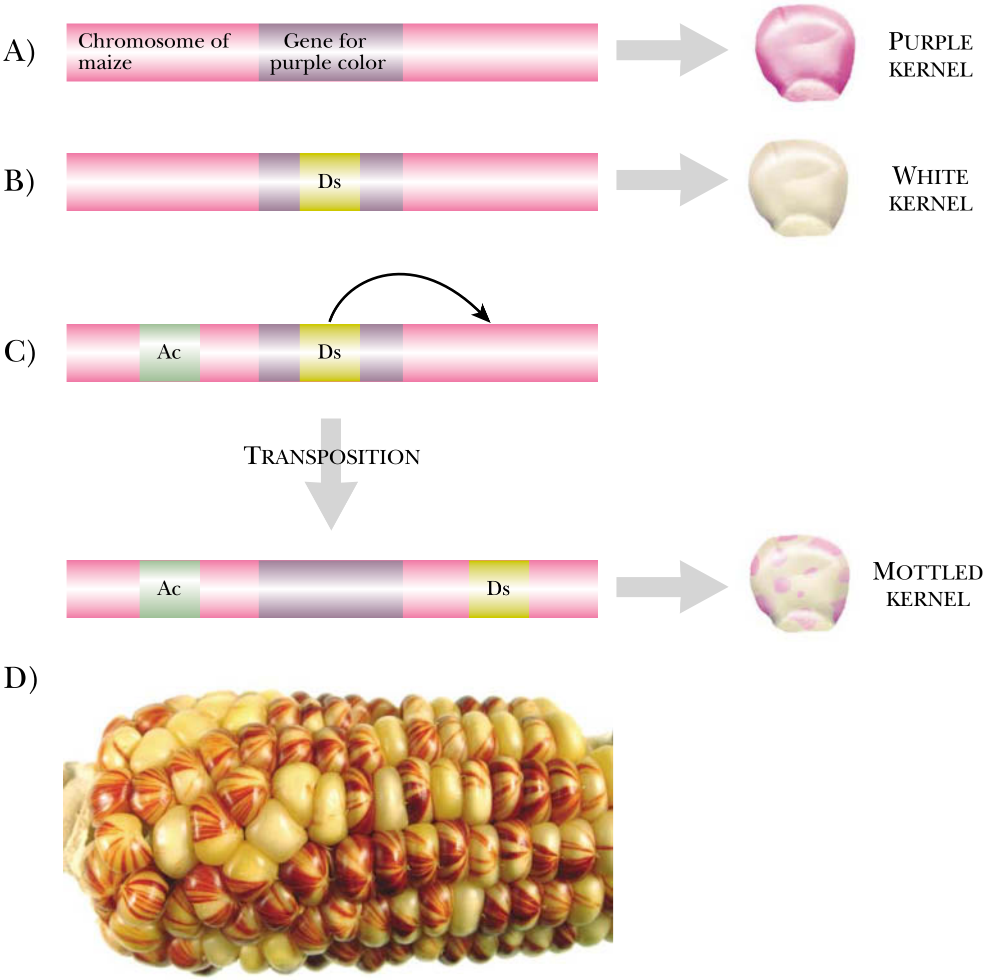


FIGURE 15.15 ***Movement of Ds Element Gives Mottled Corn***

A: The gene for purple color in maize produces a purple kernel of corn.
B: If the gene for purple color is disrupted by a Ds element, the kernel is white.
C: If both the Ac and Ds element are present in the same cell, transposase from the Ac element moves the Ds element from the purple gene, returning the cell to the original purple color. As the cell divides, the purple color is inherited in the daughter cells, and a patch of purple forms on the white kernel. Random transposition events in several such cells results in a mottled kernel.
D: Photograph of mottled corn cob showing reddish-purple streaks due to transposition of Ac/Ds. The patterns caused by transposition may be blotches, dots, irregular lines or streaks.

Caenorhabditis, and the **mariner** transposon of *Drosophila*. Members of this family are found in fungi, plants, animals (including humans) and protozoans. They range from roughly 1300 to 2500 bp in length, contain a single transposase gene and are flanked by inverted repeats. They operate by the conservative cut and paste mechanism. The breaks they leave behind are often mended by the eukaryotic double-stranded break repair systems (see Ch. 14).

Retro-Elements Make an RNA Copy

The transposons we have discussed so far move while in their DNA form. In addition, there is a vast array of mobile elements that move while in their RNA phase. These range from retroviruses (see Ch. 17), which can integrate their DNA into the host cell chromosome, to transposon-like elements. These may all be classed as **retro-elements** since they all use **reverse transcriptase** to convert an RNA transcript back into DNA. Whenever retro-elements insert into host DNA the target sequence is duplicated as for DNA-based transposons. Retro-elements are more common in higher organisms whereas DNA-based transposons predominate in bacteria.

Some transposons move via an RNA intermediate and rely on reverse transcriptase.

Retrotransposons or (for short) **retroposons** are transposable elements that rely on reverse transcriptase for movement. They are found most often in eukaryotes, especially animals. The **Ty1** (Transposon of yeast 1) retrotransposon of yeast is around 6,000 bp

Mariner elements A widespread family of conservative DNA-based transposons first found in *Drosophila*
retro-element A genetic element that uses reverse transcriptase to convert the RNA form of its genome to a DNA copy
retroposon Short for retrotransposon
retrotransposon A transposable element that uses reverse transcriptase to convert the RNA form of its genome to a DNA copy
reverse transcriptase Enzyme that synthesizes a DNA copy from an RNA template
Ty1 element Transposon yeast 1. A retrotransposon of yeast that moves via an RNA intermediate

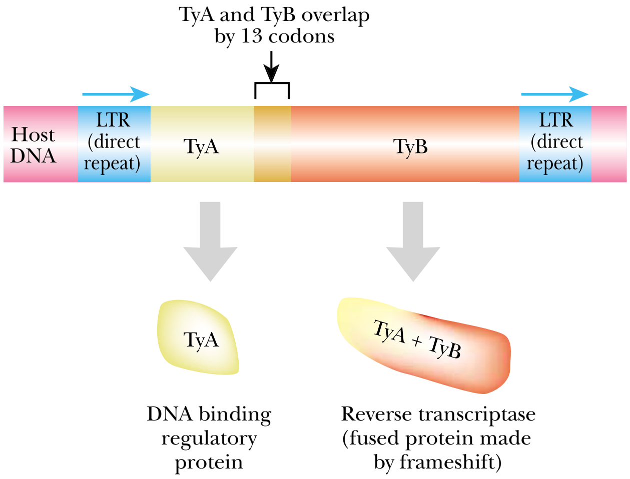

TyA and TyB overlap by 13 codons
Host DNA
LTR (direct repeat)
TyA
TyB
LTR (direct repeat)
TyA
TyA + TyB
DNA binding regulatory protein
Reverse transcriptase (fused protein made by frameshift)

FIGURE 15.16 ***Structure of Ty-1 Retrotransposon***

Ty-1 is flanked by two direct repeats (LTRs) and contains the genes for reverse transcriptase and a DNA binding regulatory protein. Reverse transcriptase is only produced if a frameshift occurs during translation. If no frameshift occurs, the truncated gene product binds to DNA and acts as a regulatory protein.

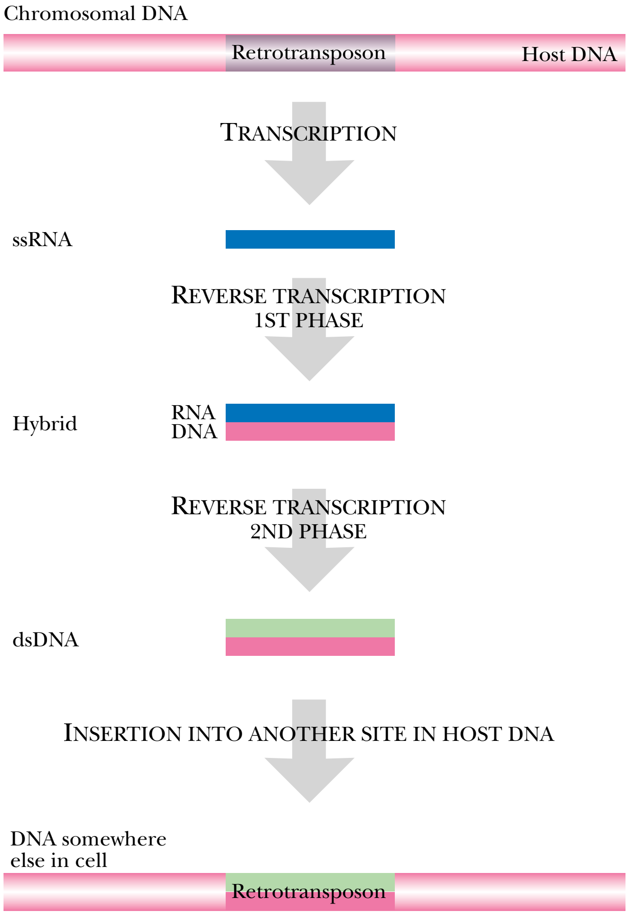

Chromosomal DNA
Retrotransposon
Host DNA
TRANSCRIPTION
ssRNA
REVERSE TRANSCRIPTION 1ST PHASE
Hybrid
RNA
DNA
REVERSE TRANSCRIPTION 2ND PHASE
dsDNA
INSERTION INTO ANOTHER SITE IN HOST DNA
DNA somewhere else in cell
Retrotransposon

FIGURE 15.17 ***Movement of Ty-1 Retrotransposon***

Ty-1 elements use host cell transcription to create a single-stranded RNA. Reverse transcriptase uses the RNA as a template to synthesize a strand of DNA. The DNA/RNA hybrid is next converted into double-stranded DNA. The dsDNA form of the retrotransposon can be inserted into a new target site within the host DNA.

long, contains a gene encoding reverse transcriptase and is flanked by direct repeats of 334 bp that correspond to the **long terminal repeats (LTRs)** of a retrovirus (Fig. 5.16).

When moving, the first step is to make a single stranded RNA copy (Fig. 15.17). The host provides the RNA polymerase II that transcribes the Ty-1 element starting in the left hand LTR. Next, reverse transcriptase makes a double stranded DNA copy. First it makes a single strand of DNA, so creating an RNA/DNA hybrid. A second round of synthesis replaces the RNA with DNA, forming double-stranded DNA. Finally, the DNA is inserted into a new site within the host cell DNA. There are 30 to 40 copies of Ty1 per yeast cell.

long terminal repeats (LTRs) Direct repeats of several hundred base pairs found at the ends of retroviruses and some other retro-elements

Human Genetic Defects due to Retroposon Insertion

Since non-coding DNA vastly outnumbers coding DNA, it is hardly surprising that most insertions of transposable elements into the genome of higher organisms occurs in the non-coding regions. Nonetheless, occasional examples are known where insertion of a retrotransposon inactivates a gene so causing a hereditary defect. The first human case to be identified is a form of muscular dystrophy known as Fukuyama-type congenital muscular dystrophy (FCMD), that is particularly common in Japan although rare elsewhere. FCMD is one of the most common autosomal recessive disorders in Japan, occurring at a frequency of approximately 0.7–1.2 per 10,000 births, with a carrier frequency estimated to be as high as 1 in 80.

This condition is caused by the insertion of a retrotransposon, which is approximately 3,000 bases long into the 3′-untranslated region of the *FCMD* gene. This could theoretically result in an altered secondary structure for the *FCMD* mRNA, possibly rendering it unstable. In any case, little of no detectable mRNA is seen in affected cells. The *FCMD* gene codes for a 461-amino-acid protein that is normally expressed in brain, skeletal muscle, and heart and is involved in muscle function.

Genetic analysis indicates that the retrotransposon insertion in the *FCMD* gene could have been derived from a single ancestor who lived 2,000 to 2,500 years ago. This was about the time that the Yayoi people migrated to Japan from Korea and China, giving rise to the possibility, as yet unproven, that these immigrants brought the defective *FCMD* allele into the Japanese population. [It is thought that humans migrated from the Asian continent to Japan in two waves. The first wave brought hunter-gatherers of the Jomon culture over 10,000 years ago. The second wave brought the Yayoi people from the Korean Peninsula about 2,300 years ago. By around 300 AD the Yayoi had spread through most of Japan. The Yayoi brought metalworking, weaving and rice growing to Japan.]

Some retro elements are closely related to retroviruses.

Retroposons have only one or two genes and lack the ability to make virus particles and infect other cells. Nonetheless, retro-elements that are intermediate between retroposons and retroviruses do exist. Some retro-elements pack their RNA into defective virus-like particles. However, these particles are not released from the cell where they were assembled and therefore cannot infect other cells. These elements are sometimes called endogenous retroviruses. Defective versions of these, known as retrovirus-like elements, are common in animal cells where they often exist in multiple copies.

Repetitive DNA of Mammals

The highly repetitive sequences known as LINEs are related to retrotransposons.

A substantial portion of the DNA of both animals and plants consists of repeated sequences as discussed in Ch. 4. Many of these were probably derived by retrotransposition events. In mammals, there are two major classes of moderate to highly repetitive DNA, **short interspersed elements**, or **SINEs**, and **long interspersed elements**, or **LINEs**. Although these were originally defined in terms of their length, the SINEs are derived from host DNA (see below) whereas the LINEs are derived from retrotransposons.

LINE Long interspersed element
long interspersed element (LINE) Long repeated sequence that makes up a significant fraction of the moderately or highly repetitive DNA of mammals
short interspersed element (SINE) Short repeated sequence that makes up a significant fraction of the moderately or highly repetitive DNA of mammals
SINE Short interspersed element

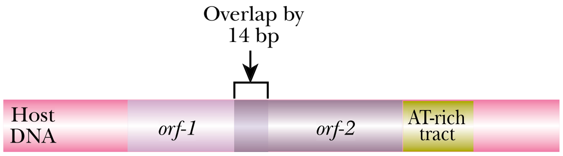


FIGURE 15.18 ***Structure of LINE-1 Element***

LINE-1 elements contain two open reading frames that encode reverse transcriptase followed by an AT-rich region. Unlike other retrotransposons, LINE-1 elements are not flanked by inverted repeats.

Most human LINE-1 sequences are defective due to deletions.

The **LINE-1 (or L1) element** of mammals is a retroposon that moves much as the Ty1 element of yeast. However, it lacks LTR sequences and at its 3′-end is a run of AT base pairs that are derived from the poly-A tail typical of eukaryotic messenger RNA (Fig. 15.18). In humans the LINE-1 element is present in over 100,000 copies and makes up 5 percent or more of the total DNA! The vast majority of the LINE-1 repeats are defective. The complete LINE-1 sequence is 6,500 bp long and contains a gene for reverse transcriptase. However, only about 3,000 of the LINE-1 sequences are full length and most of these are crippled by point mutations.

Very rarely LINE-1 makes a new copy of itself and may insert it somewhere else in the DNA. This movement causes one type of hemophilia, an inherited condition caused by a defect in blood clotting factors. A few very rare cases are due to the insertion of LINE-1 into the gene for blood clotting factor VIII on the X-chromosome. The intact LINE-1 that jumped came from chromosome 22. This still active copy of LINE-1 is found in the same location in the DNA of the gorilla, implying that it has been lurking in the same place for millions of years of primate evolution.

Retro-Insertion of Host-Derived DNA

A pseudogene is a duplicate, but functionless, copy of a gene (see Ch. 4). Pseudogenes may be generated by DNA duplications and so often lie close to the original genes. Pseudogenes made in this way will contain both the introns and exons of the original gene.

Some pseudogenes originated via reverse transcriptase acting on mRNA.

Sometimes the reverse transcriptase of a retro-element may work by accident on mRNA that derives from a host gene. Since mRNA has had the introns removed by splicing, reverse transcriptase generates a DNA copy missing the introns, i.e. **cDNA (complementary DNA)**. Occasionally, this cDNA copy may be re-integrated into the host genome (Fig. 15.19). This will give a pseudogene that lacks the introns and consists solely of the coding sequence and is therefore known as a **processed pseudogene** or a **retro-pseudogene**.

The highly repetitive Alu element is a pseudogene that contains an internal promoter.

The SINEs are a special class of processed pseudogenes that were originally derived from host DNA sequences. The most common SINE is the **Alu element**, which is derived from the **7SL RNA** gene (Fig. 15.20). This non-coding RNA is part of the machinery for protein export across membranes. It is normally transcribed from an internal promoter by RNA polymerase III (see Ch. 6). Standard pseudogenes lack a promoter (unless by chance they were integrated just next to one). Consequently, they cannot be transcribed and no additional copies are produced. In contrast, each copy of the Alu element carries an internal promoter. As long as this promoter remains intact, each Alu element can serve as the source for another round of transcription, reverse transcription and re-integration. As a result the number of Alu elements has mushroomed until there are now around a million copies in the human genome.

7SL RNA Non-coding RNA that forms part of the machinery for protein export across intracellular membranes in eukaryotic cells
Alu element The most common short interspersed element (or SINE) in the highly repetitive DNA of mammalian cells
cDNA (complementary DNA) DNA copy made by reverse transcription from messenger RNA and therefore lacking introns
LINE-1 (or L1) element A particular LINE found in many copies in the genome of humans and other mammals
processed pseudogene Pseudogene lacking introns because it was reverse transcribed from messenger RNA by reverse transcriptase
retro-pseudogene Another name for a processed pseudogene

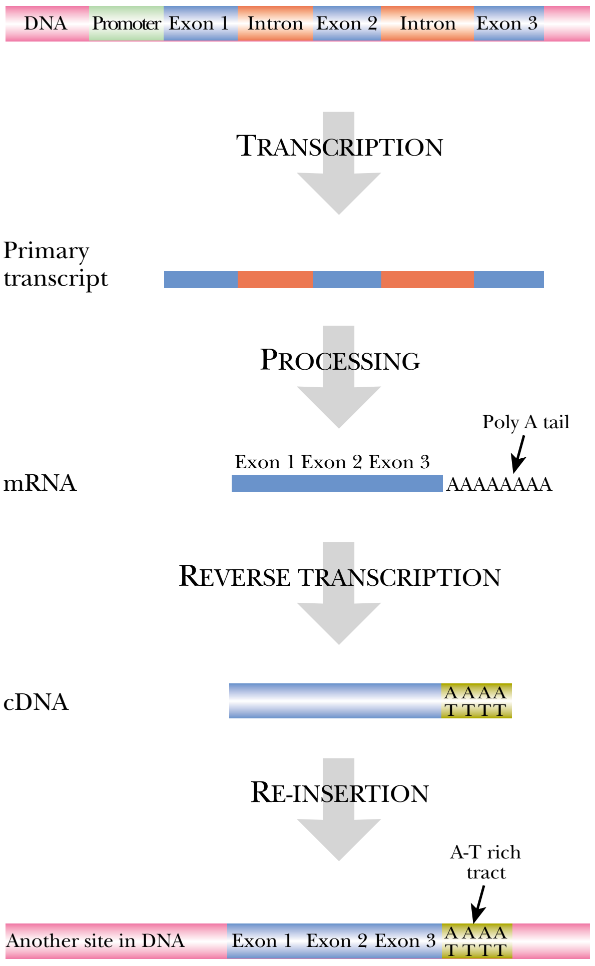


FIGURE 15.19 ***Creation of a Processed Pseudogene***

When a host gene is expressed, the primary RNA transcript is processed so the introns are removed and a poly-A tail is added. Occasionally, reverse transcriptase will convert host mRNA into a cDNA copy. The cDNA may be inserted into the host DNA (at another site) creating a processed pseudogene. This pseudogene will not be expressed as there is no promoter at its new insertion site.

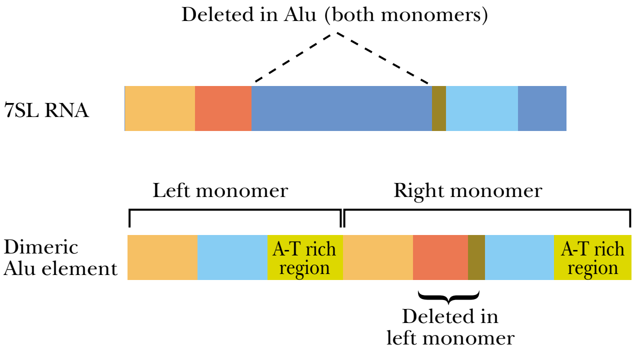


FIGURE 15.20 ***Origin of the Alu Element from 7SL RNA***

7SL RNA has given rise to the Alu element by a complex process involving reverse transcription and re-integration. The sequence of 7SL RNA is shown at the top. The Alu element below is actually derived from two fused 7SL RNA sequences that have suffered several deletions.

Retrons Encode Bacterial Reverse Transcriptase

Retrons, found in bacteria, are shorter, stranger relatives of the retroposons of higher cells. They have just one gene that is transcribed into an RNA molecule consisting of a sizeable untranslated region followed by the coding region for reverse transcriptase (Fig. 15.21).

Retrons are retro-elements found in bacteria, that generate bizarre DNA/RNA hybrid structures.

The reverse transcriptase uses the non-coding portion of the RNA molecule as both a template and a primer to make multiple copies of a bizarre branched molecule that is part RNA and part DNA (Fig. 15.22). Typically, the RNA region of the retron product is around 100 bases long and the DNA 50–75 bases. The DNA is joined to the RNA by the 2′-OH of a specific G residue. In some retrons the single-stranded DNA is cleaved

retron Genetic element found in bacteria that encodes reverse transcriptase and uses it to make a bizarre RNA/DNA hybrid molecule

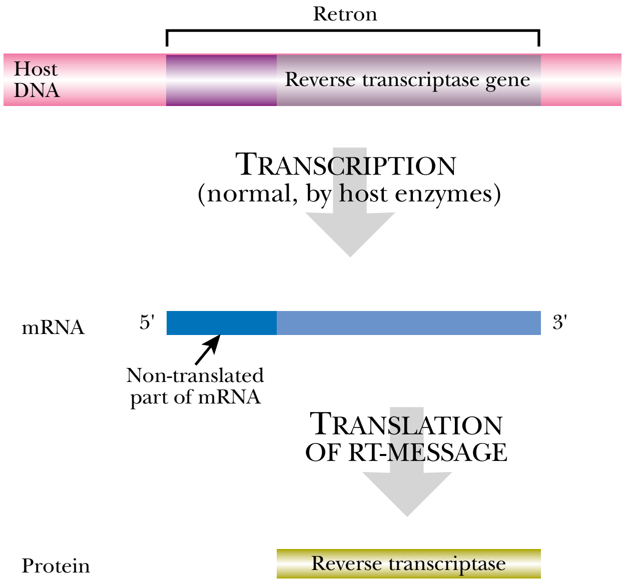

Retron
Host DNA
Reverse transcriptase gene
TRANSCRIPTION (normal, by host enzymes)
mRNA
5'
3'
Non-translated part of mRNA
TRANSLATION OF RT-MESSAGE
Protein
Reverse transcriptase

FIGURE 15.21 ***Structure of a Retron and its Gene Products***

A retron consists of a single gene for reverse transcriptase that is preceded by a long non-coding region that gives rise to untranslated RNA.

off from the RNA, in other cases, the two stay attached. The retron DNA has not been observed to reintegrate into the chromosome. However, the chromosomes of some strains of myxobacteria have repeated sequences related to retron DNA.

It is assumed that retrons move around, but it is unknown how. Perhaps the copies observed so far are defective and the mobile master copy has not yet been found. Retrons are found in relatively few bacteria and only a single copy of the retron is usually found in the bacterial chromosome. Retrons are often inserted into the DNA of bacterial viruses that have in turn inserted into the bacterial chromosome. Where the virus picked them up remains obscure.

The Multitude of Transposable Elements

We have discussed a wide range of transposable elements that vary in their mechanism of movement, and whether or not they have an RNA phase. Other transposons exist whose detailed mechanisms of action are variations on the above themes. Some larger and more complicated elements possess the ability to use multiple mechanisms, under different conditions. In others a variety of extra genes oversee the frequency of transposition or the specificity of insertion. For example, Tn7 normally inserts at only a single specific sequence on the *E. coli* chromosome. However, mutants exist that have inactive specificity proteins. These insert more or less at random like more typical transposons.

Other genetic elements exist that possess the characteristics both of transposable elements and some other life-form, such as a plasmid or virus. For example, **conjugative transposons** both transpose and promote conjugation like fertility plasmids. Again, bacteriophage Mu is both a virus and a transposon. Plasmids and viruses are discussed in Chapters 16 and 17. However, the transposon-like aspects of these hybrid elements are discussed below.

Hybrid elements that combine the properties of virus and transposon are known.

Bacteriophage Mu is a Transposon

Hybrid gene creatures exist that possess the characteristics both of transposable elements and some other genetic element. For example, **bacteriophage Mu** is both a virus and a transposon. (Note that we are not talking about a virus that carries a transposon inserted within its DNA—a frequent occurrence—but about a genetic element that behaves as both a virus and a transposon simultaneously). When Mu DNA enters *E.*

bacteriophage Mu A bacterial virus that replicates by transposition and causes mutations by insertion within host cell genes
conjugative transposon A transposon that is also capable of transferring itself from one bacterial cell to another by conjugation

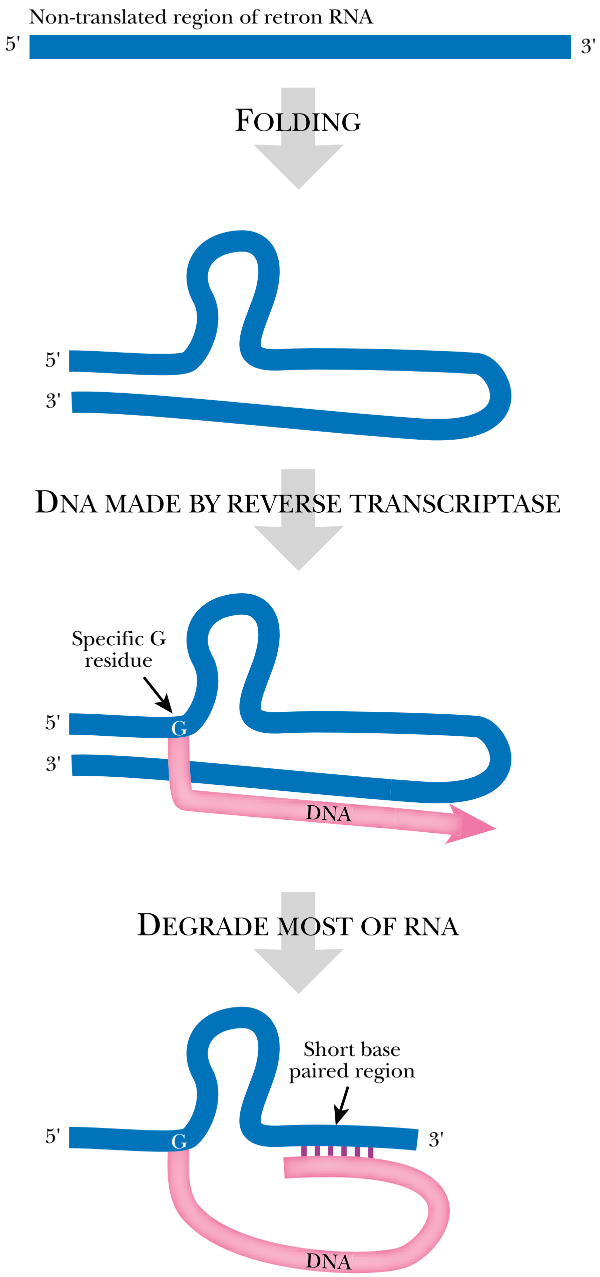


FIGURE 15.22 ***Retron RNA and RNA/DNA Hybrid***

Retron RNA only contains the gene for reverse transcriptase and a large 5′ untranslated region. This folds into a hairpin structure so that the 5′ end meets the 3′ end. Reverse transcriptase recognizes a specific guanine near the front of the hairpin and begins DNA synthesis here. The RNA is partly degraded, resulting in a strange RNA/DNA hybrid structure.

coli, its bacterial host, it integrates at random into the host chromosome by transposition (Fig. 15.23). In other words, the whole of the Mu genome is a transposon. If Mu inserts into the middle of a host gene this will be inactivated. Early investigators noticed that infection with this virus caused frequent mutations and therefore named it Mu for "mutator" phage.

Like many bacterial viruses, Mu may lie dormant as a prophage or go lytic (see Ch. 17). When Mu replicates it does so by uncontrolled replicative transposition, not by replicating as a virus. This results in multiple copies of Mu inserted into the host DNA and destroys so many host genes that it is inevitably lethal. Unlike other viruses, the DNA of Mu is never found replicating as an independent molecule, free of the chromosome. The segments of DNA packaged into the virus particle contain a whole Mu genome plus small stretches of host DNA at the ends. Thus even inside the virus particle, Mu DNA is still inserted into host DNA! When the virus DNA infects a new cell, the Mu genome transposes out of the tiny fragment of host DNA it brought with it. Thus Mu is a true transposon and its DNA is never free.

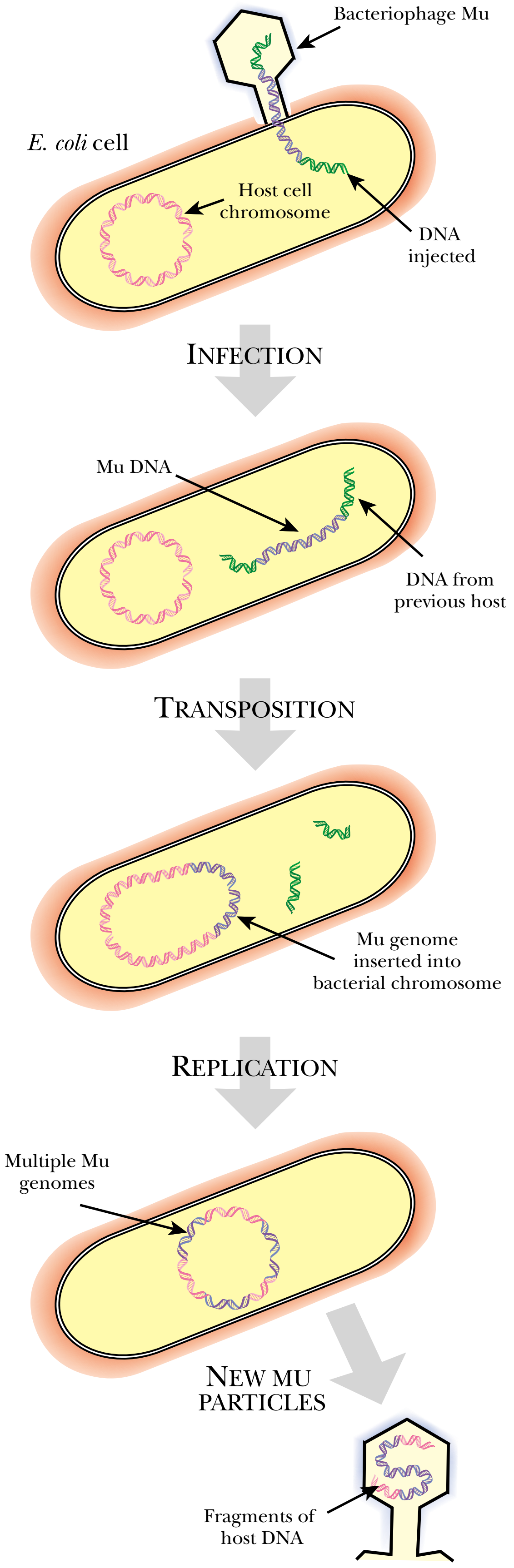


FIGURE 15.23
Bacteriophage Mu is a Transposon

Bacteriophage Mu attaches to the *E. coli* cell and injects its DNA into the cytoplasm. Once inside, the Mu DNA inserts into the host chromosome via transposition. Notice that the flanking DNA from the previous host is not inserted. Once integrated, Mu DNA may undergo so many transpositions that cellular functions are destroyed. When the Mu DNA is packaged into virus particles, short lengths of host chromosome are also packaged attached to the Mu ends; therefore, Mu DNA is always integrated into host DNA.

Note that many viruses may integrate into the host chromosome. This includes bacterial viruses such as lambda and animal viruses such as the retroviruses (see Ch. 17). Nonetheless, these viruses do not replicate by transposition and are not surrounded by host DNA when packaged into their virus particles. Thus neither lambda nor retroviruses are transposons.

Conjugative Transposons

Conjugative transposons combine the ability to move by transposition and to move from one bacterial cell to another.

Conjugative transposons, found in bacteria, are hybrid elements that can both transpose and can move from cell to cell, like a transmissible plasmid. The first of these to be discovered, Tn916, confers tetracycline resistance and was found in the bacterium *Enterococcus faecalis*. Tn916 carries several genes needed for conjugative transfer and is therefore much larger than most transposons.

Tn916 jumps by the cut-and-paste mechanism. However, it differs in two respects from typical conservative transposons. First, the target sequence is not duplicated when Tn916 inserts itself. Second, it can excise itself precisely, leaving the host cell DNA intact. When moving from one bacterial cell to another, Tn916 is thought to excise itself temporarily from the DNA of the original cell. It then transfers itself into the recipient and, once inside, it transposes into the DNA of the new host cell (Fig. 15.24). Tn916 and related elements can enter many different groups of bacteria, both gram-positive and gram-negative. Because the host range of conjugative transposons is so broad they are partly to blame in the spread of antibiotic resistance genes among diverse groups of bacteria.

Integrons Collect Genes for Transposons

Many bacteria that carry multiple drug resistance have emerged since antibiotic use has become widespread. Antibiotic resistance genes are usually carried on plasmids, many of which may be transferred between bacteria. In many cases the antibiotic resistance genes are actually carried within transposons that are inserted into the plasmids.

Novel antibiotic resistance genes often appear first in transposons of the Tn21 family found in gram-negative bacteria. This group of transposons possesses an internal element known as an **integron** that acts as a gene acquisition and expression system. An integron consists of a recognition region, the *attI* site, into which a variety of gene cassettes may be integrated, plus a gene encoding the enzyme responsible for insertion, the **integrase**. The *attI* site is flanked by two 7 bp sequences that act as recognition sites for the integrase (Fig. 15.25). Two promoters are situated upstream of the variable region. One is for the integrase gene; the other faces into the variable region and drives transcription of whatever gene has been integrated.

Integrons accumulate genes by integration of DNA modules flanked by recognition sequences.

Gene cassettes suitable for integration consist of a structural gene lacking its own promoter plus an integrase recognition sequence, the *attC* site. The *attC* sites are rather variable, except for the conserved 7 bp sequences at the ends. [The *attC* sites were originally referred to as “59 base elements” because the first examples discovered were 59 bp long. However, later examples were found that varied in length and internal sequence.] Gene cassettes may exist temporarily as free circular molecules incapable of replication and gene expression or else integrated into the *attI* site of an integron. The ultimate source of the genes on the gene cassettes is presently obscure. The reason why most known integron cassettes carry genes for antibiotic resistance is presumably due to observer bias—antibiotic resistance is clinically important and therefore noticed more often.

integrase Enzyme that integrates one segment of DNA into another DNA molecule at a specific site
integron Genetic element consisting of an integration site plus a gene encoding an integrase

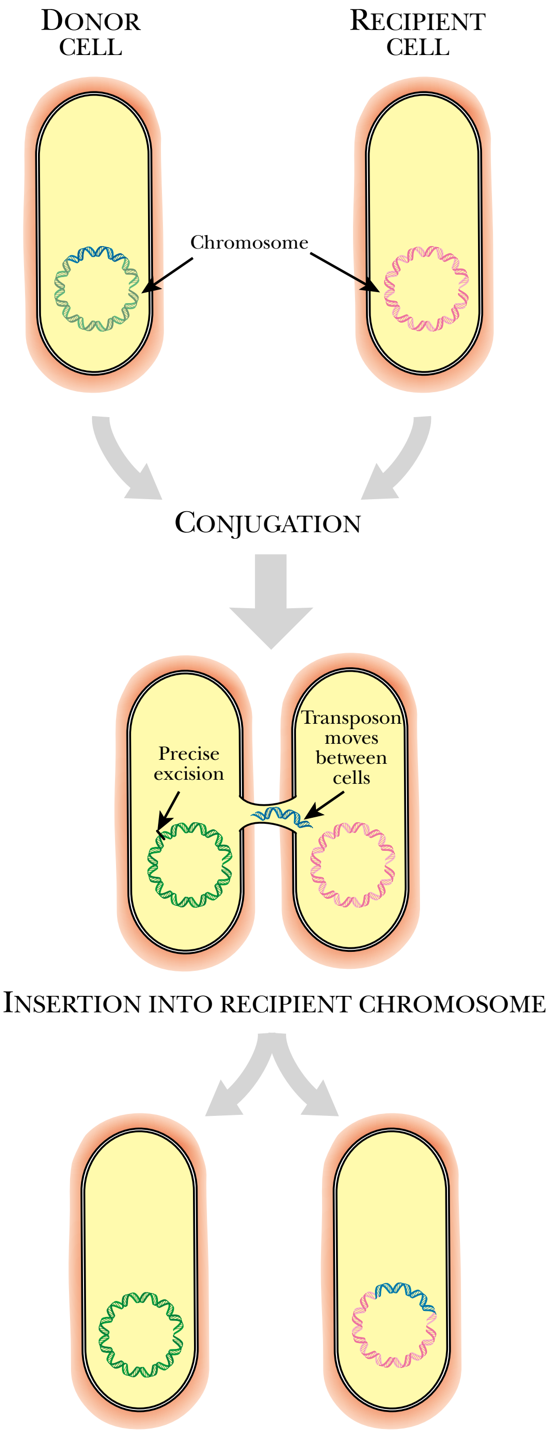

DONOR CELL
RECIPIENT CELL
Chromosome
CONJUGATION
Precise excision
Transposon moves between cells
INSERTION INTO RECIPIENT CHROMOSOME

FIGURE 15.24 ***Conjugative Transposon***

Conjugative transposons move from donor cell to recipient cell during bacterial conjugation. The transposon moves in a precise manner, leaving the donor DNA intact and integrating into the recipient DNA without duplicating the target sequence.

Transposons of the Tn21 family are widespread and frequently trade antibiotic resistance genes. The Tn2501 transposon carries no antibiotic resistance genes and appears to have an "empty" integron. It is possible for a single integron to collect multiple genes, each flanked by 7 bp boxes. Other similar integrons are found on various plasmids and transposons of other families.

Although most integrons are located on transposons and/or plasmids, a few are found in the chromosomes of gram-negative bacteria. Such chromosomal integrons may collect several hundred genes and are then known as super-integrons. The best known example is on the second chromosome of *Vibrio cholerae* (causative agent of cholera) and has collected approximately 200 genes, mostly of unknown function and unknown origin.

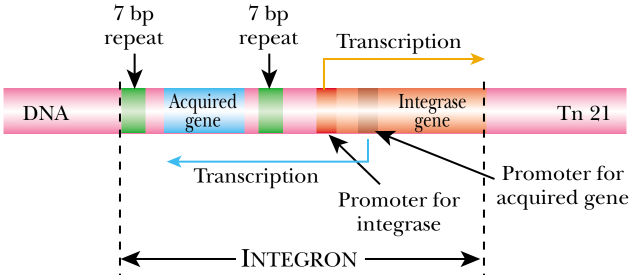


FIGURE 15.25 ***Integrons Collect Antibiotic Resistance Genes***

The structure of the integron shows two regions. The upstream region has two 7 bp repeat sequences that are recognition sites for integrase. The downstream region contains the gene for integrase. Expression of the integrase gene causes the capture of various other genes, most noticeably antibiotic resistance genes. Once integrated, these captured genes are expressed from a promoter within the integrase gene.

Junk Philosophy—A Rant from your Author: The situation in the eukaryotic genome reminds me of modern society. In the human genome a small proportion of coding DNA is outnumbered by repetitive and/or non-coding DNA, much of it junk. Similarly, most of the e-mail I get is spam and most of my regular mail is junk mail. Like selfish DNA most of "my" mail is promoting someone else's advertising agenda. Add to this urban legends, distortions generated by the media, propaganda put around by politicians, new religious cults and alternative medicine. Perhaps it is inevitable that in any complex system, whether a cellular genome or a human society, that most of the information is valueless, false or parasitic.

Junk DNA and Selfish DNA

DNA sequences that perform no useful function but merely "inhabit" the chromosomes of other organisms may be regarded as genetic parasites of a very degenerate kind. They are not merely sub-cellular but sub-molecular parasites as they are always found as constituents of other DNA molecules. These include the assorted transposable elements discussed above, both DNA-based and retro-elements.

This general type of DNA has been named "**selfish DNA**" because it behaves as if motivated by its own interests, not those of the host DNA. If the host DNA is degraded, the parasite dies with it. Thus the spread of selfish DNA is limited by the need to avoid destroying the host DNA molecule or inactivating too many host cell genes. Apart from such considerations, the selfish DNA multiplies inside its host DNA molecule just like a virus replicating inside a cell, or an infectious bacterium multiplying inside a patient. In most higher organisms, a substantial proportion of the DNA consists of multiple copies of such selfish DNA. Why doesn't the host cell purge its chromosome of these parasites? In small, efficient, fast-growing cells, like bacteria or even yeasts, there is very little selfish DNA. Presumably cells with a significant burden of extra DNA divide more slowly and are weeded out by competition. In large, slow-growing cells the parasites replicate faster relative to the host DNA and gradually increase in numbers.

Junk DNA is mostly derived from mobile selfish DNA that has degenerated.

Most selfish DNA is probably the remains of viruses or transposons that inserted into the chromosome long ago. Over long periods of time the copies diverge both by single base mutations and by deletions. Eventually most of the copies become defective and lose the ability to form virus particles or to move around; they degenerate into mere "**junk DNA**" (Fig. 15.26). The genomes of eukaryotic cells generally contain large amounts of junk DNA, indeed the great majority of mammalian DNA consists of non-coding sequences of one kind or another.

junk DNA Defective selfish DNA that is of no use to the host cell it inhabits and is no longer capable of either moving or expressing its genes
selfish DNA Any segment of DNA that replicates but which is of no use to the host cell it inhabits

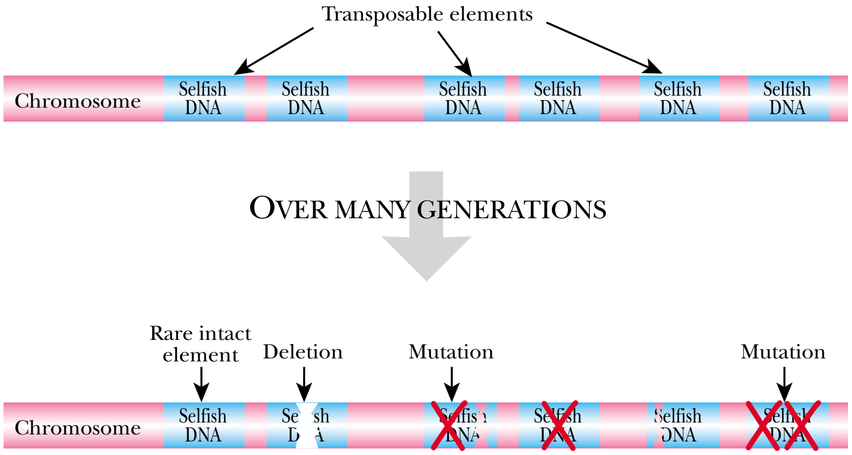


FIGURE 15.26 ***Junk DNA is Defective Selfish DNA***

Transposable elements are often termed selfish DNA because they are parasitic DNA sequences that inhabit a host genome. Over time, many copies of selfish DNA are inactivated by mutations and deletions, leaving DNA remnants called junk DNA.

Homing Introns

Homing introns are mobile DNA segments that can occupy only a single site within a host cell gene.

Mobile DNA does not only consist of transposable elements. **Homing introns** are a strange and relatively rare type of mobile DNA. As their name indicates, homing introns are intervening sequences that are inserted into genes between two exons. Each homing intron is located in one unique position within one particular gene of the cell it inhabits. This target gene can exist in two versions, with or without the homing intron inserted.

Movement of homing introns is very restricted as they can only occupy this one specific site. Mobilization can only happen when a cell contains two copies of the target gene, one with and the other without the homing intron. The homing intron will then insert itself into the target gene that lacks a copy of the intron. In eukaryotes, this situation may occur after mating when chromosomes from two different parental cells come together in the zygote. In bacteria, it may occur when DNA enters a recipient cell via any of a variety of mechanisms (see Ch. 18).

Homing introns encode an endonuclease that is responsible for their movement. The endonuclease cleaves a recognition sequence within the target gene and generates short overhanging ends. This double-stranded break triggers a gene conversion event (see Ch. 14) in which the intact version of the gene is copied and used to repair the break (Fig. 15.27). Thus the homing intron merely encodes an enzyme to cut the DNA and leaves the host cell to repair the damage. Insertion of the homing intron disrupts the recognition sequence. Thus the endonuclease only cuts the target gene if the intron is absent. The recognition sequences of homing introns may be as long as 18–20 bp and are the longest and most specific known for any nuclease. This ensures that the intron inserts only into a single unique site in the genome of each cell.

Homing introns of group I use simple endonucleases as described above. These are found in various bacteria and lower eukaryotes. Homing introns of group II are more complex. They are found in bacteria and the organelles of lower eukaryotes. Group II homing introns are retro-elements that use an RNA intermediate. The protein they encode has both endonuclease and reverse transcriptase activity. As before, the endonuclease makes a double-stranded break in the middle of the recognition sequence. Next, the reverse transcriptase generates a DNA copy of the intron for inserting into the break. For a template, it uses the primary transcript from the copy of the target gene that contains the intron (Fig. 15.28). The free 3′-OH generated by the endonuclease cleavage is used as primer for starting DNA synthesis. Consequently, the DNA copy is made already attached to the edge of the double-stranded break.

homing intron A mobile intron that encodes a protein enabling it to insert itself into a recognition sequence within a target gene

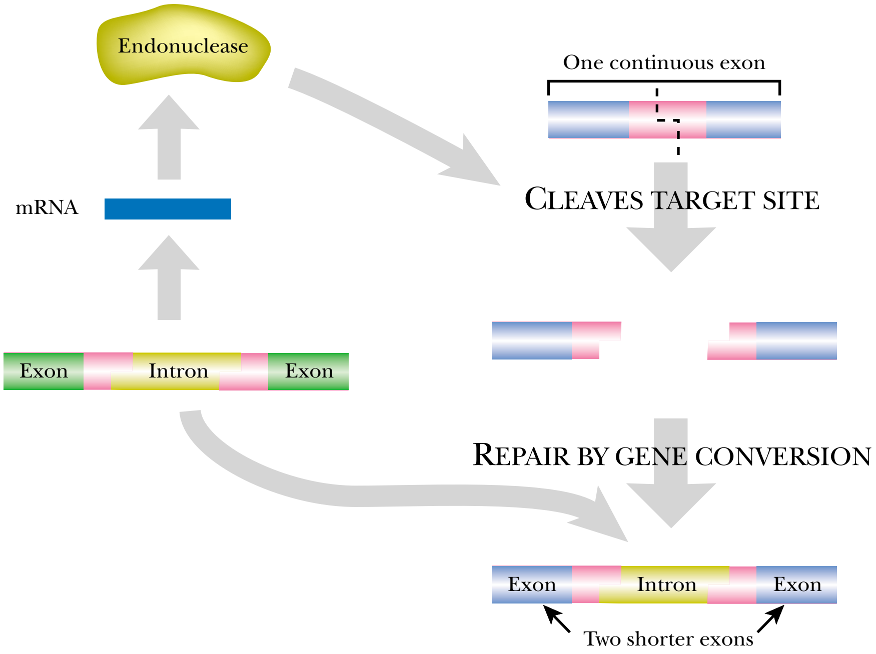


FIGURE 15.27 ***Homing Intron Inserts in a Unique Location***

A homing intron contains a single gene for an endonuclease. This enzyme cleaves a very specific target site, which is only found in a copy of the target gene not containing the homing intron. This situation only occurs when two copies of the target gene are present in a cell, one with the homing intron and one without. The double-stranded break tricks the host cell into repairing the break by gene conversion, thus duplicating the homing intron.

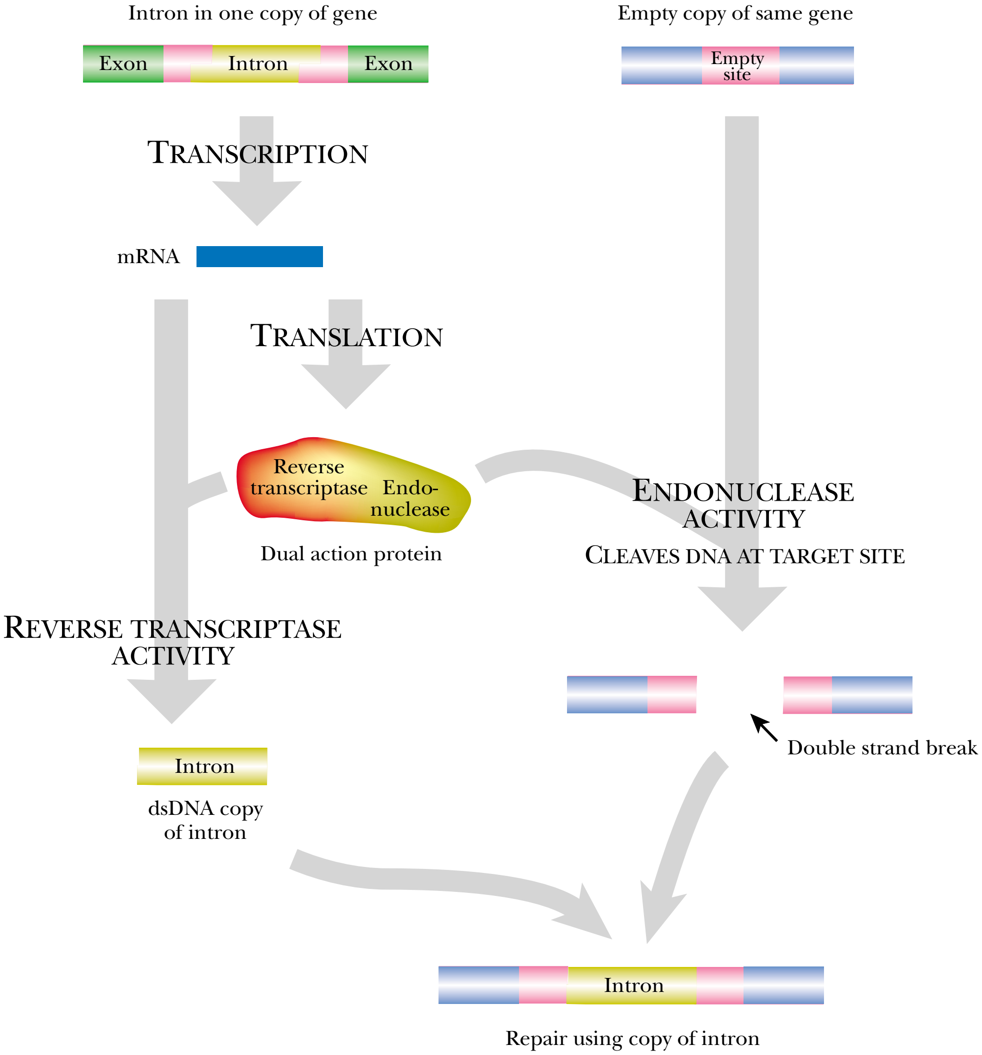


FIGURE 15.28 ***Homing Retro-Intron Inserts via RNA Intermediate***

Group II homing introns integrate themselves via reverse transcriptase rather than gene conversion. The homing intron expresses an enzyme with both reverse transcriptase and endonuclease activity. The endonuclease generates a double-stranded break at the specific target site. The reverse transcriptase generates a DNA copy of the intron using the 3′-OH of the double-stranded break as a primer for synthesis and the mRNA of the intron as a template. These two functions are shown separated in the figure for clarity. However, in real life, the dsDNA copy of the intron made by reverse transcriptase is attached to the 3′-end of the double-stranded break while it is being made.

CHAPTER **SIXTEEN**

Plasmids

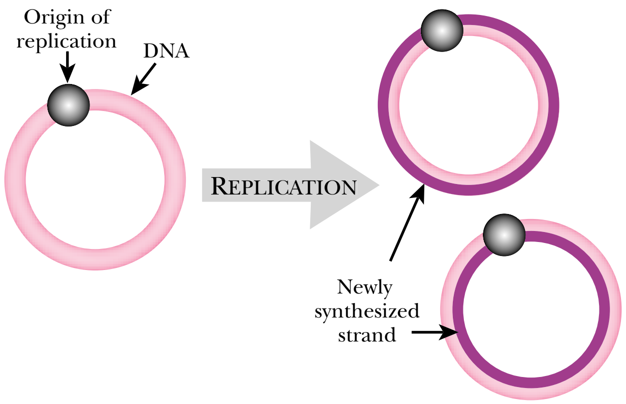


FIGURE 16.01 ***Plasmids are Self-Replicating Molecules of DNA***

Plasmids are most often rings of double-stranded DNA found inside cells but not attached to or associated with the chromosomal DNA. The plasmid carries its own origin of replication, thus it is considered a true replicon.

Plasmids as Replicons

Plasmids are "extra" self-replicating molecules of DNA that are found in many cells.

Plasmids are autonomous self-replicating molecules of DNA (or very rarely RNA) (Fig. 16.01). They are not chromosomes, although they do reside inside living cells and carry genetic information. They are not regarded as part of the cell's genome for two reasons. First, a particular plasmid may be found in cells of different species and may move from one host species to another. Second, a plasmid may sometimes be present and sometimes absent from the cells of a particular host species. Thus, although plasmids carry genetic information that may be expressed, they are not a constant part of the cell's genetic make-up nor are they needed for cell growth and division under normal conditions.

As discussed previously, **replicons** are self-replicating molecules of nucleic acid. Chromosomes, plasmids, virus genomes (both DNA and RNA) and viroids are all replicons. Strictly speaking, a replicon is defined by the possession of its own origin of replication where DNA (or RNA) synthesis is initiated. Thus, a replicon need not carry genes that encode the enzymes needed for its own replication, nor is it necessarily responsible for generating its own nucleotide precursors or energy. This means that plasmids and viruses are replicons, even though they rely on the host cell to provide energy, raw materials and many enzyme activities.

Plasmids and viruses both rely on the host cell to provide energy and raw materials but plasmids do not damage the host cell.

Plasmids may be regarded as living creatures in their own right. Just as worms wriggle through the soil and fish float in the sea, so plasmids proliferate inside their host cells. To a plasmid, the cell is its environment. So, although the plasmid is not alive in the same sense as a cell, neither is it merely part of the cell. In some ways plasmids are like domesticated viruses that have lost the ability to move from cell to cell killing as they go. Plasmids maintain some viral characteristics since the plasmid requires the host cell for replication enzymes, energy, and raw materials. Unlike viruses though, plasmids do not possess protein coats and since they cannot leave the cell they live in, they avoid damaging it. Viruses usually destroy the cell in which they replicate and are then released as virus particles to go in search of fresh victims. Plasmids replicate in step with their host cell (Fig. 16.02). When the cell divides, the plasmid divides and each daughter cell gets a copy of the plasmid.

It is easy to see how a virus that has lost the genes for its protein coat and/or for killing the host cell might evolve into a plasmid. Furthermore, certain genetic elements, such as P1 (see below), can switch between the two lifestyles and may live either as a plasmid or as a virus. It is also possible to imagine how a plasmid might pick up coat protein genes and/or killing functions and deregulate its DNA replication so evolving into a virus. Indeed, many plasmids possess host killing functions that they use to ensure that they are not lost by the host cell (see plasmid addiction, below). So which

plasmid Self-replicating genetic elements that are sometimes found in both prokaryotic and eukaryotic cells. They are not chromosomes nor part of the host cell's permanent genome. Most plasmids are circular molecules of double stranded DNA although rare linear plasmids and RNA plasmids are known

replicon Molecule of DNA or RNA that contains an origin of replication and can self-replicate

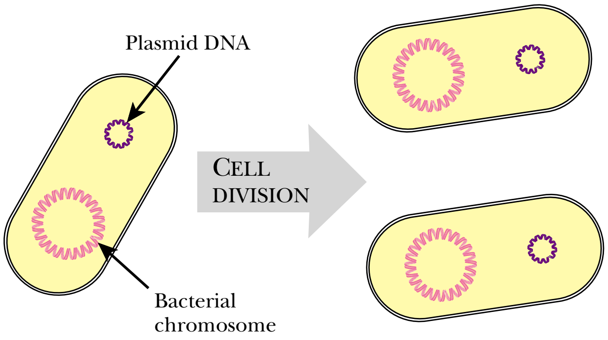


FIGURE 16.02 ***Plasmids Replicate in Step with the Host Cell***

When a bacterial cell is ready to divide, the replication machinery also duplicates the plasmid DNA. The two copies of the chromosome and two copies of the plasmid are then divided equally between the daughter cells. The replication of the plasmid does not harm the cell.

came first, plasmids or viruses? Here we have a true chicken and egg situation. Almost certainly some present day plasmids are derived from viruses and equally certainly some present day viruses are derived from plasmids. However, the ultimate origins of both kinds of element remain obscure.

General Properties of Plasmids

Most plasmids are circular, made of DNA, and much smaller than chromosomes.

Plasmids are usually circular molecules of DNA, although occasionally, plasmids that are linear or made of RNA exist. They may be found as single or multiple copies and may carry from half a dozen to several hundred genes. Plasmids can only multiply inside a host cell. Most plasmids inhabit bacteria, and indeed around 50% of bacteria found in the wild contain one or more plasmids. Plasmids are also found in higher organisms such as yeast and fungi. The 2μ circle of yeast (see below) is a well-known example that has been modified for use as a cloning vector.

Some plasmids are present in one or two copies per cell whereas others occur in multiple copies.

The **copy number** is the number of copies of the plasmid in each bacterial cell. For most plasmids it is one or two copies per chromosome but it may be as many as 50 or more for certain small plasmids such as the ColE plasmids. The number of copies influences the strength of plasmid-borne characteristics, especially antibiotic resistance. The more copies of the plasmid per cell, the more copies there will be of the antibiotic resistance genes, and therefore, higher the resulting level of antibiotic resistance.

The size of plasmids varies enormously. The F-plasmid of *E. coli* is fairly average in this respect and is about 1 percent the size of the *E. coli* chromosome. Most multicopy plasmids are much smaller (ColE plasmids are about 10 percent the size of the F-plasmid). Very large plasmids, up to 10 percent of the size of a chromosome, are sometimes found but they are difficult to work with and few have been properly characterized.

Plasmids carry genes for managing their own life cycles and some plasmids carry genes that affect the properties of the host cell. These properties vary greatly from plasmid to plasmid, the best known being resistance to various antibiotics. **Cryptic plasmids** are those that confer no identifiable phenotype on the host cell. Cryptic plasmids presumably carry genes whose characteristics are still unknown. A wide variety of plasmids, modified for different purposes, are used in molecular biology research and are often used to carry genes during genetic engineering.

The host range of plasmids varies widely. Some plasmids are restricted to a few closely related bacteria; for example, the F-plasmid only inhabits *E. coli* and related enteric bacteria like *Shigella* and *Salmonella*. Others have a wide host range; for example, plasmids of the P-family can live in hundreds of different species of bacteria. Although "P" is now usually regarded as standing for "promiscuous", due to their unusually wide host range, these plasmids were originally named after *Pseudomonas*, the bacterium in which they were discovered. They are often responsible for resistance to multiple antibiotics, including penicillins.

copy number The number of copies of a plasmid found within a single host cell
cryptic plasmid A plasmid that confers no identified characteristics or phenotypic properties

Plasmid or Chromosome?

When the genome of the Gram negative bacterium *Vibrio cholerae,* the causative agent of cholera, was sequenced it was found to consist of two circular chromosomes of 2,961,146 and 1,072,314 base pairs. Together this totals approximately 4 million base pairs and encodes about 3,900 proteins—about the same amount of genetic information as *E. coli*. Many genes that appear to have origins outside the enteric bacteria, as deduced from their different base composition, were found on the small chromosome. Many of these lack homology to characterized genes and are of unknown function. The small chromosome also carries an integron gene capture system (see Ch. 15) and host 'addiction' genes that are typically found on plasmids (see below). It seems likely that the "small chromosome" originated as a plasmid that has grown to its present size by accumulating genes from assorted external sources. The large chromosome carries almost all of the genes needed for vital cell functions such as protein, RNA and DNA synthesis as well as genes for pathogenicity.

Some plasmids can transfer themselves between bacterial cells and a few can also transfer chromosomal genes.

Certain plasmids can move themselves from one bacterial cell to another, a property known as **transferability**. Many medium sized plasmids, such as the F-type and P-type plasmids, can do this and are referred to as Tra$^+$ (transfer positive). Since plasmid transfer requires over 30 genes, only medium or large plasmids possess this ability. Very small plasmids, such as the ColE plasmids, simply do not have enough DNA to accommodate the genes needed. Nonetheless, many small plasmids, including the ColE plasmids, can be **mobilized** by self-transferable plasmids, i.e. they are Mob$^+$ (mobilization positive). However, not all transfer-negative plasmids can be mobilized. Some transferable plasmids (e.g. the F-plasmid) can also mobilize chromosomal genes. It was this observation that allowed the original development of bacterial genetics using *E. coli*. The mechanism of plasmid transfer and the conditions necessary for transfer of chromosomal genes are therefore discussed in Chapter 18, Bacterial Genetics.

Plasmid Families and Incompatibility

Plasmids are classified into families whose members share very similar replication genes.

Two different plasmids that belong to the same family cannot coexist in the same cell. This is known as **incompatibility**. Plasmids were originally classified by incompatibility and so plasmid families are often known as incompatibility groups and are designated by letters of the alphabet (F, P, I, X, etc.). Plasmids of the same incompatibility group have very similar DNA sequences in their replication genes, although the other genes they carry may be very different. It is quite possible to have two or more plasmids in the same cell as long as they belong to different families. So a P-type plasmid will happily share the same cell with a plasmid of the F-family (Fig. 16.03).

Occasional Plasmids are Linear or Made of RNA

Although most plasmids are circular molecules of DNA there are occasional exceptions. Linear plasmids of double-stranded DNA have been found in a variety of bacteria and in fungi and higher plants. The best-characterized linear plasmids are found in those few bacteria such as *Borrelia* and *Streptomyces* that also contain linear chromosomes (see Ch. 5). Linear DNA replicons in bacteria are not protected by

incompatibility The inability of two plasmids of the same family to co-exist in the same host cell
mobilizability Ability of a non-transferable plasmid to be moved from one host cell to another by a transferable plasmid
transferability Ability of a plasmid to move itself from one host cell to another

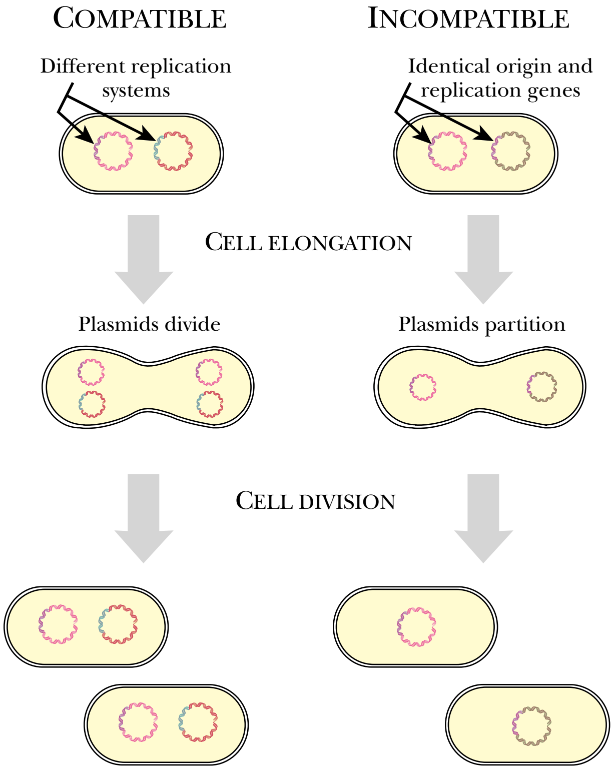


FIGURE 16.03 ***Plasmid Incompatibility***

Plasmids with different origins of replication and different replication genes are able to inhabit the same bacterial cell and are considered compatible (left). During cell division, both types of plasmid replicate; therefore, each daughter cell will inherit both plasmids, just like the mother cell. On the other hand, if two plasmids have identical origins and replication genes they are incompatible will not be replicated during cell division (right). Instead, the two plasmids are partitioned into different daughter cells. [Bacterial cells also contain circular chromosomes that divide in synchrony with cell division; however these have been omitted from this figure.]

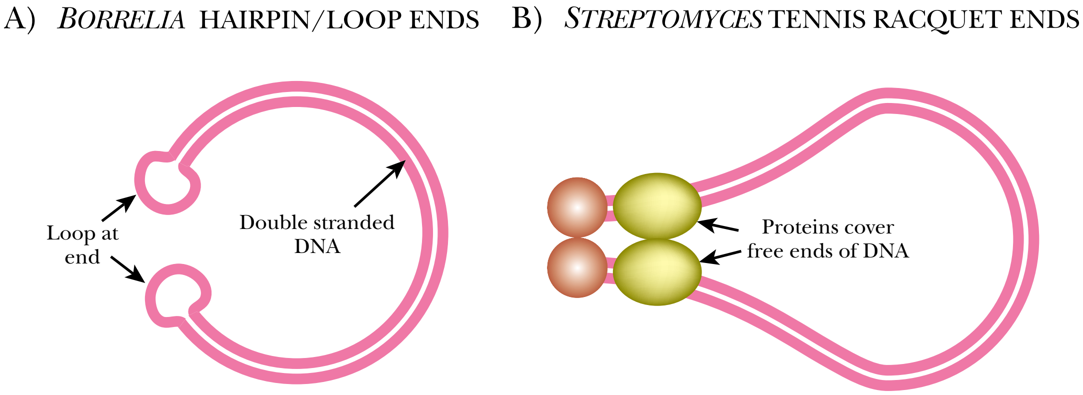

A) *BORRELIA* HAIRPIN/LOOP ENDS
B) *STREPTOMYCES* TENNIS RACQUET ENDS
Loop at end
Double stranded DNA
Proteins cover free ends of DNA

FIGURE 16.04 ***End Structures of Linear Plasmids***

A) Linear plasmids of *Borrelia* form hairpin loops at the ends. B) Linear plasmids of *Streptomyces* are coated with proteins that protect the DNA ends. If linear plasmids had exposed double-stranded ends, this might trigger recombination, repair, or degradation (Ch. 14).

Linear plasmids have special structures to protect the ends of the DNA.

telomeres like the linear chromosomes of eukaryotes. Instead a variety of individual adaptations protect the ends from endonucleases.

In *Borrelia* there are not actually any free DNA ends. Instead hairpin sequences of single-stranded DNA form loops at the ends of both linear plasmids and chromosomes (Fig. 16.04A). Some animal viruses, such as the iridovirus that causes African swine fever, have similar structures. Different species of *Borrelia* cause Lyme's disease and relapsing fever. Their linear plasmids appear to encode both hemolysins that damage blood cells and surface proteins that protect the bacteria from the host immune system. Thus, as is true of many other infectious bacteria, the virulence factors of *Borrelia* are also largely plasmid borne.

The linear plasmids of *Streptomyces* are indeed genuine linear DNA molecules with free ends. They have inverted repeats at the ends of the DNA that are held together by proteins. In addition, special protective proteins are covalently attached to the 5′-ends of the DNA. The net result is a tennis racket structure (Fig. 16.04B). The DNA of adenovirus, most linear eukaryotic plasmids and some bacterial viruses show similar structures.

Linear plasmids are also found among eukaryotes. The fungus *Flammulina velutipes,* commonly known as the enoki mushroom, has two very small linear plasmids within its mitochondria. The dairy yeast, *Kluyveromyces lactis,* has a linear plasmid that normally replicates in the cytoplasm. However, on occasion the plasmid relocates to the nucleus where it replicates as a circle. Circularization is due to site specific recombination involving the inverted repeats at the ends of the linear form of the plasmid. The physiological role of these plasmids is obscure.

RNA plasmids are rare and most are poorly characterized. Examples are known from plants, fungi and even animals. Some strains of the yeast, *Saccharomyces cerevisiae*, contain linear RNA plasmids. Similar RNA plasmids are found in the mitochondria of some varieties of maize plants. RNA plasmids are found as both single-stranded and double-stranded forms and replicate in a manner similar to certain RNA viruses. The RNA plasmid encodes RNA-dependent RNA polymerase that directs its own synthesis. Unlike RNA viruses, RNA plasmids do not contain genes for coat proteins. Sequence comparisons suggest that these RNA plasmids may have evolved from RNA viruses that have taken up permanent residence after losing the ability to move from cell to cell as virus particles.

Plasmid DNA Replicates by Two Alternative Methods

Most plasmids undergo bidirectional replication like bacterial chromosomes.

Typical plasmids made of circular dsDNA use two alternative mechanisms for replicating their DNA. Most plasmids replicate like miniature bacterial chromosomes (see Ch. 5 for details of chromosome replication). They have an origin of replication where the DNA opens and replication begins. Then two replication forks move around the circular plasmid DNA in opposite directions until they meet (Fig. 16.05). A few very tiny plasmids have only one replication fork that moves around the circle until it gets back to the origin.

Some plasmids and many viruses use the rolling circle mechanism for replication.

The other replication mechanism is **rolling circle replication**, which is used by some plasmids and quite a few viruses. At the origin of replication, one strand of the double stranded DNA molecule is nicked (Fig. 16.06). The other, still circular, strand starts to roll away from the broken strand. This results in two single stranded regions of DNA, one belonging to the broken strand and one that is part of the circular strand. DNA is now synthesized starting at the end of the broken strand, which is therefore elongated (Fig. 16.06). The circular strand is used as a template and the gap left where the two original strands rolled apart is filled in. This process of rolling and filling in continues. Eventually the original broken strand is completely unrolled and the circular strand is fully paired with a newly made strand of DNA. We now have a single strand of DNA, equal in length to the original DNA circle, dangling loose.

What happens next varies, depending on the circumstances. For simple plasmid replication, the unrolled old strand is used as a template to make a complementary strand. This double-stranded region is cut free and circularized to give a second copy of the plasmid.

Transferable plasmids use the rolling circle mechanism during transfer but bidirectional replication when dividing in step with the host cell.

Some plasmids, such as the F-plasmid of *E. coli*, can transfer themselves from one bacterium to another. Such plasmids have two separate origins of replication. They divide by bi-directional replication (also known as vegetative replication) when their host cell divides, but use the rolling circle mechanism when they move from one cell to another during conjugation (see Ch. 18). Bi-directional replication starts at *oriV*, the origin of vegetative replication, which is at a different site on the plasmid from *oriT*,

rolling circle replication Mechanism of replicating double stranded circular DNA that starts by nicking and unrolling one strand and using the other, still circular, strand as a template for DNA synthesis. Used by some plasmids and viruses

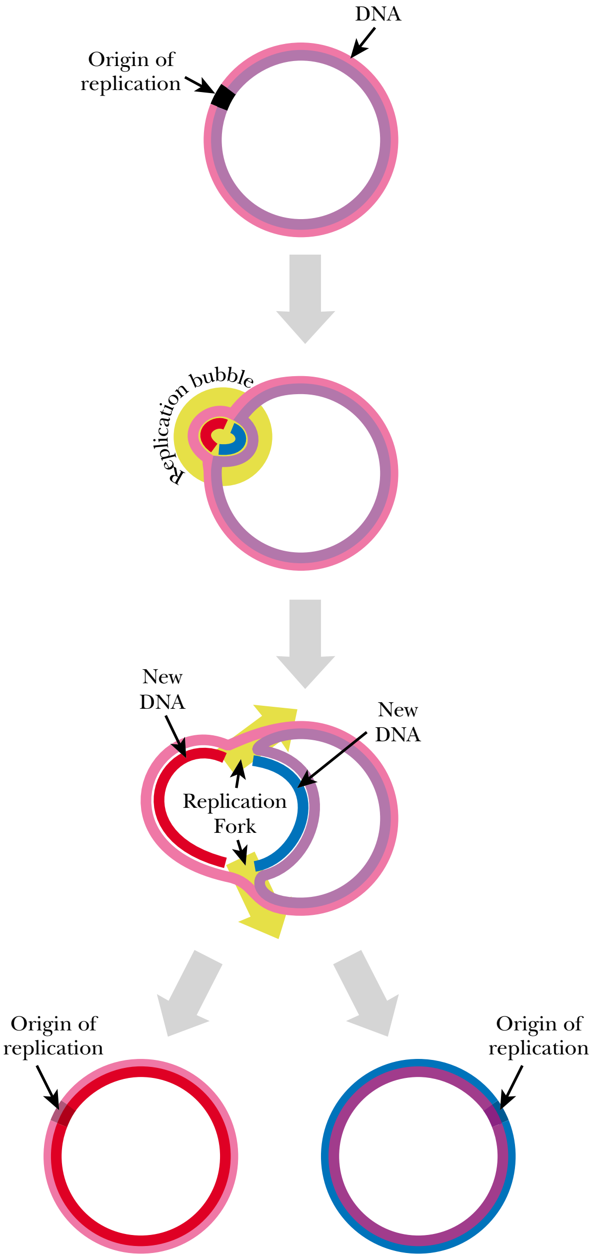


FIGURE 16.05 ***Bi-Directional Plasmid Replication***

For some circular plasmids, replication enzymes recognize the origin of replication, unwind the DNA, and start synthesis of two new strands of DNA, one in each direction. The net result is a replication bubble. As the new strands are synthesized, two distinct replication forks keep moving around the circle until they meet on the opposite side. When both DNA circles are complete, two distinct plasmids are produced.

the origin used during transfer. All plasmids must have a vegetative origin since they must all divide to survive. But only those plasmids that can transfer themselves have a special transfer origin.

The relationship between certain plasmids and viruses is illustrated by their DNA replication mechanisms. Rolling circle replication is not only used by transferable plasmids but also by many viruses (Fig. 16.07). Some manufacture many double stranded molecules of virus DNA. These viruses use the dangling strand as a template to synthesize a new strand of DNA. They just keep rolling and synthesizing and end up with a long linear double stranded DNA many times the length of the original DNA circle. This is chopped into unit genome lengths and packaged into virus particles. (Some of these viruses convert the DNA into circles before packaging, whereas others package linear DNA and only circularize their DNA after infecting a new cell when it is time to replicate again.)

Other viruses contain single stranded DNA. These viruses leave the dangling strand unpaired. They continue rolling and end up with a long linear single-stranded

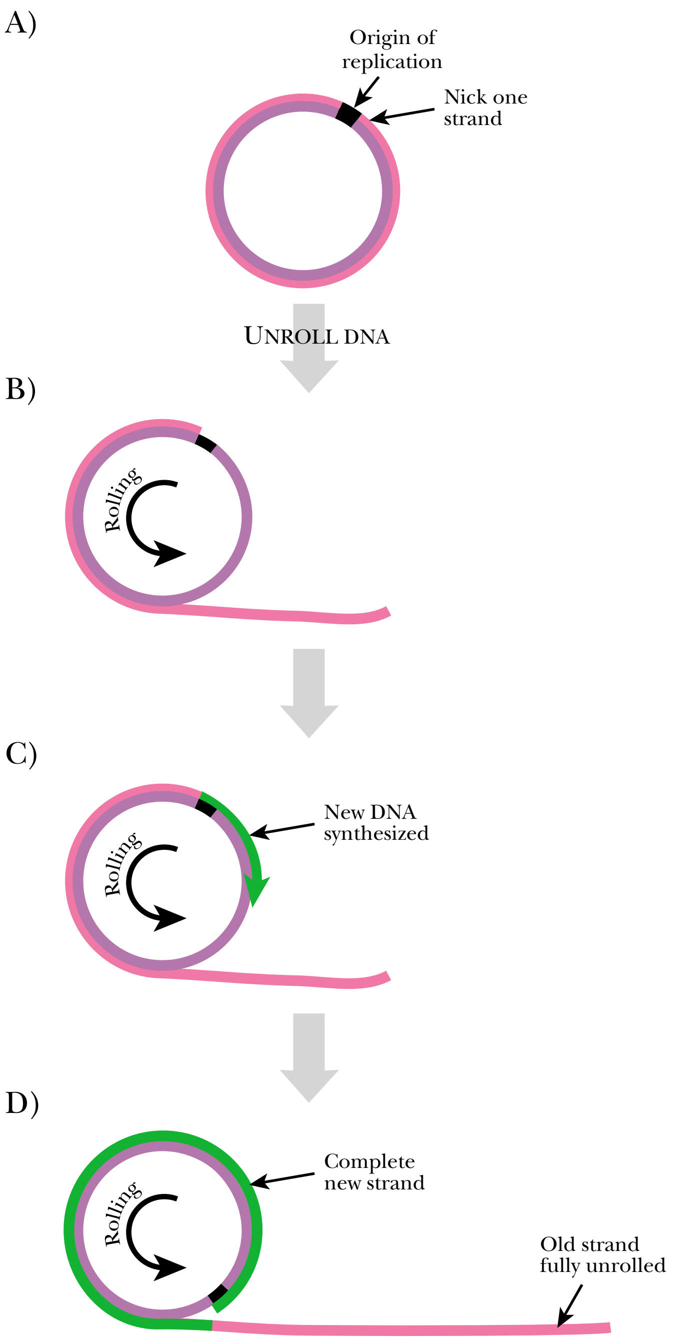


FIGURE 16.06 ***Rolling Circle Replication***

During rolling circle replication, one strand of the plasmid DNA is nicked, and the broken strand (pink) separates from the circular strand (purple). The gap left by the separation is filled in with new DNA starting at the origin of replication (green strand). The newly synthesized DNA keeps displacing the linear strand until the circular strand is completely replicated. The linear single-stranded piece is fully "unrolled" in the process.

DNA (Fig. 16.07B). This is cut into unit genome lengths and packaged as before. When these viruses infect a new cell, they synthesize the opposite strand, so converting their single strand to a double stranded DNA molecule.

Control of Copy Number by Antisense RNA

Antisense RNA is involved in regulating the copy number of many plasmids.

Both single-copy and multicopy plasmids regulate their copy number carefully. However, the regulatory mechanisms differ significantly for the two groups. High copy number plasmids limit the initiation of plasmid replication once the number of plasmids in the cell reaches a certain level. These plasmids are sometimes said to have "relaxed" copy number control. In contrast, single copy plasmids have "stringent" copy number control as their division is more tightly regulated and they replicate only once

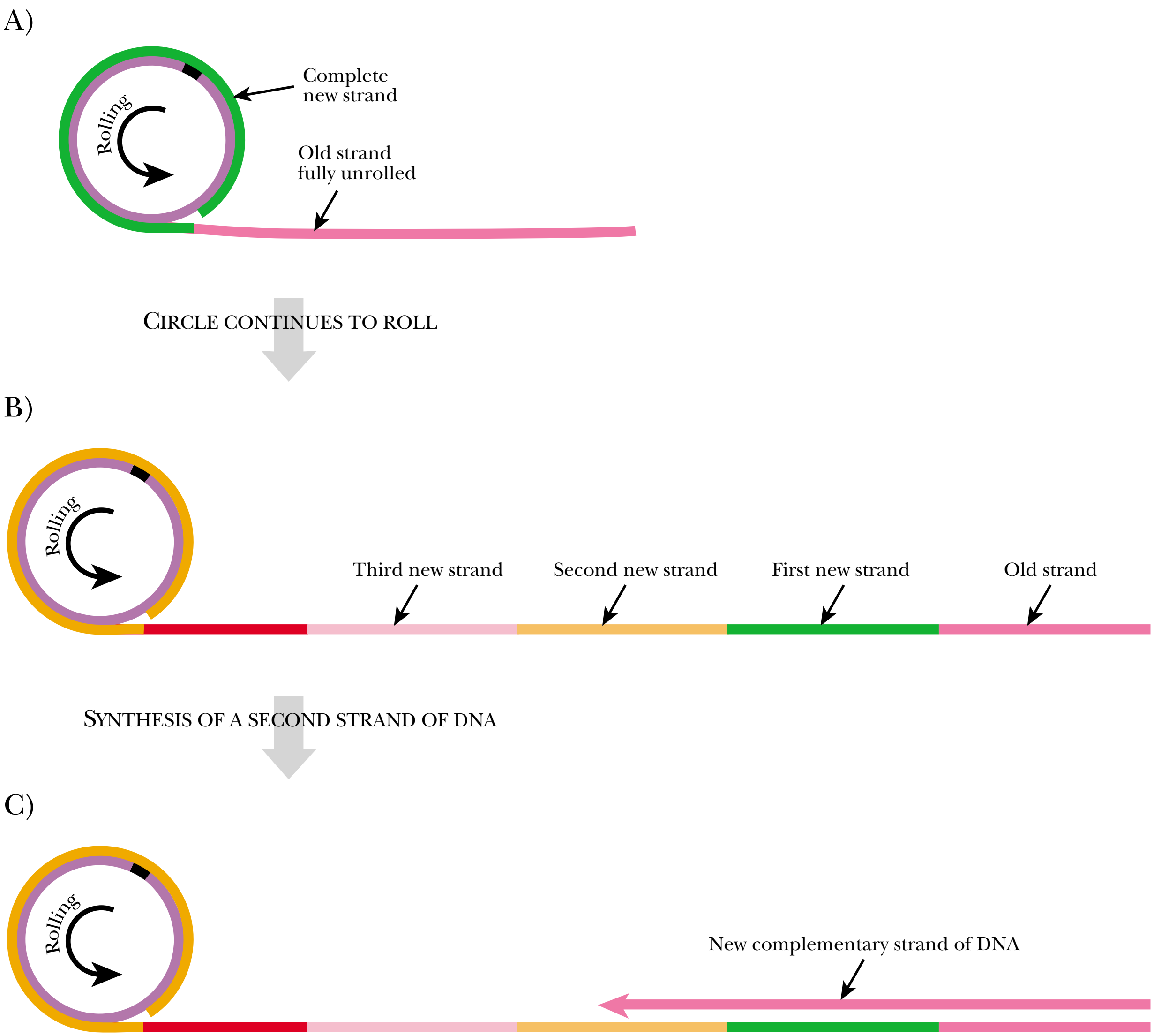


FIGURE 16.07 ***Viruses may Use Rolling Circle Replication***

Rolling circle replication occurs as described in the previous figure (A), but the replication continues around the circular DNA (purple) for many rounds (B). In some viruses, the long single-stranded piece of DNA is cut and packaged into virus particles as single-stranded DNA. In other viruses, a complementary strand is synthesized, giving double-stranded DNA (C). The double-stranded segments are then cut and packaged as single genome units.

during the cell cycle. The regulation of replication is much better understood for multicopy plasmids than for single-copy plasmids.

The most interesting aspect of copy number regulation is the involvement of **antisense RNA** to control the initiation of plasmid replication. The details are best investigated for the multicopy plasmid ColE1, but the principle of using antisense RNA applies to single-copy plasmids also.

Initiation of ColE1 replication starts with the transcription of an RNA molecule of 555 bases that can act as a primer for DNA synthesis. This pre-primer RNA (sometimes called RNAII) is cleaved by **ribonuclease H** to generate a primer with a free 3′-OH group, which can be used by DNA polymerase I (Fig. 16.08).

antisense RNA	An RNA molecule that is complementary to messenger RNA or another functional RNA molecule
ribonuclease H	A ribonuclease of bacterial cells that is specific for RNA-DNA hybrids

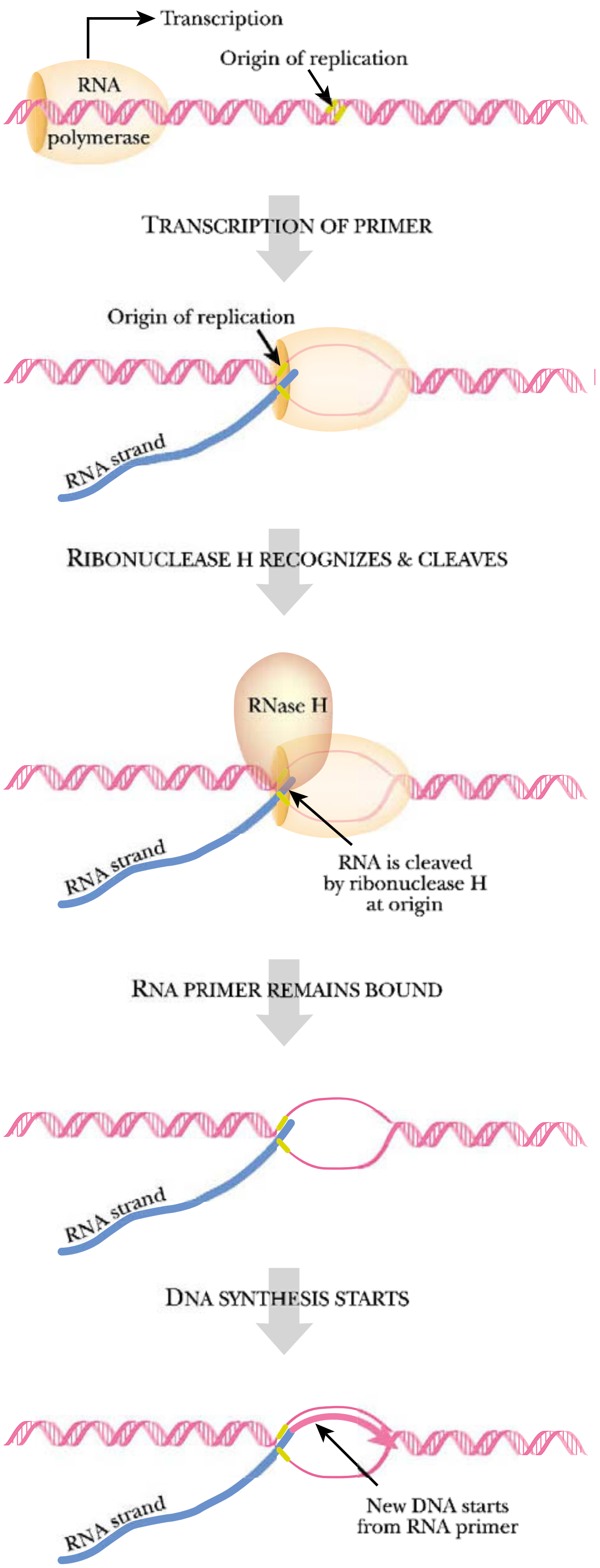


FIGURE 16.08 ***Priming of ColE1 Plasmid Replication***

RNA polymerase synthesizes a strand of RNA (RNAII) near the origin of replication. RNAII (blue strand) is recognized and cleaved by ribonuclease H. The free 3′-OH created by the cleavage primes the synthesis of DNA at the origin. The ColE1 plasmid is then replicated.

If ribonuclease H fails to cleave the pre-primer RNAII, no free 3′-end is made and replication cannot proceed. Ribonuclease H is specific for RNA-DNA hybrids. Consequently, when an antisense RNA, known as RNAI, binds to pre-primer RNAII, this prevents cleavage (Fig. 16.09). Both RNAII and RNAI are transcribed from the same region of DNA but in opposite directions. RNAI is 108 bases long and is encoded

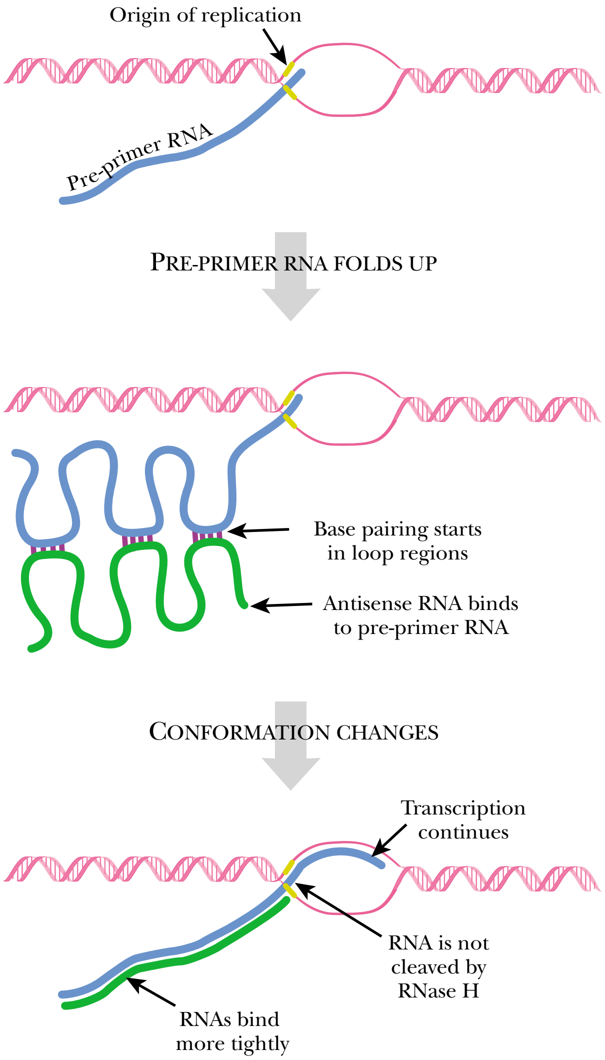


FIGURE 16.09 *Antisense RNA Prevents Primer Formation*

A second transcript, RNAI, is also made from the same region of ColE1 as RNAII. The RNA I (green) is transcribed from the opposite DNA strand and is therefore complementary to RNAII. The complementary regions of RNAII and RNAI start to base pair, forming a region of bubbles. Eventually, the entire sequence aligns and a double stranded RNA molecule is formed. Since ribonuclease H only recognizes DNA-RNA hybrid molecules, no cut is made and RNAII transcription continues. DNA synthesis fails to start and the ColE1 plasmid is not replicated.

by the opposite strand from RNAII. RNAI is complementary to the 5′-end of RNAII.

The copy number is determined by the relative strengths of binding of RNAII to the DNA at the origin of replication and of RNAI to RNAII. Mutations affecting either of these interactions will change the copy number. The Rom protein, also encoded by a gene on ColE1, increases binding between RNAI and RNAII. If the gene for Rom protein is inactivated, the copy number rises, but is still controlled.

Plasmid Addiction and Host Killing Functions

Large plasmids often make toxins that kill the host cell if, and only if, it loses the plasmid DNA.

Many larger plasmids possess genes whose function is to ensure that the host cell does not lose the plasmid. These "plasmid addiction" systems kill the bacterial cell if the plasmid is lost, so only cells that retain the plasmid survive. The details vary but the scheme involves two components that are made by the plasmid. One is lethal to the host cell and the other is the antidote. The toxin is long-lived and the antidote is short-lived. As long as the plasmid is present it continues to synthesize the antidote. If the plasmid is lost, the antidote decays but the stable toxin survives longer and kills the cell. Many plasmids actually have two or more different systems to decrease the chances of the host cell surviving after losing the plasmid.

The protein-based host killing operon of the F-plasmid consists of two genes, *ccdAB* that are expressed to give two proteins, CcdA and CcdB. CcdA is the antidote and is readily degraded by host cell proteases. As long as the plasmid is present it constantly makes a fresh supply of CcdA, which binds to CcdB and blocks its action. Consequently, the cell survives. CcdB is the toxin and in the absence of CcdA it kills the cell by inhibiting DNA gyrase and generating double-stranded breaks in the bacterial chromosome. Other plasmids have similar systems.

Both the F-plasmid and some R-plasmids also have a system where the antidote is an antisense RNA that prevents translation of the host-killing protein. This type of system was first found in plasmid R1. Here the Hok (Host Killing) protein is not translated as long as the Sok antisense RNA is present. Sok RNA binds to *hok* mRNA and prevents ribosome binding, which in turn promotes degradation of the mRNA by ribonuclease III. Sok RNA decays relatively rapidly. If the plasmid is lost Sok RNA decays and the *hok* mRNA is free to be translated. Hok protein then damages the cell membrane and kills the cell.

Many Plasmids Help their Host Cells

If plasmids are not an essential part of the cell's genome, why do cells allow them to persist? Some plasmids are indeed useless and, as discussed above, some plasmids possess special mechanisms to protect their own survival at the expense of the host cell. Nonetheless, most plasmids do in fact provide useful properties to their host cells. In principle any gene can be plasmid borne and plasmids are indeed widely used in genetic engineering to move genes between organisms. In practice certain properties are widespread among naturally occurring plasmids. A selection of these are given in Table 16.01.

Many plasmids carry genes that are beneficial to their host cells, but only under certain environmental conditions.

Plasmids often carry genes for resistance to antibiotics. This protects bacteria both from human medicine and from antibiotics produced naturally in the soil. Plasmids with genes for resistance to toxic heavy metals such as mercury, lead or cadmium protect bacteria from industrial pollution and from natural deposits of toxic mineral. Other plasmids provide genes that allow bacteria to grow by breaking down various industrial chemicals, including herbicides, or the components of petroleum. From the human perspective, such bacteria may be a nuisance or may be useful in cleaning up oil spills or other chemical pollution. Finally, some plasmids provide virulence or colonization factors needed by infectious bacteria to invade their victims and survive the countermeasures taken by the host immune system.

Antibiotic Resistance Plasmids

R-plasmids make bacteria resistant to antibiotics.

Plasmids were first discovered in Japan just after World War II, inhabiting the bacterium *Shigella*, which causes dysentery. The type of dysentery due to bacteria was originally treated with sulfonamides, the earliest type of antibiotic. Suddenly, strains of *Shigella* appeared that were resistant to sulfonamide treatment. The genes for resistance to sulfonamide proved to reside on plasmids, rather than the bacterial chromosome. Plasmids that confer antibiotic resistance are called **R-plasmids** or **R-factors** (Fig. 16.10).

Worse, the plasmids carrying the sulfonamide resistance genes were able to transfer copies of themselves from one bacterial cell to another. Consequently, the sulfonamide resistance spread rapidly from *Shigella* to *Shigella*. Although the resistance plasmid allowed the *Shigella* to survive, transferable antibiotic resistance is highly dangerous from the human medical viewpoint. By 1953, the year Watson and Crick discovered the double helix, 80 percent of the dysentery-causing *Shigella* in Japan had become resistant to sulfonamides. By 1960, 10% of the *Shigella* in Japan was resistant to four antibiotics, sulfonamides, chloramphenicol, tetracycline and streptomycin, and

R-plasmid or R-factor Plasmid that carries genes for antibiotic resistance

TABLE 16.01 Properties Conferred by Naturally Occurring Plasmids

Resistance and Defense

Antibiotic resistance against aminoglycosides, β-lactams, chloramphenicol, sulfonamides, trimethoprim, fusidic acid, tetracyclines, macrolides, fosfomycin
Resistance to many heavy metal ions including Ni, Co, Pb, Cd, Cr, Bi, Sb, Zn, Cu and Ag
Resistance to mercury and organomercury compounds
Resistance to toxic anions such as arsenate, arsenite, borate, chromate, selenate, tellurite, etc
Resistance to intercalating agents such as acridines and ethidium bromide
Protection against radiation damage by UV and X-rays
Restriction systems that degrade bacteriophage DNA
Resistance to certain bacteriophages

Aggression and Virulence

Synthesis of bacteriocins
Synthesis of antibiotics
Crown gall tumors and hairy root disease in plants caused by *Agrobacterium*
Nodule formation by *Rhizobium* on roots of legumes
R-body synthesis due to plasmids of *Caedibacter* symbionts in killer *Paramecium*
Virulence factors of many pathogenic bacteria, including toxin synthesis, protection against immune system and attachment proteins

Metabolic Pathways

Degradation of sugars e.g. lactose (in *Salmonella*), raffinose, sucrose
Degradation of aliphatic and aromatic hydrocarbons and their derivatives such as octane, toluene, benzoic acid, camphor
Degradation of halogenated hydrocarbons such as polychlorinated biphenyls
Degradation of proteins
Synthesis of hydrogen sulfide
Denitrification in *Alcaligenes*
Pigment synthesis in *Erwinia*

Miscellaneous

Transport of citrate in *E. coli*
Transport of iron
Gas vacuole production in *Halobacterium*

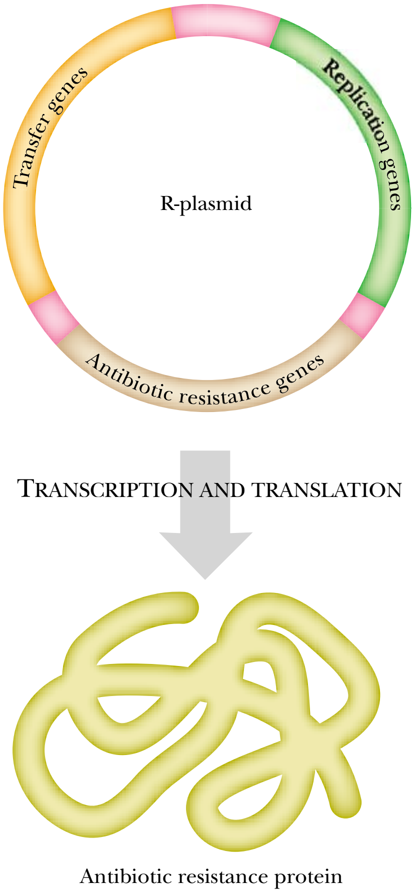

Transfer genes
Replication genes
R-plasmid
Antibiotic resistance genes
TRANSCRIPTION AND TRANSLATION
Antibiotic resistance protein

FIGURE 16.10 ***Antibiotic Resistance Plasmids***

Plasmids carry genes for replicating their DNA, transferring themselves from one host cell to another, and genes for a variety of phenotypes. Many plasmids carry genes that confer antibiotic resistance on the host cell when the genes are expressed.

by 1970 this had risen to over 30%. These strains often carry resistance genes for different antibiotics on one single plasmid. Today, the transfer of multiple antibiotic resistance plasmids between bacteria has become a major clinical problem. Patients with infections after surgery, with severe burns, or with immuno-compromised systems are at highest risk to antibiotic resistant infections.

Soil bacteria (e.g. *Streptomyces*) or fungi (e.g. *Penicillium*) produce antibiotics as a natural part of there physiology. Consequently, R-plasmids were in existence before the clinical use of antibiotics by humans, but they have spread far and wide since wide scale use of antibiotics started. A major factor in R-plasmid spread is the practice of feeding animals (e.g. pigs and chickens) antibiotics to prevent illnesses that reduce yield. Recently, some countries have banned the use of human antibiotics in animal feed and there has been a major decline in the frequency of R^+ bacteria carried by farm animals.

Most R-plasmids are of moderate to large size and present in 1–2 copies per host cell. Most are self-transmissible at a low frequency, although de-repressed mutants showing high transfer frequency are sometimes found. The original F-plasmid is such a “naturally-occurring mutant”. R-plasmids belong to a wide range of incompatibility groups. Many carry resistances to one or more antibiotics and/or toxic heavy metals and may also carry genes for colicins, virulence factors etc.

Mechanisms of Antibiotic Resistance

Plasmid borne resistance mechanisms usually inactivate or expel the antibiotic, rather than altering vital cell components.

Antibiotic resistant mutants of bacteria may be easily isolated in the laboratory. However, the mechanism of resistance in such chromosomal mutants is usually quite distinct from that of plasmid-borne resistance. The chromosomal mutations usually alter the cell component that is the target of antibiotic action, which often causes detrimental side effects. Plasmid-borne resistance generally avoids altering vital cell components. Instead the antibiotic may be inactivated or pumped out of the cell. Occasionally plasmids do provide an altered (but still functional) target component. Several of the resistance genes originally found on plasmids have been used in genetic engineering. Antibiotic resistance allows scientists to screen for cells that contain a plasmid, and kill all the cells that do not (see Ch. 22). Chloramphenicol, kanamycin/neomycin, tetracycline and ampicillin resistance genes are the most widely used in laboratories.

Resistance to Beta-Lactam Antibiotics

Penicillin and its relatives are the most widely used family of antibiotics.

The **β-lactam** family includes the **penicillins** and **cephalosporins** and is the best-known and most widely used group of antibiotics. All contain the β-lactam structure, a four-membered ring containing an amide group, which reacts with the active site of enzymes involved in building the bacterial cell wall. Cross-linking of the peptidoglycan is prevented, so causing disintegration of the cell wall and death of the bacteria. Since peptidoglycan is unique to bacteria, penicillins and cephalosporins have almost no side effects in humans, apart from occasional allergies.

Penicillin and related antibiotics are destroyed by the enzyme beta-lactamase.

Resistance plasmids carry a gene encoding the enzyme, **β-lactamase**, which destroys the antibiotic by opening the β-lactam ring (Fig. 16.11). Most β-lactamases prefer either penicillins or cephalosporins, though a few attack both antibiotics equally well. Resistance to **ampicillin**, a popular type of penicillin, is widely used in molecular

ampicillin A widely used antibiotic of the penicillin group
beta-lactams or β-lactams Family of antibiotics that inhibit cross-linking of the peptidoglycan of the bacterial cell wall; includes penicillins and cephalosporins
beta-lactamase or β-lactamase Enzyme that inactivates β-lactam antibiotics such as ampicillin by cleaving the lactam ring
cephalosporins Group of antibiotics of the β-lactam type that inhibit cross-linking of the peptidoglycan of the bacterial cell wall
penicillins Group of antibiotics of the β-lactam type that inhibit cross-linking of the peptidoglycan of the bacterial cell wall

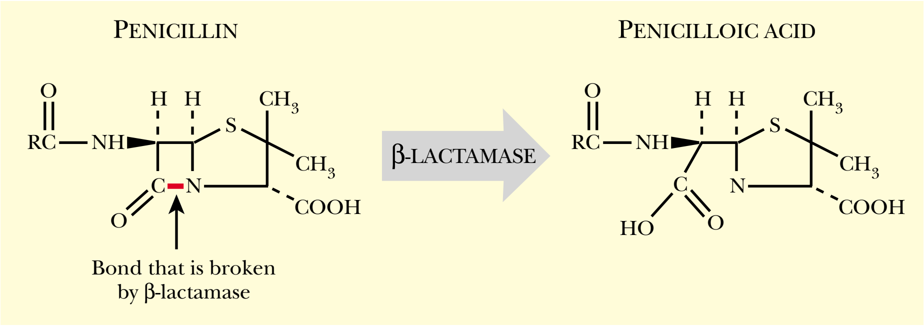


FIGURE 16.11 ***Inactivation of Penicillin by β-Lactamase***

Penicillin is an antibiotic that attacks the cell wall of bacteria, preventing the cells from growing or dividing. The antibiotic has a four-membered β-lactam ring that binds to the active site of the enzymes that assemble the cell wall. The enzyme β-lactamase cleaves the *β*-lactam ring of penicillin (red bond). The penicillin is inactivated.

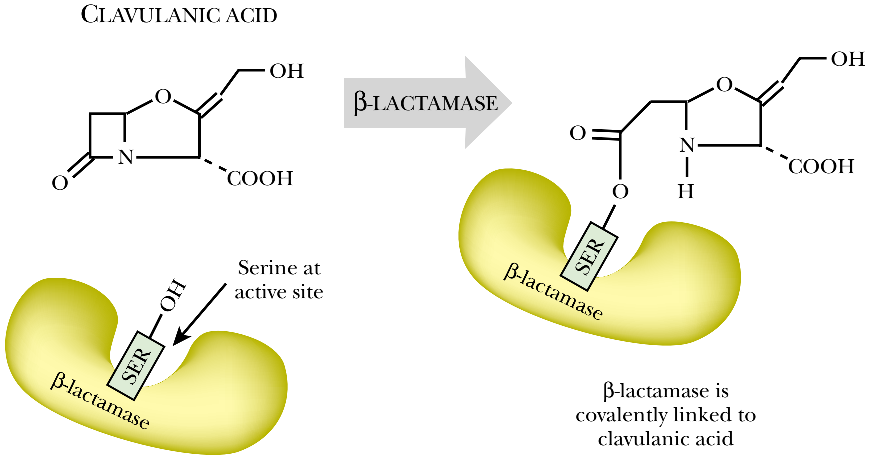


FIGURE 16.12 ***Inactivation of β-Lactamase by Clavulanic acid***

In order to inactivate β-lactamase, analogs of penicillin such as clavulanic acid are added along with the antibiotic. Clavulanic acid has a four-membered ring similar to penicillin. Consequently, β-lactamase will bind and cleave this ring. When this happens, clavulanic acid is covalently bound to β-lactamase rendering it useless against penicillin. Added penicillin can now kill the bacteria, even though they contain the resistance genes.

biology, especially for selecting plasmids during cloning procedures (see Ch. 22). The same gene is referred to as either *amp* (for ampicillin) or ***bla*** for β-lactamase. Certain strains of *Pseudomonas* carrying the R-plasmid RPl that encodes a high activity, broad-spectrum β-lactamase can actually grow on ampicillin as sole carbon and energy source!

A vast number of penicillin and cephalosporin derivatives have been made by the pharmaceutical industry. Some of these are much less susceptible to breakdown by β-lactamase. Their development has in turn led to the emergence of altered and improved β-lactamases among bacteria carrying R-plasmids. Another approach is to administer a mixture of a β-lactam antibiotic plus a β-lactam analog that inhibits β-lactamase. **Clavulanic acid** and its derivatives bind to β-lactamases and react forming a covalent bond to the amino acids in the active site that kills the enzyme (Fig. 16.12).

Resistance to Chloramphenicol

Chloramphenicol, streptomycin and kanamycin are all antibiotics that inhibit protein synthesis by binding to the bacterial ribosomes. The difference in mechanism between resistance due to chromosomal mutations as opposed to plasmid-borne genes is espe-

***bla* gene** Gene encoding β-lactamase thereby providing resistance to ampicillin. Same as *amp* gene
clavulanic acid And its derivatives bind to β-lactamases and react forming a covalent bond to the protein that kills the enzyme

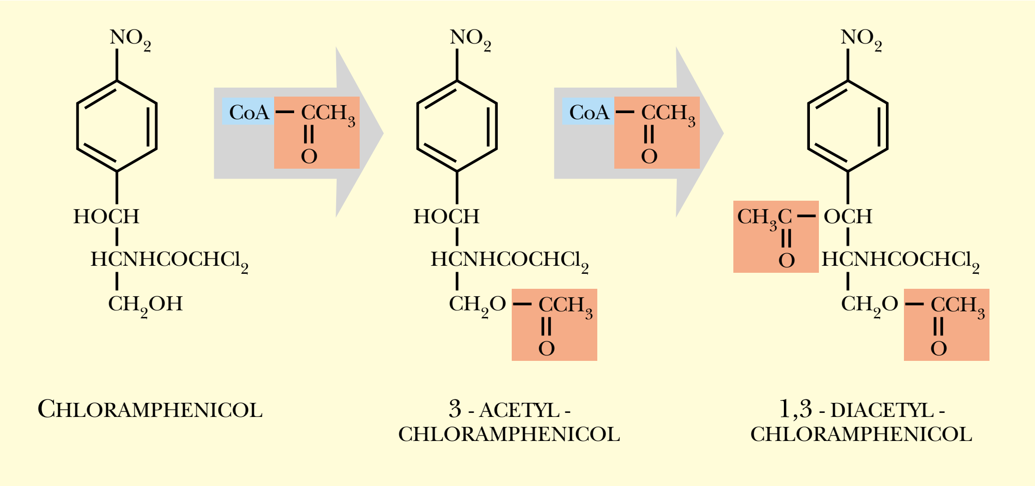


FIGURE 16.13 ***Inactivation of Chloramphenicol***

The side chain of chloramphenicol has two –OH groups that are important for binding to the bacterial ribosomes. Chloramphenicol acetyl transferase, produced by R-plasmids, catalyzes the addition of two acetyl groups to chloramphenicol. The enzyme uses acetyl-CoA as a source for the acetyl groups. The resulting 1,3-diacetyl-chloramphenicol can no longer bind to the ribosomes.

cially notable for these antibiotics. Chromosomal mutants usually have altered ribosomes that no longer bind the antibiotic. Not surprisingly such mutations often cause slower or less accurate protein synthesis and the cells grow poorly. In contrast, plasmid-borne resistance to these antibiotics usually involves chemical attack on the antibiotic itself by specific enzymes encoded by the plasmid.

Chloramphenicol is inactivated by addition of acetyl groups.

Chloramphenicol binds to the 23S rRNA of the large subunit of the bacterial ribosome and inhibits the peptidyl transferase reaction (see Ch. 8). R plasmids protect the bacteria by producing the enzyme, **chloramphenicol acetyl transferase (CAT)**. **CAT** transfers two acetyl groups from acetyl CoA to the side chain of chloramphenicol. This prevents binding of the antibiotic to the 23S rRNA (Fig. 16.13). Replacement of the terminal -OH of chloramphenicol with fluorine results in non-modifiable yet still antibacterially active derivatives. There are two major groups of chloramphenicol acetyl transferase, one from gram-positive and one from gram-negative bacteria. The two groups differ greatly from each other except for the chloramphenicol-binding region.

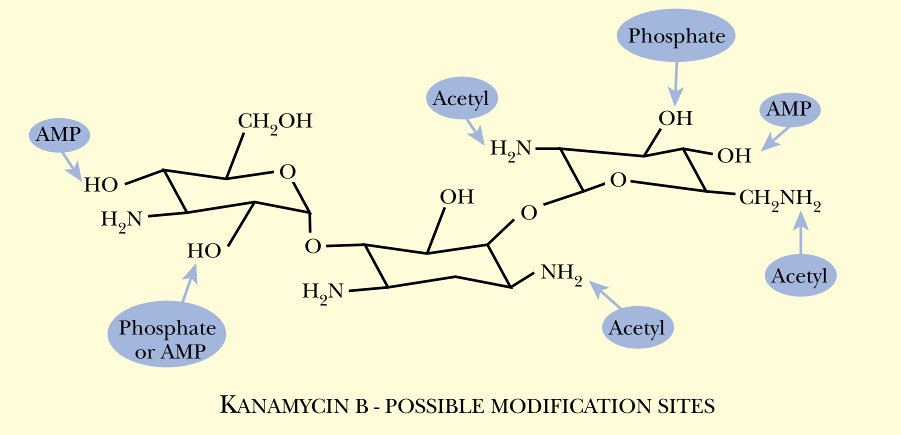
Resistance to Aminoglycosides

The **aminoglycoside** family of antibiotics includes **streptomycin**, **kanamycin**, **neomycin**, tobramycin, gentamycin and a host of others. Aminoglycosides consist of three (sometimes more) sugar rings, at least one of which (and usually two or three) has amino groups attached. They inhibit protein synthesis by binding to the small subunit of the ribosome (see Ch. 8). Plasmid-borne resistance is due to inactivation of the antibiotics. Several alternatives exist, including modification by phosphorylation of -OH groups, adenylation (i.e. addition of AMP) of -OH groups or acetylation of -NH_2 groups. ATP

aminoglycosides Family of antibiotics that inhibit protein synthesis by binding to the small subunit of the ribosome; includes streptomycin, kanamycin, neomycin, tobramycin, gentamycin and many others
chloramphenicol Antibiotic that binds to 23S rRNA and inhibits protein synthesis
CAT Chloramphenicol acetyl transferase
chloramphenicol acetyl transferase (CAT) Enzyme that inactivates chloramphenicol by adding acetyl groups
kanamycin Antibiotic of the aminoglycoside family that inhibits protein synthesis
neomycin Antibiotic of the aminoglycoside family that inhibits protein synthesis
streptomycin Antibiotic of the aminoglycoside family that inhibits protein synthesis

FIGURE 16.14 ***Inactivation of Aminoglycoside Antibiotics***

Much like chloramphenicol, members of the aminoglycoside family are inactivated by modification. One member, kanamycin B, can be modified by a variety of covalent modifications, such as phosphorylation, acetylation, or adenylation. A variety of bacterial enzymes make these modifications to prevent kanamycin B from attaching to the small ribosomal subunit.



is used as a source of phosphate and AMP groups, whereas acetyl-CoA is the acetyl donor (Fig. 16.14).

Aminoglycoside antibiotics are inactivated by addition of phosphate, AMP, or acetyl groups.

Modified aminoglycosides no longer inhibit their ribosomal target sites. There are many different aminoglycosides and a correspondingly wide range of modifying enzymes. The ***npt*** gene (**neomycin phosphotransferase**) is the most widely used and provides resistance to both kanamyin and the closely related neomycin. Aminoglycosides are made by bacteria of the *Streptomyces* group, which are mostly found in soil. These organisms need to protect themselves against the antibiotics they produce. Probably, therefore, the aminoglycoside modifying enzymes came originally from the same *Streptomyces* strains that make these antibiotics.

Amikacin is a more recent derivative of kanamycin A in which the amino group on the middle ring that gets acetylated is blocked with a hydroxybutyrate group. This made amikacin resistant to all modifying enzymes except one obscure N-acetyl transferase. However, evolution moves on and an enzyme that phosphorylates amikacin has already appeared in some bacterial strains!

Resistance to Tetracycline

Tetracycline binds to the 16S rRNA of the small subunit and also inhibits protein synthesis. However, the mechanism of resistance is quite different from chloramphenicol and aminoglycosides. Rather than inactivating tetracycline by modification, R-plasmids produce proteins that pump the antibiotic out of the bacteria. Tetracycline actually binds to both prokaryotic and eukaryotic ribosomes. Bacteria are more sensitive than animal cells because tetracyclines are actively taken up by bacterial cells, but not by eukaryotic cells. In fact, eukaryotic cells actively export tetracyclines. In tetracycline resistant bacteria, the antibiotic is actively taken into the cell, but then pumped out again. As there is no similarity between tetracycline and any known transportable nutrients, the purpose of the bacterial transport system that takes up tetracycline and its mechanism of operation are still baffling. However, the Tet resistance protein is part of a large family of sugar transporter proteins.

Tetracycline resistance is due to energy-driven export of the antibiotic.

Plasmid-encoded tetracycline resistance is typically two level. A basal constitutive level of resistance protects by 5–10 fold relative to sensitive bacteria. In addition, exposure to tetracycline induces a second higher resistance level. Both resistance levels are due to production of proteins that are found in the cytoplasmic membrane and actively expel tetracycline from the cell. Tetracycline enters the cell as the protonated form by an active transport system. Inside the cell it binds Mg^{2+}. The Tet resistance protein uses energy to expel the Tet-Mg^{2+} complex by proton antiport (Fig. 16.15).

neomycin phosphotransferase Enzyme that inactivates the antibiotics kanamycin and neomycin by adding a phosphate group
npt **gene** Gene for neomycin phosphotransferase. Provides resistance against the antibiotics kanamycin and neomycin
tetracycline Antibiotic that binds to 16S ribosomal RNA and inhibits protein synthesis

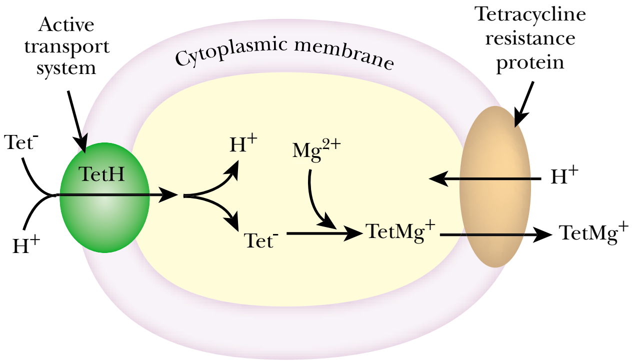


FIGURE 16.15 ***Expulsion of Tetracycline from Resistant Bacteria***

The bacterial chromosome contains the gene for TetH, a protein that takes tetracycline from the environment and actively pumps the antibiotic and a proton into the cell. Once inside the cell, tetracycline complexes with Mg$^+$, and may bind to the ribosome. In bacterial cells with an R-plasmid for tetracycline, another transport protein, called the tetracycline resistant protein, is manufactured. This protein allows a proton to enter the cell to produce energy for export of the Tet-Mg$^+$ complex.

Resistance to Sulfonamides and Trimethoprim

Both sulfonamides and trimethoprim are both antagonists of the vitamin folic acid. The reduced form of folate, tetrahydrofolate, is used as a cofactor by enzymes that synthesize methionine, adenine, thymine and other metabolites whose synthesis involves adding a one carbon fragment. **Sulfonamides** are wholly synthetic antibiotics and are analogs of *p*-aminobenzoic acid (Fig. 16.16), a precursor of the vitamin folic acid. Sulfonamides inhibit dihydropteroate synthetase, an enzyme in the synthetic pathway for folate. **Trimethoprim** is an analog of the pterin ring portion of tetrahydrofolate. It inhibits dihydrofolate reductase, the bacterial enzyme that converts dihydrofolate to tetrahydrofolate. Animal cells rely on folate in their food and so these antibiotics are effective against bacteria that normally manufacture their own tetrahydofolate.

Resistance to trimethoprim and sulfonamides is due to replacement of the target enzyme.

Plasmid mediated resistance to both sulfonamides and trimethoprim involves synthesis of folic acid biosynthetic enzymes that no longer bind the antibiotic. R-plasmid encoded dihydropteroate synthetase has the same affinity for *p*-aminobenzoic acid as the chromosomal enzyme but is resistant to sulfonamides. Similarly, R-plasmid encoded dihydrofolate reductase is resistant to trimethoprim. Sulfonamides plus trimethoprim are often used in combination for double blockade of the folate pathway. As a result, sulfonamide and trimethoprim resistance are often found together on the same R-plasmid.

Plasmids may Provide Aggressive Characters

The first plasmids drew attention because they provided their host bacteria with resistance to antibiotics. Other plasmids protect bacteria against heavy metal toxicity. However, many plasmids are known that confer aggressive, rather than defensive properties. These may be sub-divided into two broad groups. Bacteriocin plasmids encode toxic proteins used by certain strains of bacteria to kill related bacteria. **Virulence plasmids** carry genes for a variety of characters deployed by bacteria that infect higher organisms, both plants and animals, including humans.

sulfonamides Synthetic antibiotics that are analogs of *p*-aminobenzoic acid, a precursor of the vitamin folic acid. Sulfonamides inhibit dihydropteroate synthetase
trimethoprim Antibiotic that is an analog of the pterin ring portion of the folate cofactor. It inhibits dihydrofolate reductase
virulence plasmid Plasmid that carries genes for virulence factors that play a role in bacterial infection

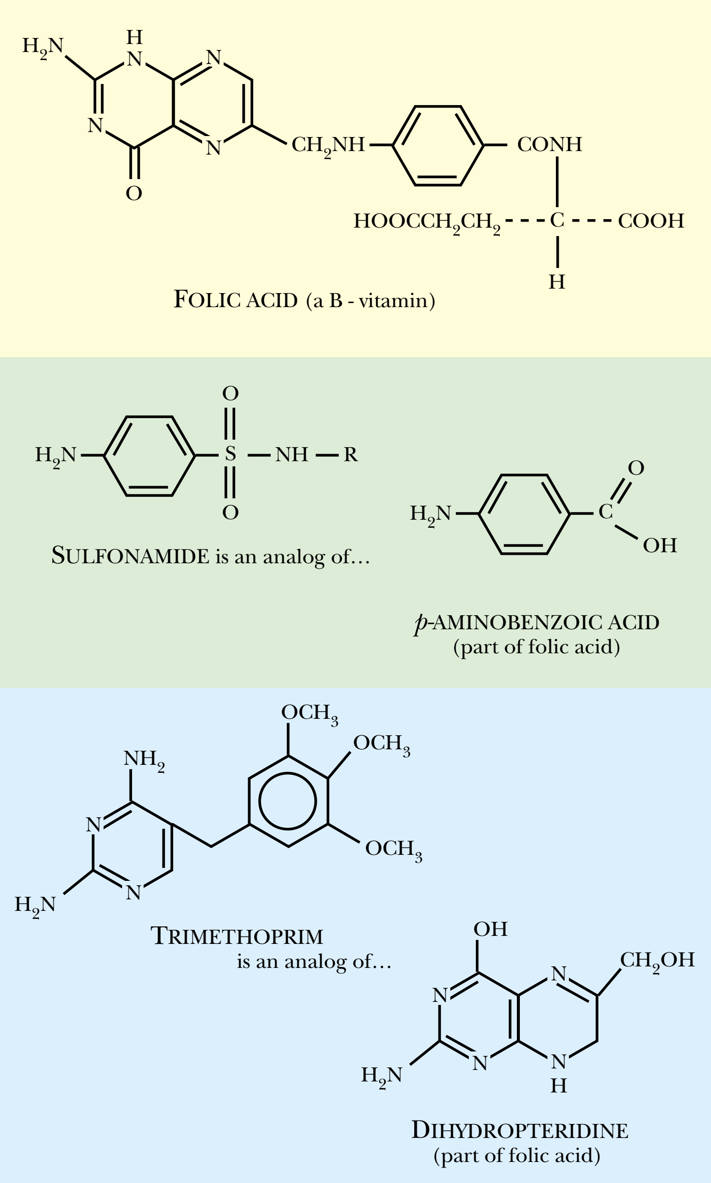


FIGURE 16.16 ***Trimethoprim, Sulfonamides and the Folate Cofactor***

Bacterial cells make folic acid, whereas, animal cells do not. The antibiotic sulfonamide is an analog of the *p*-aminobenzoic acid portion of folic acid. Trimethoprim is an analog of the dihydropteridine portion of folic acid. Both trimethoprim and sulfonamide bind to the biosynthetic enzymes and prevent synthesis of folic acid from its precursors.

Many bacteria make toxic proteins—bacteriocins—to kill closely related bacteria that compete for the same resources.

Generally speaking, bacteria are most likely to attack their close relatives. The reason is that the more closely related they are, the more likely two strains of bacteria will compete for the same resources. Proteins made by bacteria to kill their relatives are known generally as **bacteriocins**. Particular bacteriocins are named after the species that makes them. So, for example, many strains of *Escherichia coli* deploy a wide variety of **colicins**, intended to kill other strains of the same species. [Since most work has been done on *E. coli* bacteriocins from other bacteria are often referred to as colicins although this is not strictly correct.] Surveys suggest that 10–15% of enteric bacteria make bacteriocins.

Bacteriocins are usually encoded on plasmids. These provided the starting point for many genetic engineering vectors.

On several occasions, *Yersinia pestis*, the bacterium that causes Black Death (bubonic plague), has wiped out a third of the human population of Europe, and probably most of Africa and Asia. The virulence factors required for infection are carried on a series of plasmids (see below). As if this was not enough, *Yersinia pestis* also makes bacteriocins, called pesticins in this case, designed to kill competing strains of its own species.

bacteriocin Toxic protein made by bacteria to kill closely related bacteria
colicin Toxic protein or bacteriocin made by *Escherichia coli* to kill closely related bacteria

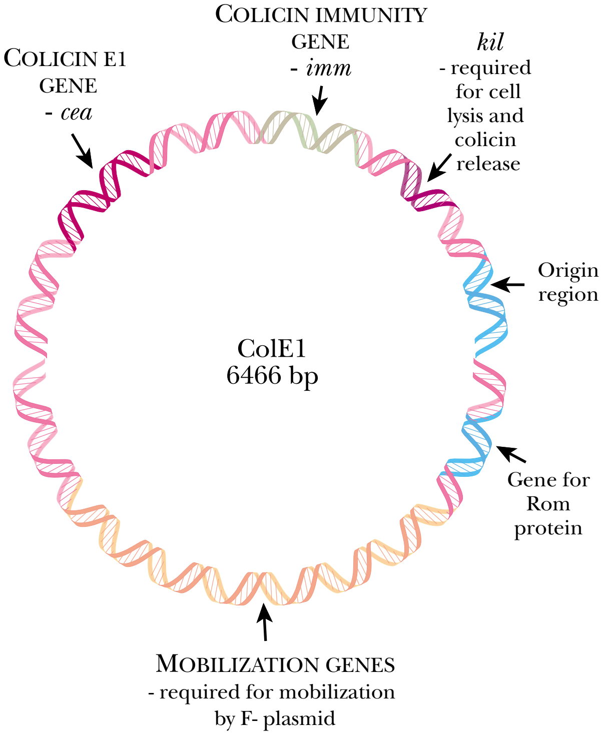

COLICIN IMMUNITY GENE - *imm*
COLICIN E1 GENE - *cea*
kil - required for cell lysis and colicin release
Origin region
ColE1
6466 bp
Gene for Rom protein
MOBILIZATION GENES - required for mobilization by F- plasmid

FIGURE 16.17 ***ColE1 is an Example of a Colicin Plasmid***

The ColE1 plasmid of *E. coli* carries genes for colicin E1 (*cea*), immunity to colicin E1 (*imm*) and the *kil* gene, required for liberation of colicin from the producer cell. The Rom gene is involved in copy number control as discussed above. ColE1 is the basis for many plasmids used in genetic engineering. The mobilization genes allow ColE1 to be transferred from cell to cell during conjugation mediated by the F plasmid.

The ability to make bacteriocins is usually due to the presence of a plasmid in the producer cell. The best known examples are the three related **ColE plasmids** of *E. coli*, ColE1 (Fig. 16.17), ColE2 and ColE3. These are small plasmids that exist in 50 or more copies per cell and have been used to derive many of the cloning vectors used in genetic engineering (see Ch. 22). These cloning vectors have the actual colicin genes removed. A variety of other colicin plasmids also occur, including the ColI and ColV plasmids. These are large single-copy plasmids and are usually transferable from one strain of *E. coli* to another. Many ColI and ColV plasmids also carry genes for antibiotic resistance.

Most Colicins Kill by One of Two Different Mechanisms

The two most popular modes of action for bacteriocins are:
a) damaging the cell membrane or
b) destroying nucleic acids.

The Col plasmids allow the strains of *E. coli* that possess them to kill other related bacteria. There are two basic approaches to this. The first is to damage the victim's cell membrane. A gene on the ColE1 plasmid encodes the colicin E1 protein that inserts itself through the membrane of the target cell and creates a channel allowing vital cell contents, including essential ions to leak out and protons to flood into the cell (Fig. 16.18). The influx of protons collapses the proton motive force. The energy derived from the proton motive force drives the production of ATP and the uptake of many nutrients, without which the bacteria quickly die. A single molecule of colicin E1 that penetrates the membrane is enough to kill the target cell. Colicin I and colicin V operate by a similar mechanism.

Colicin M and Pesticin A1122 destroy the peptidoglycan of the cell wall rather than puncturing the cytoplasmic membrane. These colicins need to penetrate only as far as the outer surface of the cytoplasmic membrane, i.e. the site of peptidoglycan assembly. Without the peptidoglycan, the bacterial cell loses shape and eventually bursts. Pesticin A1122 is made by *Yersinia pestis* and kills *Y. pseudotuberculosis, Y. enterocolitica*, plasmid free *Y. pestis* and many strains of *E. coli* (although curiously not *E. coli* K12).

ColE plasmid Small multicopy plasmid that carries genes for colicins of the E group. Used as the basis of many widely used cloning vectors

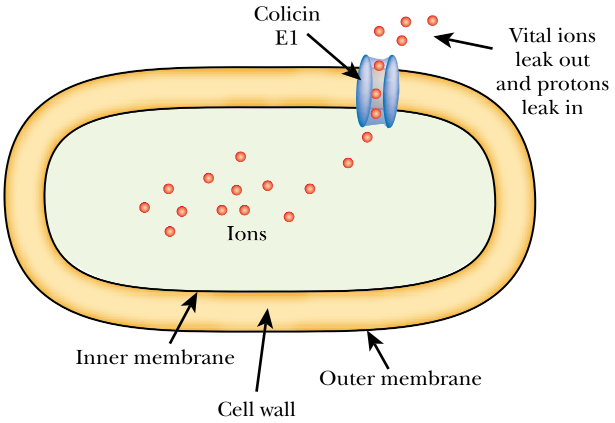


FIGURE 16.18 ***Some Colicins Damage the Cell Membrane***

When colicin E1 protein attacks a bacterial cell, it punctures a hole through the outer membrane, cell wall, and inner membrane. The hole allows protons to leak into the bacteria and vital ions to leak out. A single channel abolishes energy generation.

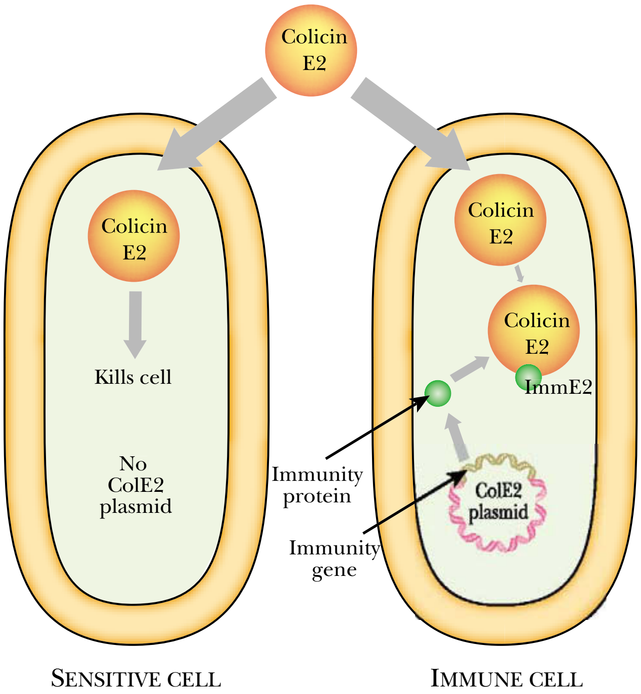


FIGURE 16.19 ***Colicin Immunity System***

In order to protect itself, a colicin-making cell also produces an immunity protein (right). This protein is also encoded by the colicin plasmid. It blocks the active site of the colicin thus preventing the cell from killing itself. The immunity protein is specific and only inhibits one type of colicin. If a cell lacks immunity protein, the colicin is able to kill the cell (left).

The second approach is to degrade the nucleic acids of the victim. The ColE2 and ColE3 plasmids both encode nucleases, enzymes that degrade nucleic acids. The colicin E2 and E3 proteins are very similar over their N-terminal 75% and as a result they share the same receptor on the surface of sensitive bacteria. They differ in the C-terminus and have different nucleic acid targets. Colicin E2 is a deoxyribonuclease that cuts up the chromosome of the target cell. Colicin E3 is a ribonuclease that snips the 16S rRNA of the small ribosomal subunit at a specific sequence, releasing a fragment of 49 nucleotides from the 3′ end. This abolishes protein synthesis and though much more specific than colicin E2, is just as lethal. Again, a single colicin molecule that enters the victim is enough to kill the target cell.

Bacteria are Immune to their own Colicins

Those bacterial cells producing a particular colicin are immune to their own brand, but not to other brands of colicin. Immunity is due to specific **immunity proteins** that bind to the corresponding colicin proteins and cover their active sites (Fig. 16.19). For

immunity protein Protein that provides immunity. In particular bacteriocin immunity proteins bind to the corresponding bacteriocins and render them harmless

Bacteria that make bacteriocins also make immunity proteins to protect themselves.

example, the ColE2 plasmid carries genes for both colicin E2 and a soluble immunity protein that binds colicin E2. This immunity protein does not protect against any other colicin, including the closely related colicin E3. Immunity to membrane active colicins is due to a plasmid-encoded inner membrane protein that block the colicin from forming a pore in the host cell. For example, the Ia immunity protein protects membranes against colicin Ia but not against the closely related colicin Ib even though colicins Ia and Ib share the same receptor, have the same mode of action, and have extensive sequence homology. Although the immune systems of animals are much more complex, the concept of immunity is based on the ability of immune system proteins to recognize and neutralize specific alien or hostile molecules.

Colicin Synthesis and Release

In a population of ColE plasmid-carrying bacteria, most cells do not produce colicin. Every now and then an occasional cell goes into production and manufactures large amounts of colicin. It then bursts and releases the colicin into the medium. This kills the producer cell. Note that the burst and release mechanism kills the producer cell, not the colicin. All sensitive bacteria in the area are wiped out, but those with the ColE plasmid have immunity protein and survive.

Bacteriocin production is often a suicidal process.

About 1 in 10,000 cells actually produce colicin in each generation. Thus release of colicin E is a communal action in the sense that a small minority of producer cells sacrifice themselves so that their relatives carrying the same ColE plasmid can takeover the habitat. Colicin E production involves expression of two plasmid genes, *cea* (colicin protein) and *kil* (lysis protein). LexA, the repressor of the SOS DNA repair system (Chapter 14), normally represses these genes. Thus colicin production is induced by DNA damage and those cells that sacrifice themselves were probably injured anyway. Note that many lysogenic bacteriophage are also induced by DNA damage monitored via the SOS system (see Ch. 17).

Not all colicins are produced by the suicidal mechanism. Many colicins made by large single-copy plasmids (e.g. colicin V, colicin I) are apparently made continuously in smaller amounts. These colicins tend to remain attached to the surface of the producer cell rather than being released as freely soluble proteins, like the E colicins. When the producer bumps into a sensitive bacterium the colicin may be transferred, with lethal results.

Virulence Plasmids

Virulence plasmids help bacteria infect humans, animals or even plants, by a variety of mechanisms. Some **virulence factors** are toxins that damage or kill animal cells, others help bacteria to attach to and invade animal cells (Fig. 16.20), whereas yet others protect bacteria against retaliation by the immune system.

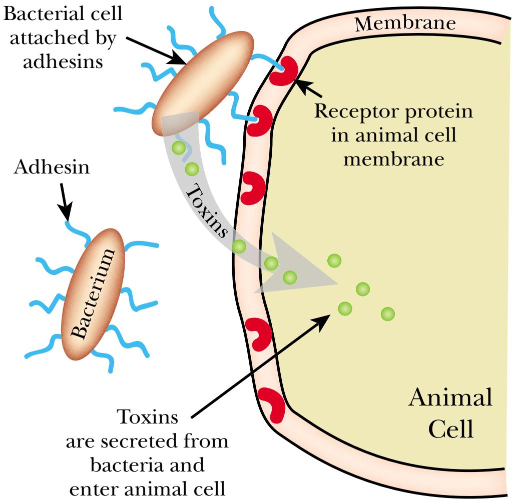
Many pathogenic bacteria carry genes for virulence on plasmids or other mobile genetic elements.

Although most strains of *Escherichia coli* are harmless, occasional rogue strains cause disease. These pathogenic *E. coli* generally rely on plasmid-borne virulence factors. Wide ranges of toxins are found in different pathogenic *E. coli* strains, including heat-labile **enterotoxin** (resembles **choleratoxin**), heat-stable enterotoxin, **hemolysin** (lyses red blood cells) and Shiga-like toxin (similar to the toxin of dysentery-causing *Shigella*). There is a similar variety of **adhesins** or "**colonization factors**",

adhesin Protein that enables bacteria to attach themselves to the surface of animal cells. Same as colonization factor
choleratoxin Type of toxin made by *Vibrio cholerae* the cholera bacterium
colonization factor Protein that enables bacteria to attach themselves to the surface of animal cells. Same as adhesin
enterotoxins Types of toxin made by enteric bacteria including some pathogenic strains of *E. coli*
hemolysin Type of toxin that lyses red blood cells
virulence factors Proteins that promote virulence in infectious bacteria. Include toxins, adhesins and proteins protecting bacteria from the immune system

FIGURE 16.20 *Toxins and Adhesins*

Bacteria are able to attack animal cells by attaching to the cellular membrane and releasing toxins. The bacteria contain plasmids that encode adhesins, which are protein filaments able to recognize and attach to cell-surface receptors found on animal cells. One attached, the bacteria secrete toxins, which can penetrate the animal cell membrane and kill the cell.


Bacterial cell attached by adhesins
Membrane
Receptor protein in animal cell membrane
Adhesin
Toxins
Bacterium
Animal Cell
Toxins are secreted from bacteria and enter animal cell

proteins that enable bacteria to stick to the surface of animal cells. Adhesins form filaments that vary in length and thickness, but generally resemble pili. Consequently the symptoms and severity of infection by *E. coli* vary greatly.

Other enteric bacteria, such as *Salmonella typhi* (typhoid) and *Yersinia pestis* (bubonic plague) cause severe infections. They also carry virulence plasmids. In *Salmonella* the majority of the virulence genes are on the chromosome, but a handful are plasmid-borne. In contrast, in *Yersinia* several plasmids carry the bulk of the virulence genes. In addition to toxins and adhesins, these "professional" pathogens possess more sophisticated virulence factors that protect against host defenses. Although plasmids have been investigated most intensively in enteric bacteria, it is clear that virulence in many other bacteria often depends on at least some plasmid-borne genes.

Ti-Plasmids are Transferred from Bacteria to Plants

Although the F-plasmid of *E. coli* is limited in its host range to a few enteric bacteria, it can actually promote DNA transfer between *E. coli* and yeast! Similarly, broad host range plasmids of the IncP, IncQ and IncW incompatibility groups can mobilize DNA from gram-negative bacteria into both gram-positive bacteria and yeast. In both cases, the range of species in which the plasmid can survive and replicate is much smaller than the range of species to which DNA may be transferred. Therefore, many plasmids are degraded or destroyed after they are transferred to an incompatible cell. Some DNA mobilized in this manner may survive if it is recombined with the host chromosome or resident plasmids.

The Ti-plasmids can mediate transfer of DNA from bacteria to plant cells.

The greatest versatility in plasmid transfer is shown by the highly specialized **Ti-plasmids** (Ti = tumor-inducing) that allow certain bacteria to insert DNA into the nucleus of plant cells. The Ti-plasmid is carried by soil bacteria of the *Agrobacterium* group, in particular *A. tumefaciens*, and confers the ability to infect plants and produce tumors, inside which the bacteria grow and divide happily. This results in tumor-like swellings on the stems of infected plants, a condition known as **"crown gall disease"**. The related Ri-plasmid is carried by *Agrobacterium rhizogenes,* which infects roots and causes hairy root disease.

Agrobacterium is attracted by chemicals, such as acetosyringone, which are released by wounded plants. It then enters via the wound and transfers a portion of

crown gall Type of tumor formed on plants due to infection by *Agrobacterium* carrying a Ti-plasmid
Ti-plasmid Tumor-inducing plasmid. Plasmid that is carried by soil bacteria of the *Agrobacterium* group and confers the ability to infect plants and produce tumors

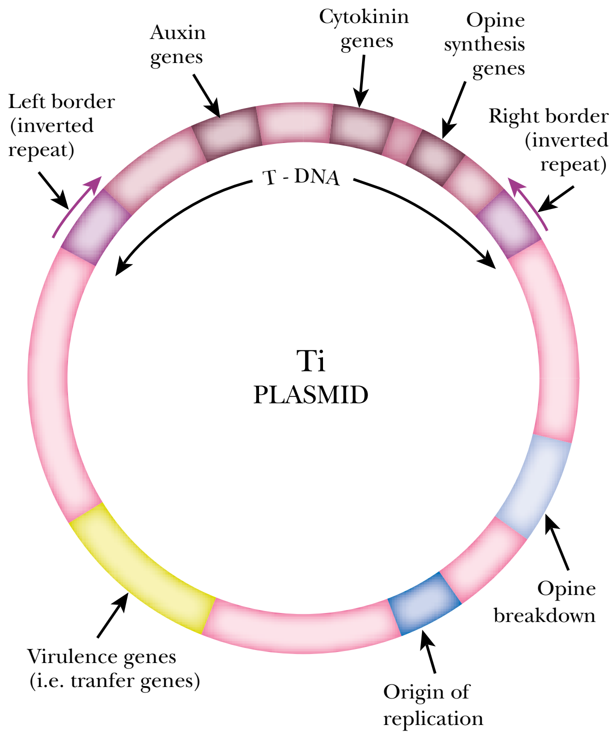


FIGURE 16.21 ***Structure of the Ti-Plasmid***

The Ti plasmid of *Agrobacterium* has several regions. The T-DNA region is flanked by two inverted repeats and contains genes for auxin and cytokinin, which induce the plant cells to grow, and genes for opine synthesis, a carbon source for *Agrobacterium*. This region is transferred into the plant cell by the expression of the transfer genes found on the other part of the Ti plasmid. The Ti plasmid also has an origin of replication and genes for opine breakdown.

Bacteria carrying Ti-plasmids infect plants and cause the formation of tumors.

the Ti plasmid into the plant cell by a mechanism similar to bacterial conjugation. A slight abrasion that is trivial to the health of the plant is of course sufficient for the entry of a microorganism. The result is a crown gall tumor that provides a home for the *Agrobacterium* at the expense of the plant.

The Ti-plasmid consists of several regions (Fig. 16.21), but only one segment, the **T-DNA** (tumor-DNA), is actually transferred into the plant cell, where it enters the nucleus. The T-DNA is flanked by 25 bp inverted repeats. Any DNA included within these repeats will be transferred into the plant cell. Consequently, Ti-plasmids have been widely used in the genetic engineering of plants. The virulence genes on the plasmid are responsible for cell-to-cell contact and transfer of the T-DNA but do not themselves enter the plant cells.

Only part of the Ti-plasmid enters the plant cell, where it integrates into the plant chromosomes.

Acetosyringone, which attracts the bacteria to the wounded plant, also induces the virulence genes, thus facilitating the transfer of the T-DNA region (Fig. 16.22). Acetosyringone binds to VirA protein in the *Agrobacterium* membrane. This activates VirG, which in turn switches on the other *vir* genes, including *virD* and *virE*. VirD makes a single-stranded nick in the Ti-plasmid at the left border of the T-DNA and the T-DNA unwinds from the cut site. The single-stranded T-DNA is bound by VirE protein and unwinding stops at the right border. The Ti-plasmid then replicates by a rolling circle mechanism as the single stranded T-DNA region enters the plant cell. Overall this results in DNA transfer from the bacteria into the plant cells. The mechanism resembles bacterial conjugation and the "virulence" genes of the Ti-plasmid are equivalent to the *tra* genes of other plasmids. The T-DNA then integrates at random into a chromosome in the plant cell nucleus.

Once inserted, the genes in the T-DNA are switched on. The enzymes they encode synthesize two plant hormones, **auxin** and **cytokinin**. Auxin makes plant cells grow bigger and cytokinin makes them divide. When this happens rapidly in the absence of normal cell differentiation, the result is a tumor (Fig. 16.23 and Fig. 16.24).

auxin Plant hormone that induces plant cells to grow bigger
cytokinin Plant hormone that induces plant cells to divide
T-DNA (tumor-DNA) Region of the Ti-plasmid that is transferred into the plant cell nucleus

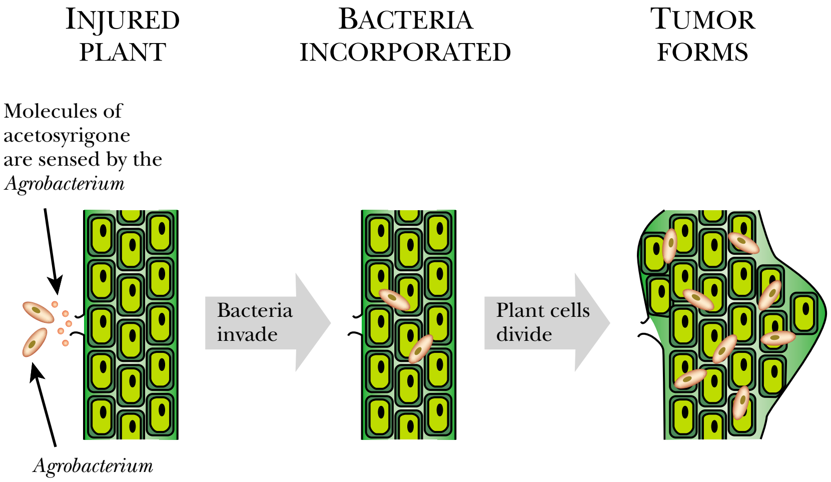


FIGURE 16.22 ***Formation of Tumor by Agrobacterium***

Agrobacterium are attracted to an injured region of a plant by sensing molecules of acetyosyrigone. The bacteria enter the plant through the open wound, and begin colonizing the area. The plant cells are stimulated to divide and a tumor forms around the bacteria.

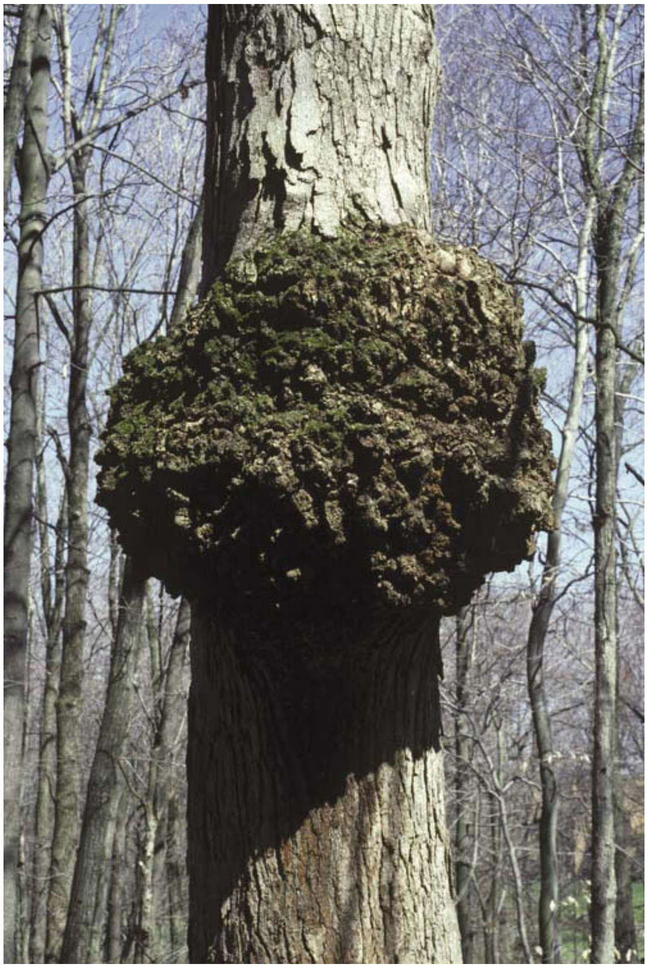


FIGURE 16.23 ***Crown Gall Tumor Caused by*** **Agrobacterium**

A crown gall tumor formed by *Agrobacterium* is shown on a tree trunk. © E. R. Degginger, Photo Researchers Inc.

The T-DNA also carries genes that subvert the plant cell into making opines. These are unusual nutrient molecules that are made at the expense of the plant cell but can only be used by bacteria that carry special genes for opine breakdown. The genes for opine degradation are found on the part of the Ti-plasmid that does not enter the plant cell. So the *Agrobacterium* can grow by using the opines but the plant cannot use them. Other bacteria that might infect the plant are also excluded as they do not have the opine breakdown genes either.

Modified Ti-plasmids are widely used in genetic engineering of plants.

Modified Ti-plasmids are widely used in the genetic engineering of plants. The genes for plant hormones and opine synthesis are removed and the genes to be trans-

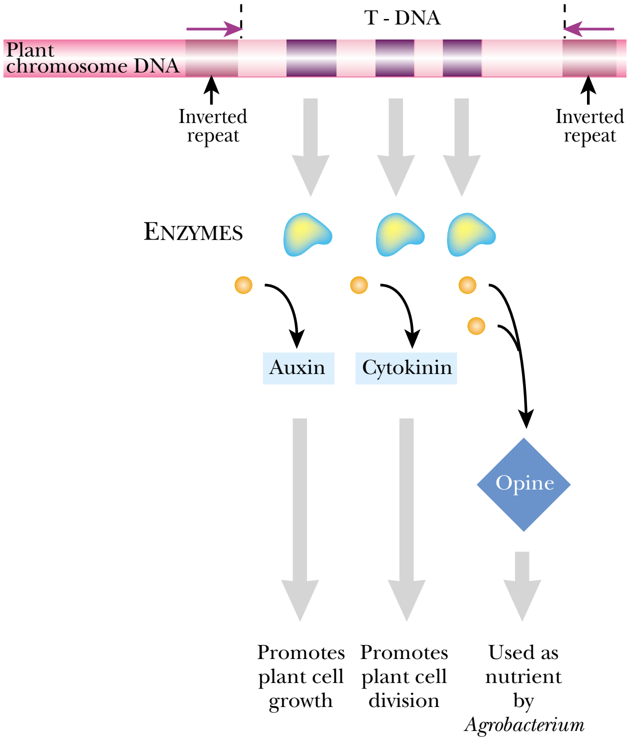

T - DNA
Plant chromosome DNA
Inverted repeat
Inverted repeat
ENZYMES
Auxin
Cytokinin
Opine
Promotes plant cell growth
Promotes plant cell division
Used as nutrient by *Agrobacterium*

FIGURE 16.24 *Expression of Genes on T-DNA*

Survival of *Agrobacteria* in the plant requires space for them to grow and a carbon source to provide energy. The genes of the T-DNA region trick the plant cell into providing these factors. The genes for auxin and cytokinin are growth factors that induce the plant cells to grow at the site of infection, providing the space. The opine is a carbon source for the bacteria, providing a constant food supply.

ferred into the plant are inserted in their place. In practice, *Agrobacterium* carrying an engineered Ti-plasmid is used to transfer genes of interest into plants using plant tissue culture.

In addition to inserting external genes into plants the Ti-plasmid system may be used for analysis of plant gene function. Insertion of T-DNA into the plant chromosome may disrupt a plant gene if insertion occurs into the coding sequence (or essential regulatory sequences). The model plant, *Arabidopsis thaliana*, has been used to generate a set of gene knockouts by random insertion of T-DNA. The locations of nearly 90,000 such insertions has been determined (as of 2003). These include insertions into about 22,000 of the estimated 29,500 genes of *Arabidopsis*. These insertions may be used to investigate the functions of the inactivated genes by comparing the knockout mutants with the parental wild type plant.

The 2-Micron Plasmid of Yeast

Many yeast strains contain a small multicopy plasmid, the "2 micron circle".

Plasmids are found in higher organisms, although they are less common than in bacteria. The yeast, *Saccharomyces cerevisiae*, has been used as a model organism for the investigation of eukaryotic molecular biology. Most strains of yeast harbor a plasmid known as the **2μ circle** or **2μ plasmid**. This is a circular molecule consisting of 6318 bp of double-stranded DNA. It is present at 50–100 copies per haploid genome and is located in the nucleus of the yeast cell, where it is bound by histones and forms nucleosomes like chromosomal DNA. The 2μ plasmid has been widely used in genetic engineering as the basis for multicopy eukaryotic cloning vectors. Similar plasmids are found in other species of yeast.

Flippase catalyses inversion of the DNA located between its recognition sites.

The 2μ plasmid contains two perfect inverted repeats of 599 bp that separate the plasmid into two regions of 2774 and 2346 bp respectively (Fig. 16.25). The plasmid encoded Flp protein (**Flp recombinase or "flippase"**) catalyzes recombination between

2 micron plasmid See 2μ plasmid
2μ circle Same as 2μ plasmid
2μ plasmid (or 2μ circle) A multicopy plasmid found in the yeast, *Saccharomyces cerevisiae*, whose derivatives are widely used as vectors
Flp recombinase (or flippase) Enzyme encoded by the 2μ plasmid of yeast that catalyzes recombination between inverted repeats (FRT sites)

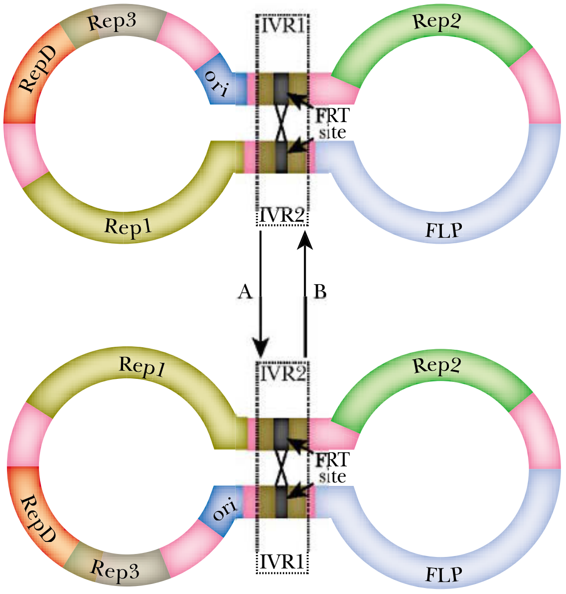


FIGURE 16.25 *The 2μ Plasmid of Yeast*

Two alternate forms of the 2μ plasmid are inter-converted by recombination. The plasmid has two inverted repeats (IVR1 and IVR2), which can align. The enzyme, Flp recombinase, recognizes the FRT sites (flip recombination target) and makes a crossover that inverts one half of the plasmid relative to the other. Notice the top plasmid has origin (*ori*) close to the Rep2 sequence, whereas, the bottom plasmid has the origin on the other side (close to the *FLP* gene). The Rep1 and Rep2 proteins regulate both the *FLP* gene and the replication of the plasmid itself.

the inverted repeats. Flp recognizes a 48 bp target site (**Flp recombination target, or FRT site**) located within the inverted repeats. The result is the inversion of one half of the plasmid relative to the other. The two forms of the plasmid are found in roughly equal proportions. The Rep1 and Rep2 proteins regulate the expression of the *FLP* gene and also bind to the origin of replication (*ori*) and the *REP3* DNA sequence.

The Flp recombinase is used in genetic engineering to control the expression of a variety of genes by inverting segments of DNA. Flp is functional in bacteria, plants and animals provided the correct recognition sites are present. In addition to the inversion reaction, Flp recombinase will promote site-specific insertion and deletion reactions of segments flanked by **FRT sites**. The Flp/FRT system is similar to the widely used Cre/*loxP* recombinase system of bacterial virus P1.

Certain DNA Molecules may Behave as Viruses or Plasmids

P1 combines the properties of plasmid and virus and can choose either lifestyle depending on the circumstances.

There are several similarities between the behavior of plasmids and viruses. In fact, some circles of DNA can choose to live either as a plasmid or as a virus. The bacterial virus P1 is a good example. It can indeed behave as a virus, in which case it destroys the bacterial cell, replicates by rolling circle mode and manufactures large numbers of virus particles to infect more bacterial cells. This is known as **lytic growth** since the host cells are "lysed" (derived from the Greek for broken).

Alternatively, P1 can choose to live as a plasmid and divide in step with the host cell. In this case, the circular P1 DNA uses bi-directional replication like a typical plasmid (Fig. 16.26). Each descendant of the infected bacterial cell gets a single copy of P1 DNA. The cell is unharmed and no virus particles are made. This state is known as **lysogeny** and a host cell containing such a virus in its plasmid mode is called a **lysogen**.

Changing conditions may stimulate a lysogenic virus to return to destructive virus mode. This tends to happen if the host cell is injured, in particular if there is severe

Flp recombination target (or FRT site) Recognition site for Flp recombinase
FRT site Flp recombination target, the recognition site for Flp recombinase
lysogen Host cell containing a lysogenic virus
lysogeny State in which a virus replicates its genome in step with the host cell without making virus particles or destroying the host cell. Same as latency, but generally used to describe bacterial viruses
lytic growth Growth of virus resulting in death of cell and release of many virus particles

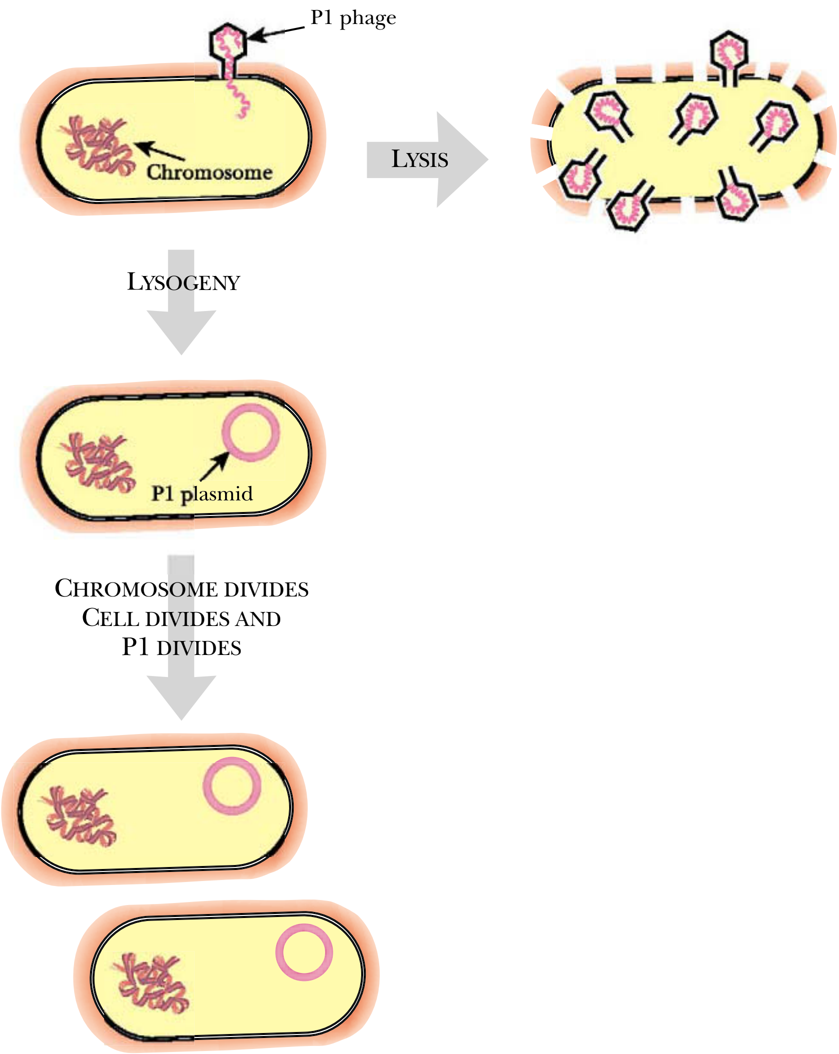


FIGURE 16.26 ***Lysis versus Lysogeny***

Some plasmids, such as the P1 plasmid of bacteria, have a dual personality. P1 can exist in a lysogenic state as a plasmid, using bi-directional replication to divide when the host cell divides. P1 can also grow as a virus and destroy the cell. During such lytic growth, P1 divides by the rolling circle mechanism, creating a large number of copies. It then packages genome-sized units into new virus particles and lyses the bacterial cell.

damage to the host cell DNA. The virus decides to make as many virus particles as possible before the cell dies. If, on the other hand, the host cell is growing and dividing in a healthy manner, the virus will most likely decide to lie dormant and divide in step with its host. Further aspects of virus behavior are covered in the following chapter, Ch. 17.

CHAPTER SEVENTEEN

Viruses

Viruses are Infectious Packages of Genetic Information

Life Cycle of a Virus

Bacterial Viruses are Known as Bacteriophage

Lysogeny or Latency by Integration

The Great Diversity of Viruses

Small Single-Stranded DNA Viruses of Bacteria

Complex Bacterial Viruses with Double Stranded DNA

DNA Viruses of Higher Organisms

Viruses with RNA Genomes Have Very Few Genes

Bacterial RNA Viruses

Double Stranded RNA Viruses of Animals

Positive-Stranded RNA Viruses Make Polyproteins

Strategy of Negative-Strand RNA Viruses

Plant RNA Viruses

Retroviruses Use both RNA and DNA

Genome of the Retrovirus

Subviral Infectious Agents

Satellite Viruses

Viroids are Naked Molecules of Infectious RNA

Prions are Infectious Proteins

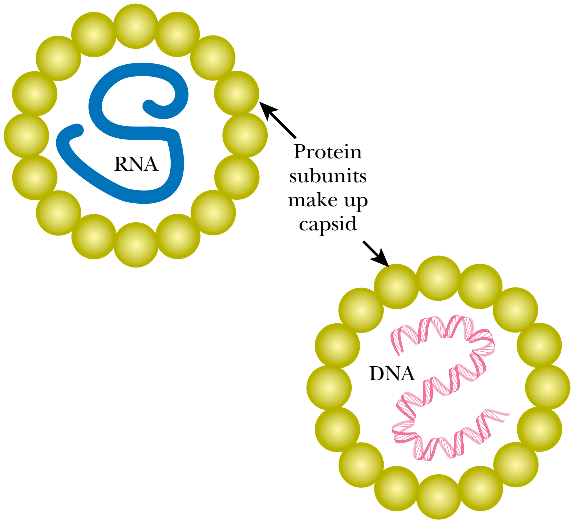


FIGURE 17.01 ***Viruses Consist of Protein plus DNA or RNA***

A simple virus contains a genome of RNA (blue) or DNA (pink) surrounded by a protein coat.

Viruses are Infectious Packages of Genetic Information

Viruses are sub-cellular parasites that rely on a cell to provide energy and raw material.

Viruses are packages of genes inside protective shells of protein. Viruses cannot grow or divide alone. In order to replicate, a virus must first infect a host cell. Only then are the virus genes expressed and the virus components manufactured using the host cell machinery. Viruses are not merely pieces of nucleic acid like plasmids or transposons and neither are they true living cells with the ability to generate energy and make protein. They lie in the gray area between. Viruses cannot make their own proteins or generate their own energy. They can only multiply when they have entered a suitable host cell and taken over the cellular machinery. Despite this a virus is certainly not inert; it does replicate if it can subvert a host cell.

Virus particles contain DNA or RNA protected by a shell of protein.

Virus particles contain proteins plus genetic information in the form of DNA or RNA (Fig. 17.01). The virus particle, or **virion**, consists of a protein shell, known as a **capsid**, surrounding a length of nucleic acid, either RNA or DNA, which carries the virus genes and is often referred to as the **viral genome**. Many simple viruses have only these two components.

Trying to define precisely what is living and what is non-living can be quite confusing. Here we will sidestep the issue of defining life by noting that being alive and being a **living cell** are not necessarily the same. To qualify as a genuine living cell, a structure must send genetic messages (RNA) from its genes (DNA) to its own ribosomes to make its own proteins. A living cell generates the energy to produce these proteins and maintain cellular integrity (Fig. 17.02).

A key property of a living cell is that it possesses its own ribosomes for making proteins. Viruses do not contain ribosomes but use the host cell ribosomes to make proteins.

In contrast to self-sufficient living cells, viruses rely upon their host for many functions. Living cells store information as DNA and make the messages out of RNA, whereas, viruses can store their genetic information as either RNA or DNA and rely upon the host cell to make the messages out of RNA. Genuine cells possess ribosomes that are capable of making proteins. Viruses are parasitic and rely on the host cell to provide the ribosomes for translating virus mRNA into protein. Cells transport nutrients and metabolize them to generate energy and make a variety of metabolic intermediates. Viruses rely on host cell metabolism for energy and precursors.

capsid Shell or protective layer that surrounds the DNA or RNA of a virus particle
living cell A unit of life that possesses a genome made of DNA and sends genetic messages (RNA) from its genes (DNA) to its own ribosomes to make its own proteins with energy it generates itself
viral genome Molecule of DNA or RNA that carries the genes of a virus
virion Virus particle
virus Infectious agent, consisting of DNA or RNA inside a protective shell of protein, that must infect a host cell in order to replicate

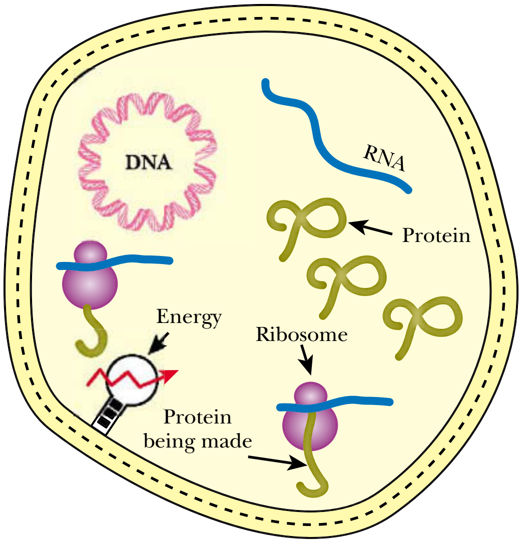


FIGURE 17.02
Characteristics of a Living Cell

This simplified cell shows all the essential characteristics of a living cell: an energy source to provide ATP (ATP synthetase), genetic information (chromosomal DNA and messenger RNA), ribosomes to convert genetic information into proteins, and a biological membrane that maintains cellular integrity.

Living cells have membranes whereas most viruses do not.

Living cells are surrounded by metabolically active cell membranes. Most simple viruses have only a protein shell and no true membrane, although, some complex virus particles have a membrane stolen from the previous host. However, the membranes around virus particles are not active metabolically either in energy generation or nutrient transport. Nonetheless, virus particles do have an outer covering and can survive on their own outside their host cells (admittedly without multiplying).

Viruses are all **parasites** that cannot multiply without a host cell. Furthermore, viruses are **intracellular parasites**; that is to say that they must actually enter the cells of the host organism to replicate. Note that not all intracellular parasites are viruses. Certain disease-causing bacteria and protozoans may enter the cells of higher organisms and live inside them as parasites. However, these parasites are nonetheless living cells themselves and contain their own ribosomes to make their own proteins. This chapter does not attempt to cover the realm of virology systematically. Rather, examples are given to illustrate novel aspects of molecular biology found among the viruses.

Life Cycle of a Virus

Virus genes subvert the host cell into manufacturing more virus particles.

A virus alternates between two forms, an inert virus particle, the virion, which survives outside the host cell, and an active intracellular stage. The life cycle of a typical virus goes through the following stages (Fig. 17.03):

a. Attachment of virion to the correct host cell
b. Entry of the virus genome
c. Replication of the virus genome
d. Manufacture of the virus proteins
e. Assembly of new virus particles (virions)
f. Release of new virions from the host cell

Attachment of a virus requires a protein on the virus particle to recognize a molecule on the surface of the target cell. Sometimes this receptor is another protein; sometimes it is a carbohydrate. Often it is a glycoprotein, that is a protein with carbohydrate groups attached. On some virus particles, the recognition proteins form

intracellular parasite Parasite that lives inside the cells of its host organism
parasite An organism or infectious agent that uses the resources of another organism in order to grow and multiply

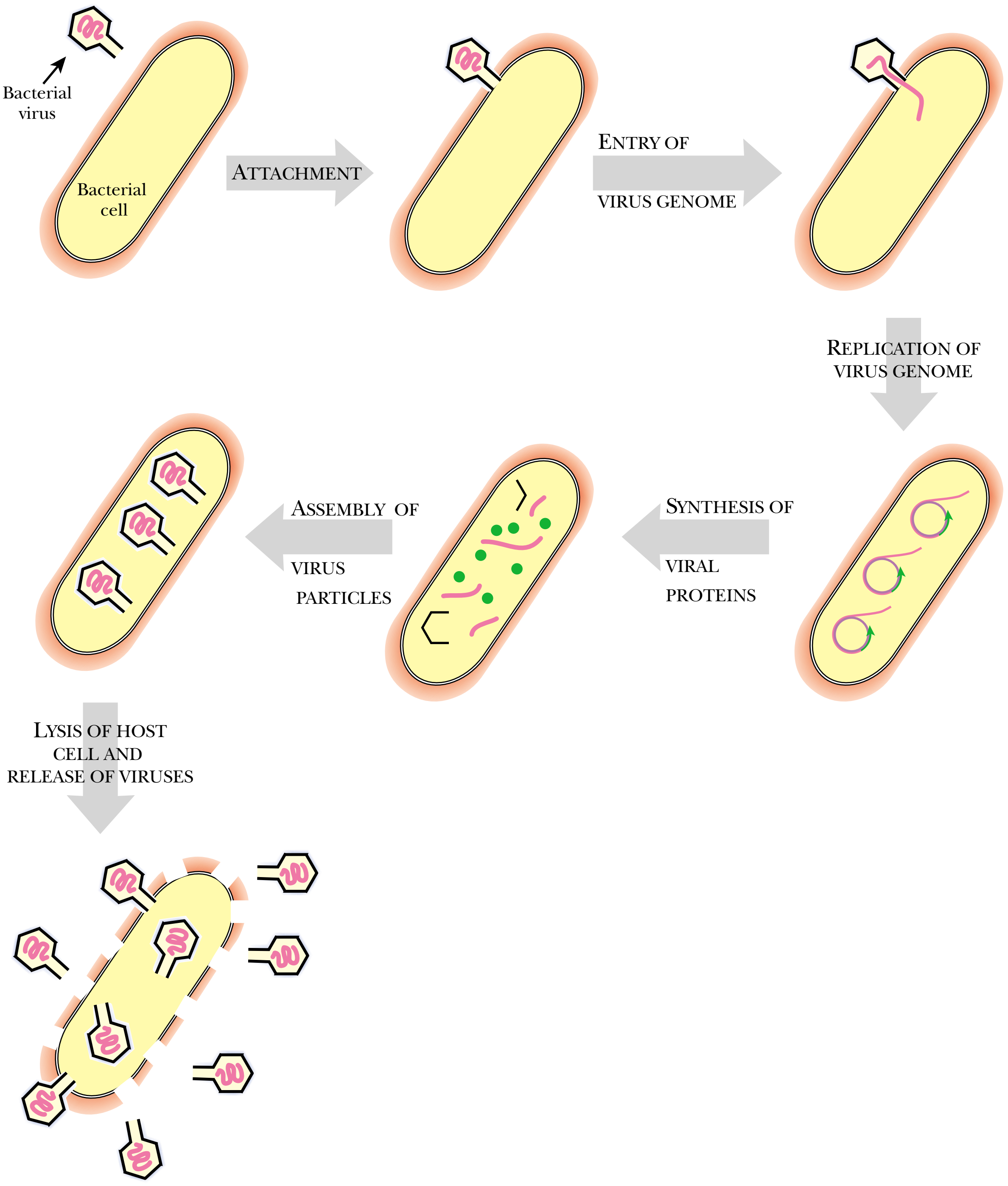

Bacterial virus
Bacterial cell
ATTACHMENT
ENTRY OF VIRUS GENOME
REPLICATION OF VIRUS GENOME
SYNTHESIS OF VIRAL PROTEINS
ASSEMBLY OF VIRUS PARTICLES
LYSIS OF HOST CELL AND RELEASE OF VIRUSES

FIGURE 17.03 ***Virus Life Cycle***

The life cycle of a virus starts when the viral DNA or RNA enters the host cell. Once inside, the virus uses the host cell to manufacture more copies of the virus genome and to make the protein coats for assembly of virus particles. Once multiple copies of the virus have been assembled, the host cell is burst open to allow the viral progeny to escape and find new host cells to infect.

Before entering the host cell, a virus must bind to a receptor on the cell surface.

spikes or prongs sticking out from the surface. Most bacterial and plant viruses abandon their protein coat when they infect a new host cell. Only the genetic material (DNA or RNA) enters the cell. Animal viruses vary in regard to when exactly they disassemble their protein coat.

Many animal viruses have an extra envelope outside the protein shell. This is made of membrane stolen from the previous host cell into which virus proteins have been

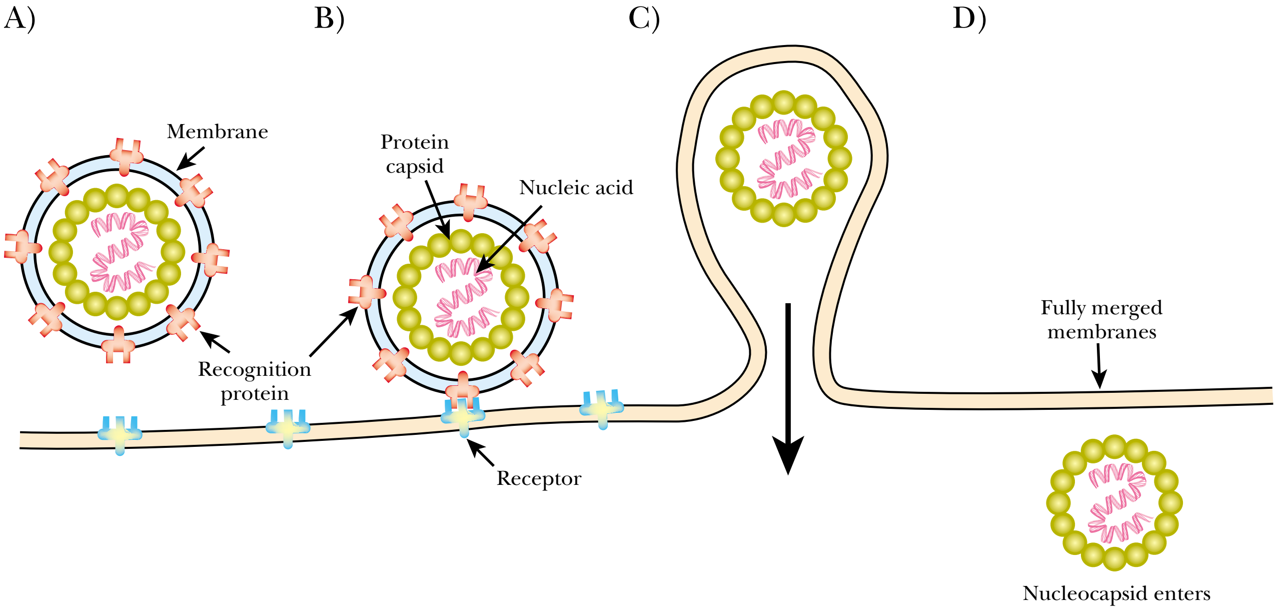


FIGURE 17.04 ***Enveloped Viruses Merge with the Animal Cell Membrane***

When the virus particle left the previous host cell, it surrounded itself with a layer of the host cell membrane. This outer layer contains viral recognition proteins previously inserted into the host cell membrane during virus infection. The recognition proteins bind to the cell membrane receptors of another animal cell. The protein complex triggers the animal cell to take in the particle by fusing the two membranes. The nucleocapsid structure enters the animal cell.

inserted. These virus-encoded proteins detect and bind to receptors on the next target cell. When an enveloped virus enters a new animal cell, its envelope layer merges with the cell membrane and the inner protein shell containing the nucleic acid (the "**nucleocapsid**") enters (Fig. 17.04). Once inside, the protein shell disassembles, exposing the genome.

Once inside the host cell, the virus genome has two major functions. First, it must replicate to produce more virus genomes. Second, it must subvert the cell to manufacture lots of virus proteins for the assembly of new virus particles. Note that viruses do not divide like cells. They are assembled from components manufactured by the host cell using genetic information in the virus genome (Fig. 17.05).

Viruses do not divide. Instead their genes code for components which are assembled into new virus particles.

The genes of viruses are often divided into "**early genes**" and "**late genes**". The early genes have promoters that resemble those of the host cell and encode for proteins responsible for replicating the virus genome. Consequently, they are transcribed by host cell RNA polymerase and are expressed immediately after infection. In very small viruses, host enzymes are largely responsible for replicating the virus genome so there may be very few "early genes" involved in replication. Conversely, in viruses that have large numbers of genes, such as bacteriophage T4 or the poxviruses of animals, regulation is obviously more complex and there may be several sub-categories of genes such as "immediate early", "delayed early" etc.

The late genes have promoters that are not recognized by host polymerase alone. These genes are expressed late in infection and encode the structural proteins of the virus particle together with proteins involved in the assembly and packaging processes and in lysing the host cell. Some viruses, such as bacteriophage T7, encode their own RNA polymerase in order to express late genes, others, such as bacteriophage T4, modify the host RNA polymerase. For example, T4 gene 55 encodes an alternative sigma factor that recognizes the promoters of T4 late genes.

early genes Genes expressed early during virus infection and that mainly encode enzymes involved in virus DNA (or RNA) replication
late genes Genes expressed later in virus infection and that mainly encode enzymes involved in virus particle assembly
nucleocapsid Inner protein shell of a virus particle that contains the nucleic acid

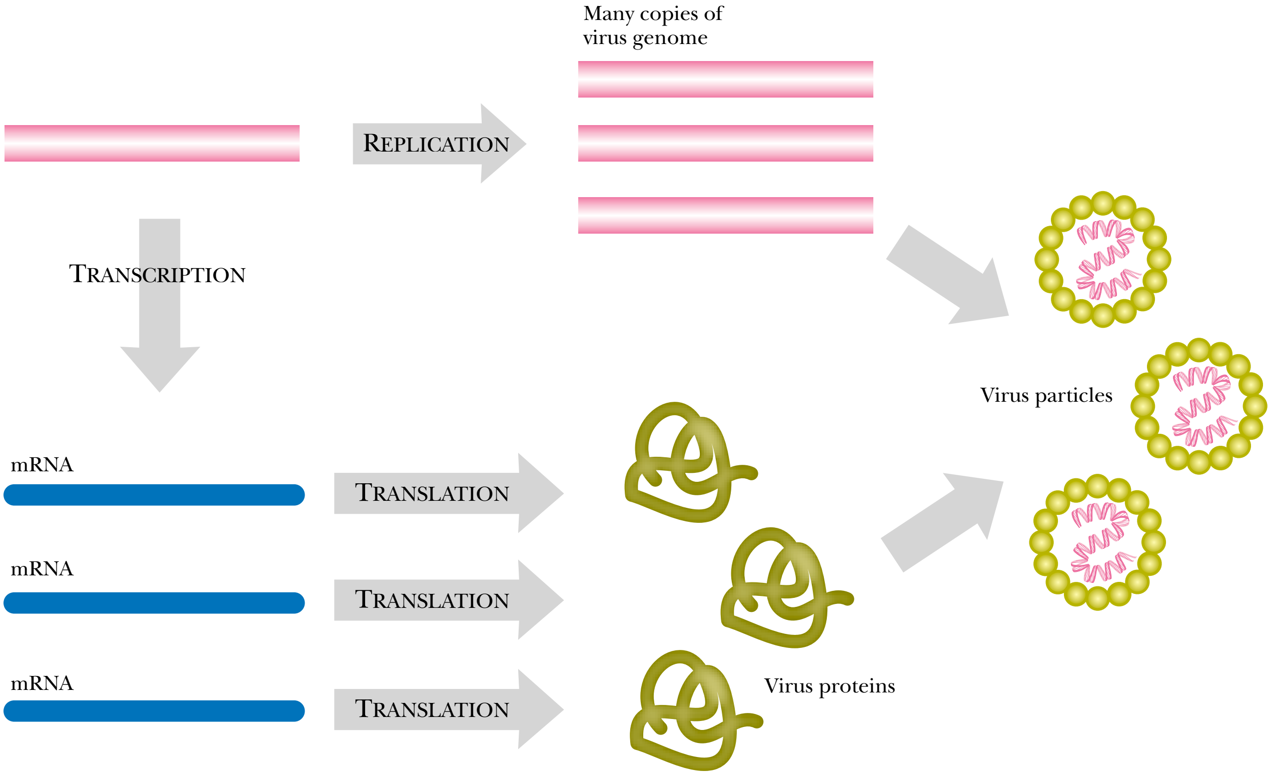


FIGURE 17.05 ***Synthesis and Assembly of Virus Components***

The viral genome (in pink) directs the host cell to replicate many copies of the virus genome. The viral genome is also transcribed and the mRNA is translated, giving viral proteins. The viral genome carries all the genes needed for making the protein coat. Finally, the coat proteins and the viral genomes are assembled to give new virus particles.

Bacterial Viruses are Known as Bacteriophage

Viruses that infect bacteria are often called **bacteriophage** or **phage** for short. Phage is derived from the Greek for "eat" and refers to the way in which bacterial viruses eat holes or "**plaques**" in a lawn of bacteria growing on the surface of agar (Fig. 17.06). Bacterial viruses were heavily used in the early days of molecular biology to investigate the nature of the gene. Because viruses only contain DNA or RNA surrounded by a protein coat and because bacteriophage infect the simplest of all cells, bacteria, they proved highly convenient.

Bacteria have a cell wall protecting their cell membrane and so bacterial viruses cannot simply merge with the membrane, as do animal viruses. Therefore, bacterial viruses do not bother with an outer envelope layer. They just have a protein shell surrounding the DNA or RNA. After binding to the cell surface, they inject their nucleic acid into the bacterial cell and the outer protein coat of the virus particle is left behind (Fig. 17.07). Many well-known bacterial viruses have a complex capsid that resembles a miniature moon-lander. The capsid has an eicosahedral head, a tail and six-landing legs with attachment proteins at the tips. The tail contracts and injects the DNA through the bacterial cell envelope.

Bacterial viruses leave their protein shell behind and only their genome enters the host cell.

In 1952 Hershey and Chase performed a classic experiment using bacteriophage T4 to demonstrate that only the DNA entered the host cell, *E. coli* (Fig. 17.08). The protein and the DNA of the virus particles were radioactively labeled with two dif-

bacteriophage (phage) Virus that infects bacteria
phage Short for bacteriophage, a virus that infects bacteria
plaque (When referring to viruses) A clear zone caused by virus destruction in a layer of cultured cells or a lawn of bacteria

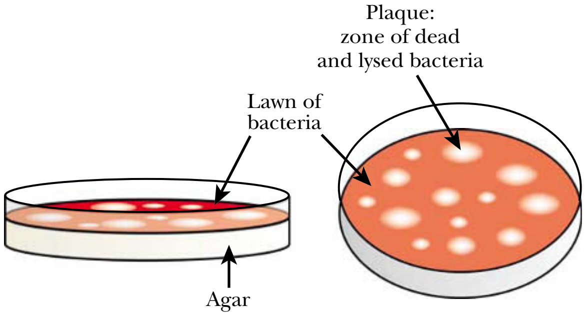


FIGURE 17.06 ***Formation of Plaques in Lawn of Bacteria***

Bacteriophage are viruses that infect bacteria. To isolate individual types of bacteriophage, plaques are made. A mixture of bacteriophage is added to a large number of bacteria, and poured onto a nutrient agar plate. The bacteria grow quickly, covering the agar with a cloudy layer of bacteria, known as a lawn (red). Wherever a bacteriophage infects a bacterial cell, it destroys the cell and produces many more bacteriophage. These spread out to infect neighboring bacteria, forming a clear zone in the lawn that is called a plaque. Each plaque contains descendents of the single original bacteriophage that landed in that region of the lawn. If needed, purified lines of bacteriophage can be isolated from individual plaques.

Phages are the Most Numerous Life Form

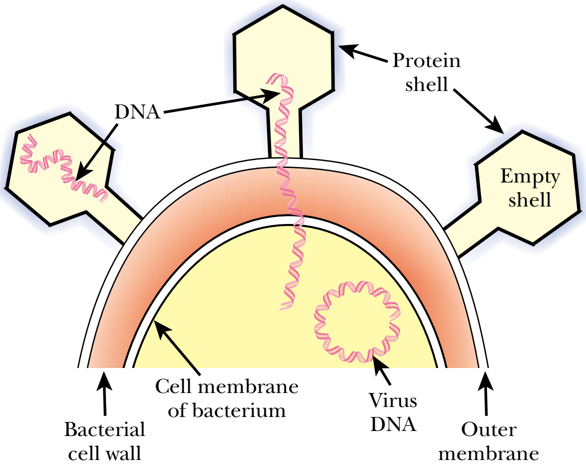
It is likely that there are more bacteriophages on our planet than any other life form. There are an estimated 10^{30} bacteria on the planet and it is estimated that there are probably about 10 phages for every living bacterial cell, which would give a total of around 10^{31} phage. Virus particles, including bacteriophage, are ubiquitous on earth. Examination of seawater under the electron microscope has shown typical counts of 50×10^6 virus particles per ml. It has been estimated that phages destroy up to 40% of the bacteria in the ocean every day. Remnants of these lysed bacteria add significant amounts of organic matter to the ocean water and may affect global carbon cycling. A colossal amount of novel genetic material is present in the vast number of phages present in natural habitats. Preliminary surveys have indicated that around 75% of the genes carried by phages are unrelated to anything presently in the DNA data banks.

Many of the virus particles in the environment are probably orphaned in the sense that susceptible host cells are no longer available in their habitats. In some cases the host cell may be extinct or perhaps only mutants resistant to the virus have survived. Conversely, many virus particles are inherently defective and are incapable of successfully infecting host cells, even if available. This is especially true of RNA viruses where the mutation rate is extremely high. The benefit of a high mutation rate is that the virus constantly changes and so evades recognition by the host defense systems. The downside is that most mutations are deleterious and a high percentage of defective virus genomes are made. Indeed, for some RNA viruses the majority of virus particles released are defective mutants and only a minority are infectious particles.

ferent isotopes. The labeled viruses were added to bacterial cells and the fate of the two radioactive labels was followed. The protein, labeled with ^{35}S, was left outside and the DNA, labeled with ^{32}P, entered the cells. Moreover, some of the ^{32}P labeled DNA was found in the new generation of virus particles liberated when the infected cells burst. Since only the virus DNA enters the host cell and the other components are abandoned outside, this provides further evidence that nucleic acids rather than proteins carry genetic information. Historically, the Hershey and Chase experiment

FIGURE 17.07 ***Bacterial Viruses Inject their Nucleic Acid***

To enter a bacterial cell, bacterial viruses must get their genomes through three layers, the bacterial outer membrane, the cell wall, and the inner membrane. The bacterial wall structure prevents the virus from simply merging membranes, as in animal cells. To overcome the defenses, the virus punches a hole through the three layers and injects its DNA (or RNA) into the cytoplasm.



demonstrated for the first time that DNA alone was the carrier of the genetic information and that the associated protein was not required. These findings prompted other researchers to investigate the structure of DNA and its role as the genetic material.

Lysogeny or Latency by Integration

Viruses may replicate aggressively, killing the host cell. Alternatively, they may limit themselves to duplicating their genome in step with cell division.

When an infecting virus generates many virus particles and destroys the cell, this is known as **lytic growth** because the cell is burst or lysed. When instead, the virus divides in step with the host chromosome, this is known as **lysogeny** and a cell containing such a virus is a **lysogen**. The term **latency** means the same as lysogeny but is usually used when referring to animal cells. In Chapter 16 we discussed the close relationships between plasmids and viruses. Some gene creatures can choose to live either as a plasmid or as a virus. Some plasmids are probably derived from viruses that have lost the ability to grow lytically. Conversely, some viruses may have evolved from plasmids that obtained the genes for lytic growth, either from another virus or, over a longer period, from the host cell.

Lysogeny or latency means that the virus has decided to divide in step with the host cell instead of killing it. It does not necessarily mean the virus has decided to live as a plasmid. Many cases of lysogeny or latency are caused by integration of the virus DNA into a host cell chromosome. Such an integrated virus is known as a **provirus** (or **prophage** in the case of bacterial viruses). The virus DNA becomes a physical part of the chromosome and is replicated when the chromosome divides.

The bacterial virus **lambda (λ)**, which infects the bacterium *E. coli*, recognizes and integrates into a special sequence of DNA on the chromosome of its host cell, known as ***att*λ (λ attachment site)**. Integration occurs by site-specific recombination as described in Ch. 14. This allows lambda to occasionally pick up and carry bacterial genes as described in the chapter on bacterial genetics (Ch. 18). Some animal viruses, such as the herpes viruses, also insert themselves into the chromosomes of their host cells. Some have special recognition sites, while others insert at random. Retroviruses,

***att*λ (λ attachment site)** Recognition sequence on the chromosome of *Escherichia coli* where bacteriophage lambda integrates
lambda (λ) Virus that infects the *Escherichia coli* and may integrate into a special sequence of DNA on the bacterial chromosome
latency Type of virus infection in which the virus becomes largely quiescent, makes no new virus particles and duplicates its genome in step with the host cell. Same as lysogeny but used of animal viruses
lysogen A cell containing a lysogenic virus
lysogeny Type of virus infection in which the virus becomes largely quiescent, makes no new virus particles and duplicates its genome in step with the host cell. Same as latency but used of bacterial viruses
lytic growth Type of infection in which a virus generates many virus particles and destroys the cell
prophage Bacteriophage genome that is integrated into the DNA of the bacterial host cell
provirus Virus genome that is integrated into the host cell DNA

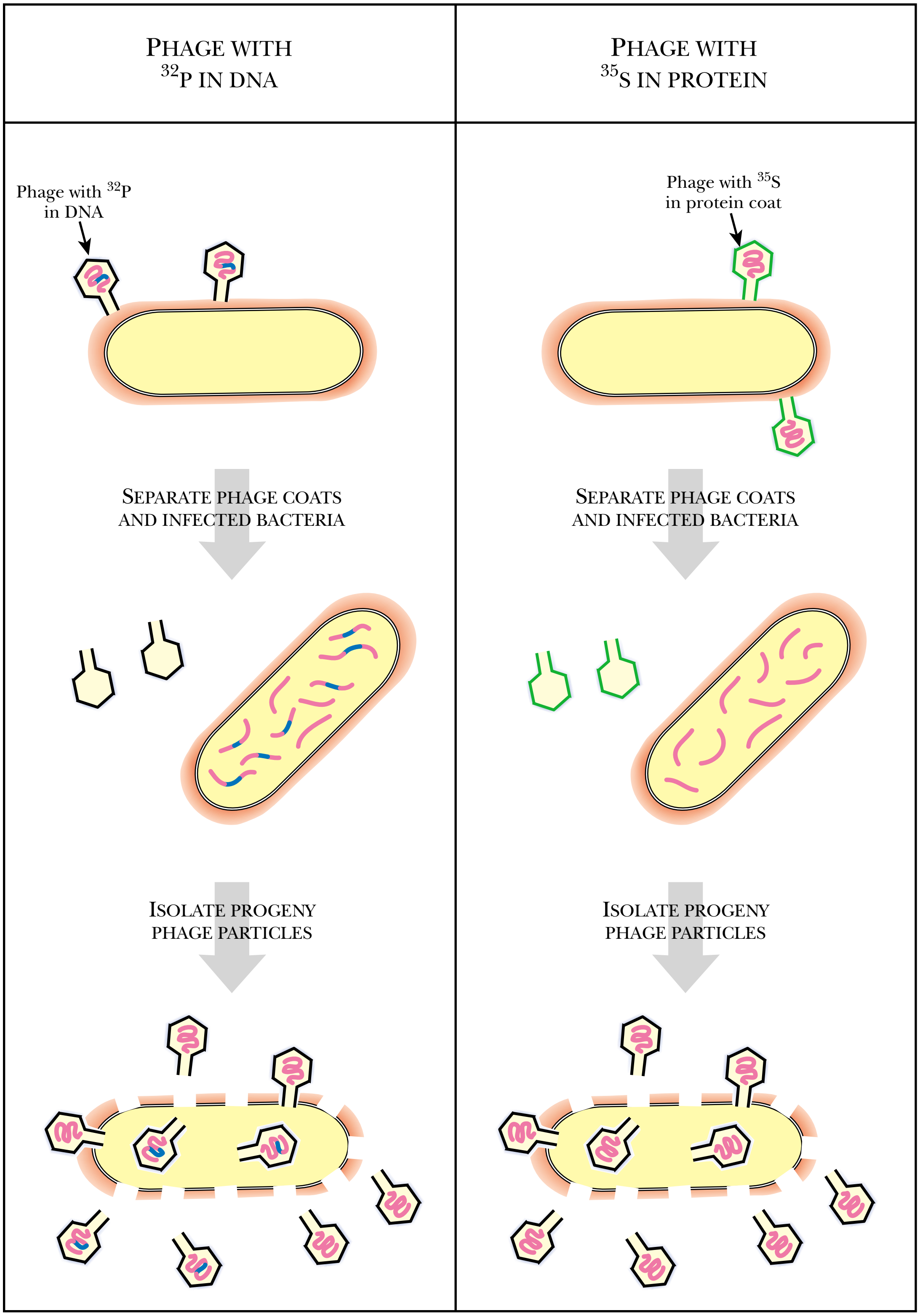

PHAGE WITH
^{32}P IN DNA
PHAGE WITH
^{35}S IN PROTEIN
Phage with ^{32}P in DNA
Phage with ^{35}S in protein coat
SEPARATE PHAGE COATS AND INFECTED BACTERIA
SEPARATE PHAGE COATS AND INFECTED BACTERIA
ISOLATE PROGENY PHAGE PARTICLES
ISOLATE PROGENY PHAGE PARTICLES

FIGURE 17.08 ***Only the DNA of Bacteriophage T4 Enters the Cell***

In a classic experiment by Hershey and Chase, the viral DNA was found to enter the bacterial cell but the protein coat did not. To trace the DNA and protein, bacteriophage T4 was first grown with radioactive precursors (^{32}P-dATP and ^{35}S-methionine, respectively). The radioactively labeled T4 were isolated and purified. Phage labeled with ^{32}P (blue DNA) were added to new host *E. coli*, and the *E. coli* were separated from remaining virus particles (left panel). After viral replication, the progeny viruses were isolated. The ^{32}P labeled DNA was found in these viral particles, implying that the original virus DNA entered the host *E. coli*, and was eventually packaged in new particles. The same experiment was performed with ^{35}S-methionine labeled T4 (green coats), but the final viral particles did not contain any ^{35}S-methionine (right panel). This suggested that none of the original protein coat entered the host *E. coli*.

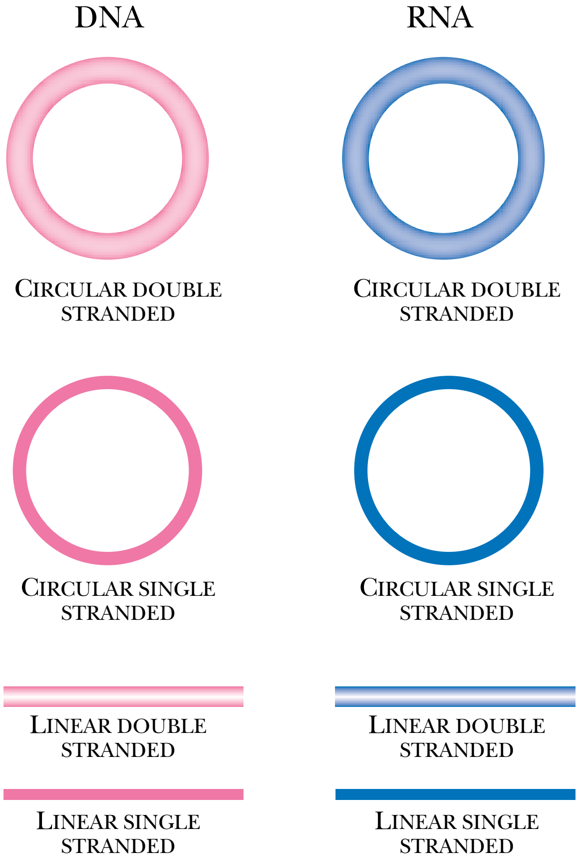


FIGURE 17.09 ***Variety of Virus Genomes***

Viral genomes come in all shapes and sizes. They may be comprised of DNA or RNA. They may be circular or linear, and they may be double-stranded or single-stranded.

which carry their genes as RNA in the virus particle, must first make a DNA copy of themselves to insert into the host chromosome (see below).

The Great Diversity of Viruses

Different viruses have from three to several hundred genes.

Viruses have been found that attack animal cells, plant cells and bacterial cells. There is colossal variation in the structure of viruses and the detailed way in which they take over the cells they invade. The smallest viruses have only three genes; the largest have two or three hundred and carry out some extremely complicated genetic maneuvers to outwit their host cells. The largest known virus genome is that of Bacteriophage G, which infects *Bacillus megaterium*. It has nearly 700 genes, more than some bacteria. Some viruses have DNA genomes while others have RNA. Furthermore, the nucleic acid may be either single or double stranded and either linear or circular. All of these possibilities exist (Fig. 17.09; Table 17.01), though some are more common than others. Some viruses even have segmented genomes made up of several pieces of DNA or RNA. We shall survey a range of viruses, partly to illustrate their genetic diversity. We will also include some viruses that are widely used in molecular biology or are especially notorious for causing disease.

Many viruses replicate their genomes by some form of rolling circle mechanism.

The details of virus replication also vary considerably. Nonetheless, many viruses use versions of the rolling circle scheme for replication that has been discussed in Chapter 16, "Plasmids". Viruses whose particles contain linear DNA or RNA will often circularize after infection in order to perform rolling circle replication. Viruses with single-stranded DNA or RNA synthesize a second strand upon infection making a double-stranded circular version of the genome, known as the **replicative form (RF)**. DNA viruses may or may not rely on host enzymes for replication, but RNA viruses

replicative form (RF) Circular double-stranded version of a virus genome used for rolling circle replication

TABLE 17.01 Families of Viruses

Virus Family	Typical Examples
Double-Stranded DNA Viruses	
Myoviridae	T4-like bacteriophages
Siphoviridae	lambda-like bacteriophages
Fuselloviridae	bacteriophages of *Sulfolobus*
Poxviridae	cowpox, *Vaccinia*, smallpox (= *Variola*), ectromelia
Baculoviridae	baculoviruses of insects
Herpesviridae	herpes, chickenpox (*Varicellavirus*)
Adenoviridae	adenovirus
Single-Stranded DNA Viruses	
Inoviridae	small filamentous bacteriophage e.g. fd
Microviridae	small spherical bacteriophage e.g. ΦX174
Geminiviridae	assorted plant viruses, e.g. beet curly top virus
Parvoviridae	parvoviruses, adeno-associated virus
Reverse Transcribing Viruses	
Hepadnaviridae	hepatitis B virus, cauliflower mosaic virus
Retroviridae	human immunodeficiency virus (HIV)
Double-Stranded RNA Viruses	
Reoviridae	reoviruses, bluetongue virus of sheep, rotaviruses
Birnaviridae	*Drosophila* X virus
Totiviridae	*Saccharomyces cerevisiae* virus L-A
Negative Single-Stranded RNA Viruses	
Paramyxoviridae	parainfluenza, measles, mumps
Rhabdoviridae	vesicular stomatitis virus, rabies
Filoviridae	Marburg virus, Ebola virus
Orthomyxoviridae	influenza
Bunyaviridae	hantavirus, tomato spotted wilt virus
Positive Single-Stranded RNA Viruses	
Leviviridae	bacteriophages MS2 and Qβ
Picornaviridae	polio, common cold (*Rhinovirus*), hepatitis A, foot-and-mouth virus
Tombusviridae	tobacco necrosis virus (TNV)
Flaviviridae	yellow fever, hepatitis C
Togaviridae	rubella (German measles), tobacco mosaic virus (TMV)

must provide a special RNA polymerase to replicate their RNA genomes, an **RNA replicase**.

Virus particles come in a wide range of sizes and shapes.

Viruses may also be classified according to the structure of the virus particle, or virion. The three major shapes seen are spherical, filamentous and complex. Spherical viruses are not truly spherical but have 20 triangular faces and are thus icosahedrons (Fig. 17.10). Cross sections through an icosahedron may be five- or six-sided depending on where the cut is made. As with most other biological filaments, filamentous viruses consist of helically arranged protein subunits forming cylindrical shells. These may be either open or closed off at the ends. Complex viruses are large viruses with multiple structural components that do not fit neatly into the spherical versus filamentous classification. In addition, outer envelopes, partly derived from the cytoplasmic membrane of the previous host cell, surround some virus particles (Fig. 17.11).

Small Single-Stranded DNA Viruses of Bacteria

Bacteriophage ΦX174 is a small simple spherical virus that contains 5,386 bases of circular single-stranded DNA. At each of the 12 vertices of the virion is a spike made of

bacteriophage ΦX174 A small spherical virus that contains circular single-stranded DNA and infects *Escherichia coli*
RNA replicase Special RNA polymerase used by RNA viruses to replicate their RNA genomes

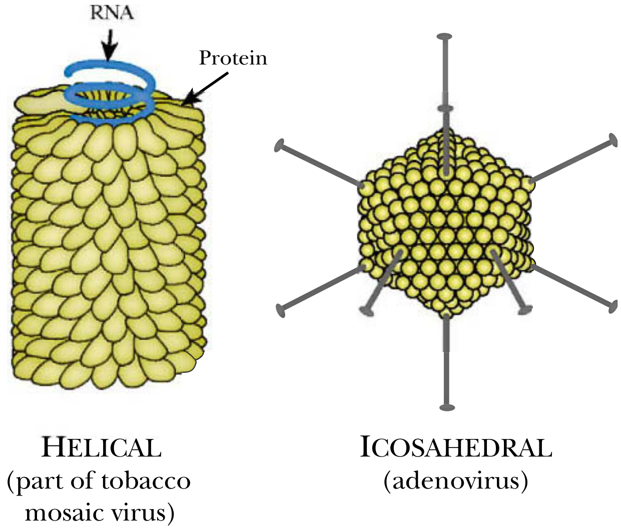


FIGURE 17.10 ***Filamentous and Spherical Virus Structures***

Filamentous viruses are actually built form helical arrays of protein. The proteins of tobacco mosaic virus coat the helical RNA molecule to form a cylinder. Spherical viruses, such as adenovirus, are actually icosahedral. They have a 20-sided form that may or may not have protein spikes sticking out from the protein coat at the vertexes. These protruding knobs carry the proteins that recognize virus receptors on the host cell surface.

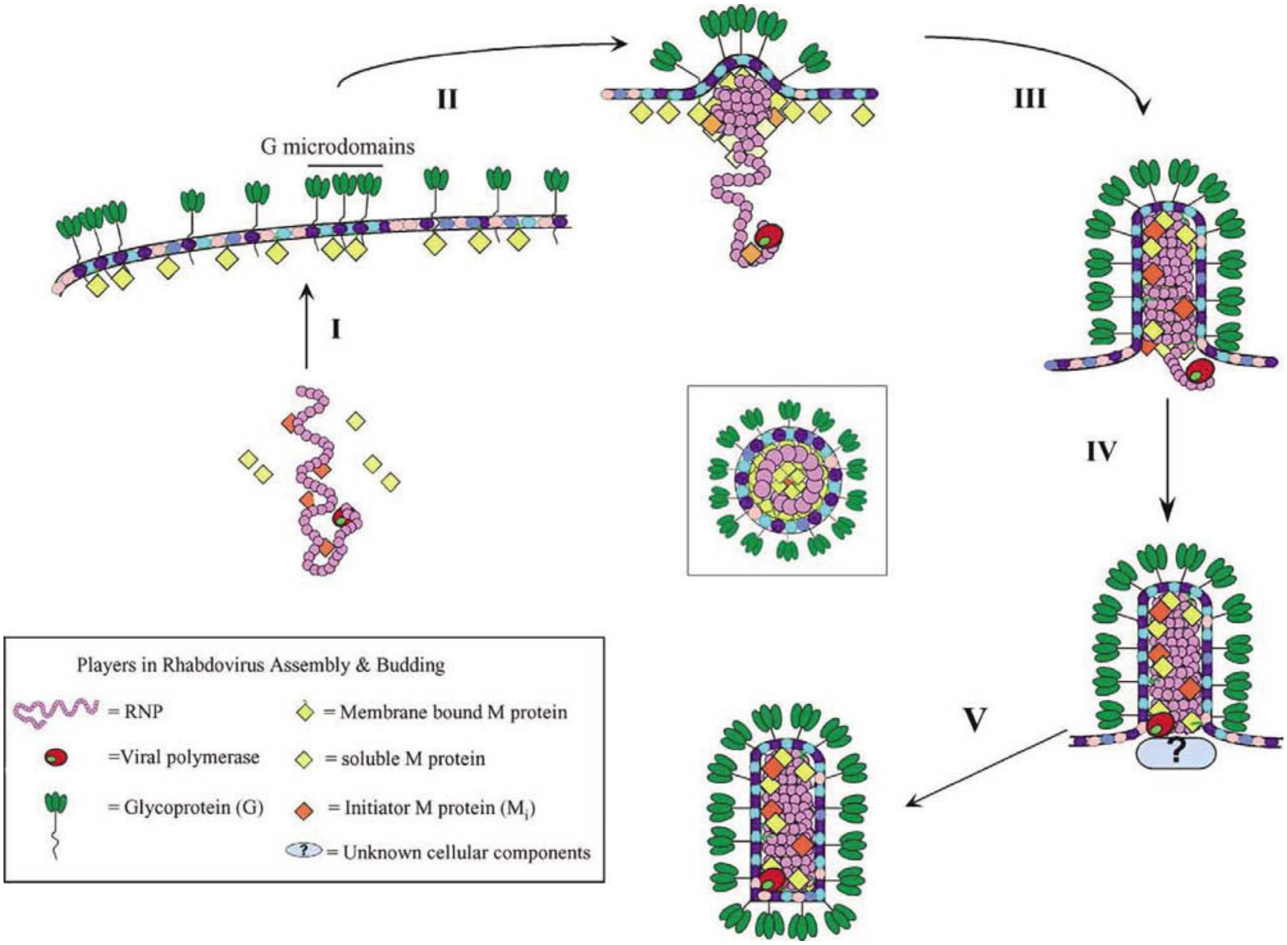

II
III
G microdomains
I
IV
V
?
Players in Rhabdovirus Assembly & Budding
= RNP
=Viral polymerase
= Glycoprotein (G)
= Membrane bound M protein
= soluble M protein
= Initiator M protein (M_i)
? = Unknown cellular components

FIGURE 17.11 ***Rhabdovirus is an Enveloped Virus***

Rhabdoviruses are examples of enveloped RNA viruses. The group includes vesicular stomatitis, rabies and related viruses. The envelope is generated by budding of the virus from the cytoplasmic membrane of the infected host cell, as shown. (I) The viral ribonucleoprotein (RNP) associates with the inside of the membrane via M protein. (II) RNP condensation begins and bud sites form. Budding occurs preferentially where the viral glycoprotein (G) has clustered. (III) Assembly and extrusion of the virion. (IV) Cellular proteins of unknown identity (?) complete the budding and promote release of virions (V) via membrane fission. The inset shows a virion in cross-section with its central M protein scaffold and glycoprotein spikes protruding from the viral envelope. From Jayakar et al, Rhabdovirus assembly and budding. Virus Research (2004) 106:117.

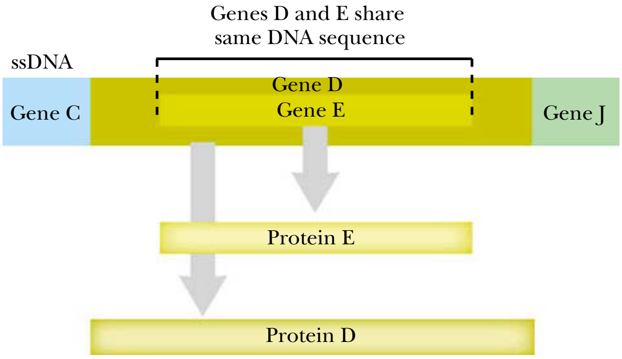


FIGURE 17.12
Bacteriophage ΦX174 has Overlapping Genes

Since many viral genomes are very small, the same region of the genome sometimes encodes two different genes. In the ΦX174 genome, the gene for E protein is entirely embedded within the gene for D protein. The two gene sequences are not identical, even where they overlap, since D uses a different reading frame than E.

Small viruses quite often have overlapping genes.

two different proteins involved in recognizing the host cell. Overall, ΦX174 looks rather like a World War II naval mine. The most interesting property of ΦX174 is that it has so little DNA that five of its 11 genes overlap others. For example, gene E is completely inside the DNA for gene D. Genes D and E are read in two different reading frames so they produce two totally different proteins (Fig. 17.12). D-protein helps assemble the virus capsid, though it does not form part of the final structure, and E-protein destroys the cell wall of the host bacterium to allow the newly made virus particles out. A mutation in the DNA for gene E will also alter gene D. Although overlapping saves on DNA, the two genes are no longer free to evolve separately. Although overlapping genes are quite often found in small viruses, they are only found in real cells under exceptional circumstances.

Some other small bacterial DNA viruses are filamentous rather than spherical. For example, **bacteriophage M13** has single stranded circular DNA like ΦX174 but the virus particle is a long thin filament. M13 is **"male-specific"** which means that it only infects bacteria carrying the F-plasmid i.e. "male" bacteria (Ch. 18). M13 does not kill its host when it releases the progeny, therefore, scientists use the bacteriophage to make single-stranded DNA for sequencing (see Ch. 24). Since the host is not dead more DNA can be isolated if needed.

Complex Bacterial Viruses with Double Stranded DNA

Complex bacterial viruses have a head, containing DNA, plus a tail that contracts for injecting their genome into the host.

There is a large family of bacterial viruses that all have a complex form made up of a head, tail, and tail fibers (Fig. 17.13). The head of the virus particle contains a large molecule of linear dsDNA. They include bacteriophages T4, Lambda, P1 and Mu, which are all used in bacterial genetics (see Ch. 18) and molecular biology. The number of genes ranges from Mu with approximately 40 genes to T4 with nearly 200. T4 and its close relatives are some of the most complex types of virus known.

The head of the Mu virus particle is more or less spherical (i.e., it is a symmetrical icosahedron). However, the head of T4 is relatively elongated in order to accommodate the much greater amount of DNA. Attached to the head is a tail with tail fibers that act as landing legs. These viruses bind to bacterial cells by means of recognition proteins on the end of their tail fibers. After setting down like lunar landers, their tails contract and they inject their DNA like miniature hypodermic syringes.

bacteriophage M13 A small male-specific filamentous virus that contains circular single-stranded DNA and infects *Escherichia coli*
male-specific phage Virus that only infects "male" bacteria, i.e., those bacteria carrying the F-plasmid

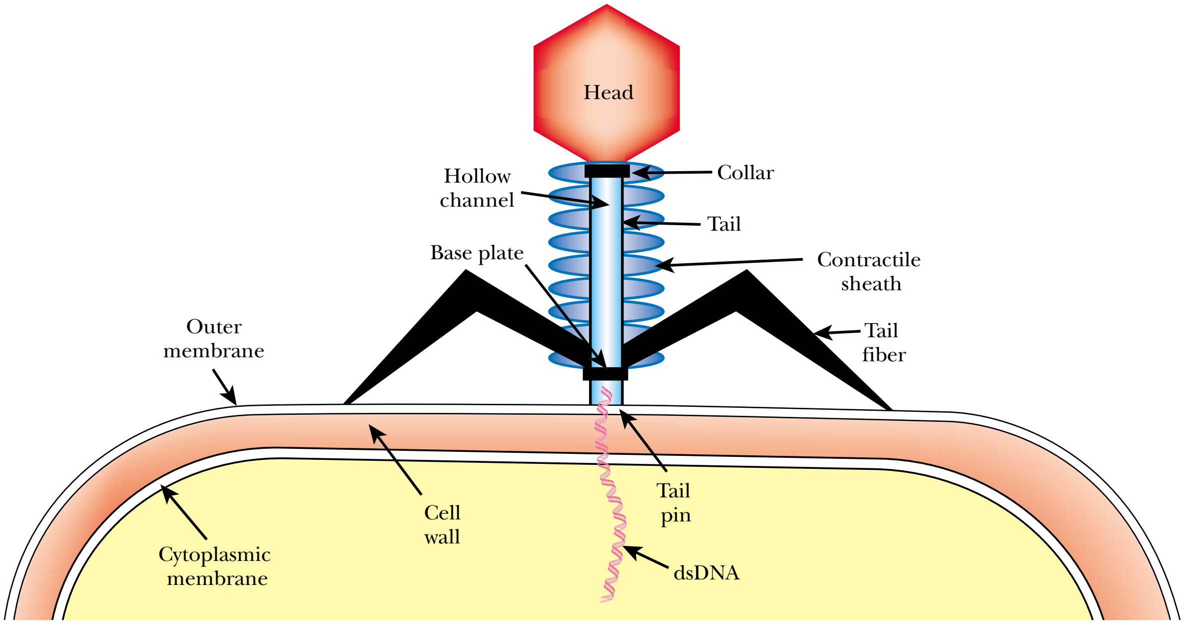


FIGURE 17.13 ***Bacteriophage with Heads and Tails***

Complex bacterial viruses such as Mu and T4 have a head region that holds the dsDNA, a tail region that injects the DNA into the bacteria, and tail fibers that recognize specific proteins on the bacterial outer membrane and so facilitate attachment. The contractile sheath helps inject the DNA from the virus head into the inside of the bacterium.

DNA Viruses of Higher Organisms

Herpesviruses can remain latent for long periods of time.

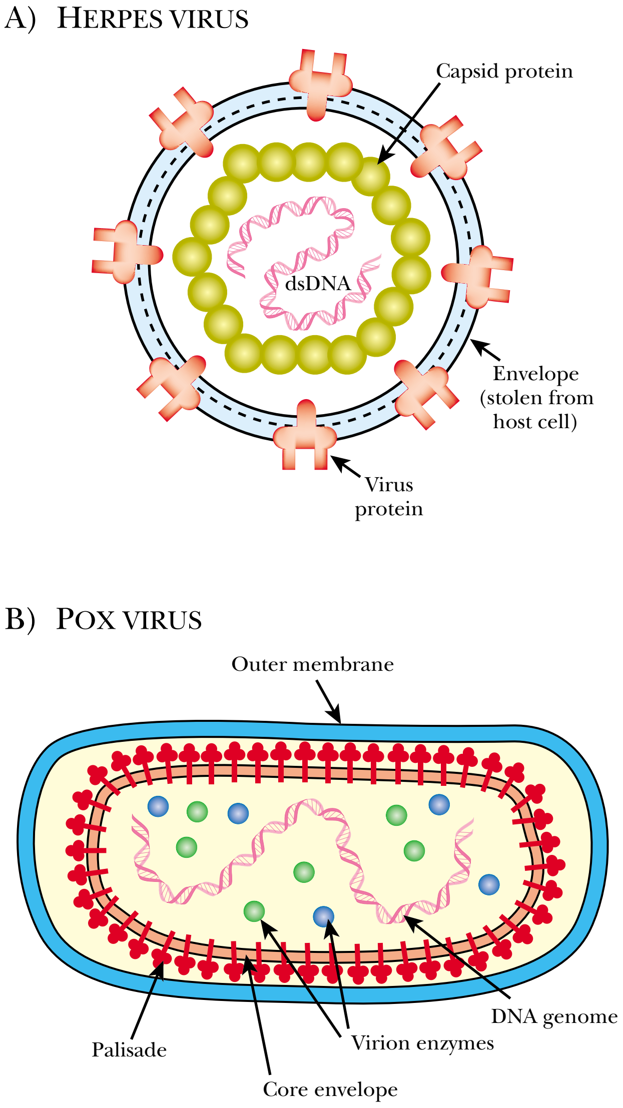
Most DNA viruses of animals contain double stranded DNA. For example, **Simian Virus 40 (SV40)** is a smallish, spherical virus that causes cancer in monkeys by inserting its DNA into the host chromosome. Another double stranded DNA virus, **Herpes virus**, is spherical with an extra outer envelope of material stolen from the nuclear membrane of the host cell (Fig. 17.14). The internal nucleic acid with its protein shell is referred to as the nucleocapsid. This family includes viruses that cause cold sores and genital herpes as well as chickenpox and infectious mononucleosis. The herpes viruses are difficult to cure completely as they are capable of remaining in a latent state where they cause no damage but merely replicate in step with the host cell. Active infections may then break out again after a long period of quiescence, due to stress or other factors.

Poxviruses are the largest animal viruses.

Poxviruses are the most complex animal viruses and are so large they may just be seen with a light microscope (Fig. 17.14). They are approximately 0.4 by 0.2 microns, compared to 1.0 by 0.5 microns for bacteria like *E. coli*. Unlike other animal DNA viruses which all replicate inside the cell nucleus, poxviruses replicate their dsDNA in the cytoplasm of the host cell. Virus particles are manufactured inside subcellular factories known as inclusion bodies. Poxviruses have 150 to 200 genes, about the same number as the T4 family of complex bacterial viruses.

Plant viruses containing DNA are relatively rare. One example is **cauliflower mosaic virus**, which has circular DNA inside a small spherical shell and kills cauliflower and its relatives such as cabbages and Brussels sprouts. This virus is of note because the promoters from some of its genes are extremely strong and have been used in plant

cauliflower mosaic virus A small spherical virus of plants with circular DNA. Some of its promoters are used in plant genetic engineering
herpes viruses A family of spherical animal DNA viruses with an outer envelope of material stolen from the nuclear membrane of the host cell
poxviruses A family of large and complex dsDNA animal viruses with 150 to 200 genes
simian virus 40 (SV40) A small, spherical dsDNA virus that causes cancer in monkeys by inserting its DNA into the host chromosom


A) Herpes virus

B) Pox virus

FIGURE 17.14 ***Herpes Viruses and Pox Viruses***

Many animal viruses use dsDNA for their genomes. The Herpesvirus is a simple virus that has a protein coat and outer envelope surrounding the double-stranded DNA genome. The Poxvirus has two envelope layers. A protein layer, known as the palisade, is embedded within the core envelope. In addition, premade viral enzymes are also packaged with the genome to allow replication immediately upon infection. These viruses infect animals and in both cases the outermost viral membrane is derived from the membrane of the previous host cell.

genetic engineering to express insect-killing toxins or other transgenes. Most plant viruses have RNA genomes, as will be discussed below.

Strong promoters from cauliflower mosaic virus are used to express genes in the genetic engineering of plants.

Viruses with RNA Genomes Have Very Few Genes

The smallest genomes are those of RNA-containing viruses. As explained in Ch. 13, “Mutations”, this is related to their high mutation rates. RNA-based genomes have rates of mutation that are 1000-fold higher than DNA. Since each mutant gene product must continue to interact with other virus components, this limits the number of genes to no more than a dozen or so. The advantage of a high mutation rate to these viruses is that it allows them to change their proteins rapidly so evading recognition by host defense systems. The overall success of this strategy is shown by the fact that the majority of the best-known viral diseases such as flu are due to RNA viruses.

RNA viruses have small genomes and very high mutation rates.

Given an RNA genome, there are three alternatives: double-stranded RNA, single-stranded positive RNA and single-stranded negative RNA. The terms positive and negative refer to the coding strand and the non-coding strand, respectively. Remember that when messenger RNA is made only one of the two DNA strands is used as the template (see Fig. 17.15). The mRNA will be complementary in sequence to the template strand and identical to the non-transcribed strand of the DNA (except for using U instead of T in RNA). The non-transcribed strand of DNA and the mRNA are both

The story of how Jenner invented the technique of **vaccination** against infectious diseases is well known. He observed that milkmaids who worked in close contact with cattle often caught cowpox, a mild disease. After recovering, they became immune to the much more severe disease, smallpox. So Jenner then tested this connection experimentally in 1796. After deliberately infecting patients with cowpox he found that they had indeed become immune to smallpox. Jenner called the material he used "vaccine" after the cows where it originated (vacca = cow in Latin) and the novel technique was named vaccination.

For a long time it was believed that cowpox virus and the vaccine virus were identical. However, in 1939 it was found that cowpox virus (re-isolated from cows) and the virus cultures maintained for use in vaccination were actually quite distinct. The vaccine cultures were then named *Vaccinia* to distinguish them from cowpox. It seems clear that Jenner himself did indeed vaccinate his patients with cowpox. Since viruses could not be cultured or purified in the laboratory until much later, those who followed Jenner constantly re-isolated new strains of what they thought was cowpox virus from cows and horses. At some time the original cowpox was replaced by a different virus, the present *Vaccinia* virus. No virus presently circulating in the wild corresponds to *Vaccinia* so its origin is unknown. Furthermore, the differences between *Vaccinia* and cowpox are too large for mutation to be responsible for cowpox evolving into *Vaccinia*. Presumably another poxvirus was circulating among cattle and horses during the nineteenth century and was eventually kept because it is even milder than cowpox.

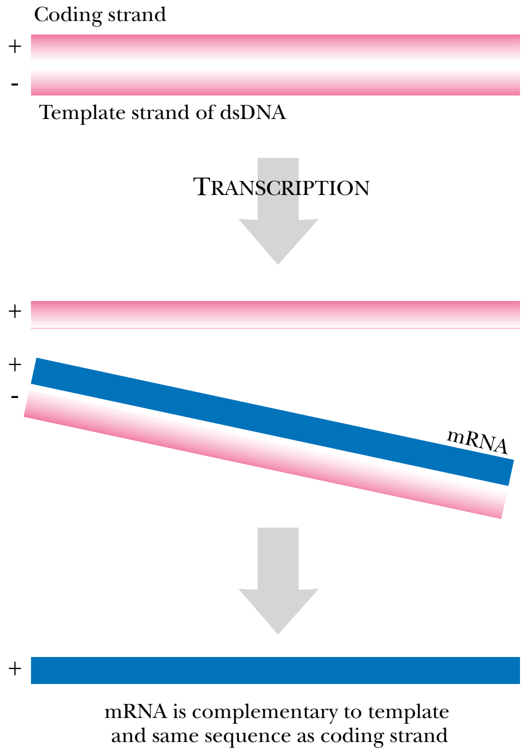

Coding strand
+
-
Template strand of dsDNA
TRANSCRIPTION
+
+
-
mRNA
+
mRNA is complementary to template and same sequence as coding strand

FIGURE 17.15 ***Plus and Minus Strands***

Double stranded DNA has two strands, the template strand and the coding strand. The coding strand can also be referred to as the plus strand and the template strand as the minus strand. The sequence of messenger RNA is by definition the coding sequence and is identical to the plus strand and complementary to the minus strand.

coding strands and are therefore **positive or "plus" strands** whereas the template strand is the **negative or "minus" strand**.

If a virus particle contains single-stranded RNA (ssRNA), there are two alternatives. The virus RNA can be either the plus strand or the minus strand. If the virus con-

negative or "minus" strand The non-coding strand of RNA or DNA
positive or "plus" strand The coding strand of RNA or DNA
vaccination Immunization of a patient by introducing a milder or inactivated form of the disease-causing agent

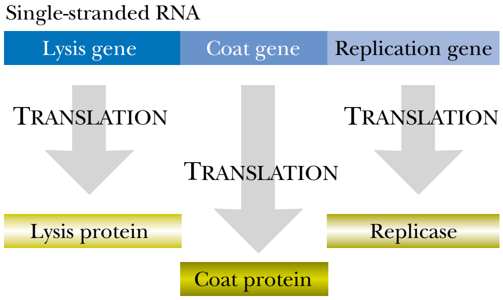


FIGURE 17.16 ***Genome of Bacteriophage Qβ***

Qβ is a small spherical virus that contains a short single-stranded RNA genome. The virus only makes three proteins, a replicase to replicate its RNA, a coat protein, and lysis protein to lyse the infected bacterial cell. Although the genome is very small, the three proteins are all that the virus needs to successfully infect bacteria and replicate itself.

tains the plus strand it can use this RNA directly as a messenger RNA to make proteins immediately upon entering the host cell. Such viruses are called positive-strand RNA viruses. Conversely, if the virus particle contains the minus strand, it must first make the complementary, positive strand before moving on to manufacture proteins. Despite this apparent disadvantage, it is the negative-stranded RNA viruses that appear to be the most successful and widespread.

Bacterial RNA Viruses

Some of the smallest virus genomes are found among the RNA-containing bacteriophages. A minimalist virus can get by with only three genes (Fig. 17.16). These encode the protein coat, the RNA replicase that replicates the viral genome and the lysis protein that destroys the wall of the host cell so the newly made virus particles can get out. An example is bacteriophage Qβ that has approximately 3,500 bases of single stranded RNA and infects the bacterium *Escherichia coli*. Qβ and its relatives, such as MS2, are small spherical viruses with single stranded RNA and only three or four genes. They are "male-specific" as they only infect bacteria carrying the F-plasmid (i.e. "male" bacteria) because they attach to the sex pilus, which is only found on the surface of F^+ cells (see Ch. 18). Very few bacterial viruses have double stranded RNA.

Double Stranded RNA Viruses of Animals

Double-stranded RNA viruses are relatively uncommon. Among animal viruses, the reoviruses are the best-known dsRNA viruses. They are spherical, with two concentric protein shells but no envelope and contain a dozen or so separate dsRNA molecules, each coding for a single virus protein. Their most famous member is the rotavirus (Fig. 17.17) which causes infant diarrhea, a disease infecting vast numbers of babies and small children.

Positive-Stranded RNA Viruses Make Polyproteins

Positive strand RNA viruses contain RNA that can be directly translated.

Picornaviruses are small, spherical single-stranded RNA viruses. They include polio, common cold, hepatitis A, and foot and mouth disease. Their genome is long enough for about a dozen genes. Since they are positive-strand RNA viruses, their RNA can be directly used as mRNA. The viral RNA does have a poly(A) tail attached to its 3′-end but has no cap. Instead a protein, the Vpg protein, is attached covalently to the 5′-

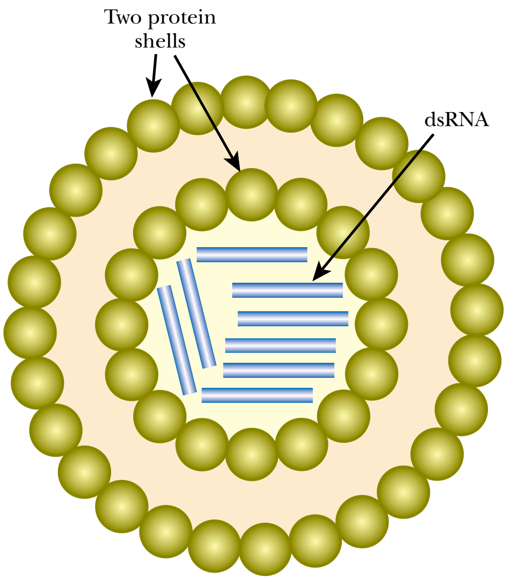


FIGURE 17.17 ***Rotaviruses Have Multiple Double-Stranded RNA Molecules***

Rotavirus particles have two protein coats that encapsulate about a dozen double-stranded RNA molecules. Each of the dsRNA pieces encodes one of the proteins necessary for viral survival and propagation.

Polyproteins are made from a single giant gene and then cut up to give several final proteins.

end of the virus RNA. However, there is a technical problem. Unlike bacteria where a single mRNA molecule may code for several proteins (an operon; see Ch. 6), in higher organisms each molecule of mRNA only encodes a single protein. Eukaryotic ribosomes will only translate the first reading frame on an RNA message, even if it carries several. Picornaviruses do indeed use their positive ssRNA molecule directly as a messenger RNA. They avoid the problem by using the RNA to code for a single giant polypeptide that uses all of their genetic information (Fig. 17.18). This "**polyprotein**" is then chopped up into 10 to 20 smaller proteins.

Strategy of Negative-Strand RNA Viruses

The RNA in negative strand RNA viruses is the antisense stand.

The negative-strand RNA viruses are divided into several families and include the agents of well-known diseases such as rabies, mumps, measles and influenza as well as more exotic emerging pathogens such as Ebola virus. In all of these, the ssRNA in the virus particle is complementary to the messenger RNA and is therefore the minus strand. These viruses vary in shape and structure but are similar in having an outer envelope derived from the membrane of the host cell where they were assembled.

After infiltrating the cell, the first mission of a negative-strand RNA virus is to make its RNA double stranded by synthesizing the corresponding positive RNA strand. Once it has the two strands it uses them both as templates. The plus strand is used as a template to manufacture more negative strands for the next generation of virus particles. The minus strand is used as a template to manufacture multiple positive strands that act as mRNA molecules (Fig. 17.19). This strategy is not only an effective division of labor but also avoids the problem of translating multiple reading frames on a single incoming virus RNA molecule.

Plant RNA Viruses

Most plant viruses are small with just a few genes and contain single-stranded RNA. Some, like cucumber mosaic virus, are spherical, while others are rod-shaped, like **tobacco mosaic virus (TMV)**, the most widespread plant virus. TMV attacks lots of plants including vegetables like the tomato, pepper, beet and turnip as well as tobacco.

polyprotein A long polypeptide that is cut up to generate several smaller proteins
tobacco mosaic virus A filamentous single-stranded RNA virus that infects a wide range of plants

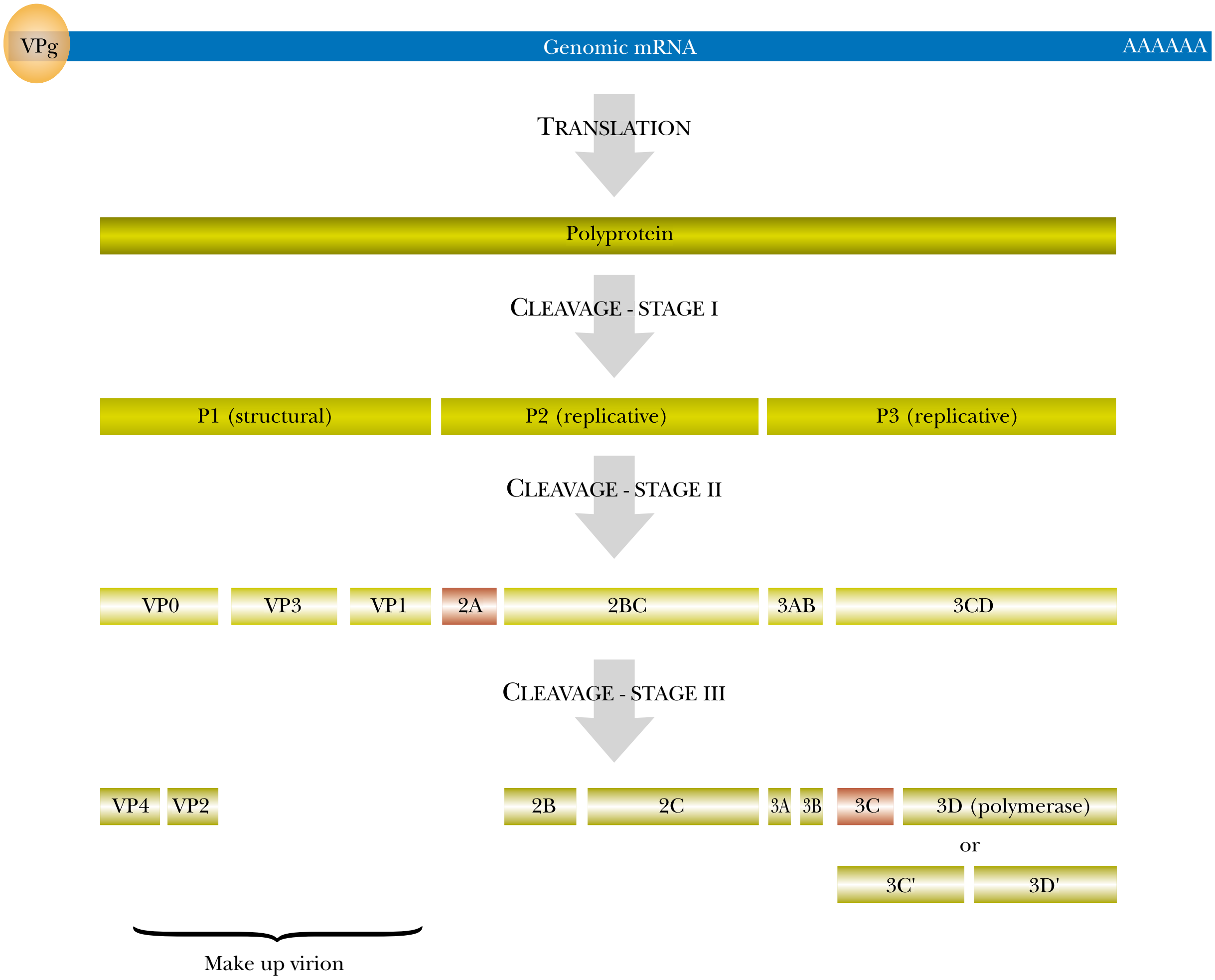


FIGURE 17.18 ***Polyprotein Strategy of Plus Strand RNA Viruses***

Rather than having many positive single-stranded RNA molecules, the picornaviruses have one large positive ssRNA molecule (top). Translation of the ssRNA produces a large polyprotein that is cleaved into successively smaller pieces. In Stage I the structural proteins are cleaved from the replicative proteins. In Stage II the segments are cleaved into separate proteins. For example, the P1 structural region is cleaved into VP0, VP3, and VP1 and thenVP0 is cleaved into VP4 and VP2. The four proteins VP1, VP2, VP3 and VP4 combine to form the coat of the virion. The proteins shown in red are proteases. The Vpg protein is attached to the 5′-end of the virus RNA.

The name "mosaic" refers to the yellowish blotches on the leaves of infected plants (Fig. 17.20). The virus coat consists of 2,130 identical protein molecules arranged in a helix with the RNA in the center (Fig. 17.21). TMV and its relatives have around 10,000 bases of RNA, enough for about 10 genes.

TMV is notable for its use in early structural investigations. X-ray crystallography and other physical methods were used to show that cylindrical viruses are in fact made of helically arranged protein subunits. In addition, the capsids of TMV show **self-assembly**. Purified capsid proteins plus virus RNA will spontaneously form virus particles when mixed. The virus may be disassembled by cooling and reassembled by gentle warming. This illustrates that the capsid proteins are held together largely by the hydrophobic force (see Ch. 7) which (unlike other forms of chemical bonding) grows weaker at lower temperatures.

self-assembly The spontaneous assembly of a biological structure from its subunits

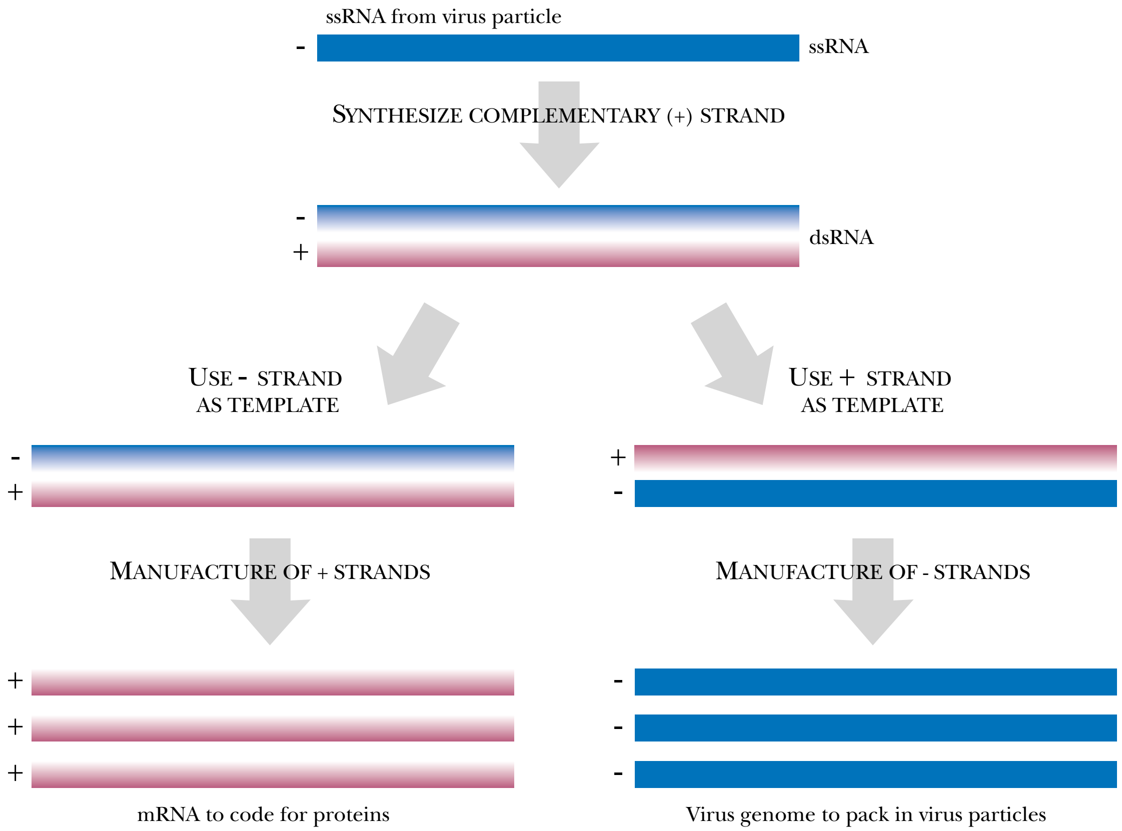

ssRNA from virus particle
-
ssRNA
SYNTHESIZE COMPLEMENTARY (+) STRAND
-
+
dsRNA
USE - STRAND AS TEMPLATE
USE + STRAND AS TEMPLATE
-
+
+
-
MANUFACTURE OF + STRANDS
MANUFACTURE OF - STRANDS
+
+
+
-
-
-
mRNA to code for proteins
Virus genome to pack in virus particles

FIGURE 17.19 ***Strategy of Negative-Strand RNA Virus***

The negative strand of RNA has a sequence complementary to the coding strand. Therefore, viruses that use this type of genome must synthesize the complementary plus strand upon entry into the host cell. The plus RNA strand can then be used as a template to manufacture more viral genomes (right side). The negative RNA strand is then free to manufacture more copies of the plus strand (left side). These act as mRNA and direct viral protein synthesis.

Retroviruses Use both RNA and DNA

Retroviruses carry RNA in the particle but make a DNA copy after entering the host cell.

The **retroviruses** infect animals and include the notorious **AIDS** virus. They are unique in having both RNA and DNA versions of their genome. The virus particle contains single-stranded RNA (Fig. 17.22) inside two protein shells surrounded by an outer envelope layer. The retrovirus envelope is made from the cell membrane of its previous victim. This membrane layer has retrovirus proteins both inserted through it and covering its surface (Fig. 17.23).

What makes a retrovirus "retro" is that upon entering a host cell it reverses the normal flow of genetic information by making a DNA copy of its RNA genome. Upon infecting an animal cells the retrovirus converts the ssRNA into a double-stranded DNA copy by using the enzyme **reverse transcriptase** (Fig. 17.24). The retrovirus DNA is then inserted into the host cell DNA. Once integrated, the retrovirus DNA is never excised but remains permanently inserted in the host genome. Consequently, retroviruses are impossible to get rid of completely after infection and integration, at least using current medications and procedures.

AIDS (aquired immunodeficiency syndrome) Disease caused by human immunodeficiency virus (HIV) that damages the immune system
retroviruses A family of animal viruses with single-stranded RNA inside two protein shells surrounded by an outer envelope. They possess reverse transcriptase which is used to convert the ssRNA version of the genome to a dsDNA copy
reverse transcriptase An enzyme that uses single-stranded RNA as a template for making double-stranded DNA

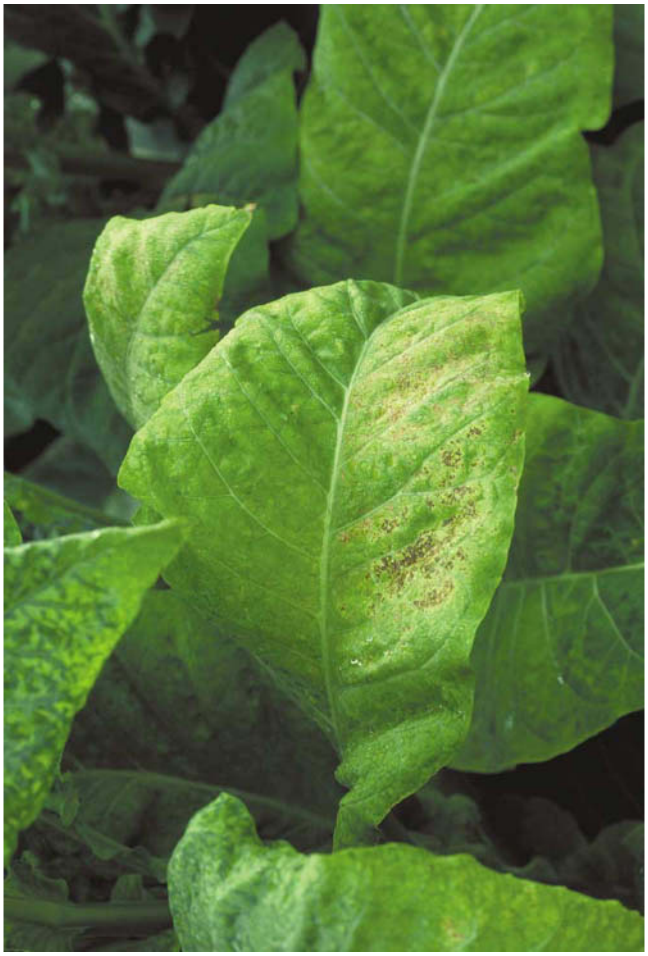


FIGURE 17.20 ***Plants with Tobacco Mosaic Virus***

Tobacco plant infected with tobacco mosaic virus. Lafayette County, Florida. Credit: Norm Thomas, Photo Researchers, Inc.

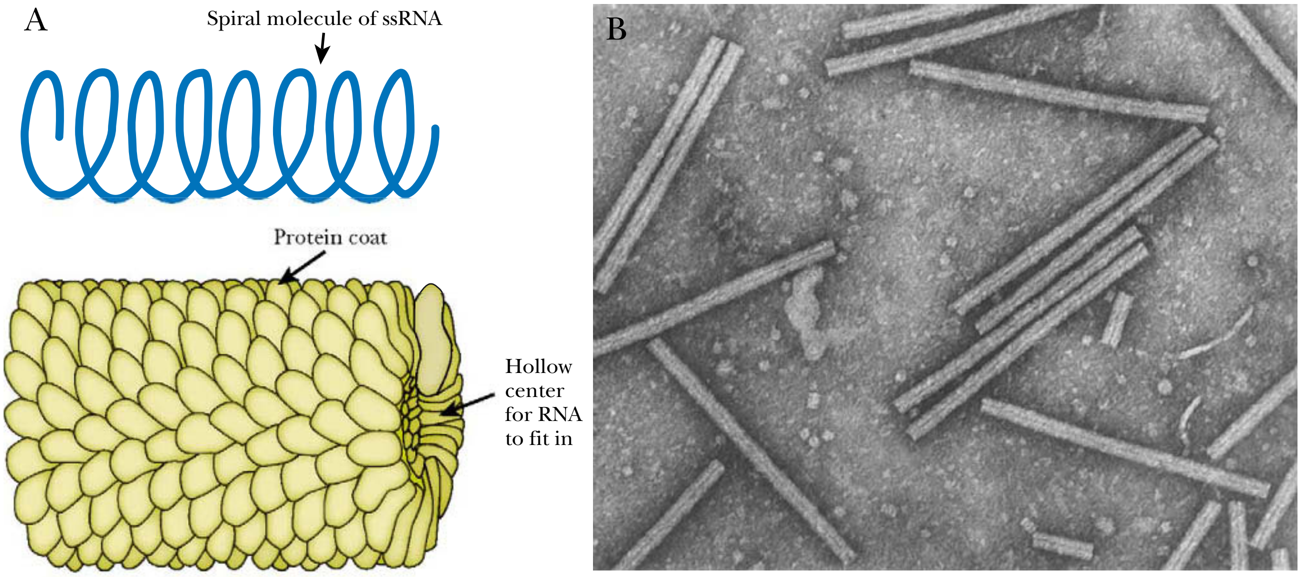


FIGURE 17.21 ***Tobacco Mosaic Virus Structure***

Tobacco mosaic virus has a simple cylindrical structure. (A) Its single-stranded RNA genome is packaged inside a shell of proteins that are arranged as a helix. (B) Electron micrograph of Tobacco mosaic virus rod-shaped particles. Copyright 1994, Rothamsted Experimental Station.

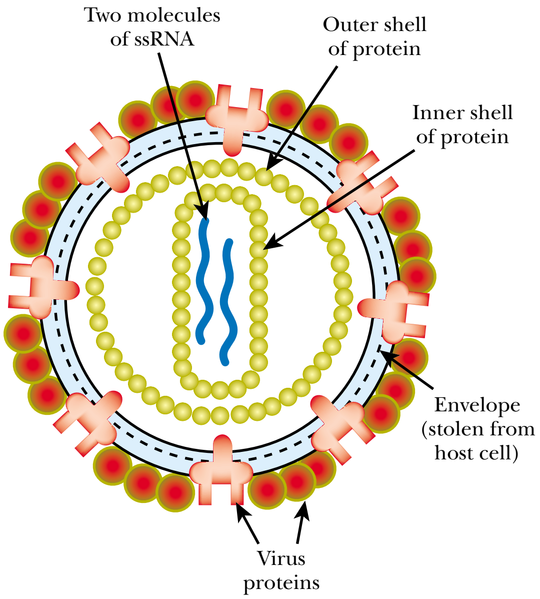


FIGURE 17.22 ***Retrovirus Particle***

Retroviral particles such as HIV have two single-stranded RNA molecules surrounded by a double protein shell. Like many other animal viruses, the retrovirus has an outer membrane derived from the host cell. This envelope contains viral proteins that coat the entire surface of the virus particle. Some of the proteins are embedded in the outer envelope while others sit upon the surface.

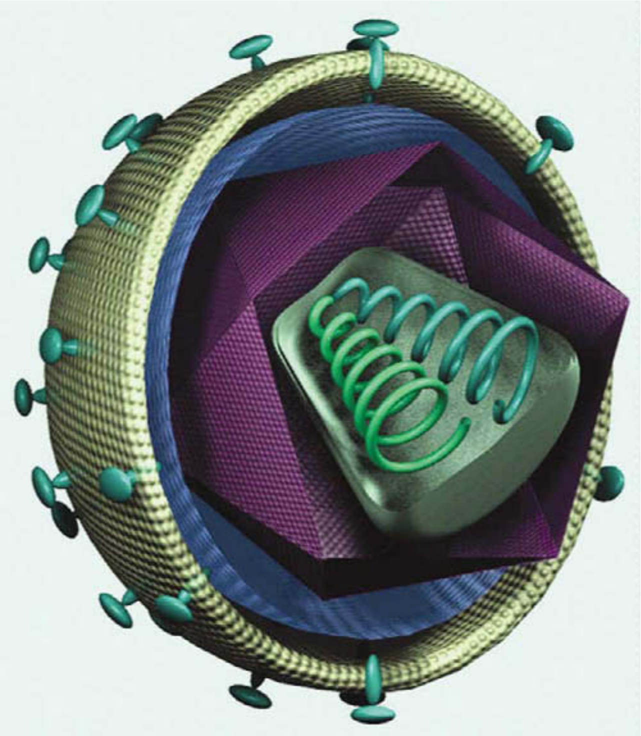


FIGURE 17.23 ***Structure of Retrovirus Particle***

Three dimensional cut-away view of a retrovirus particle. The two coils at the center (green and pale blue) are the copies of the virus genomic RNA. The purple and dark blue layers surrounding the RNA are protein shells. The outermost fawn layer is the viral envelope, with protruding proteins. Copyright 2002, The Universal Virus Database of the International Committee on Taxonomy of Viruses.

Reverse transcriptase synthesizes DNA from an RNA template.

Retrovirus particles carry ssRNA of the plus conformation. Although this RNA has a sequence identical to a messenger RNA, it is not used as an mRNA. Instead it is used as a template for reverse transcriptase to make a DNA copy of the retrovirus genome (Fig. 17.24). Reverse transcriptase first uses ssRNA to make a complementary DNA strand. It then degrades the RNA and uses the first DNA strand as a template to make a second DNA strand. The process of making a double-stranded DNA copy from an RNA sequence is known as **reverse transcription** and its discovery forced the first major revision to the Central Dogma of Molecular Biology. Previously it was believed that information flowed only from DNA to RNA to protein, never in reverse.

reverse transcription The process in which single-stranded RNA is used as a template for making double-stranded DNA

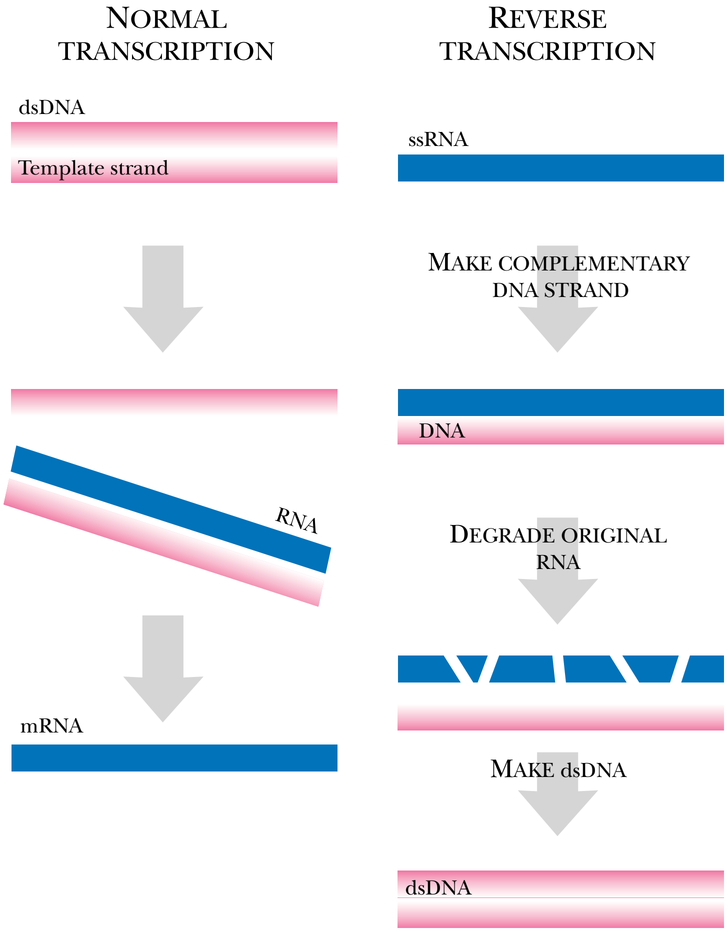


FIGURE 17.24 ***Reverse Transcriptase***

During normal transcription, mRNA is made using the template strand of DNA (left). Retroviruses convert their ssRNA to a double-stranded DNA molecule using an enzyme called reverse transcriptase. Making double-stranded DNA from ssRNA is a two step process. First a complementary strand of DNA is made forming an RNA–DNA hybrid molecule. The original RNA is then degraded and a second DNA strand is made.

Although a retrovirus genome consists of single-stranded RNA, the virus particle actually contains two identical ssRNA molecules (Fig. 17.25). These are bound together by base pairing with two molecules of transfer RNA stolen from the previous host cell. In addition to binding the two virus RNA molecules together in the virion, the tRNA has another role. It is used as a primer by the reverse transcriptase when starting a new strand of DNA.

The DNA version of a retrovirus integrates into the chromosomes of the host cell.

When a retrovirus enters a new cell the outer envelope merges with the host cell membrane and the core particle or nucleocapsid is released into the cytoplasm and disassembles, liberating the ssRNA. One of the two ssRNA molecules is then used by the reverse transcriptase to make the double stranded DNA form of the retrovirus. This dsDNA now enters the nucleus of the host cell. The retrovirus dsDNA has repeated sequences at each end, the **long terminal repeats (LTRs)**. These are direct not inverted repeats and are required for integration of the retrovirus DNA into the host cell DNA (Fig. 17.26A). The site of integration is more or less random and once integrated, the retrovirus DNA is there to stay. It has become a permanent part of the host cell chromosome.

The integrated retrovirus DNA is transcribed to give messenger RNA molecules that are capped and tailed just like the mRNA of a typical eukaryotic cell (Fig. 17.26B; see Ch. 12 for processing of mRNA). The retrovirus RNA molecules exit the nucleus to the cytoplasm. Some are translated to produce viral proteins and others are packaged into virus particles. Thus the virus particle actually contains mRNA molecules. However, when infecting a new cell, this mRNA is used as a template to make DNA instead of being used as a message. Because the incoming viral RNA does not get trans-

long terminal repeats (LTRs) Direct repeats found at the ends of the retrovirus genome which are required for integration of the retrovirus DNA into the host cell DNA

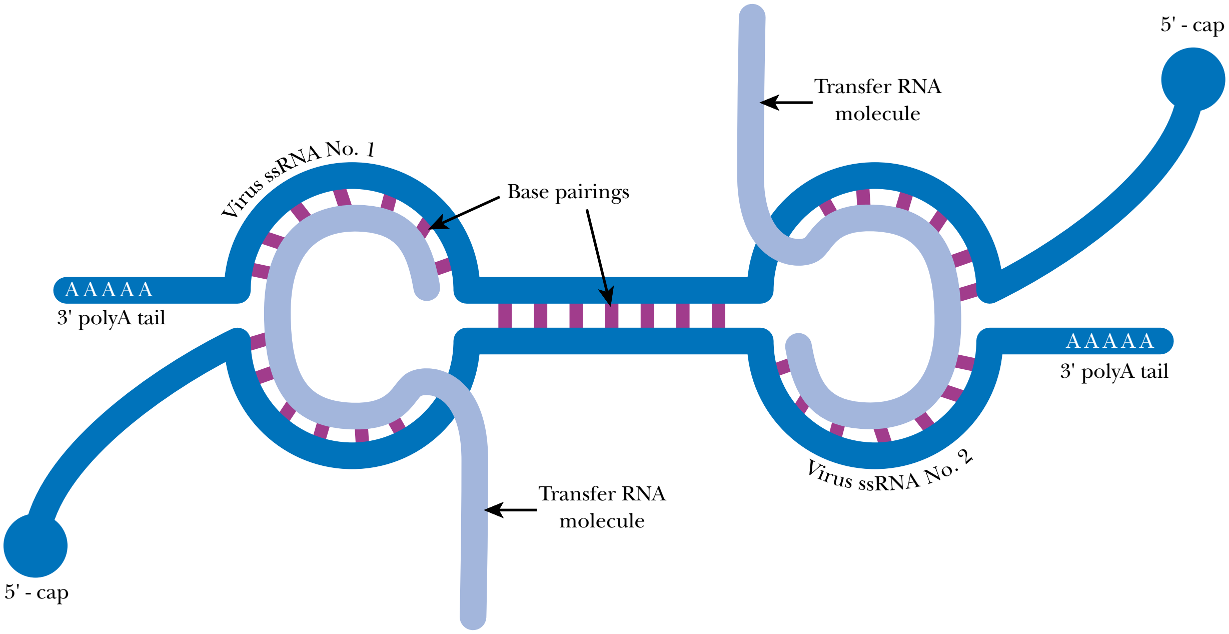


FIGURE 17.25 ***Retrovirus Particle Contains Two ssRNA Molecules***

The genome of retroviruses has a unique structure. Two molecules of ssRNA are held together by base pairing. In addition, two tRNA molecules from the previous host are also base-paired with the two ssRNA molecules. The tRNA acts as a primer for reverse transcriptase.

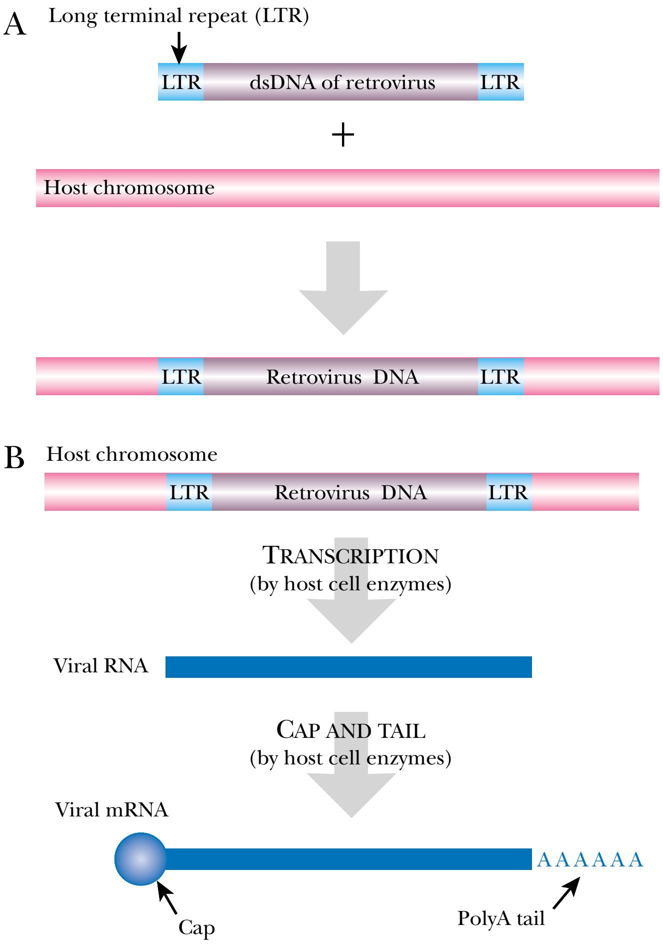


FIGURE 17.26 ***Integration and Transcription of Retrovirus DNA***

(A) After double-stranded retrovirus DNA is made by reverse transcriptase, the dsDNA integrates into the host chromosome. The DNA is flanked by two long terminal repeats (LTR) that facilitate the insertion. (B) Once integrated into the host chromosome, the retrovirus DNA is transcribed and expressed as any other gene in the host cell. The retrovirus RNA is processed by addition of the 7-methyl-G-cap and the poly(A) tail.

Use of Reverse Transcriptase to Make cDNA

Reverse transcriptase is now widely used in genetic engineering for making DNA copies of RNA (see Ch. 22 for details). It is especially useful in obtaining copies of eukaryotic genes that lack the non-coding intervening sequences. Many genes from higher animals have more non-coding DNA than coding sequence and for many purposes the coding sequence alone is easier to handle. Such **complementary DNA (cDNA)** copies of eukaryotic genes are made using the messenger RNA (which lacks introns due to processing) as template. Although the cDNA lacks the introns found in the original natural DNA gene it still encodes the same protein.

lated, several molecules of reverse transcriptase must be packaged along with the RNA in the retrovirus particle.

Retroviruses sometimes pick up host genes and carry them to another animal.

As with integrated bacterial viruses, retroviruses can transduce host cell genes. Chromosomal DNA close to the retrovirus integration site may be fused to virus sequences by deletion of intervening DNA. The fused DNA may then be transcribed as a unit. The RNA may be processed (removing any introns present in the original host genes) and packaged into virus particles. As with bacteriophage lambda, this first gives a defective virus, in which host genes have replaced virus genes. However, recombination with a wild-type retrovirus can generate a functional virus that carries host genes as well as a complete retrovirus genome. Occasional retroviruses carry oncogenes and may cause cancer. Oncogenes are genes that are involved in regulating cell division in animals and when mutated cause cancer. The oncogenes carried by viruses are originally of animal origin and were been picked up by the virus from some previous host animal. Although these cancer-causing viruses have attracted most notice, in principle any host gene close to the site of integration could be moved by retrovirus transduction.

Genome of the Retrovirus

A typical retrovirus has three major genes, *gag*, *pol* and *env*. The **Human Immunodeficiency Virus (HIV)**, which causes AIDS, and its relatives have several extra small genes (Fig.17.27). Two of these, *tat*, and *rev*, are regulatory in function and the others are involved in various aspects of virus maturation and/or modulating host cell metabolism (Table 17.02). The products of the three major genes, *gag*, *pol*, and *env*, are polyproteins that are cleaved to generate several shorter proteins (Table 17.02).

Subviral Infectious Agents

A variety of **subviral agents** are found, most of which have RNA genomes. These may be self-replicating but lack a protein coat (**viroids**) or they may be defective and depend on a helper virus to provide replication genes and/or a protein coat. These entities are listed in Table 17.03. We will discuss only the **satellite viruses** and the viroids in more detail.

complementary DNA (cDNA) Copies of eukaryotic genes lacking introns that are made by reverse transcriptase using messenger RNA as template
human immunodeficiency virus (HIV) The retrovirus that causes AIDS
satellite virus A defective virus that needs an unrelated helper virus to infect the same host cell in order to provide essential functions
subviral agent Infectious agents that are more primitive than viruses and encode fewer of their own functions
viroid Naked single-stranded circular RNA that forms a stable highly base-paired rod-like structure and replicates inside infected plant cells. Viroids do not encode any proteins but possess self-cleaving ribozyme activity

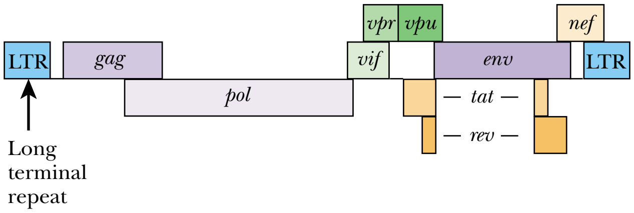


FIGURE 17.27 ***Genome of AIDS Retrovirus***

The genome of a retrovirus is very compact and only encodes nine genes. It is flanked by two long terminal repeats (LTR) that are required for insertion of the DNA version of the virus genome into host DNA. Many of the genes overlap in sequence with each other and are shown side-by-side in the diagram. The regulatory *tat* and *rev* genes are each encoded as two segments that must be spliced together at the RNA level during gene expression. The protein products of *gag*, *pol*, and *env* are actually polyproteins that give rise to two or three proteins after cleavage.

TABLE 17.02 Retrovirus Gene Products and Proteins

Gene Product	**Functions**	
Major Proteins	Cleavage Products	
Gag	MA	matrix protein (between nucleocapsid and envelope)
	CA	capsid protein (major structural protein of virion)
	NC	nucleocapsid protein (protects RNA)
Pol	PR	protease (cleaves precursor proteins)
	RT	reverse transcriptase (makes DNA copy of viral genome)
	IN	integrase (integrates virus DNA into host genome)
Env	SU	surface protein (forms spikes on virus surface that recognize host cell receptors, in particular **CD4 protein**)
	TM	transmembrane protein (fuses virus envelope with host cell membrane)
Regulatory Proteins		
Tat	Increases transcription of virus mRNA—binds to TAR (trans-activator response element).	
Rev	Regulates alternative splicing of virus RNA and export of RNA from nucleus into cytoplasm—binds to RRE (Rev response element)	
Accessory Proteins		
Nef	Decreases surface expression of the CD4 protein and has other effects on host cell immune system.	
Vif	a) Prevents premature cleavage of Gag and Pol polyproteins b) Attaches ubiquitin to APOBEC3G, which is then degraded by the proteasome. (APOBEC3G is a host anti-viral protein that deaminates C to U in retrovirus DNA.)	
Vpr	a) Prevents host cell entering mitosis. b) Promotes entry of viral genome into nucleus.	
Vpu	a) Forms an ion channel and is needed for maturation and release of virions. b) Newly synthesized host CD4 protein that has not yet reached the cell surface is capable of binding the Env protein and blocking virus release. Vpu binds to and promotes degradation of this internal CD4.	

CD4 protein A protein found on the surface of T-cells that acts as a receptor during the immune response

Defunct Retroviruses Make Up 7% of the Human Genome

Endogenous retroviruses are the remains of integrated retroviruses that are no longer in circulation as virus particles. About 7% of human DNA is derived from these defunct sequences. The vast majority of such sequences are defective and many have large internal deletions. The DNA versions of the genomes of about 20 groups of endogenous retroviruses have been found located in our chromosomes. From one to several thousand copies of the genomes of each type of these retroviruses may be present. About 10 times as many single LTR elements have been found as endogenous retroviruses. These are created when homologous recombination between two LTR sequences deletes out all the internal retroviral genes, leaving behind a single LTR.

Based on sequence comparison, most endogenous retroviruses occupied their present sites before divergence of humans and other primates. A few have entered the human genome since the human/chimpanzee split (approximately 5 million years ago). These are all in the same place in different human groups, implying that new additions happen very infrequently and that once integrated the endogenous retroviruses rarely move from one position to another.

Whether these endogenous retroviruses are a health hazard remains disputed. For example, virus particles related to endogenous retrovirus sequences have been isolated from patients with multiple sclerosis, but whether these contribute to the disease or are a side effect of tissue damage is unknown. There is also evidence that some human promoters, enhancers and alternative splice sites may have been recruited from the remains of endogenous retroviruses. For example, alternative splicing of the receptor for the hormone leptin, which controls fat metabolism, depends on sequences within a defunct retrovirus LTR.

TABLE 17.03 Subviral Infectious Agents

Satellite Virus. A defective DNA or RNA virus that needs an unrelated helper virus to infect the same host cell. The helper virus provides the essential functions that it lacks, such as supplying a replicase, capsid proteins, or other functions that allow the satellite virus to survive inside the host cell.

Virusoid. An RNA molecule that does not encode any proteins and depends on a helper virus for replication and capsid formation. The virusoid genome resembles a viroid and consists of circular ssRNA with self-cleaving ribozyme activity.

Satellite RNA. Any small RNA molecule that requires a helper virus for replication and capsid formation. Satellite RNAs range from 200 to 1700 nucleotides and the larger ones may encode a protein. Virusoids are sometimes regarded as a sub-type of satellite RNA.

Defective Interfering RNA (DI-RNA). Shorter RNA molecule derived from viral RNA by deletions that remove essential functions. DI-RNA depends on the original parental virus for replication. The presence of DI-RNA usually reduces the yield of the parental virus.

Viroid. Self-replicating pathogen of plants that consists solely of naked single-stranded circular RNA. The viroid RNA forms an extremely stable highly base-paired rod-like structure. Viroids do not code for any proteins. Most viroid RNAs have self-cleaving ribozyme activity.

Satellite Viruses

Some viruses are parasitic on other viruses. Such satellite viruses are only capable of replication and/or packaging when another virus (the **helper virus**) is present to provide the necessary gene products. Note that satellite viruses are not merely deleted variants of the helper virus. Deletion mutants of viruses do of course exist and can sometimes replicate in the presence of wild-type virus as helper. Satellite viruses are distinct entities that rely on helper virus for a variety of functions. For example, satellite tobacco necrosis virus (STNV) has 1221 nucleotides of ssRNA and encodes a single

Satellite viruses are incomplete and rely on a helper virus to provide essential genes.

helper virus A virus that provides essential functions for defective viruses, satellite viruses and satellite RNA protein—the virus coat protein. For replication it relies on tobacco necrosis virus (TNV), which contains 3759 nucleotides of ssRNA and encodes six proteins. The two viral RNAs share no homology and have quite distinct coat proteins. Replication of STNV requires only the RNA polymerase of TNV; the other TNV gene products are not used by the satellite.

Two other satellite viruses are hepatitis delta virus (HDV) and adeno-associated virus. HDV is a small single-stranded RNA satellite of hepatitis B virus. Despite replicating in human cells, HDV resembles the smaller plant **satellite RNA** viruses. Adeno-associated virus is a satellite of adenovirus that is being used as a eukaryotic cloning vector in gene therapy.

The level of degeneracy varies from one satellite virus to another. For example bacteriophage P4 of *E. coli* has 11.6 kb of dsDNA encoding several genes. Alone, P4 can replicate as a plasmid or integrate into the host chromosome but cannot form virus particles. P4 relies on phage P2 as helper for structural components and assembly of the phage particle. To switch on the P2 genes it needs, P4 deploys an anti-repressor protein (E protein). This prevents the repressor protein of P2 switching off the genes of the P2 virus genome. The structural genes of P2 are therefore expressed and the proteins they code for can be used by P4. The genome of P4 is much shorter than that of its helper P2. Therefore P4 makes smaller capsids although it uses P2 components. The P4 *sid* gene product controls capsid *s*ize *d*etermination. When *sid* is expressed by P4, the capsids become too small for the P2 genome and P2 can no longer be packaged.

Viroids are Naked Molecules of Infectious RNA

Viroids have no protective coat, just single-stranded RNA.

Viroids are infectious agents that consist only of naked RNA without any protective layer such as a protein coat. Viroids infect plants where they are replicated at the expense of the host cell. Viroid genomes are small single-stranded circles of RNA that are only 250 to 400 bases long. For example, the coconut cadang-cadang viroid has only 246 bases of RNA (Fig. 17.28).

Although viroid RNA is single-stranded, base pairing occurs between bases on opposite halves of the circle to produce a rod-like structure (Fig. 17.29). Because viroids have no protein coat, they lack attachment proteins and cannot recognize and penetrate healthy cells as can a true virus. Viroids can infiltrate a plant cell only when its surrounding membrane is already damaged. Once inside, viroids may be passed from one plant cell to another via cellular junctions.

Viroids have no protein coding genes, but the viroid RNA itself acts as a ribozyme.

Viruses all encode at least one protein needed for replication of the virus genome. However, viroid RNA does not contain any genes that encode proteins; it merely carries signals for its own replication by the host machinery. Although the viroid encodes no protein enzymes, the viroid RNA itself acts as a **ribozyme**, that is, the RNA catalyzes an enzymatic reaction. Whether or not a viroid has any genes depends on whether we count the sequence of RNA that possesses ribozyme activity as a gene.

Viroids replicate by a rolling circle mechanism (Fig. 17.30). The viroids own ribozyme activity is used for self-cleavage of the multimeric RNA generated during replication. Host enzymes provide all other functions. First, the circular plus strand is copied by host RNA polymerase to form a multimeric minus strand. Site-specific cleavage of this strand by the ribozyme gives monomers that are circularized by a host RNA ligase. The minus stranded circles are the templates for a second round of rolling circle replication by RNA polymerase. The resulting multimeric plus strand undergoes ribozyme cleavage to create monomers. These are circularized to produce the progeny viroids (circular, positive ssRNA).

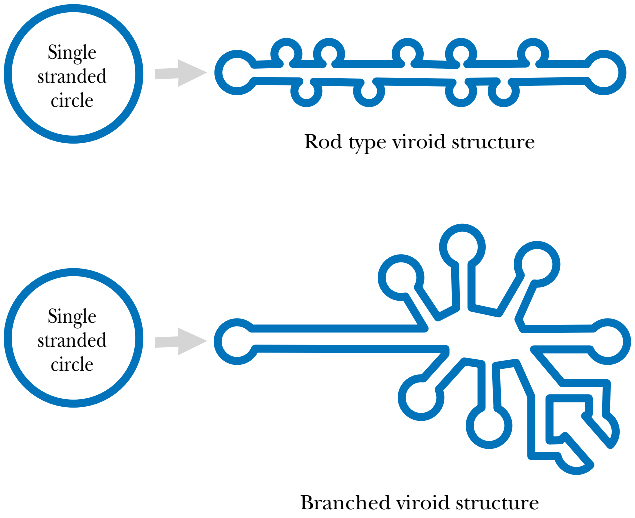
Most viroids replicate in the plant cell nucleus and rely on RNA polymerase II for RNA synthesis. A smaller group of viroids (e.g. chrysanthemum chlorotic mottle

ribozyme An RNA enzyme, that is an RNA molecule with catalytic activity
satellite RNA Parasitic RNA molecule that requires a helper virus for replication and capsid formation

```
CUGGGGAAAU CUACAGGGCA CCCCAAAAAC CACUGCAGGA GAGGCCGCUU
GAGGGAUCCC CGGGGAAACG UCAAGCGAAU CUGGGAAGGG AGCGUACCUG
GGUCGAUCGU GCGCGUUGGA GGAGACUCCU UCGUAGCUUC GACGCCCGGC
CGCCCCUCCU CGACCGCUUG GGAGACUACC CGGUGGAUAC AACUCACGCG
GCUCUUACCU GUUGUUAGUA AAAAAAGGUG UCCCUUUGUA GCCCCU
```

FIGURE 17.28 ***Coconut Cadang-Cadang Viroid, Variant CCCVd.1***

Complete sequence (246 bases) of coconut cadang-cadang viroid



Rod type viroid structure

Branched viroid structure

FIGURE 17.29 ***Viroid RNA Forms a Rod-Like Structure***

Viroids are naked pieces of RNA, which can only infiltrate an already damaged plant cell. The viroid is a single-stranded piece of circular RNA that has an unusual structure due to complementary base pairing. Some form a simple rod-like structure, whereas other viroids have a complex branched structure.

viroid) have a highly branched structure, rather than a rod with bulges, and replicate in the chloroplast.

Prions are Infectious Proteins

Until recently, all infectious diseases were thought to be caused by germs that carry at least some of their own genes. Some diseases are due to living cells like bacteria, while others are due to viruses or viroids, but all contain their own genetic information in the form of DNA or RNA. However, several bizarre diseases of the nervous system are caused by infectious agents containing absolutely no nucleic acid. These diseases are due to rogue proteins known as **prions**. The first to be described was a disease of sheep known as **scrapie**.

The **prion protein (PrP)** is actually encoded by a gene (*Prnp*) belonging to the victim. This gene is transcribed and translated normally and produces a protein found attached to the outside surface of nerve cells, especially in the brain. The prion protein is a glycoprotein, with one or two attached carbohydrate groups, and is attached to the cell membrane by a covalently attached phospholipid molecule (phosphatidyl inositol). Its proper function in the brain is still unknown, although it binds copper ions and may be involved in protection against oxidative stress.

The critical property of the prion protein is that it can fold into two alternative conformations. Occasionally the normal, properly folded form (**cellular PrP or PrP^C**)

A few diseases of the nervous system are caused by infectious proteins, known as prions.

prion A protein that can mis-fold into an alternative pathological form that then promotes its own formation auto-catalytically. Misfolded prion proteins are responsible for the neurodegenerative diseases known as spongiform encephalopathies that include scrapie, kuru and BSE
prion protein (PrP) The prion protein found in the nervous tissue of mammals and whose misfolded form is responsible for prion diseases
PrP^C (cellular PrP) The healthy, normal form of the prion protein
scrapie An infectious disease of sheep that causes degeneration of the brain and is caused by mis-folded prion proteins

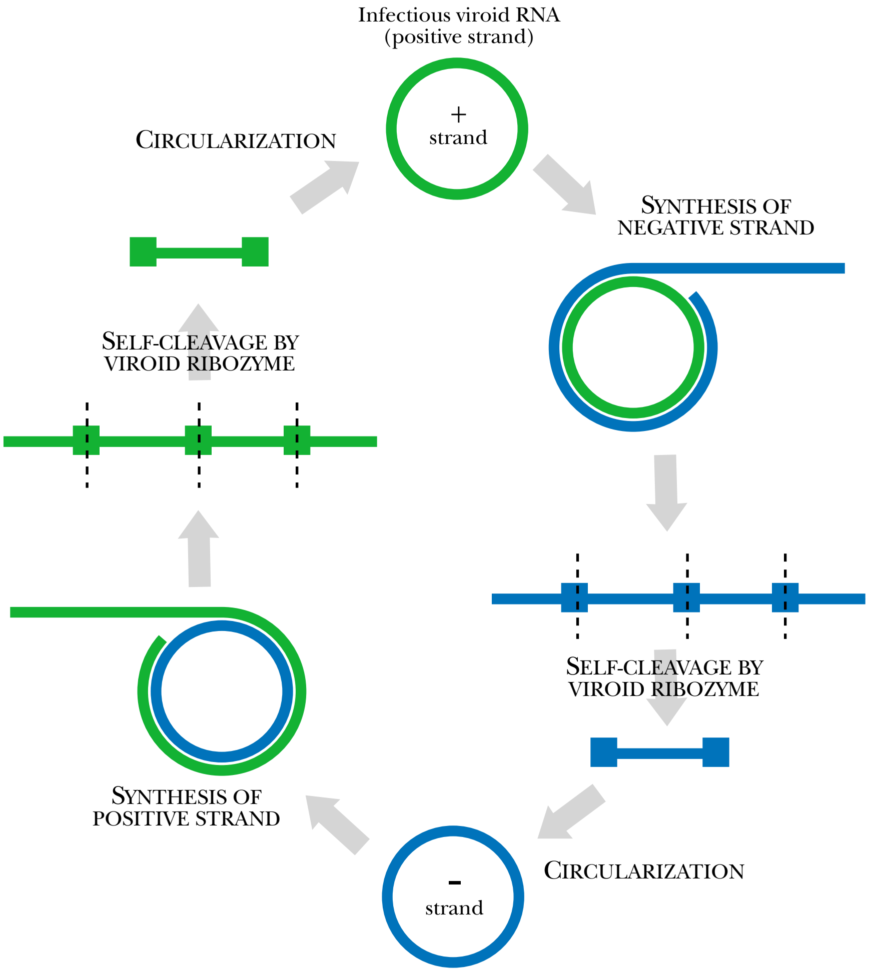

Infectious viroid RNA (positive strand)
+ strand
CIRCULARIZATION
SYNTHESIS OF NEGATIVE STRAND
SELF-CLEAVAGE BY VIROID RIBOZYME
SELF-CLEAVAGE BY VIROID RIBOZYME
SYNTHESIS OF POSITIVE STRAND
- strand
CIRCULARIZATION

FIGURE 17.30 ***Viroids Replicate by a Rolling Circle Mechanism***

Two rounds of rolling circle replication are used by viroids to replicate themselves. Upon entry into a plant cell, the circular, positive ssRNA uses the plant RNA polymerase to make a minus strand. The polymerase continues to make multiple copies using the rolling circle mechanism. The linear, negative ssRNA uses its own catalytic activity to cut itself into genome-sized units that are circularized. The circular, negative ssRNA then undergoes another round of rolling circle replication and self-cleavage to produce multiple copies of the linear plus strand. Finally these are circularized to give the infectious circular, positive ssRNA form.

Prion proteins exist in two alternative forms. The pathogenic form provokes the normal form to change conformation so making more of the pathogenic version.

rearranges to produce the pathological form of the protein (**scrapie PrP or PrP^{Sc}**), which then polymerizes to form fibrillar aggregates. The prion protein is not chemically altered; it merely changes shape. Healthy prions consist largely of α-helical segments, whereas rogue prions have less α-helix and lots of β-sheet regions instead. PrP^{C} is estimated to have 42% α-helix and 3% β-sheet whereas PrP^{Sc} has 30% α-helix and 43% β-sheet (see Ch. 7 for protein structures). The presence of the misfolded PrP^{Sc} form induces the normal PrP^{C} proteins to change conformation also (Fig. 17.31). Thus, once a few molecules of PrP^{Sc} are present, they propagate themselves by catalyzing the conversion of PrP^{C} to more of the PrP^{Sc} isoform. Precisely how these changes in protein conformation and aggregation damage nerve cells is still obscure. Nonetheless, once the change to the scrapie agent has been initiated, slow nervous degeneration and eventual death are inevitable.

Prion disease occurs by three mechanisms, all of which lead to a similar final result. In infectious prion disease, the pathological form of the prion protein is passed from

PrP^{Sc} (scrapie PrP) The pathological form of the prion protein, sometimes known as the scrapie agent

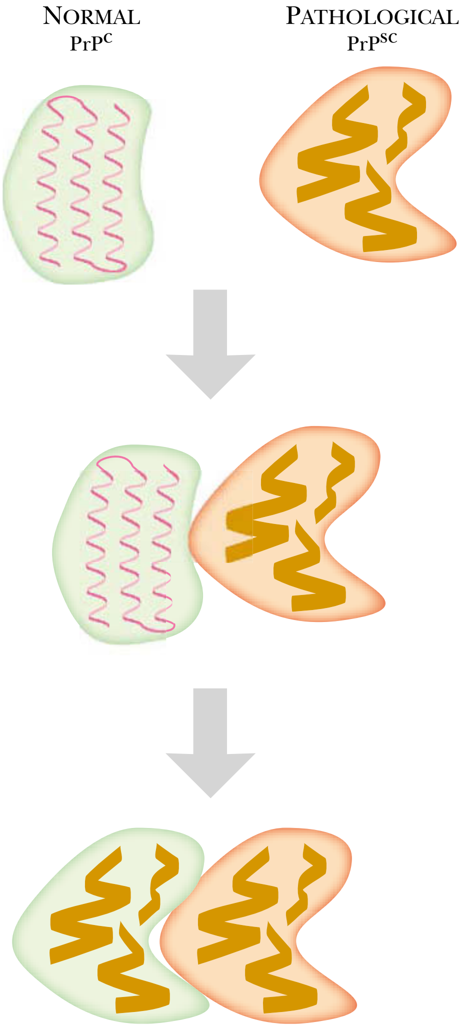


FIGURE 17.31 ***Re-Folding of Normal Prion is Triggered by Rogue Prion***

Normal prions can be induced to change their conformation by contact with a misfolded prion. When the pathological misfolded form PrP^{sc} comes in contact with the normal prion, PrP^c, the normal alpha-helical structure is converted into beta-sheet. The alternate conformation has a tendency to aggregate, forming clumps that damage nerve cells.

Prion disease may be infectious, inherited or spontaneous.

one individual to another. In sporadic prion disease, a spontaneous chance misfolding occurs in a PrP molecule within a brain cell of a normal, uninfected individual. This change is propagated as described above, and over many years the pathological PrP^{Sc} form of the prion protein gradually accumulates until symptoms appear, usually in old age. This happens to about one person in a million. In the inherited form of prion disease a mutation in the prion gene results in a mutant prion protein that changes more often into the disease-causing form. This scenario is little different than a typical inherited disease. Several different mutations are known in the *Prnp* gene and the symptoms of the resulting disease vary slightly. Three main sub-categories are known, Creutzfeldt-Jacob disease (CJD), Gerstmann-Straussler-Scheinker syndrome and Fatal Familial Insomnia. These account for about 15% of the human cases of prion disease.

CHAPTER EIGHTEEN

Bacterial Genetics

Reproduction versus Gene Transfer

Sex and reproduction are not at all the same thing in all organisms. In animals, reproduction normally involves sex, but in bacteria, and in many lower eukaryotes, these are two distinct processes. Bacteria divide by **binary fission**. First they replicate their single chromosome and then the cell elongates and divides down the middle. No re-sorting of the genes between two individuals (that is, no sex) is involved and so this is known as **asexual or vegetative reproduction**.

In bacteria, cell division and the re-assortment of genetic information are completely separate processes.

From a biological perspective, **sexual reproduction** serves the purpose of reshuffling genetic information. This will sometimes produce offspring with combinations of genes superior to those of either parent. Although bacteria normally grow and divide asexually, gene transfer may occur between bacterial cells. During sexual reproduction in higher organisms, germ line cells from two parents fuse to form a zygote that contains equal amounts of genetic information from each parent. In contrast, in bacteria gene transfer is normally unidirectional and cell fusion does not occur. Genes from one bacterial cell are donated to another. We thus have a **donor cell** that donates DNA and a **recipient cell** that receives the DNA.

The transfer of genes between bacteria fulfils a similar evolutionary purpose to the mingling of genes during sexual reproduction in higher organisms. However, mechanistically it is very different. Consequently, some scientists regard bacterial gene transfer as a primitive or aberrant form of sex, whereas others believe that it is quite distinct and that use of the same terminology is misleading.

Gene transfer between bacteria may involve uptake of naked DNA, transport of DNA inside virus particles or transfer of DNA from cell to cell.

Molecular biologists use bacteria together with their plasmids and viruses to carry most cloned genes, whether they are originally from cabbages or cockroaches. Consequently, an understanding of bacterial gene transfer is needed to understand the genetic engineering of plants and animals. Gene transfer in bacteria occurs by three basic mechanisms. Only in **conjugation** are genes transferred by cell-to-cell contact. In **transduction**, genes are transferred inside virus particles, and in **transformation**, free molecules of DNA are taken up by a bacterial cell. Before considering these three mechanisms in detail, we will discuss what happens to the DNA after uptake, as similar considerations apply in all three cases.

Fate of the Incoming DNA after Uptake

Irrespective of its mode of entry, DNA that enters a bacterial cell has one of three possible fates. It may survive as an independent DNA molecule, it may be completely degraded, or part may survive by integration or recombination with the host chromosome before the rest is degraded.

Linear fragments of DNA will be destroyed by cells that receive them.

For incoming DNA to survive inside a bacterial cell as a self-replicating DNA molecule, it must be a replicon. In other words it must have its own origin of replication and lack exposed ends. For survival in the vast majority of bacteria, this means that it must be circular. In those few bacteria, such as *Borrelia* and *Streptomyces* (see Ch. 5) with linear replicons, the ends must be properly protected. In eukaryotes, long-term survival of a linear DNA molecule requires a replication origin, a centromere sequence, and telomeres to protect the ends (see Ch. 5).

A linear fragment of double-stranded DNA that enters a bacterial cell will normally be broken down by exonucleases that attack the exposed ends. For any of its

asexual or vegetative reproduction Form of reproduction in which there is no reshuffling of the genes between two individuals
binary fission Simple form of cell division in which the cell elongates and divides down the middle after replication of the DNA
conjugation Process in which genes are transferred by cell to cell contact
donor cell Cell that donates DNA to another cell
recipient cell Cell that receives DNA from another cell
sexual reproduction Form of reproduction that involves reshuffling of the genes between two individuals
transduction Process in which genes are transferred inside virus particles
transformation Process in which genes are transferred into a cell as free molecules of DNA

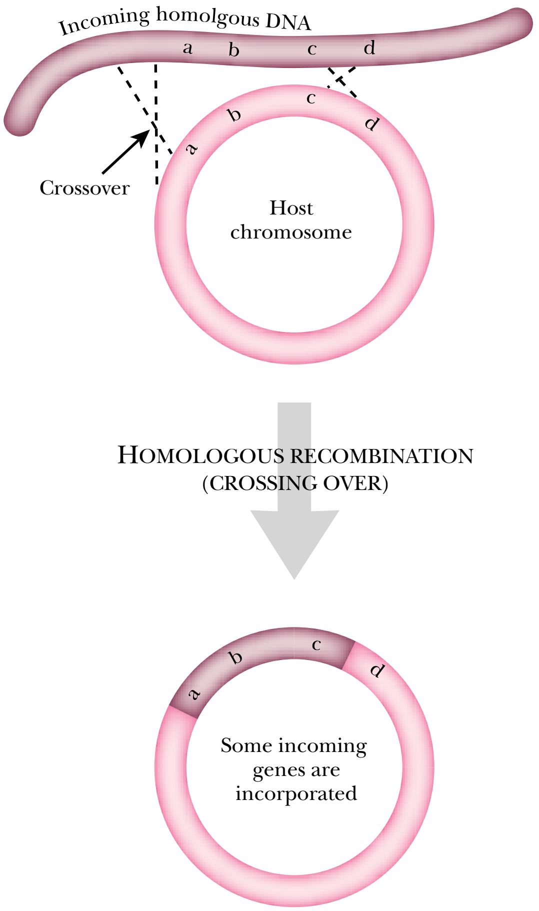

Incoming homolgous DNA
a
b
c
d
c
b
d
a
Crossover
Host chromosome
HOMOLOGOUS RECOMBINATION (CROSSING OVER)
b
c
a
d
Some incoming genes are incorporated

FIGURE 18.01
Recombination allows survival of transformed DNA

In most cases, incoming linear DNA molecules are degraded by the host cell exonucleases. If there are homologous regions between incoming DNA and the host chromosome, crossing over may replace regions of the host chromosome with part of the incoming DNA.

Incoming genes may be preserved from destruction by recombination onto the host chromosome.

genes to survive, they must be incorporated into the chromosome of the recipient cell by the process of recombination (see Ch. 14). For recombination to occur, crossovers must form between regions of DNA of similar sequence—i.e. homologous sequences. DNA between two crossover points will be swapped by the two DNA molecules (Fig. 18.01). Consequently, if genes from incoming DNA are incorporated, the corresponding original genes of the recipient cell are lost.

Such homologous recombination normally only occurs between closely related molecules of DNA—e.g. DNA from two strains of the same bacterial species. Unrelated DNA may be incorporated by recombination provided it is surrounded by sequences that are related (Fig. 18.02). Another possibility is that the incoming DNA contains a transposon that can function in the recipient cell. If so, then the transposon may survive by abandoning the incoming DNA molecule and jumping into the chromosome of the new host cell.

Incoming circular DNA with its own origin of replication can survive without recombination.

If the incoming DNA is a plasmid that can replicate on its own, recombination into the chromosome is not necessary for survival. For genetic engineering purposes, it is usually more convenient to avoid recombination. Consequently, molecular biologists often put the genes they are working with onto plasmids (see Ch. 22).

Restriction enzymes degrade foreign DNA, whether linear or circular.

In addition to exonuclease attack, incoming DNA is often susceptible to restriction. This is a protective mechanism designed to destroy incoming foreign DNA. Most bacteria assume that foreign DNA is more likely to come from an enemy, such as a virus, than from a harmless relative and they cut it into small fragments with restriction enzymes. This applies to both linear and circular DNA, as the degradative enzymes are endonucleases that cut DNA molecules in the middle (see Ch. 22 for details). Only DNA that has been modified by methylating the correct recognition sequences is accepted as friendly. In genetic engineering, restriction negative host strains are used to surmount this obstacle.

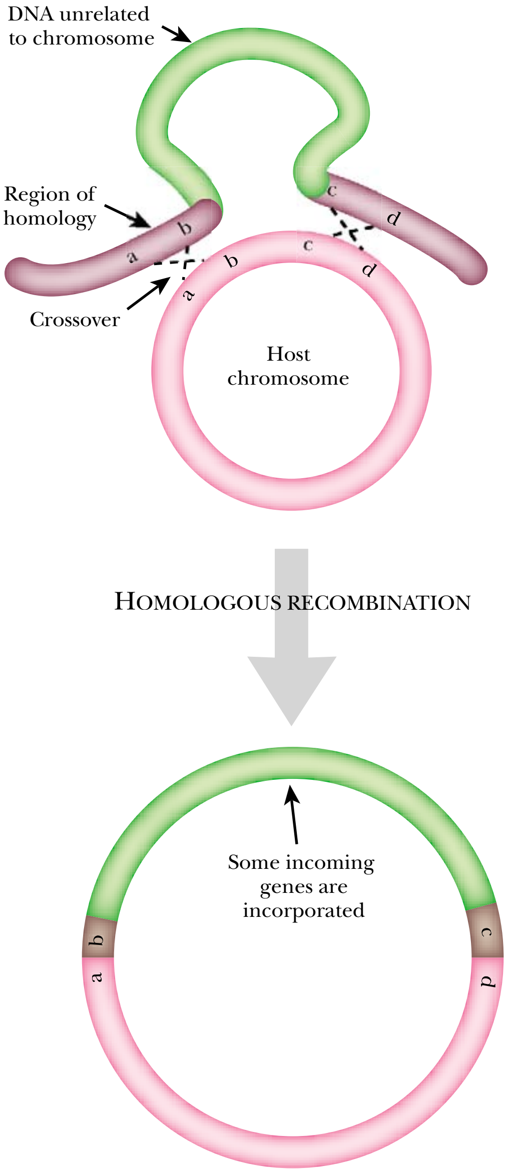


FIGURE 18.02
Incorporation of Unrelated DNA

Incoming DNA does not have to be entirely related to the host in order for recombination to occur. In some instances, the incoming DNA has regions that are related (purple) and regions that are totally unrelated (green). The regions of homology may be large enough to allow recombination, thus integrating an unrelated piece of DNA into the host chromosome. Receiving new genetic material may provide the host cell with a new trait that is desirable to changing environments. In organisms that make identical clones during reproduction, this strategy is critical to evolutionary survival.

Transformation is Gene Transfer by Naked DNA

The simplest way to transfer genetic information is for one cell to release DNA into the medium and for another cell to import it. The transfer of "pure" or "naked" DNA from one cell to another is known as transformation (Fig. 18.03). By "naked", is meant that no other biological macromolecules, such as protein, are present to enclose or protect the DNA. No actual cell-to-cell contact is involved in transformation, nor is the DNA packaged inside a virus particle. Bacterial cells can often take up naked DNA molecules and may incorporate the genetic information they carry.

In practice, transformation is mostly a laboratory technique. The DNA is extracted from one organism by the experimenter and offered to other cells in culture. Cells able to take up DNA are said to be "**competent**." Some species of bacteria readily take up external DNA without any pre-treatment. Probably they use this ability to take up DNA under natural conditions. From time to time, bacteria in natural habitats die and disintegrate. In doing so they release DNA that nearby cells may import.

competent cell Cell that is capable of taking up DNA from the surrounding medium

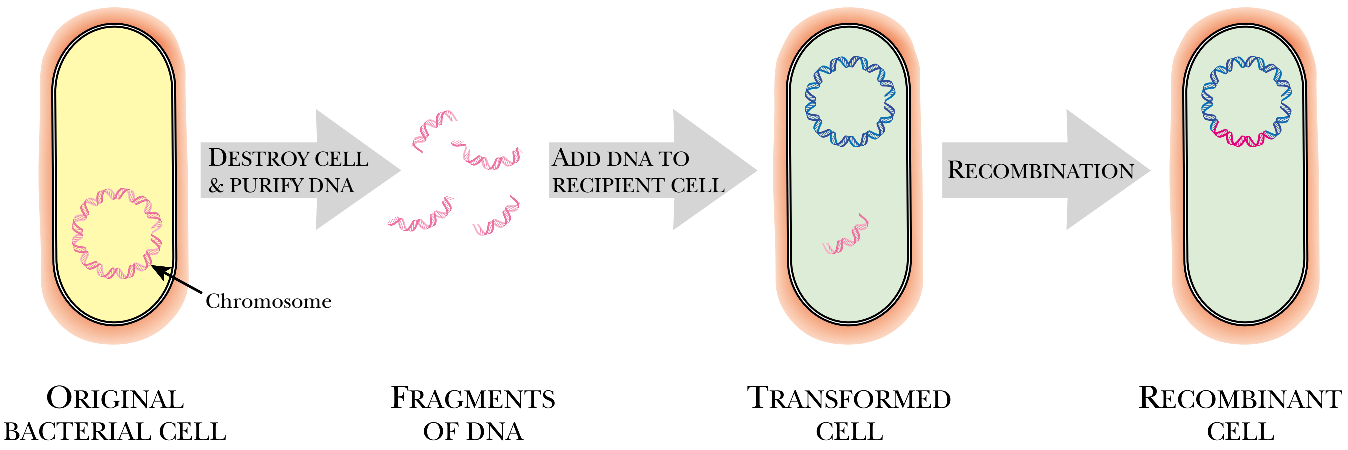


FIGURE 18.03 ***Gene Transfer by Transformation***

Under the right conditions, bacteria can take up pieces of naked DNA from the external environment. The fragment of DNA may pass through the outer cell layers without the aid of a protein or virus. Once inside the bacteria, the fragment of DNA must recombine with the chromosome to prevent degradation by exonucleases or restriction enzymes.

Cells that have cell walls usually need some sort of treatment before they can take up DNA.

Other bacteria must first be treated in the laboratory to make them competent. There are two ways of doing this. The older method is to chill the bacterial cells in the presence of metal ions, especially high concentrations of Ca^{2+}, that damage their cell walls and then to heat shock them briefly. This loosens the structure of the cell walls and allows DNA, a huge molecule, to enter. A more modern method is electroshock treatment. Bacteria are placed in an "**electroporator**" and zapped with a high voltage discharge. After genes or other useful segments of DNA have been cloned in the test tube, it is almost always necessary to put them into some bacterial cell for analysis or manipulation. Thus, laboratory transformation techniques are an essential tool in genetic engineering. *E. coli* is normally treated by some variant of the Ca^{2+}/cold-shock treatment and does not require electroshock. Yeast cells may also be transformed. Since yeast has a very thick cell wall, electroshock is used. Conversely, animal cells, which lack cell walls, often take up DNA readily without any pretreatment, both when grown in culture and in the body.

Transformation as Proof that DNA is the Genetic Material

The transfer of inherited characteristics due to the uptake of pure DNA was part of the original proof that DNA was the genetic information.

Transformation was first observed by Oswald Avery in 1944 and provided the earliest strong evidence that purified DNA carries genetic information and, therefore, that genes are made of DNA. *Pneumococcus pneumoniae* (now renamed *Streptococcus pneumoniae*) has two variants, one forms smooth colonies when plated on nutrient agar, the other has a rough appearance. The smooth variant has a capsule that surrounds the bacterial cell wall, whereas the bacteria in the rough colonies are missing the capsule. The ability to make a capsule affects both colony shape and virulence as the capsule protects bacteria from the animal immune system. Thus, if smooth isolates of *S. pneumoniae* are injected into a live mouse, it dies of bacterial pneumonia. In contrast, rough strains are non-virulent. Avery exploited this difference to prove that DNA from one strain could "transform" or change the other strain. Avery used DNA extracted from virulent strains of *S. pneumonia*. He purified it and added it to harmless strains of the same bacterial species. Some of the harmless bacteria took up the DNA and were transformed into virulent strains. Hence Avery named this process transformation (Fig. 18.04). [Strictly speaking, Avery's transforming DNA could have

electroporator Device that uses a high voltage discharge to make cells competent to take up DNA

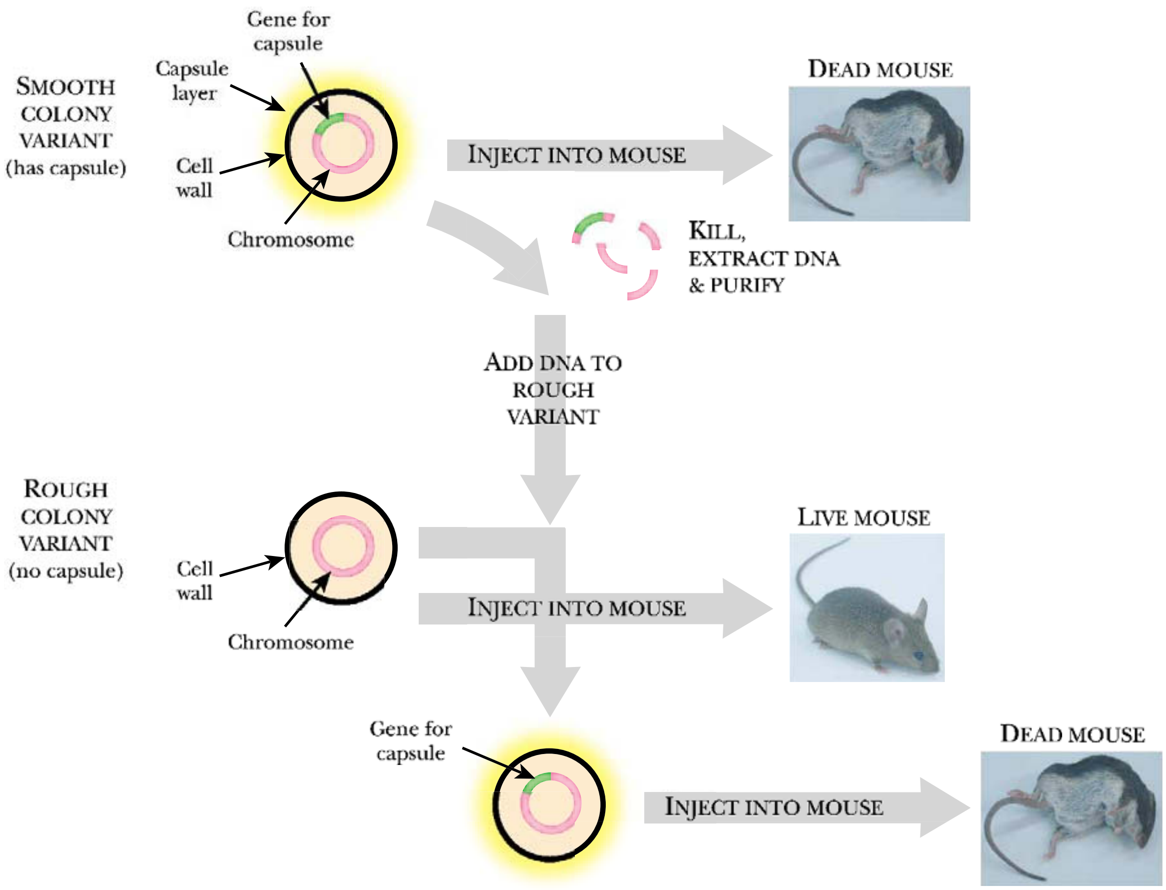


FIGURE 18.04 ***Avery's experiment***

Avery isolated DNA from the virulent variant and added it to the rough variant of *S. pneumonia*. He noticed that the virulent DNA "transformed" or changed the rough variant into a smooth variant. To confirm that the bacteria were truly transformed, he exposed mice to the newly created smooth variants, and the mice died. Thus the transformed bacteria had gained both the smooth appearance and virulence by taking up DNA from the original virulent strains.

interacted in some unknown way with the host chromosome to promote a genetic change. His experiment was therefore not absolute proof that DNA is the genetic material. Nonetheless, this is the most obvious interpretation and this observation convinced many scientists that genes were very likely made of DNA.]

Special terminology is used when scientists use naked viral DNA during transformation. In a viral *infection*, the virus punctures a hole in the bacterial cell wall and injects DNA from the viral particle into the cytoplasm. The viral DNA induces the host to manufacture new viral particles. When viruses infect cells naturally, they often leave their protein coats behind and only the viral genome enters (see Ch. 17). The term **transfection** (a hybrid of transformation with infection) refers to the use of purified viral DNA in transformation. In this case the experimenter purifies the viral genome from the virus particle and offers it to **competent cell**s (Fig. 18.05). If taken up, purified viral DNA induces the cell to synthesize virus, illustrating that the virus coat is only necessary to protect the viral DNA outside the host cell and does not carry any of the virus genetic information. The fact that virus infection may be caused by the DNA alone is further evidence that DNA is the genetic material.

Transformation and transfection have two other meanings, also. Cancer specialists use the term "**transformation**" to refer to the changing of a normal cell into a cancer

competent cell Cell that is capable of taking up DNA from the surrounding medium
transfection Process in which purified viral DNA enters a cell by transformation. Often used to refer to entry of any DNA, even if not of viral origin, into an animal cell
transformation (As used of cancer) Changing a normal cell into a cancer cell, even if no extra DNA enters the cell

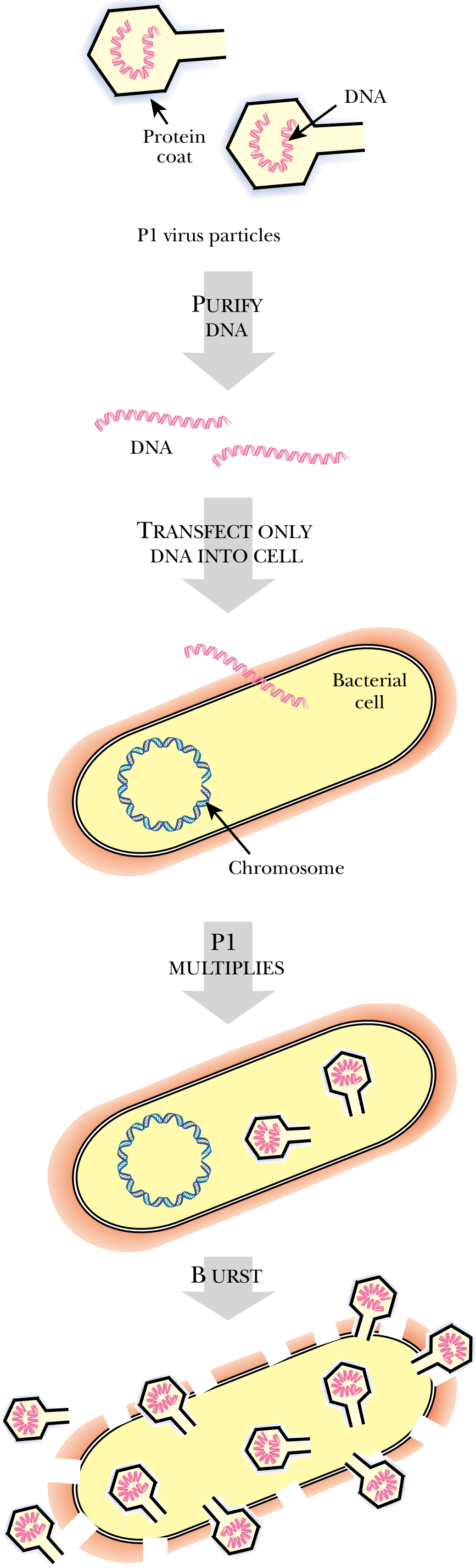


FIGURE 18.05 ***Transfection***

During viral *transfection*, an experimenter first isolates pure viral DNA from virus particles. In the diagram, DNA is isolated from P1 virus. Next, the bacterial cell wall is made competent to take up naked DNA (usually by treating with calcium ions or by electroshock). The isolated DNA and the competent bacteria are mixed. If the bacteria take up the P1 DNA, the bacteria will start producing viral particles and burst to release the viral progeny. Thus, viral DNA alone can give the same end result as infection with whole virus particles.

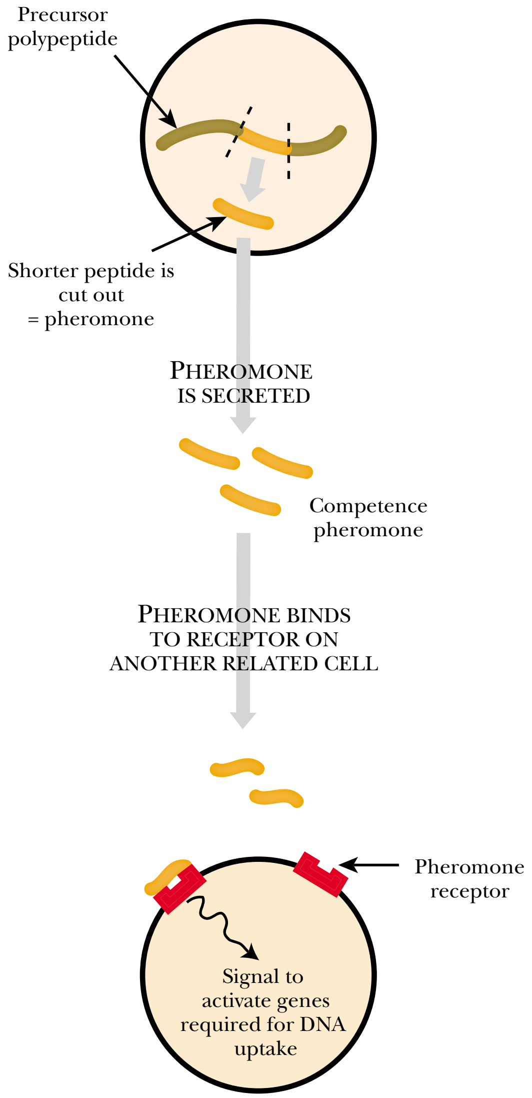

Precursor polypeptide
Shorter peptide is cut out = pheromone
PHEROMONE IS SECRETED
Competence pheromone
PHEROMONE BINDS TO RECEPTOR ON ANOTHER RELATED CELL
Pheromone receptor
Signal to activate genes required for DNA uptake

FIGURE 18.06 ***Competence Pheromones***

Dense cultures of *S. pneumoniae* start producing competence pheromones that induce nearby cells to take up DNA. First, certain cells of the culture produce polypeptide precursors that are digested into a small peptide, or competence pheromone. The small peptide is secreted from the producer cell and binds to a receptor on a nearby cell. The receptor then signals that cell to make proteins used in DNA uptake.

cell, even though in most cases no extra DNA enters the cell. [Note that alterations in the DNA are indeed involved in creating cancer cells, but as a result of mutation.] Supposedly to avoid ambiguity, researchers who use animal cells often use the term "transfection" to refer to the uptake of DNA (by transformation!) whether it is of viral origin or not.

Transformation in Nature

Transformation occurs among certain bacteria in the natural environment.

More detailed investigation of *Streptococcus pneumoniae* and other gram-positive bacteria, including *Bacillus*, shows that they develop natural competence in dense cultures. Competence is induced by competence **pheromones**. (A pheromone is a hormone that travels between organisms, rather than circulating within the same organism). Competence pheromones are short peptides that are secreted into the culture medium by dividing bacteria (Fig. 18.06). Only when the density of bacteria is high will the pheromones reach sufficient levels to trigger competence. This mechanism is presumably meant to ensure that any DNA taken up will come from related bacteria as competence is only induced when there are many nearby cells of the same species.

pheromone Hormone or messenger molecule that travels between organisms, rather than circulating within the same organism

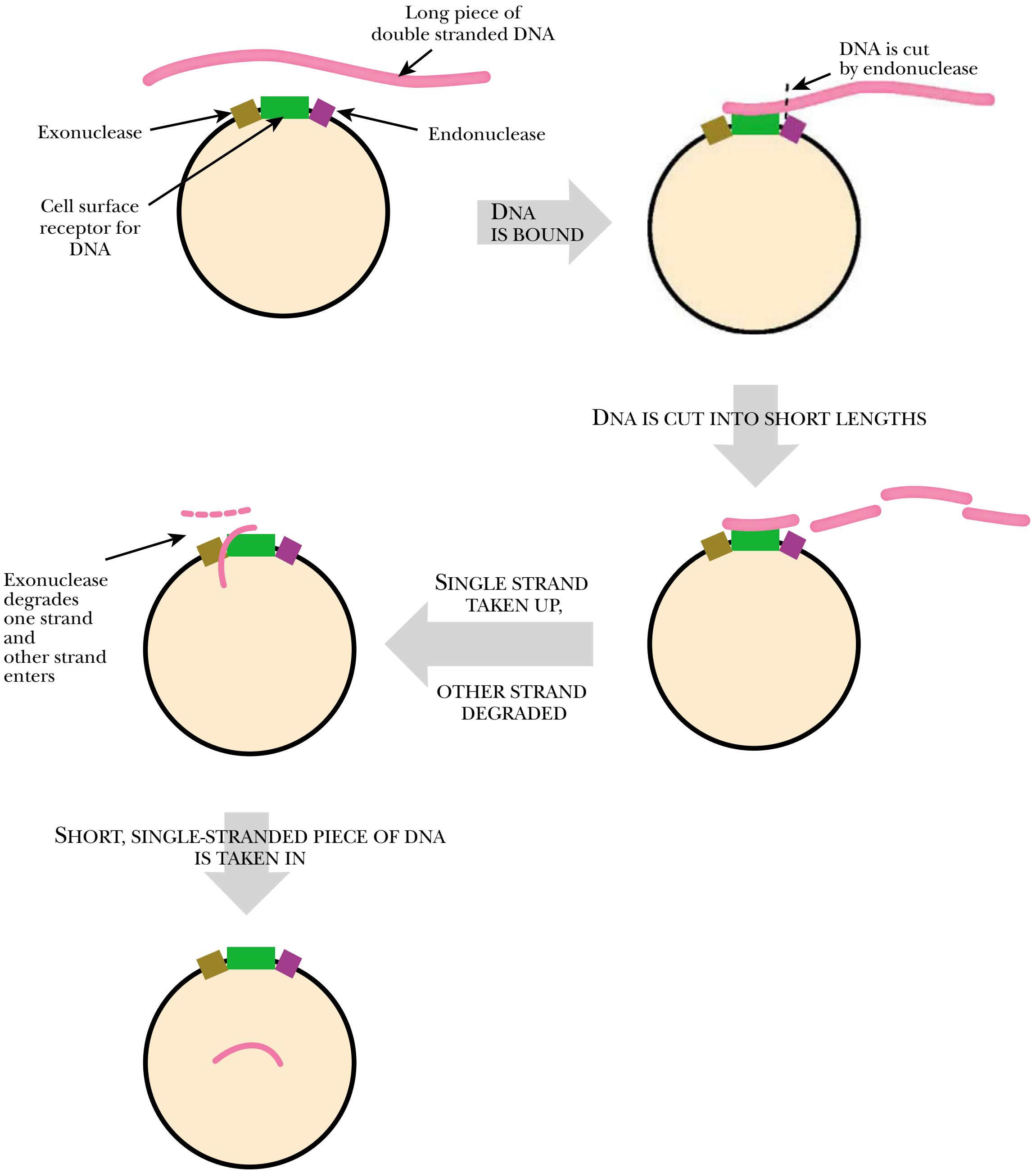

Long piece of double stranded DNA
Exonuclease
Endonuclease
Cell surface receptor for DNA
DNA IS BOUND
DNA is cut by endonuclease
DNA IS CUT INTO SHORT LENGTHS
SINGLE STRAND TAKEN UP,
OTHER STRAND DEGRADED
Exonuclease degrades one strand and other strand enters
SHORT, SINGLE-STRANDED PIECE OF DNA IS TAKEN IN

FIGURE 18.07 ***Mechanism of Natural Competence***

A cell that is naturally competent takes DNA into its cytoplasm by a protein-mediated process. First, the long molecule of double-stranded DNA is recognized by a receptor on the surface of the competent cell. A cell-surface endonuclease digests the DNA into small fragments. An exonuclease then degrades one strand of the DNA. The remaining single-stranded fragment is taken into the cytoplasm of the bacterium.

Natural competence is not merely due to random entry of DNA but involves the induction of a variety of genes whose products take part in DNA uptake. First DNA is bound by cell surface receptors (Fig. 18.07). Then the bound DNA is cut into shorter segments by endonucleases and one of the strands is completely degraded by an exonuclease. Only the resulting short single-stranded segments of DNA enter the cell. Part of the incoming DNA may then displace the corresponding region of the host chromosome by recombination.

Note that in the case of artificially induced competence, the mechanism is quite different. Double-stranded DNA enters the cell through a cell wall that is seriously

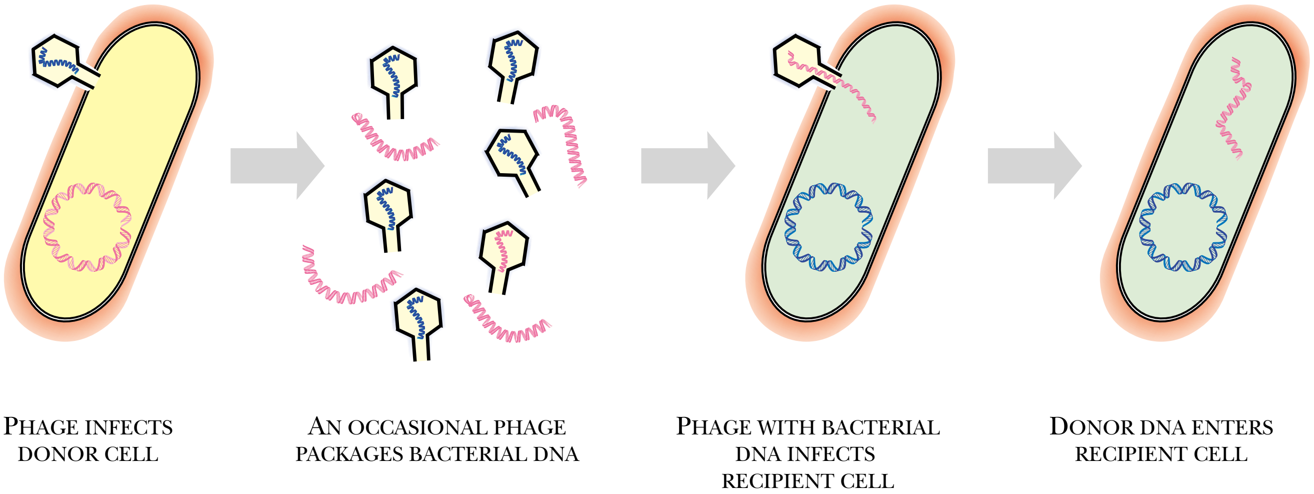

PHAGE INFECTS DONOR CELL
AN OCCASIONAL PHAGE PACKAGES BACTERIAL DNA
PHAGE WITH BACTERIAL DNA INFECTS RECIPIENT CELL
DONOR DNA ENTERS RECIPIENT CELL

FIGURE 18.08 ***Principle of Transduction***

Occasionally, when a phage infects a bacterium, one of the virus coats will be packaged with host bacterial DNA. The defective phage particle still infects a nearby cell where it injects the bacterial DNA. This cell will survive since it is not injected with viral DNA. The incoming DNA may be recombined with the host chromosome, thus this cell may gain new genetic information.

damaged. Indeed, many, perhaps the majority, of the cells that are made artificially competent are killed by the treatment. It is the few survivors who take up the DNA.

Gene Transfer by Virus—Transduction

Viruses may pick up fragments of host DNA and carry them to another host cell.

When a virus succeeds in infecting a bacterial cell it manufactures more virus particles, each of which should contain a new copy of the virus genome. Occasionally, viruses make mistakes in packaging DNA, and fragments of bacterial DNA get packaged into the virus particle. From the viewpoint of the virus, this results in a defective particle. Nonetheless, such a virus particle, carrying bacterial DNA, may infect another bacterial cell. If so, instead of injecting viral genes, it injects DNA from the previous bacterial victim. This mode of gene transfer is known as transduction.

Bacterial geneticists routinely carry out gene transfer between different but related strains of bacteria by transduction using bacterial viruses, or bacteriophages (phages for short). If the bacterial strains are closely related the incoming DNA is accepted as "friendly" and is not destroyed by restriction. In practice, transduction is the simplest way to replace a few genes of one bacterial strain with those of a close relative.

To perform transduction, a bacteriophage is grown on a culture of the donor bacterial strain. These bacteria are destroyed by the phage, leaving behind only DNA fragments that carry some of their genes and are packaged inside phage particles. If required, this phage sample can be stored in the fridge for weeks or months before use. Later, the phage are mixed with a recipient bacterial strain and the DNA is injected. Most recipients get genuine phage DNA and are killed. However, others get donor bacterial DNA and are successfully transduced (Fig. 18.08).

Generalized Transduction

Some viruses carry random fragments of host DNA.

There are two distinct types of transduction. In **generalized transduction** fragments of bacterial DNA are packaged more or less at random in the phage particles. This is the case for bacteriophage P1 as described above (Fig. 18.08). Consequently all genes have

generalized transduction Type of transduction where fragments of bacterial DNA are packaged at random and all genes have roughly the same chance of being transferred roughly the same chance of being transferred. In **specialized transduction** certain regions of the bacterial DNA are carried preferentially—see below.

For a bacterial virus to transduce, several conditions must be met. In particular, the phage must not degrade the bacterial DNA. For example, phage T4 normally destroys the DNA of *E. coli* after infection. However, mutants of T4 that have lost the ability to degrade host cell DNA work well as transducing phages. The packaging mechanism is also critical. Some phages, such as lambda (see below) use specific recognition sequences when packaging their DNA into the virus particle and so will not package random fragments of DNA. In other cases, packaging depends on the amount of DNA the head of the virus particle can hold. Such "**headful packaging**" is essential for generalized transduction.

Two examples of generalized transducing phages are **P1**, which works on *Escherichia coli*, and **P22**, which infects *Salmonella*. The ratio of transducing particles to live virus is about 1 : 100 in both cases, that is, for every 100 virus particles made, one will be packaged with bacterial host DNA. The likelihood of the transduced DNA recombining into the recipient chromosome is roughly 1 to 2 in 100. P1 can package approximately 2% of the *E. coli* chromosome (about 90 kb of DNA), whereas P22 is smaller and can carry only 1% of the *Salmonella* chromosome. Taken all together, about 1 in 500,000 P1 particles will successfully transduce any particular gene on the *E. coli* chromosome. This may seem a low probability but as both typical bacterial cultures and preparations of P1 contain about 10^9 cells per ml, transduction happens at useful frequencies in practice. P1 can also transduce DNA from *E. coli* into certain other gram-negative bacteria, such as *Klebsiella*.

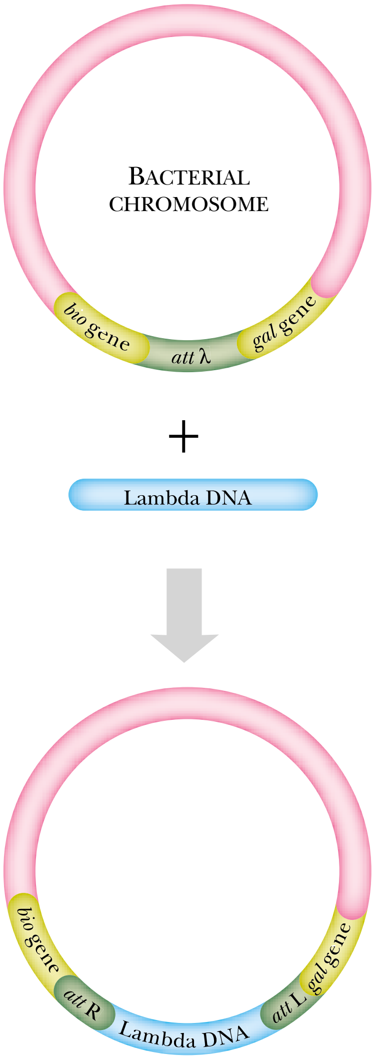


FIGURE 18.09 *Integration of Lambda into the E. coli Chromosome*

When the bacteriophage lambda infects a host *E. coli* cell, it can integrate its phage DNA into the chromosome. The phage DNA will only integrate at a site called *att*λ, which is found between the *bio* gene and *gal* gene of the chromosome. Once integrated, the phage is referred to as a prophage.

Viruses that integrate into host DNA pick up fragments of host DNA that lie next to the integration site.

Specialized Transduction

During specialized transduction, certain specific regions of the bacterial chromosome are favored. This is due to integration of the bacteriophage into the host chromosome (see Ch. 17). If the virus enters a lytic cycle and manufactures virus particles, those bacterial genes nearest the integration site are most likely to be incorrectly packaged into the viral coats. As discussed in Ch. 17, when bacteriophage **lambda (or λ)** infects *E. coli*, it sometimes inserts its DNA into the bacterial chromosome (Fig. 18.09). This occurs at a single specific location, known as the **lambda attachment site (*att*λ)**, which lies between the *gal* and *bio* genes. The integrated virus DNA is referred to as a **prophage**.

When lambda is induced, it excises its DNA from the chromosome and goes into lytic mode. The original donor cell is destroyed, and several hundred virus particles containing lambda DNA are produced. Just like generalized transducing phages, a small fraction of lambda virus particles contain bacterial DNA. There are, however, two major differences. First, only chromosomal genes next to the attachment site are transduced by lambda. Second, the specialized transducing particles contain a hybrid DNA molecule comprising both lambda and chromosomal DNA (Fig. 18.10). This hybrid molecule results from mistakes during excision of the lambda prophage from the chromosome. Chromosomal DNA to the right or to the left of the prophage, but not both, may be included in the transducing phage. In practice this means that either the *gal* or *bio* genes are picked up.

Mistakes in excision of lambda only occur at a rate of 1 in a million relative to correct excision. Furthermore, the defective excision must generate a segment of

headful packaging Type of virus packaging mechanism that depends on the amount of DNA the head of the virus particle can hold (as opposed to using specific recognition sequences)
lambda (or λ) Specialized transducing phage of *Escherichia coli* that may insert its DNA into the bacterial chromosome
lambda attachment site (*att*λ) Site where lambda inserts its DNA into the bacterial chromosome
P1 Generalized transducing phage of *Escherichia coli*
P22 Generalized transducing phage of *Salmonella*
prophage Virus DNA that is integrated into the host chromosome
specialized transduction Type of transduction where certain regions of the bacterial DNA are carried preferentially

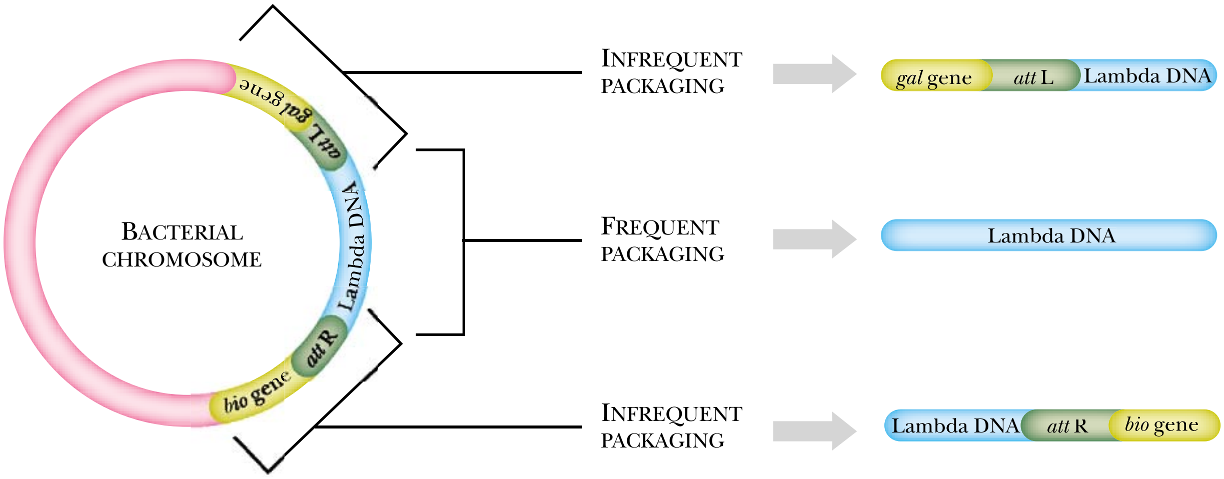


FIGURE 18.10 ***Packaging of Host DNA During Transduction by Lambda***

When lambda phage enters its lytic cycle and makes phage particles, it usually packages the lambda DNA between the *att*L and *att*R sites. Occasionally, a mistake will occur, and part of the bacterial chromosome DNA will be packaged. Since lambda DNA normally integrates between the *gal* and *bio* genes of the *E. coli* chromosome, the defective lambda particles will most likely contain one or other of these genes.

DNA approximately the same length as the lambda genome in order for it to fit into the phage head. Consequently, specialized transducing particles arise only at extremely low frequency. However, once a lambda transducing phage has been created, it may re-integrate its DNA into the chromosome of another host cell. This may occur either in the ***att*λ** site or into the chromosomal copy of the gene (usually *gal* or *bio*) carried by the lambda transducing phage. Inducing this defective prophage DNA will give a second generation of transducing phage particles at a much higher frequency.

The properties of lambda transducing phages depend on which lambda genes were lost in exchange for chromosomal DNA. The λd*gal* transducing phages lack lambda genes needed for making head and tail components and are therefore "defective" (hence the "d" in λd*gal*), but instead, the virus contains the *E. coli gal* gene. **Defective phage** may be grown together with a wild-type lambda as a **helper phage** to provide the missing functions. In the case of λd*gal*, helper phage would make the head and tail components. Another example, the λp*bio* transducing phages lack the Lambda *int* gene, which integrates the phage DNA into the *att*λ site, and instead contains the *bio* gene from *E. coli*. Since the phage cannot integrate, λp*bio* must enter the lytic phase, and therefore, are obligate plaque formers (hence the "p" in λp*bio*). If wild-type helper phage are added, the *int* function is restored, and the phage can form lysogens. Cloning vectors derived from lambda are widely used in genetic engineering (see Ch. 22). Since the cloned DNA replaces many of the lambda genes, such vectors need to be grown in the presence of helper phages.

Transfer of Plasmids between Bacteria

Transferability is the ability of certain plasmids to move themselves from one bacterial cell to another. Many medium sized plasmids, such as the F-type and P-type plas-

***att*λ** Lambda attachment site—site where lambda inserts its DNA into the bacterial chromosome
defective phage Mutant phage that lacks genes for making virus particles
helper phage Phage that provides the necessary genes so allowing a defective phage to make virus particles
transferability Ability of certain plasmids to move themselves from one bacterial cell to another

mids, can do this and are referred to as **Tra**$^+$ (transfer positive). For transfer to occur, the two bacterial cells must come into physical contact and move the DNA by a process known as bacterial conjugation. DNA is transferred in one direction only, from the plasmid-carrying donor to the recipient (Fig. 18.11). The donor cell manufactures a **sex pilus** that binds to a suitable recipient and draws the two cells together. Next, a **conjugation bridge** forms between the two cells and provides a channel for DNA to move from donor to recipient. In real life, mating bacteria actually tend to cluster together in groups of five to ten (Fig. 18.12).

The genes for formation of the sex pilus and conjugation bridge and for overseeing the DNA transfer process are known as ***tra* genes** and are all found on the plasmid itself. Since plasmid transfer requires over 30 genes, only medium or large plasmids possess this ability. Very small plasmids, such as the ColE-plasmids, do not have enough DNA to accommodate the genes needed. Plasmids that enable a cell to donate DNA are called **fertility plasmids** and the most famous of these is the **F-plasmid** of *E. coli*, which is approximately 100 kbp long. Donor cells are sometimes known as F$^+$ or "male" and recipient cells as F$^-$ or "female" and conjugation is sometimes referred to as bacterial mating. Note however, that the "sex" of a bacterial cell is determined by the presence or absence of a plasmid and that DNA transfer is unidirectional, from donor to recipient. When a recipient cell has received the F-plasmid it becomes F$^+$. From a human perspective it has been transmuted from "female" into a "male"! Thus bacterial mating is not at all equivalent to sexual reproduction among higher organisms.

Transferable plasmids move from one cell to another via the conjugation bridge.

Plasmid transfer actually involves replication by the rolling circle mechanism (Fig. 18.13). First one of the two strands of the double stranded DNA of the F-plasmid opens up at the origin of transfer. This linearized single strand of DNA moves through the conjugation bridge from the donor into the recipient cell. An unbroken single stranded circle of F-plasmid DNA remains inside the donor cell. This is used as a template for the synthesis of a new second strand to replace the one that just left. As the linear single strand of F-plasmid DNA enters the female cell, a new complementary strand of DNA is made using the incoming strand as template. Thus only one strand of the F-plasmid DNA is transferred from the donor to the recipient.

A single strand of newly made DNA is transferred from the donor to the recipient cell.

Although ColE and other small plasmids are not self-transferable, they are often mobilizable (Mob$^+$). A transferable plasmid, such as the F-plasmid, can mobilize the ColE plasmid if they both inhabit the same cell. The F-plasmid oversees conjugation and forms the conjugation bridge and the ColE plasmid is transferred through this. The *mob* (mobilization) genes of the ColE plasmid are responsible for making a single-stranded nick at the origin of transfer of ColE and for unwinding the strand to be transferred.

Plasmids unable to transfer themselves may be able to hitch-hike using the transfer systems of other plasmids.

Transfer of Chromosomal Genes Requires Plasmid Integration

Although many plasmids allow the cells carrying them to conjugate, usually only the plasmid itself is transferred through the conjugation bridge. Much less often, plasmids mediate transfer of the host chromosome when they move from one bacterial cell to another. In order to transfer chromosomal genes, a plasmid must first physically integrate itself into the chromosome of the bacterium. This event involves pairs of identical (or nearly identical) DNA sequences, one on the plasmid and the other on the chromosome. In practice, **insertion sequences** (see Ch. 15) are used for integration of the F-plasmid into the chromosome of *E. coli* (Fig. 18.14).

Transferable plasmids sometimes move chromosomal DNA from one cell to another.

conjugation bridge Junction that forms between two cells and provides a channel for DNA to move from donor to recipient during conjugation
fertility plasmid Plasmid that enables a cell to donate DNA by conjugation
F-plasmid Fertility plasmid that allows *E. coli* to donate DNA by conjugation
insertion sequence A simple transposon consisting only of inverted repeats surrounding a gene encoding transposase
sex pilus Protein filament made by donor bacteria that binds to a suitable recipient and draws the two cells together
Tra$^+$ Transfer positive (refers to a plasmid capable of self-transfer)
***tra* genes** Genes needed for plasmid transfer

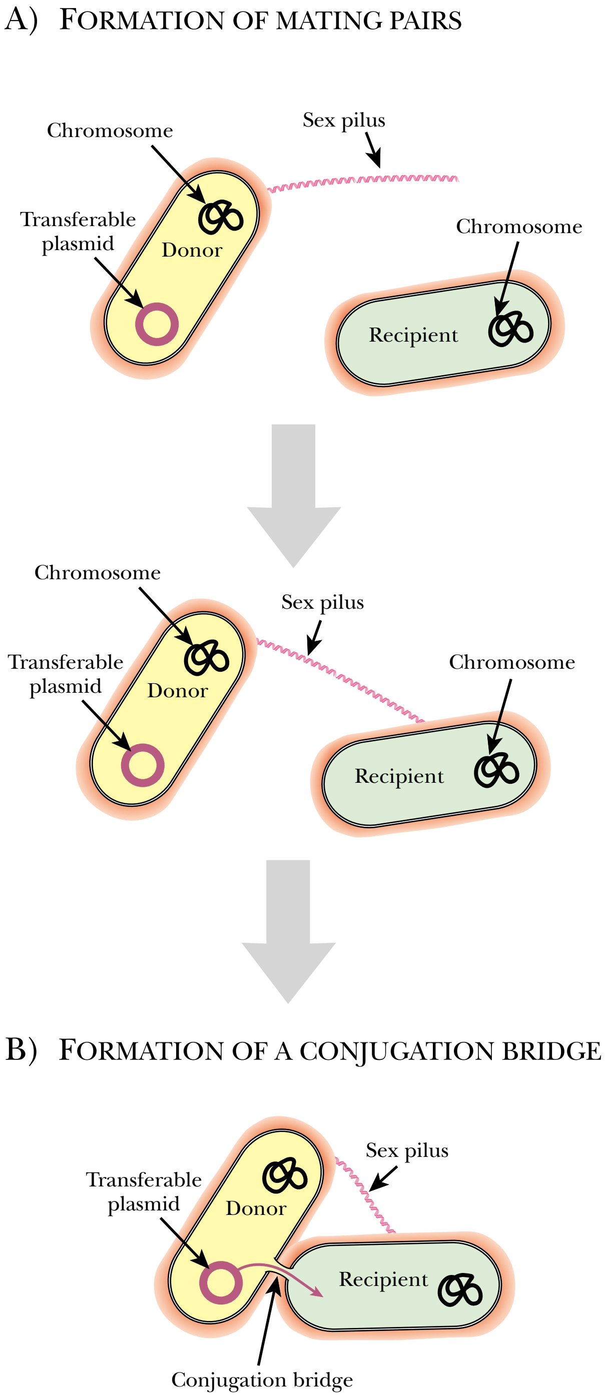


FIGURE 18.11 ***Bacterial Conjugation***

Certain plasmids, called Tra^+ or transfer positive, are able to move a copy of their DNA into a different cell through a mechanism called bacterial conjugation. First the cell containing a Tra^+ plasmid manufactures a rod-like extension on the surface of the outer membrane, called a sex pilus. The sex pilus binds to a nearby cell and pulls the two cells together. Once the cells are in contact, a connection is made between the two cells called the conjugation bridge. This connects the cytoplasm of the two cells, so the plasmid can transfer a copy of itself to the recipient cell.

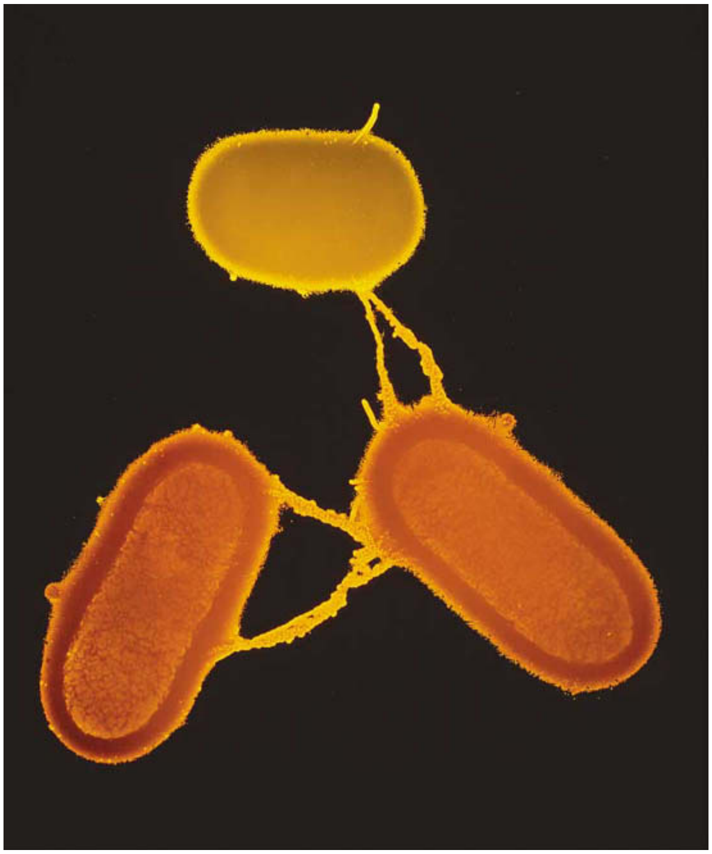


FIGURE 18.12 ***Conjugating Cells of* Escherichia coli**

False-color transmission electron micrograph (TEM) of a male *Escherichia coli* bacterium (bottom right) conjugating with two females. This male has attached two F-pili to each of the females. The tiny bodies covering the F-pili are bacteriophage MS2, a virus that attacks only male bacteria and binds specifically to F-pili. Magnification: ×11,250. Credit: Dr. L. Caro, Photo Researchers, Inc.

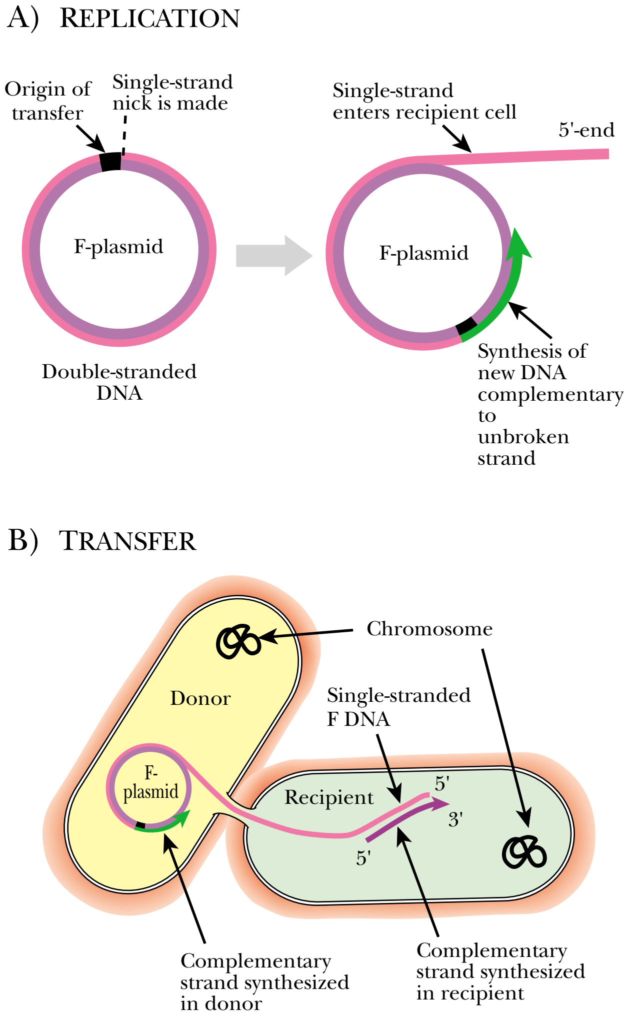

A) REPLICATION
Origin of transfer
Single-strand nick is made
F-plasmid
Double-stranded DNA
Single-strand enters recipient cell
5'-end
F-plasmid
Synthesis of new DNA complementary to unbroken strand
B) TRANSFER
Chromosome
Donor
Single-stranded F DNA
F-plasmid
Recipient
5'
3'
5'
Complementary strand synthesized in donor
Complementary strand synthesized in recipient

FIGURE 18.13 Plasmid Transfer Involving Rolling Circle Replication

A) During bacterial conjugation, the F-plasmid of *E. coli* is transferred to a new cell by rolling circle replication. First, one strand of the F-plasmid is nicked at the origin of transfer. The two strands start to separate and synthesis of a new strand starts at the origin (green strand). B) The single-strand of F-plasmid DNA that is displaced (pink strand) crosses the conjugation bridge and enters the recipient cell. The second strand of the F-plasmid is synthesized inside the recipient cell. Once the complete plasmid has been transferred, it is re-ligated to form a circle once again.

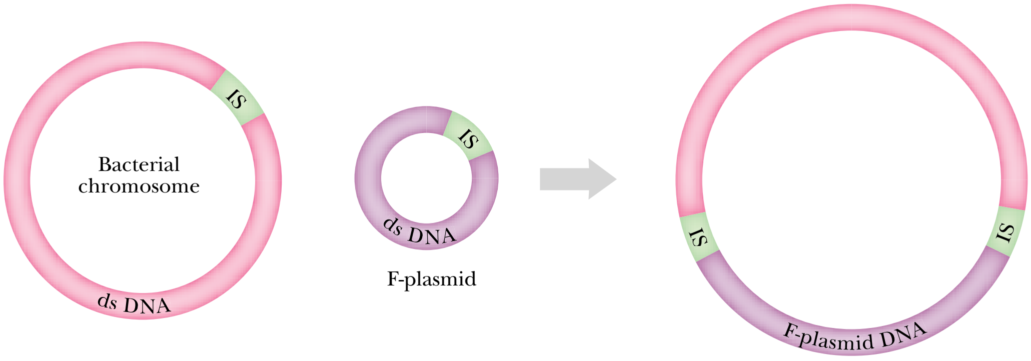


FIGURE 18.14 ***Integration of F-Plasmid into Chromosome***

If recombination occurs between two insertion sequences, one on the F-plasmid and one on the host bacterial chromosome, the entire F-plasmid is integrated into the chromosome.

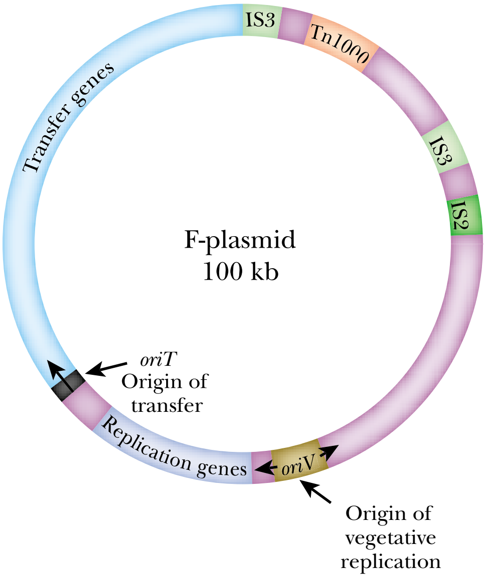


FIGURE 18.15 ***Insertion Sequences on F-Plasmid and Chromosome***

Insertion sequences are scattered throughout the F-plasmid and chromosome of *E. coli*. The F-plasmid has two IS3 elements and one IS2 element. Even more copies of IS2 and IS3 are found on the chromosome (not shown). A recombination event between any of the chromosomal IS2 or IS3 elements and the corresponding element on the F-plasmid will integrate the entire F-plasmid into the chromosome. Tn1000 (also known as γδ) is another insertion sequence, although not generally involved in F-plasmid integration in *E. coli*.

A variety of different insertion sequences are found on the chromosome of *E. coli* and in its plasmids and viruses. The F-plasmid, has three insertion sequences (Fig. 18.15); two copies of IS3 and a single copy of IS2. The chromosome of *E. coli* has 13 copies of IS2 and six copies of IS3 scattered around more or less at random. Integration of the F-plasmid may occur in either orientation at any of these 19 sites.

In order to mobilize chromosomal DNA, the plasmid must first integrate into the chromosome.

When an F-plasmid that is integrated into the chromosome transfers itself by conjugation, it drags along the chromosomal genes to which it is attached (Fig. 18.16). Just like the unintegrated F-plasmid, only a single strand of the DNA moves and the recipient cell has to make the complementary strand itself. Bacterial strains with an F-plasmid integrated into the chromosome are known as **Hfr-strains** because they transfer chromosomal genes at high frequency. A prolonged mating of 90 minutes or so is needed to transfer the whole chromosome of *E. coli*. More often, bacteria break off after a shorter period of, say, 15 to 30 minutes, and only part of the chromosome is transferred. Since different Hfr-strains have their F-plasmids inserted at different sites on the bacterial chromosome, transfer of chromosomal genes begins at different points. In addition, the F-plasmid may be inserted in either orientation. Consequently, gene transfer may be either clockwise or counterclockwise for any particular Hfr strain.

Hfr strains are useful for identifying the order of genes on the *E. coli* chromosome. In order to determine whether the recipient cell has received the gene in question, the donor and recipient strains must have different alleles of this gene that can be distin-

Hfr-strain Bacterial strain that transfers chromosomal genes at high frequency due to an integrated fertility plasmid

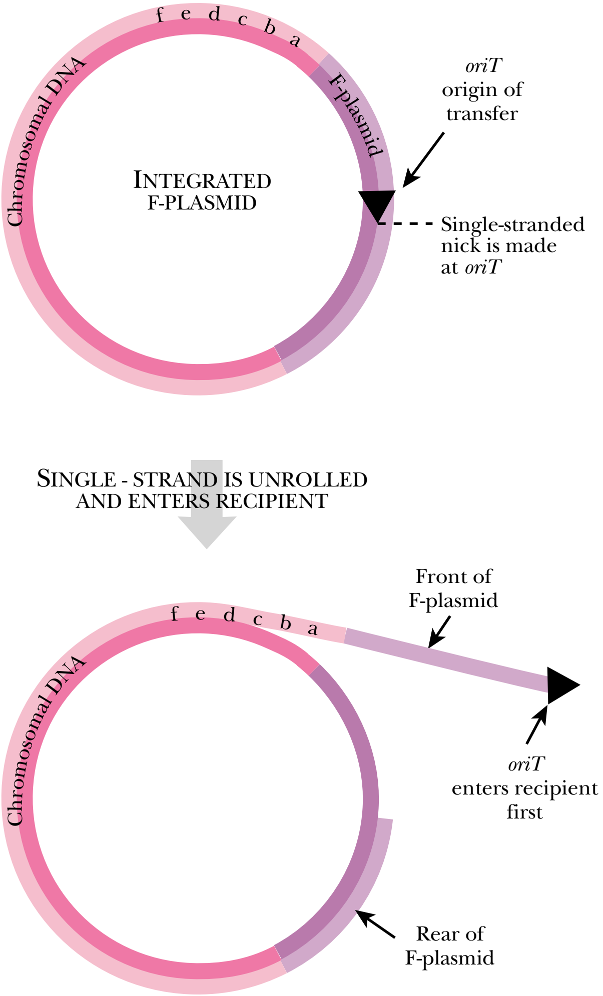


FIGURE 18.16 *Transfer of Chromosomal Genes by F-Plasmid*

An integrated F-plasmid can still induce bacterial conjugation and rolling circle transfer of DNA into another bacterial cell. Since rolling circle replication does not stop until the entire circle is replicated, the attached chromosome is also transferred into the recipient cell. First, a single-stranded nick is made at the *ori*T, or transfer origin of the integrated plasmid. The free 5′ end (black triangle) enters the recipient cell through the conjugation bridge. Once inside the recipient, the second strand of DNA is synthesized. Notice that the transfer of the single stranded DNA does not end with the F-plasmid DNA and continues into the chromosomal DNA. Genes closest to the site of plasmid integration are transferred first (in the order a, b, c, d, e, f, in this example). The amount of chromosomal DNA that is transferred depends on how long the two bacteria remain attached by the conjugation bridge.

guished phenotypically, usually by their growth properties. For example, the recipient may have a mutation that makes the bacteria unable to grow with only lactose as a carbon source. The donor Hfr strain would have the gene that restores the ability to grow on lactose. Using this method, genetic maps may be constructed by two major approaches. First, the **cotransfer frequency** of two genes may be measured. For example, if gene a and b are close to each other, the donor Hfr strain would transfer these two genes together at a high frequency. Alternately, if gene a and b were on opposite sides of the chromosome, the donor Hfr strain would usually only transfer gene a, and the cotransfer frequency would be low.

Secondly, time of entry measurements may be made. Genes are transferred starting at the site where the F-plasmid is integrated and proceeding sequentially around the circular chromosome (Fig. 18.17). The length of time it takes for a series of genes to enter the recipient cell gives an estimate of their relative distance from the origin of transfer of the Hfr strain used. In order to determine the time of entry by conjugation the site and orientation of the F-plasmid must be known. In addition, mutations in the genes being studied (a, b, c, and d) must give recognizable phenotypes, such as the ability to grow using lactose as a carbon source or a requirement for some nutrient. Finally, the recipient must be resistant to some antibiotic (e.g., streptomycin) so that it can be selected on medium that prevents growth of the Hfr strain. Different Hfr strains will transfer the same genes in different orders and at different times, depending on their location relative to the site of integration of the F plasmid.

cotransfer frequency Frequency with which two genes remain associated during transfer of DNA between cells

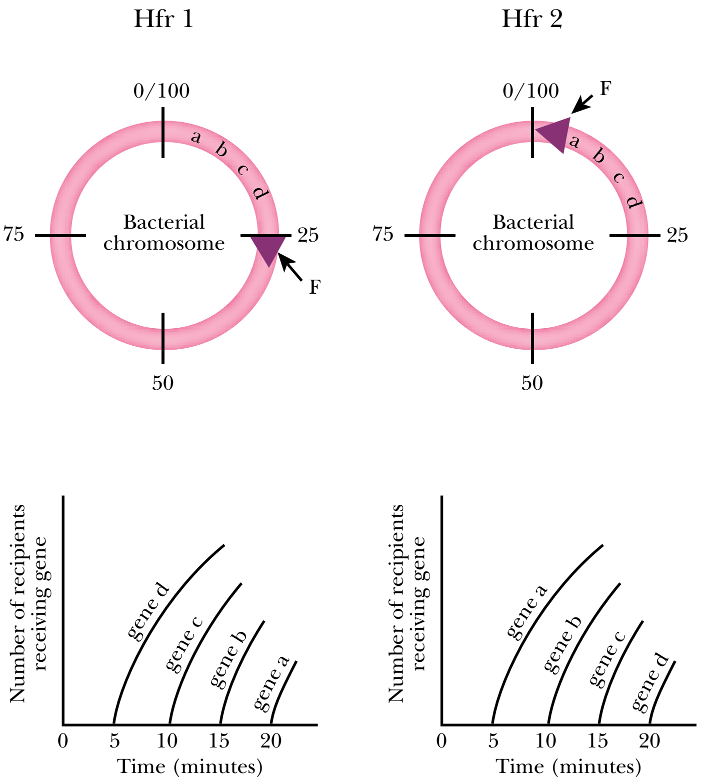


FIGURE 18.17 ***Time of Entry by Conjugation***

To determine the time of entry by conjugation the Hfr strain is mixed with a recipient strain carrying a defective copy of a particular gene, "a". After conjugation has proceeded for a specific time, a sample of the mixture is removed. This is plated on agar that prevents growth of the Hfr and only allows growth of strains carrying the wild type version of gene "a". Survivors are derivatives of the recipient that have gained the wild type version of gene "a" from the Hfr. This is repeated for several time points. The whole procedure is then repeated for the other genes. In strain Hfr 1 (left panel), the integrated F-plasmid is closest to gene "d" and only begins transferring gene "a" after about 20 minutes. In strain Hfr 2 (right panel), the F-plasmid is integrated closer to gene "a" which therefore begins to appear in the recipient as early as five minutes after transfer begins.

Integrated F-plasmids may excise themselves from the chromosome by reversing the integration process. Sometimes they excise with pieces of chromosomal DNA, much like the way lambda forms specialized transducing phages. Recombination may occur between homologous sequences in the chromosome outside the F-DNA. The example shown in (Fig. 18.18) uses the IS1 sequences that are found in multiple copies in the *E. coli* chromosome but are absent from the F-plasmid. This creates F′- or F-prime plasmids which retain all of the F-plasmid and gain extra chromosomal DNA. F-prime plasmids may be transferred to recipient cells, carrying with them the chromosomal genes from their original host cell. F-primes are often used to carry part of the *lacZ* gene in the alpha-complementation method used for screening recombinant plasmids (see Ch. 22).

Gene Transfer among Gram-Positive Bacteria

Traditionally, the Eubacteria are divided into two major groups, the **gram-negative** and the **gram-positive bacteria**. This division was originally based on their response to the

gram-negative bacteria Major division of Eubacteria that possess an extra outer membrane lying outside the cell wall
gram-positive bacteria Major division of Eubacteria that lack an extra outer membrane lying outside the cell wall

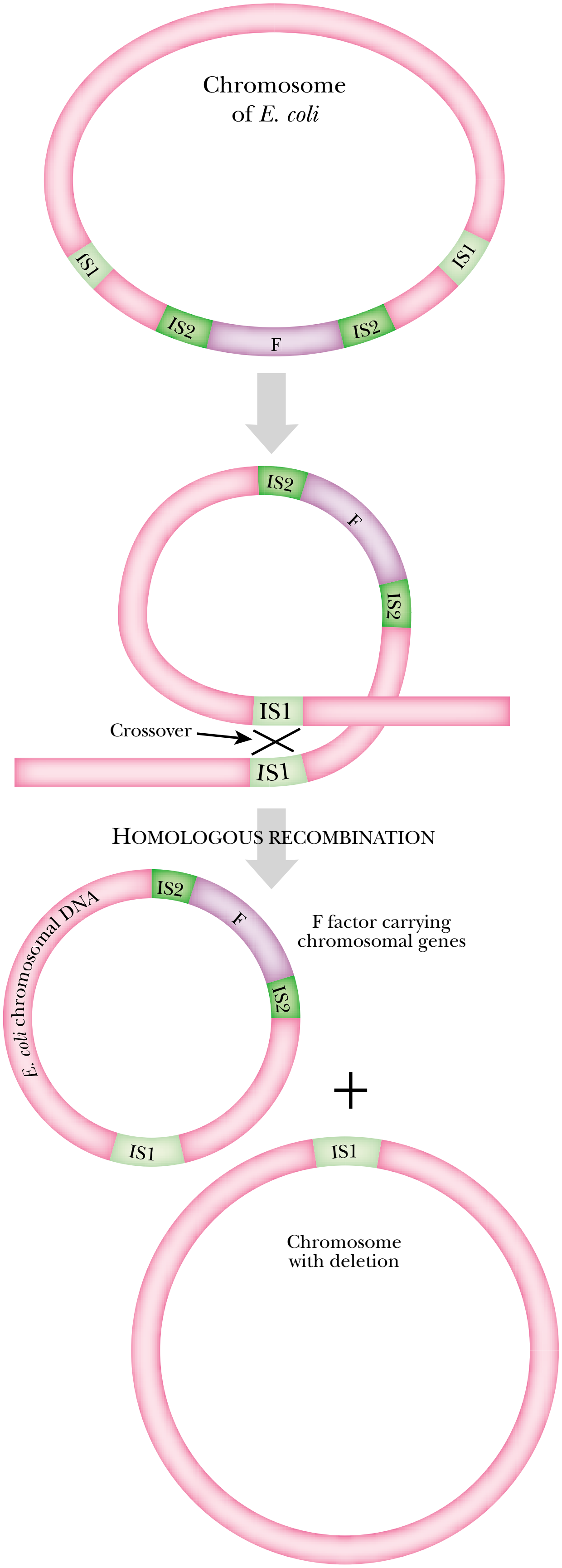


FIGURE 18.18 ***Formation of F-prime Plasmid***

Integrated F-plasmids can excise from the chromosomal DNA to re-form independent plasmids. In some instances, the plasmids capture pieces of chromosomal DNA during excision. In the example shown the F plasmid (purple) originally integrated using IS2 (shown flanking F). If different insertion sequences are used during excision, a novel plasmid is created. Thus, recombination between the two IS1 sequences creates a large plasmid that includes the original F-plasmid plus some chromosomal DNA. The chromosome now contains a deletion and the F-plasmid contains the corresponding chromosomal DNA. The new plasmid is termed an F-prime or F′-plasmid.

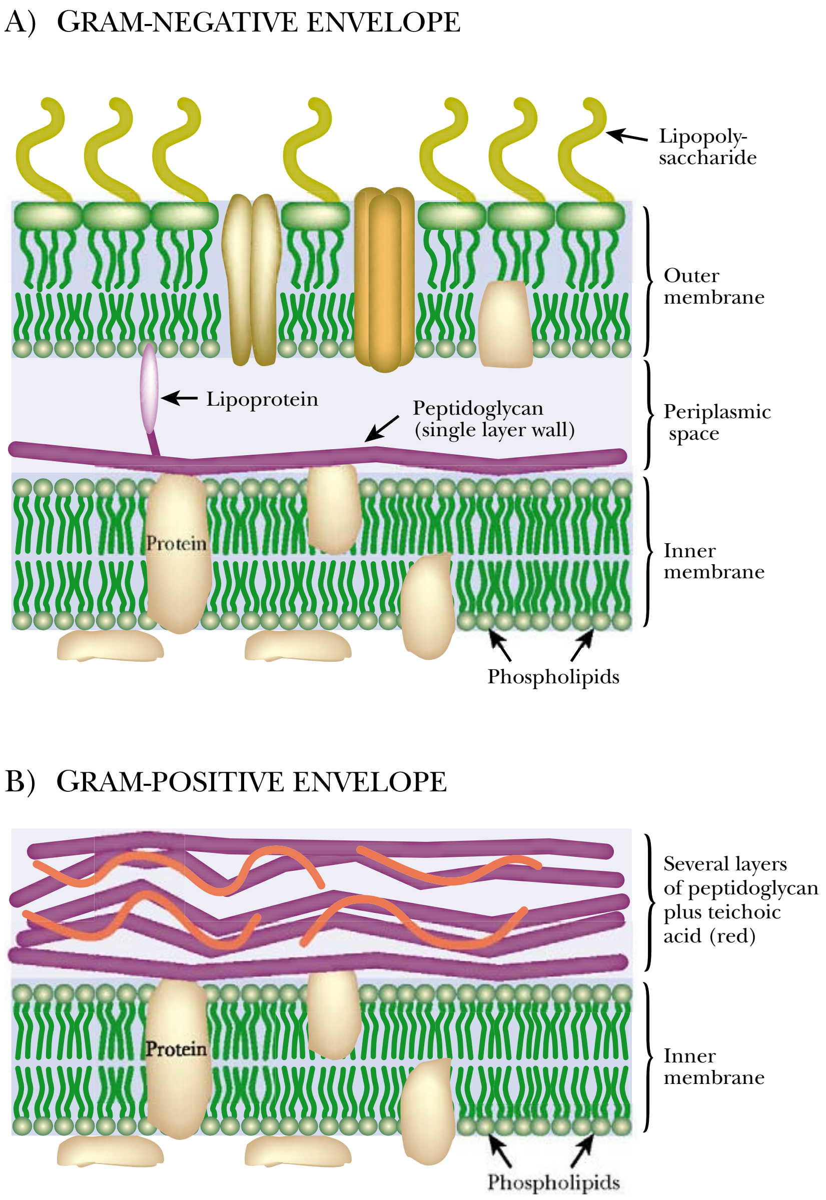

A) GRAM-NEGATIVE ENVELOPE
Lipopoly-saccharide
Outer membrane
Lipoprotein
Peptidoglycan (single layer wall)
Periplasmic space
Protein
Inner membrane
Phospholipids
B) GRAM-POSITIVE ENVELOPE
Several layers of peptidoglycan plus teichoic acid (red)
Protein
Inner membrane
Phospholipids

FIGURE 18.19 ***Differences in Envelopes of Gram-negative and Gram-positive Bacteria***

The outer surfaces of gram-positive and gram-negative bacteria have different structures. A) In gram-negative bacteria, such as *E. coli*, there are three surface layers. The outermost layer, called the outer membrane, is a lipid-bilayer that contains various proteins embedded within the lipids, and an outer coating of lipopolysaccharide. Next, within the periplasmic space, the cell wall contains a single layer of peptidoglycan. Lipoproteins connect this cell wall to the outer membrane. The layer closest to the cytoplasm, called the inner membrane, is a lipid bilayer embedded with various proteins. B) The outer surface of gram-positive bacteria only has two layers, a thick coating of peptidoglycan and teichoic acid surrounding the inner membrane.

Gram-negative bacteria, including *Escherichia coli*, have an extra outer membrane.

Gram stain. These differences in staining are in fact due to differences in the chemical composition and structure of the cell envelope. The envelope of gram-negative bacteria consists of the following layers (from inside to outside): cytoplasmic membrane, cell wall (peptidoglycan) and **outer membrane** (Fig. 18.19). The envelope of gram-positive bacteria is simpler and lacks the outer membrane. Both kinds of bacteria sometimes have an extra protective layer, the capsule, on the very outside.

The gram-negative bacterium *E. coli is* widely used as a host for cloning and expressing genes from a variety of other organisms. The synthesis of large amounts of a purified recombinant protein, such as a human growth factor or hormone is often desirable. Secretion of a recombinant protein into the culture medium would be very convenient since this avoids purifying it away from all the other proteins inside the bacterial cell. However, the complex envelope of gram-negative bacteria is a major hindrance in the excretion of proteins into the culture medium. In contrast, secretion across the simpler gram-positive envelope is easier. Indeed, many gram-positive bacteria, such as *Bacillus*, excrete proteins into the culture medium naturally. As a result,

outer membrane Extra membrane lying outside the cell wall in gram-negative but not gram-positive bacteria

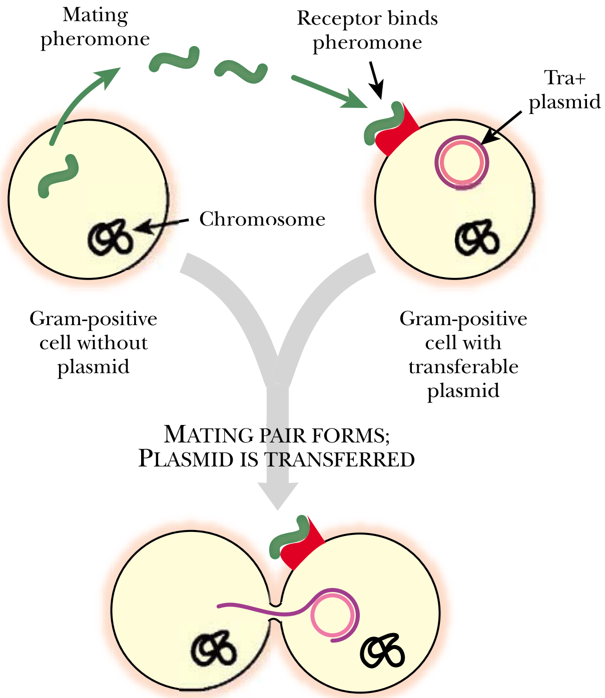

Mating pheromone
Receptor binds pheromone
Tra+ plasmid
Chromosome
Gram-positive cell without plasmid
Gram-positive cell with transferable plasmid
MATING PAIR FORMS; PLASMID IS TRANSFERRED

FIGURE 18.20
Pheromones Induce Mating in Gram-positive Bacteria

In gram-positive bacteria such as *Enterococcus*, cells without plasmids secrete pheromones to attract bacteria with transferable (Tra$^+$) plasmids. Mating pheromones bind to receptors on the surface of cells containing Tra$^+$ plasmids. Binding the receptor activates the transfer genes to form a conjugation bridge and transfer the plasmid by rolling circle replication. Each pheromone is specific and only attracts bacteria with certain plasmids.

there is considerable interest in using gram-positive bacteria as hosts in genetic engineering. Unfortunately, the genetics of gram-positive bacteria is far behind that of the intensively studied *E. coli* and its relatives. Nonetheless, mechanisms of gene-transfer are available in gram-positive bacteria.

Transfer of plasmids between gram-positive bacteria is often promoted by pheromones.

Self-transmissible plasmids are widespread among gram-positive bacteria and many of these plasmids are rather promiscuous. Since the cell envelope is simpler in gram-positive bacteria, plasmid transfer is also simpler and a sex pilus is not needed. Apparently only half-a dozen genes are required to encode the transfer functions. Some gram-positive bacteria, such as *Enterococcus*, secrete mating pheromones into the culture medium. These are short peptides that induce the *tra* genes of plasmids in neighboring bacteria. This results in aggregation and plasmid transfer (Fig. 18.20). Different pheromones are specific for different plasmids. Only bacteria that lack a particular plasmid secrete the corresponding pheromone. Furthermore, the plasmid only expresses its transfer genes when a suitable recipient is nearby.

Gram-positive bacteria also harbor **conjugative transposons** (e.g. Tn916 of *Enterococcus*). These can transfer themselves from one bacterial cell to another (see Ch. 15). Apparently these elements excise themselves temporarily from the chromosome of the donor cell before conjugation. Once inside the recipient, they re-insert themselves into the bacterial chromosome.

Archaebacterial Genetics

There are two genetically distinct lineages of prokaryotes, the "normal" bacteria or **Eubacteria** and the **Archaebacteria (or Archaea)**. Although both have a prokaryotic cell without a nucleus, the Eubacteria and Archaebacteria are no more related to each other than either is to the eukaryotes. [See Ch. 20, Molecular Evolution,

Archaebacteria (or Archaea) Type of bacteria forming a genetically distinct domain of life. Includes many bacteria growing under extreme conditions
conjugative transposon Type of transposon that can transfer themselves from one bacterial cell to another by conjugation
Eubacteria Bacteria of the normal kind as opposed to the genetically distinct Archaebacteria

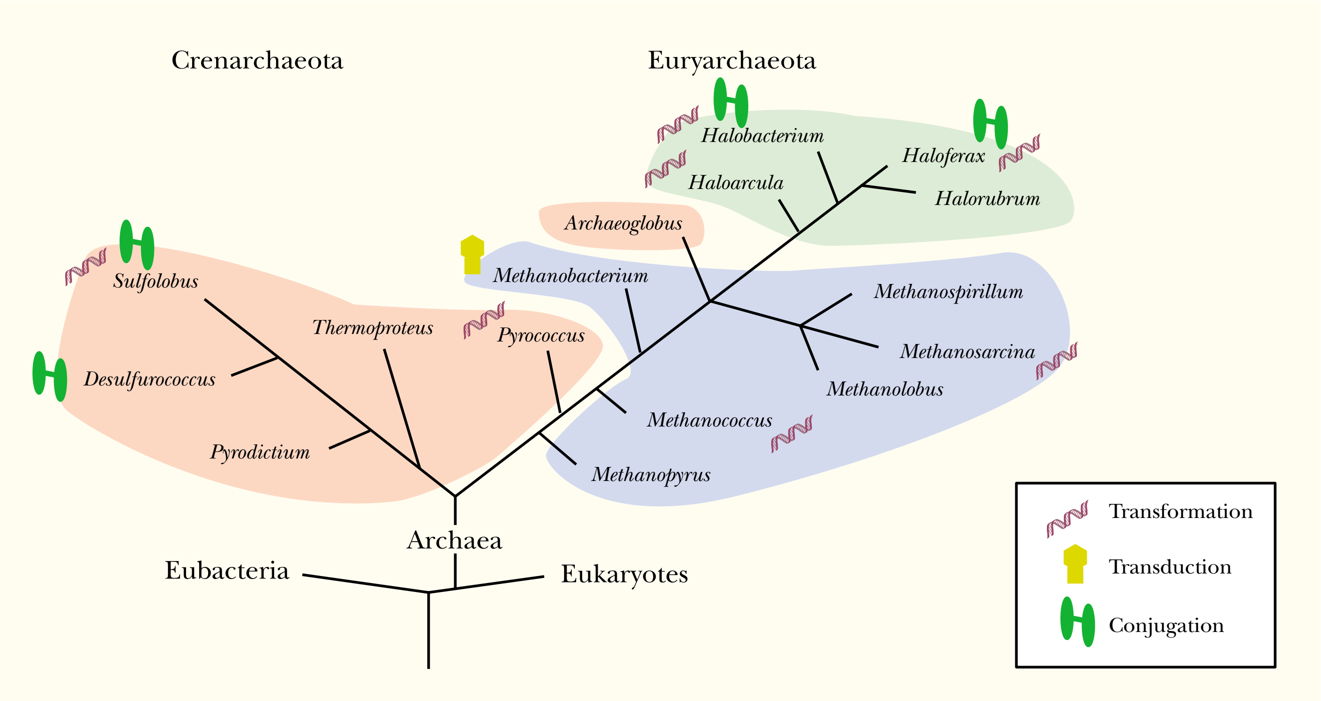


FIGURE 18.21 ***Groups of Archaebacteria and their Gene Transfer Mechanisms***

Phylogenetic tree of the Archaea lineage illustrating that different types of gene transfer can occur. The green zone contains salt tolerant organisms, the blue zone indicates methane producers and the red zone contains Archaea that grow at extremely high temperatures. Some Archaea use transformation whereas others use conjugation. Rare cases of viral transduction also occur. The modes of gene transfer seen within each family do not correlate well with either lifestyle or evolutionary relationships. The Crenarchaeota and the Euryarchaeota are the two major branches of the Archaea.

for further discussion of these relationships.] The Eubacteria include most bacteria found in normal environments, including both the gram-negative and gram-positive bacteria discussed above. The Archaebacteria include the methane bacteria and a variety of less well-known bacteria found in extreme environments. Many have strange biochemical pathways and are adapted to extremes of temperature, pH or salinity. This makes Archaebacteria an attractive source of novel enzymes or proteins with unusual properties and/or resistance to extreme conditions. There are many possible industrial uses for enzymes capable of withstanding extreme temperatures, for example.

Archaebacteria are genetically distinct and often live under unusual or extreme conditions.

Although several complete genome sequences are available for members of the Archaebacteria, development of systems for gene transfer has lagged way behind the Eubacteria. There are many practical problems, including the need to grow many Archaebacteria under extreme conditions. For example, some extreme thermophiles grow at temperatures high enough to melt agar. Obtaining colonies on solid media has required the development of alternative materials.

Plasmids have been found in several Archaebacteria and some have been developed into cloning vectors (Fig. 18.21). Transformation procedures now exist for getting DNA into several Archaebacteria. They rely on removal of divalent cations, especially Mg^{2+}, which results in the disassembly of the glycoprotein layer surrounding many Archaebacterial cells. (Note the contrast with the corresponding procedures for eubacteria, which involve cold-shock in the presence of divalent cations!) It has been possible to express the *lacZ* reporter gene in methane bacteria under control of an archaebacterial promoter. However, staining of β-galactosidase with Xgal requires exposure to air, which kills methane bacteria! Consequently, colonies must first be replicated and one set sacrificed for analysis.

Gene transfer in Archaebacteria is widespread but still poorly understood.

A major problem is choice of a selectable marker. Most standard antibiotics do not affect Archaebacteria due to their unusual biochemistry. For example, Archaebacteria do not have cell walls made of peptidoglycan and are therefore not susceptible to penicillins. In addition, many resistance proteins from normal organisms are denatured at the extremes of temperature, salinity or pH under which many Archaebacteria grow. Novobiocin (a DNA gyrase inhibitor—see Ch. 5) and mevinolin (an inhibitor of the isoprenoid pathway) have been used to inhibit halophiles, and puromycin and neomycin (both protein synthesis inhibitors—see Ch. 16) will inhibit methane bacteria.

Viruses have been discovered that infect many Archaebacteria. So far only one, the ΨM1 phage of *Methanobacterium thermoautotrophicum*, has been shown to transduce genes of its host bacterium. Unfortunately this is of no practical use because of the low burst size—about six phage are liberated per cell after infection. The SSV1 phage of *Sulfolobus solfataricus* integrates into the bacterial chromosome and may be of future use.

Conjugation in Archaebacteria is of two types. Self-transferable plasmids that promote conjugation are found in *Sulfolobus*. In contrast, some halobacteria form conjugation bridges without the involvement of fertility plasmids. Moreover, in these cases DNA transfer is bi-directional. Neither of these phenomena has so far been developed into routine gene-transfer systems.

Whole Genome Sequencing

The bacterium, *Hemophilus influenzae*, was the first organism to have its DNA completely sequenced.

The techniques for gene transfer described in this chapter have allowed the construction of genetic maps for *E. coli* and a few other well investigated bacteria. However, for the vast majority of microorganisms, no "classical" genetics exists. Nowadays these are largely being investigated by more modern techniques, such as gene cloning and DNA sequencing.

Since the development of rapid automated techniques for sequencing DNA (see Ch. 24) many whole genomes have been totally sequenced. The first genome sequence to be finished was *Hemophilus influenzae*, in 1995. Since 1995, nearly 50 complete bacterial genomes have been sequenced, and another 100 are partially sequenced. Sequence comparison with genes of well-investigated organisms allows provisional identification of many genes. However, even in *E. coli,* the function of about a third of the genes remains uncertain.

Virulence genes are often clustered together forming "islands".

Differences in GC/AT ratios reveal segments of chromosomes with foreign origins.

Whole genome sequencing of pathogenic bacteria and comparison with their harmless relatives may reveal extra blocks of genes responsible for causing disease. Many virulence genes are carried on plasmids as discussed in Chapter 16. Others are found clustered together in regions of the chromosome known as "**pathogenicity islands**". Most genes of *Salmonella*, as well as their order around the chromosome, correspond to those of its close relative *E. coli*, as would be expected. However, extra segments of DNA are found in *Salmonella* that are lacking in *E. coli*. Some of these are pathogenicity islands (Fig. 18.22). Such extra regions are often flanked by inverted repeats, implying that the whole region was inserted into the chromosome by transposition at some period in the evolutionary past. In agreement with this idea, such islands are often found in some strains of a particular species but not others. In addition, these islands tend to have different GC to AT ratios and/or codon usage frequencies from the rest of the chromosome, suggesting their origin in some other organism. Conversely, *E. coli* possesses a few DNA segments missing in *Salmonella*. Interestingly, one of these is the area including the *lac* operon and a few surrounding genes. Thus the classic *lac* operon, the most-studied "typical" gene of the "standard organism" is probably a relatively recent intruder into the *E. coli* genome!

Pathogenicity islands are simply the best known case of "specialization islands". These are blocks of contiguous genes, presumed to have a "foreign" origin, that con-

pathogenicity island Region of bacterial chromosome containing clustered genes for virulence

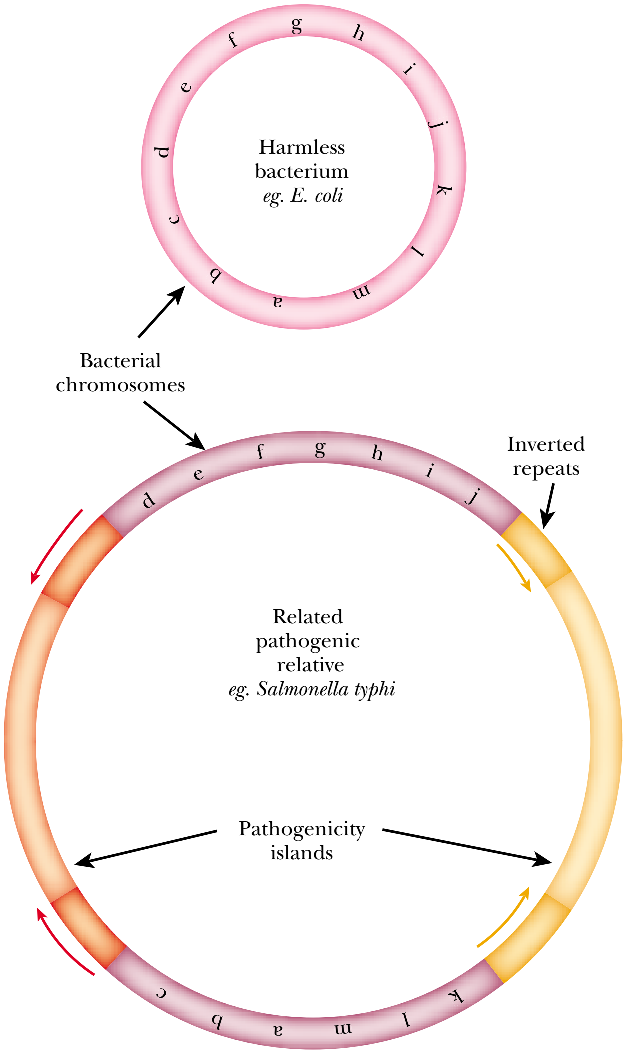


FIGURE 18.22
Pathogenicity Islands of Salmonella

Comparison of the *E. coli* genome with its close relative, *Salmonella*, reveals large regions of DNA that have no homology (orange). The remaining regions have similar genes that are in identical order. For example, *Salmonella* genes d through j are clustered together in the exact same order as *E. coli* genes d through j. Since *Salmonella* is pathogenic and *E. coli* is not, the regions of no homology probably encode the genes required for pathogenicity; therefore, they are termed pathogenicity islands. The islands are flanked by inverted repeats, suggesting the DNA may have been acquired through transposition. Note: this figure is not drawn to scale; the pathogenicity islands are greatly exaggerated relative to the rest of the chromosome for purposes of illustration.

tribute to some specialized function that is not needed for simple survival. Not surprisingly, medical relevance has drawn most human interest. Other examples include genes encoding pathways for the biodegradation of aromatic hydrocarbons, herbicides and other products of human industry and pollution.

Horizontal transfer of genes is especially significant in bacteria.

Movement of genes "sideways" is designated **lateral or horizontal gene transfer** in distinction to the "vertical" transfer of genes from ancestors to their direct descendents. Horizontal gene transfer can occur by natural transformation, viral transduction, or transposon jumping. Horizontal gene transfer may occur between closely related organisms or those far apart taxonomically. Estimates suggest that in typical bacteria around 5% of the genes have been obtained by lateral gene transfer, and in rare cases up to 25%. *Thermotoga* is a eubacterium adapted to life at very high temperatures and which consequently shares its habitat with several archaebacteria. *Thermotoga* has apparently gained around 25% of its genes by transfer from thermophilic archaebacteria such as *Archaeoglobus* and *Pyrococcus*. When we remember that the F-plasmid of *E. coli* can mediate DNA transfer into yeast (see Ch. 16), these results are perhaps not so surprising.

horizontal gene transfer Movement of genes sideways between unrelated organisms. Same as lateral gene transfer
lateral gene transfer Movement of genes sideways between unrelated organisms. Same as horizontal gene transfer

CHAPTER NINETEEN

Diversity of Lower Eukaryotes

Origin of the Eukaryotes by Symbiosis

The Genomes of Mitochondria and Chloroplasts

Primary and Secondary Endosymbiosis

Is Malaria Really a Plant?

Symbiosis: Parasitism versus Mutualism

Bacterial Endosymbionts of Killer *Paramecium*

Is *Buchnera* an Organelle or a Bacterium?

Ciliates have Two Types of Nucleus

Trypanosomes Vary Surface Proteins to Outwit the Immune System

Mating Type Determination in Yeast

Multi-Cellular Organisms and Homeobox Genes

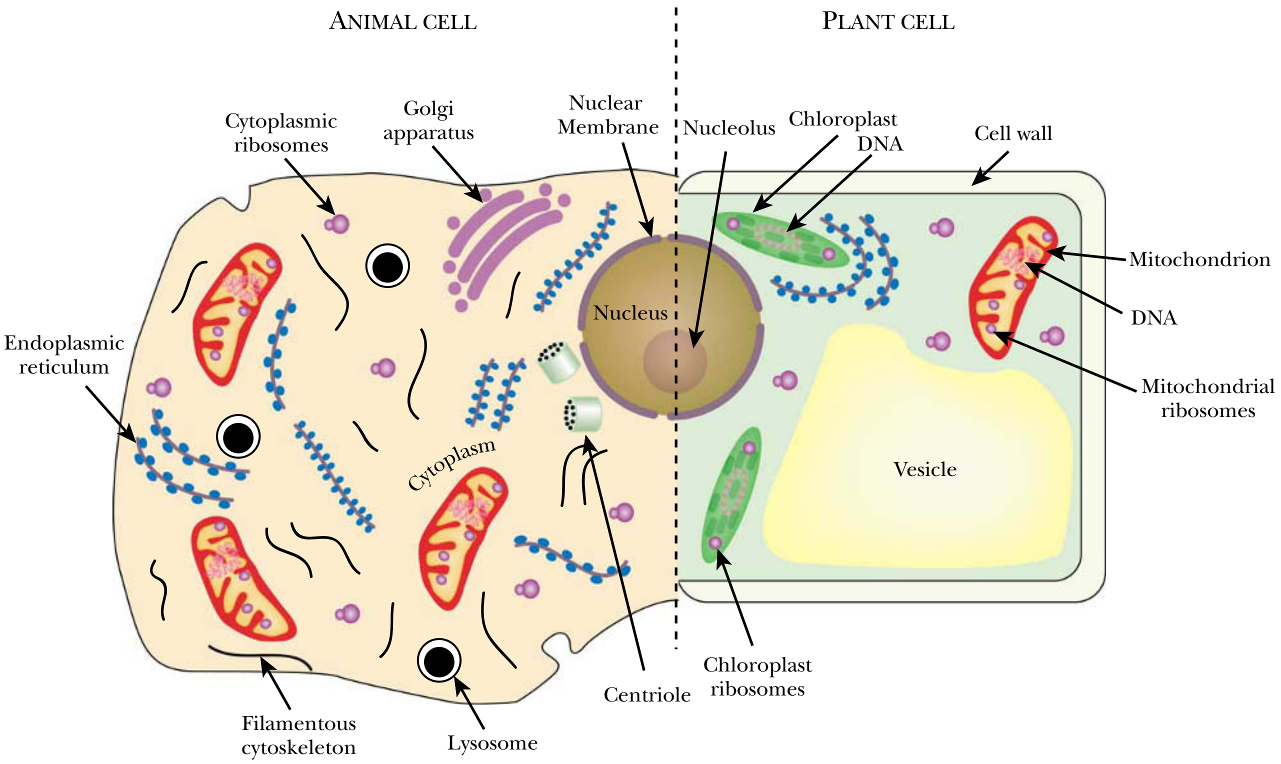


FIGURE 19.01 ***Defining Features of Eukaryotic Cells***

Eukaryotic cells have membrane bound compartments that are not found in prokaryotes. In a typical animal cell, these compartments include the nucleus, mitochondria, endoplasmic reticulum, golgi apparatus, and lysosomes. In addition a typical plant cell also has chloroplasts, which harvest light energy and convert it into ATP. Animal cells maintain their three-dimensional shape with an internal cytoskeleton composed of microtubules and microfilaments. In contrast, plant cells maintain their shape by a rigid cell wall surrounding the cytoplasm.

Origin of the Eukaryotes by Symbiosis

Eukaryotic cells are divided into compartments, including the nucleus, by membranes.

Unlike the cells of the more primitive prokaryotes, eukaryotic cells are divided into compartments by membranes (Fig. 19.01). The most important of these is the nucleus, where the chromosomes reside. Eukaryotes are defined by the possession of a nucleus that typically contains several linear chromosomes. Higher eukaryotes are normally diploid and have pairs of homologous chromosomes, although this is not always the case with the less advanced eukaryotes.

Eukaryotes are derived from the merger of two or more ancestral organisms.

In addition to a nucleus, almost all eukaryotic cells (animal, plant or fungus) contain mitochondria. Plant cells contain chloroplasts as well as mitochondria. Both of these organelles provide the majority of energy for all cellular processes. These organelles contain their own genomes and encode at least a few of their own proteins. The organelle genome is prokaryotic in nature. It consists of a circular DNA molecule that is not bound by histones. Mitochondria and chloroplasts synthesize their own ribosomes, which are more closely related to those of bacteria than to those of the eukaryotic cytoplasm. Both mitochondria and chloroplasts are roughly the same size and shape as bacterial cells, and the organelles grow and divide in the same manner as bacteria (Fig. 19.02). When a eukaryotic cell divides, each daughter cell inherits some of its parent's mitochondria and chloroplasts. If either organelle is lost, it cannot be reconstructed because the nucleus does not have all of the genetic information needed to synthesize the entire organelle.

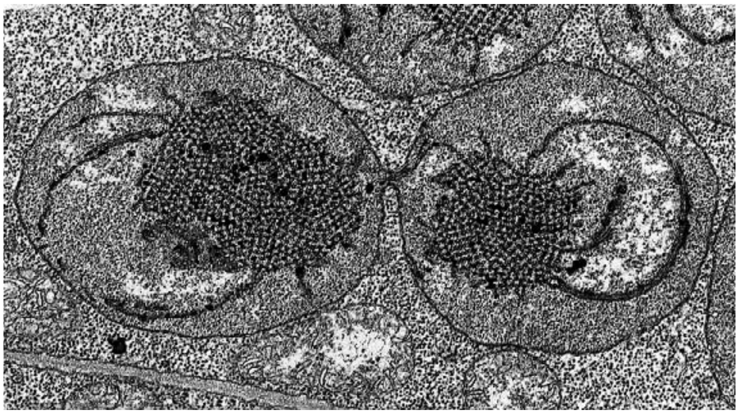


FIGURE 19.02 ***Chloroplasts arise by Division***

Transmission electron micrograph of a dividing chloroplast in a bean seedling. Chloroplasts and mitochondria divide independently of the eukaryotic cell in which they reside. The organelles divide by binary fission in a manner reminiscent of prokaryotic cells. The plastid shown here is technically an "etioplast", a precursor chloroplast that has not yet developed any green pigment. From: Biochemistry and Molecular Biology of Plants by Buchanan, Gruissem and Jones, 2000, American Society of Plant Physiologists.

The **symbiotic theory** proposes that the complex eukaryotic cell arose by a series of symbiotic events in which organisms of different lineages merged. The cells of higher organisms are thus not individuals but symbiotic associations. The word "**symbiosis**" is from the Greek sym, meaning together, and bios, meaning life. The nuclear genes of a eukaryotic cell are sometimes referred to as derived from the "**urkaryote**." The urkaryote is the hypothetical ancestor that provided the genetic information found in the present day eukaryotic nucleus.

Mitochondria are derived from ancestral bacteria that specialized in respiration whereas chloroplasts are descended from ancestral photosynthetic bacteria.

According to the symbiotic theory, mitochondria are descended from bacteria that were trapped long ago by the ancestors of modern eukaryotic cells. These bacteria received shelter and nutrients, and in return, devoted themselves to generating energy by respiration. During the eons following their capture, these bacteria became narrowly specialized for energy production, lost the ability to survive on their own, and evolved into mitochondria. The term **endosymbiosis** is sometimes used to indicate those symbiotic associations where one partner is physically inside the other (endo is from the Greek for inside), as in the present case.

Plant cells contain chloroplasts that perform photosynthesis with light harvesting pigments such as chlorophyll. The rRNA from chloroplasts matches rRNA from photosynthetic bacteria better than rRNA from the plant cell nucleus. Thus, the prevailing theory is that chloroplasts descended from photosynthetic bacteria trapped by the ancestors of modern-day plants. Some plants have lost the ability to photosynthesize but still contain defective chloroplasts. The term **plastid** refers to all organelles that are genetically equivalent to chloroplasts, whether functional or not. Since fungi do not contain chlorophyll, the green light-absorbing pigment of all other plants, it was once thought that fungi were degenerate plants that had lost their chlorophyll through evolution. However, fungi contain no trace of a plastid genome and rRNA analysis implies that the ancestral fungus was never photosynthetic, but split off from the ancestors or green plants before the capture of the chloroplast. If anything, rRNA sequencing implies that fungi are more closely related to animals than plants.

The Genomes of Mitochondria and Chloroplasts

Chloroplasts and mitochondria possess small circular genomes.

Both mitochondria and chloroplasts contain a genome consisting of a circular DNA molecule that is presumably derived from the ancestral bacterial chromosome. Over evolutionary time, these organelle genomes have lost many genes that were unnecessary for life as an organelle inside a host cell. In addition, many genes that are still necessary have been transferred to the chromosomes in the nucleus. As a consequence, the mitochondria of animals have very little DNA left. For example, human mito-

endosymbiosis Form of symbiosis where one organism lives inside the other
plastid Any organelle that is genetically equivalent to a chloroplast, whether functional in photosynthesis or not
symbiosis Association of two living organisms that interact
symbiotic theory Theory that the organelles of eukaryotic cells are derived from symbiotic prokaryotes
urkaryote Hypothetical ancestor that provided the genetic information of the eukaryotic nucleus

chondrial DNA has only 13 protein encoding genes, together with genes for several rRNA and tRNA molecules (Fig. 19.03). However, the mitochondrial proteome, that is the complete set of proteins expressed in mitochondria, numbers approximately 400. The genes for most of these reside in the nucleus and these polypeptides must be imported into the organelle after synthesis on the ribosomes of the eukaryotic cytoplasm.

The chloroplasts of higher plants retain rather more DNA than mitochondria—approximately enough for a hundred genes—but this is still much less than their bacterial ancestors. It is estimated that 1,000 or more genes from the ancestral photosynthetic prokaryote have been transferred to the plant cell nucleus.

Defects due to mutations in mitochondrial genes are inherited maternally.

During sexual reproduction, mitochondria and chloroplasts are inherited maternally. When a sperm fertilizes an egg cell to create a zygote, the organelles of the sperm are lost. The new individual retains the organelles from the egg cell, i.e. those from the female parent only. Certain inherited defects of humans are due to mutations in the mitochondrial DNA. These affect the generation of energy by respiration and affect the function of muscle and nerve cells in particular. These defects are passed on through the maternal line as all children with the same mother inherit the same mitochondria.

Partial exceptions to the rule of maternal inheritance for organelles occur in a few single-celled eukaryotes. *Chlamydomonas* is a single-celled green alga whose cells contain a single chloroplast. During mating, about 5% of the zygotes have two chloroplasts rather than one. In these cells recombination can occur between the two different chloroplast genomes. Division of the zygote gives cells with only a single chloroplast each. These may be examined to determine the outcome of the genetic crosses.

Primary and Secondary Endosymbiosis

When organisms of different species live together the situation is referred to as symbiosis and when one organism lives inside the other it is known as endosymbiosis. Especially in the latter case, both organisms will show major adaptations to the symbiotic state. **Primary endosymbiosis** refers to the original internalization of a prokaryote by an ancestral eukaryotic cell, resulting in the formation of an organelle. This process gave rise to the mitochondria and chloroplasts of most eukaryotes. Two membranes surround mitochondria and chloroplasts. The inner one is derived from the bacterial ancestor and the outer "mitochondrial" or "chloroplast" membrane is actually derived from the host cell membrane. However, several lineages of protozoans appear to have engulfed other single-celled eukaryotes, in particular algae. Several groups of algae therefore have chloroplasts acquired at second-hand by what is termed **secondary endosymbiosis**.

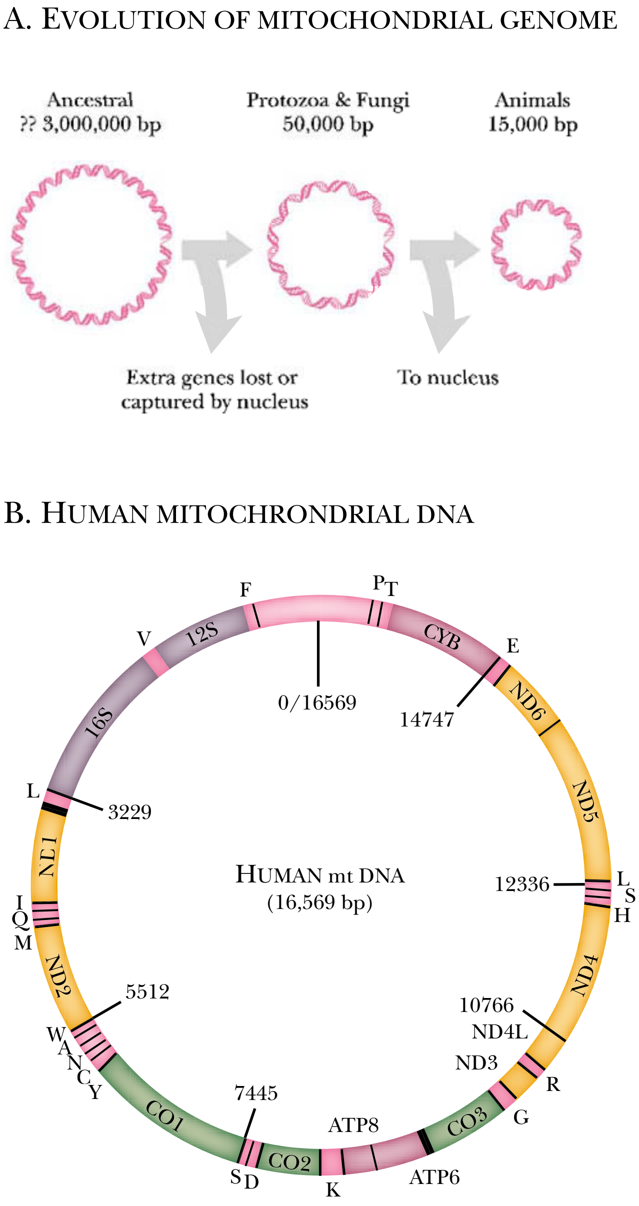
In contrast to the typical two membranes of primary organelles, four membranes surround chloroplasts obtained by secondary endosymbiosis. In most cases, the nucleus of the engulfed eukaryotic alga has disappeared without trace. Occasionally, the remains of this nucleus are still to be found lying between the two pairs of membranes (Fig. 19.04). This structure is termed a **nucleomorph** and can be seen in cryptomonad algae where it represents the remains of the nucleus of a red-alga that was swallowed by an amoeba-like ancestor. The nucleomorph contains three vestigial linear chromosomes totaling 550 kb of DNA. These carry genes for rRNA that is incorporated into a few eukaryotic type ribosomes that are also located in the space between the two pairs of membranes.

Cells due to secondary endosymbiosis are composites of four or five original genomes. These include the primary ancestral eukaryote nucleus and its mitochondrion

nucleomorph Degenerate remains of the nucleus of a symbiotic eukaryote that was incorporated by secondary endosymbiosis into another eukaryotic cell

primary endosymbiosis Original uptake of prokaryotes by the ancestral eukaryotic cell, giving rise to mitochondria and chloroplasts

secondary endosymbiosis Uptake by an ancestral eukaryotic cell of another single-celled eukaryote, usually an alga, thus providing chloroplasts at second hand



A. EVOLUTION OF MITOCHONDRIAL GENOME

B. HUMAN MITOCHRONDRIAL DNA

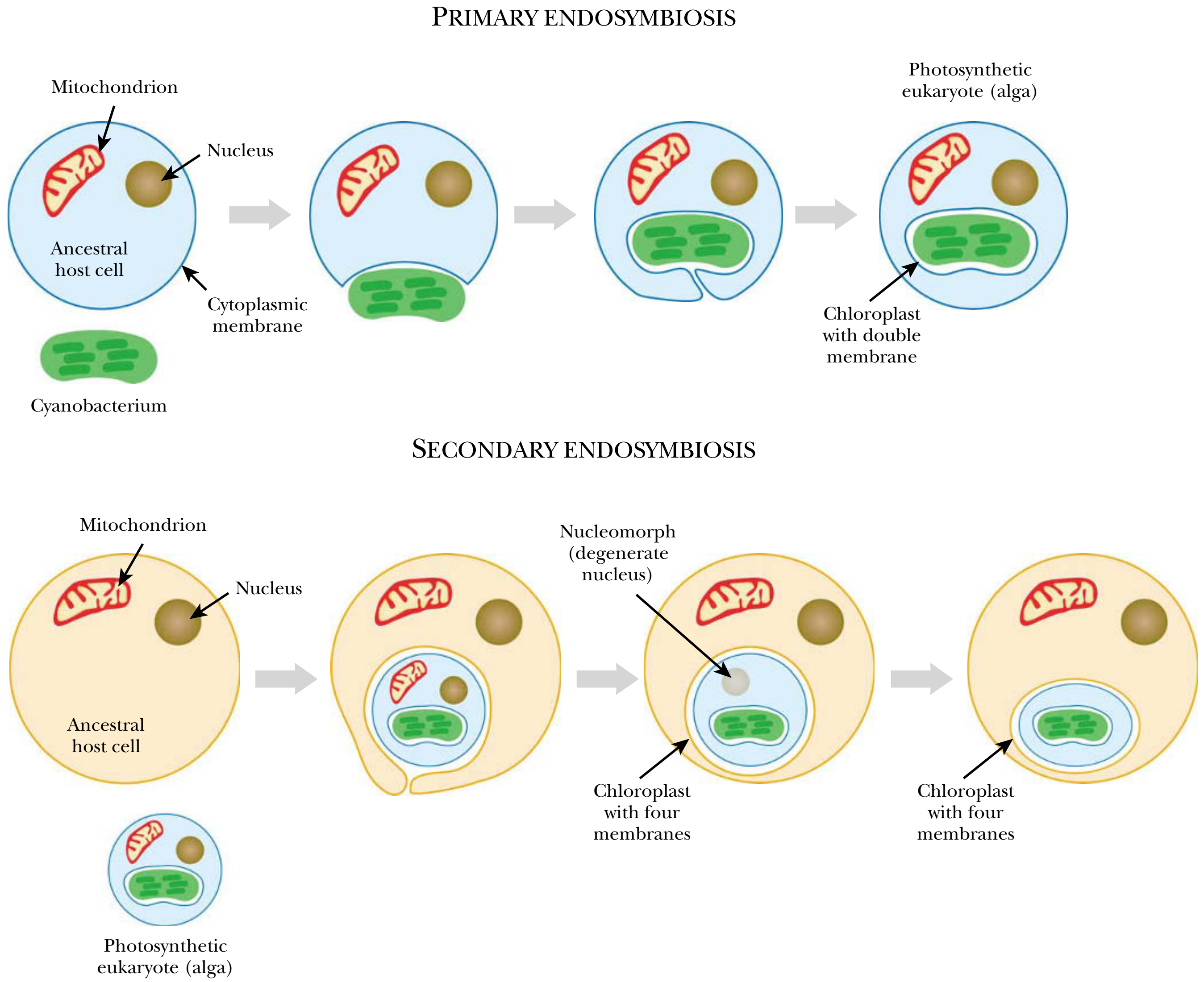
FIGURE 19.03 ***Genetic Map of Human Mitochondrial DNA***

A: During evolution, the mitochondrial genome has been streamlined. Many of the genes necessary for mitochondrial function have moved to the nucleus, cause the mitochondrial genome to shrink in size.
B: The mitochondrial DNA of humans contains the genes for ribosomal RNA, transfer RNA, and some proteins of the electron transport chain.

plus the chloroplast, mitochondrion and nucleus from the secondary endosymbiont. Many of the genes of the subordinate genomes have been lost during evolution and no trace has ever been found of the secondary mitochondrion. Some genes from the secondary endosymbiont nucleus have been transferred to the primary eukaryotic nucleus. The protein products of about 30 of these are made on ribosomes belonging to the primary nucleus and shipped from the primary eukaryotic cytoplasm back into the nucleomorph compartment. In turn the nucleomorph contains genes for proteins that are made on the 80S ribosomes in the nucleomorph compartment and transported across the inner two membranes into the chloroplast. Finally, there are proteins now encoded by the primary nucleus that must be translocated across both sets of double membranes from the primary cytoplasm into the chloroplast!

Is Malaria Really a Plant?

Malaria is a disease that affects many millions of people world wide and is responsible for two or three million deaths each year, mostly in Africa. Malaria is caused by a



PRIMARY ENDOSYMBIOSIS

SECONDARY ENDOSYMBIOSIS

Mitochondrion
Nucleus
Ancestral host cell
Cytoplasmic membrane
Cyanobacterium
Photosynthetic eukaryote (alga)
Chloroplast with double membrane
Mitochondrion
Nucleus
Ancestral host cell
Photosynthetic eukaryote (alga)
Nucleomorph (degenerate nucleus)
Chloroplast with four membranes
Chloroplast with four membranes

FIGURE 19.04 *Primary versus Secondary Endosymbiosis*

Primary endosymbiosis yields organelles with two membranes. In this example, the original independent cyanobacterium has a cytoplasmic membrane, which is retained, and an outer membrane, which is lost during symbiosis. When the two cells associate, the host cell cytoplasmic membrane surrounds the cyanobacterium, which is therefore left surrounded by two membranes. In contrast to primary endosymbiosis, secondary endosymbiosis occurs when an ancestral host cell engulfs a photosynthetic eukaryotic alga. The alga already has a chloroplast with two membranes as well as a nucleus and other organelles. Since the host cell only needs the energy from the chloroplast, the other captured organelles degenerate and eventually disappear. However, the membranes often remain and the chloroplast is left with four membranes, rather than two.

single-celled eukaryote, known as ***Plasmodium***. The malaria parasite and other related single-celled eukaryotes are members of the phylum **Apicomplexa**. Although these single-celled eukaryotes live inside humans and mosquitoes, far from the sunlight, they possess plastids as well as mitochondria. These plastids are degenerate, non-photosynthetic chloroplasts that contain a circular genome. In *Plasmodium* this plastid DNA is 35 kb and encodes rRNA, tRNA and a few proteins, mostly involved in translation (Fig. 19.05).

The malarial parasite contains degenerate non-photosynthetic chloroplasts that are involved in synthesizing fatty acids.

Apicomplexa Phylum of parasitic single-celled eukaryotes that contain both mitochondria and degenerate non-photosynthetic chloroplasts
Plasmodium The malaria parasite, a protozoan belonging to the Apicomplexa

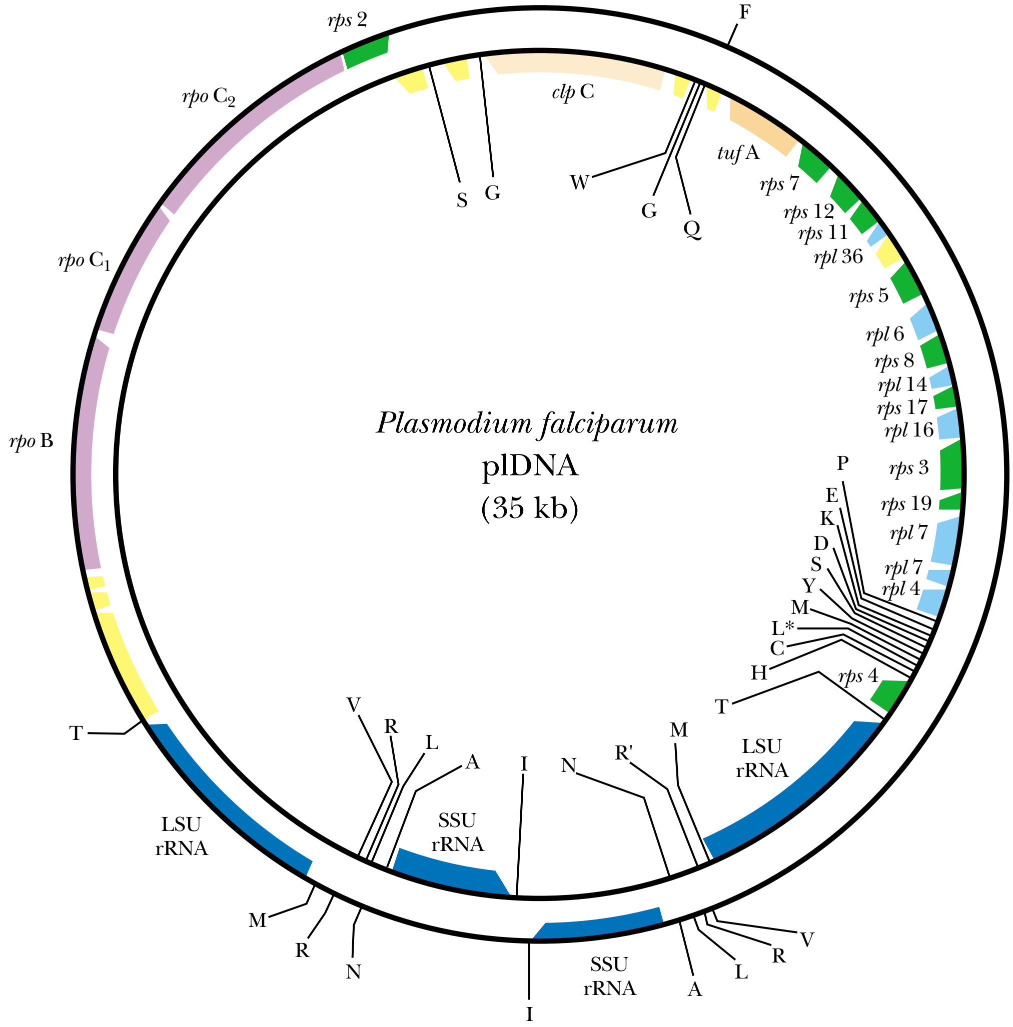


FIGURE 19.05 *Plastid Genome of* Plasmodium

The circular genome of *Plasmodium* has genes for rRNA, tRNA and protein synthesis. The tRNA genes are denoted by the single-letter amino acid code, for example, S for the tRNA for serine.

The malarial plastid or "**apicoplast**" is thought to derive from secondary endosymbiosis. The ancestor of the Apicomplexa appears to have swallowed a single-celled eukaryotic alga that already possessed a chloroplast. The algal nucleus has been completely lost, but the plastid was kept and is still surrounded by four membranes. Certain sequence similarities suggest the malarial apicoplast may be most closely related to the chloroplast of red algae.

The apicoplast is essential for the survival of *Plasmodium* and other Apicomplexans. Since the apicoplast does not produce chlorophyll and convert light into energy, why is this organelle maintained? It turns out that the apicoplast plays a vital role in lipid metabolism. Several enzymes of fatty acid synthesis are encoded in the nucleus but translocated into the apicoplast where fatty acid synthesis occurs. As a result, certain herbicides that prevent fatty acid synthesis in the chloroplasts of green plants are effective against *Plasmodium* and other pathogenic apicomplexans such as *Toxoplasma* and *Cryptosporidium*. For example, the herbicide clodinafop targets the acetyl-CoA carboxylase of chloroplasts and triclosan inhibits the enoyl ACP reductase of plants and bacteria but has no effect on fatty acid synthesis in animals or fungi. In addition, the herbicide fosmidomycin inhibits the isoprenoid pathway of plants and bacteria, which differs from that of animals. Fosmidomycin inhibits growth of *Plasmodium* and cures malaria-infected mice. *Plasmodium* and other apicomplexans are also inhibited by chloramphenicol, rifamycin, macrolides and quinolones, all of which are

apicoplast Degenerate non-photosynthetic chloroplast found in members of the Apicomplexa, including the malaria parasite regarded as anti-prokaryotic antibiotics. These antibiotics are also thought to act via the apicoplast.

Symbiosis: Parasitism versus Mutualism

Symbiotic relationships can be view as beneficial to both organisms or harmful to one of the organisms. **Parasitism** involves one organism living completely at the expense of the other, whereas **mutualism** involves mutual benefit to both partners. Parasitism has a gradation from **pathogens** that seriously incapacitate or kill their hosts to microorganisms that consume so little of the host's resources that their presence is scarcely significant. Mutualism also shows a gradation, from relationships in which both organisms can survive without each other to relationships that are completely co-dependent.

Many varied examples of bacteria inhabiting other organisms are known.

Many different examples of mutualistic and parasitic symbiotic relationships exist. Mutualistic associations were perhaps responsible for the origin of mitochondria and chloroplasts from free-living prokaryotes. *Rhizobium* invades the plant roots of the pea family where it builds a nodule or small bump to inhabit. *Rhizobium* receives nutrients from the plant to survive, and in exchange, it converts nitrogen from the soil into a form the plant can readily use. Consequently, this is an example of a mutualistic symbiosis in which both partners benefit. We have already discussed *Agrobacterium* (Ch. 16), which infects plant cells and transfers T-DNA into the plant cell nucleus. Much like *Rhizobium*, the *Agrobacterium* induces the plant to make a mass of cells for the bacteria to inhabit. Unlike *Rhizobium*, the *Agrobacterium* does not provide any service for the plant, and simply uses the plant to provide food. This is a simple example of parasitism.

We shall consider two more examples where bacteria live inside eukaryotic cells to the benefit of both partners. In the case of *Paramecium* and *Caedibacter*, the bacteria are only slightly modified and some can still live independently. In the case of *Buchnera* and the insects they inhabit, the bacteria are so modified that they may almost be regarded as a specialized biosynthetic organelle.

Bacterial Endosymbionts of Killer Paramecium

Killer *Paramecium* contains symbiotic bacteria that make the toxin for killing sensitive strains of *Paramecium.*

Paramecium is a protozoan that cruises around in freshwater and feeds by swallowing bacteria whole and digesting them. This lifestyle has led to the development of killer strains of *Paramecium* that use bacteria as a biological weapon to kill other, sensitive strains of *Paramecium*. Such **killer *Paramecium*** produce a toxin that kills all sensitive *Paramecium* that get too close. Killers are immune to their own toxin, but different brands of killer *Paramecium* exist that kill each other. Killer *Paramecium* contain **kappa particles** in their cytoplasm (Fig. 19.06). These are actually endosymbiotic bacteria (*Caedibacter*) that grow and divide inside the larger, eukaryotic cell. The kappa particles are inherited by cytoplasmic or maternal inheritance, like mitochondria. Unlike mitochondria, kappa particles are still obviously bacteria. Some strains of *Caedibacter* can still grow and divide outside the Paramecium, although most types are obligate symbionts.

Besides harboring *Caedibacter*, the ability to kill also depends on a gene found in the nucleus of the *Paramecium*. The gene has two alternate forms: K and k. *Paramecium* is diploid, so any individual has two copies of each gene and may be KK, Kk or

kappa particle Endosymbiotic bacteria (*Caedibacter*) that grow and divide inside killer *Paramecium*
killer *Paramecium* *Paramecium* that contains kappa particles in the cytoplasm, which it uses to kill other strains of *Paramecium*
mutualism Form of symbiosis where both partners benefit
Paramecium A type of free-living protozoan that feeds on bacteria
parasitism Form of symbiosis where one organism lives at the expense of the other
pathogen Parasite that seriously incapacitates or kills its host

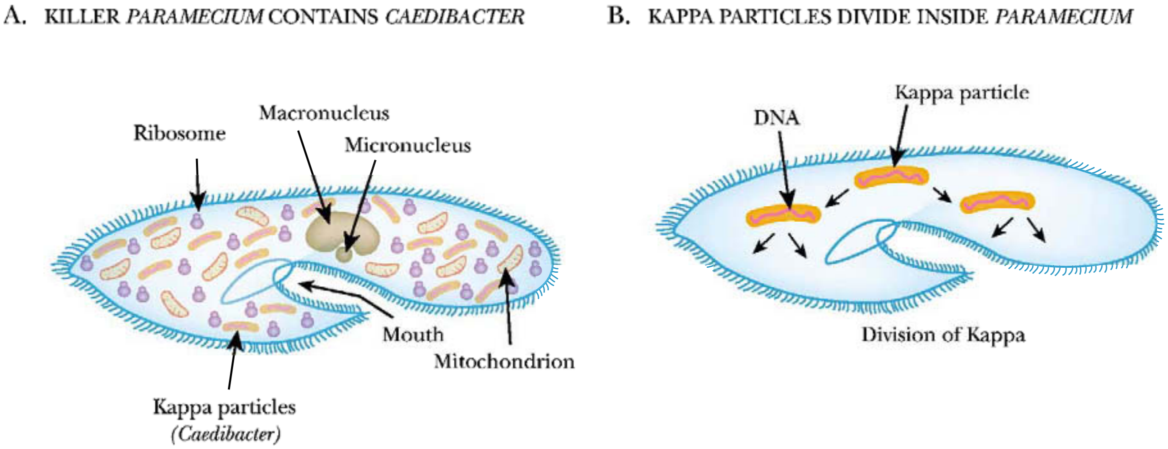


FIGURE 19.06 ***Killer* Paramecium *contains* Caedibacter**

(A) The *Caedibacter* or kappa particles are found in the cytoplasm of the *Paramecium*. (B) Kappa particles are symbiotic in many strains of *Paramecium*, yet they have their own DNA and divide like typical bacteria.

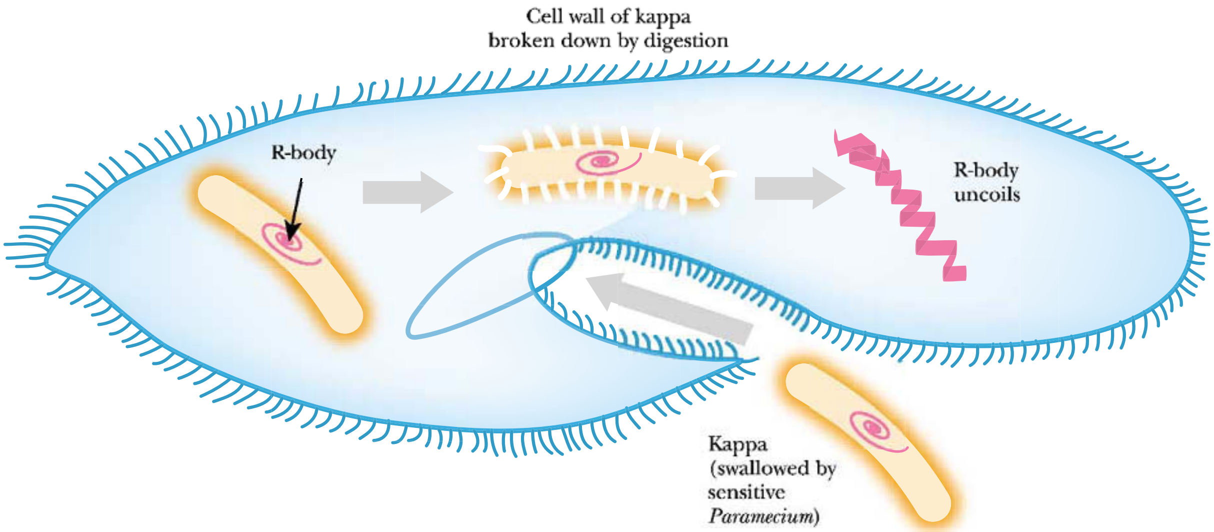


FIGURE 19.07 ***Sensitive* Paramecium *is Killed by R-Bodies***

Kappa particles contain a protein toxin coiled into a crystal known as an R-body. When the membrane of the kappa particle is digested, the toxin is free to uncoil and releases the toxin protein. Once released inside its victim, the toxin kills the *Paramecium*.

kk. Since the K allele is dominant, possession of one or two K alleles (i.e., KK or Kk) allows the kappa particles to grow and divide inside the *Paramecium*. However, kk individuals lose the kappa particles. Thus a killer must have both the cytoplasmic kappa particle and at least one copy of the nuclear K allele.

All *Paramecium* cells without kappa particles are sensitive, whether they are KK, Kk or kk. Killing occurs when a few kappa particles are liberated from killer strains into the culture medium. The sensitive *Paramecium* strains swallow the kappa particles, believing them to be harmless digestible bacteria. Inside the kappa particle are toxic proteins coiled into crystals known as **R-bodies**. When the kappa particle is digested by a sensitive *Paramecium*, the R-body is liberated, uncoils, releases the toxin and kills the *Paramecium* (Fig. 19.07).

R-bodies Toxic proteins that form crystals inside the kappa particles of killer *Paramecium*

The R-body protein is actually encoded by a gene belonging to a bacterial plasmid or a defective bacterial virus, not by the *Caedibacter* chromosome. So a toxin made by a virus infecting the symbiotic *Caedibacter* has been domesticated and diverted to the purpose of killing sensitive *Paramecium*.

Is *Buchnera* an Organelle or a Bacterium?

Symbiotic bacteria living inside insects provide their hosts with essential amino acids.

Many insects contain symbiotic bacteria. Aphids live by sucking the sap of plants. Although it contains plenty of carbohydrate, this diet is deficient in the so-called "**essential amino acids**", i.e. those that animals are unable to synthesize for themselves (Arg, Val, Ile, Leu, Phe, Trp, Thr, Met, Lys, His). This deficiency is made up by symbionts of the genus ***Buchnera***. These are gram-negative bacteria, closely related to *E. coli* but have genomes only one seventh the size. The symbionts live inside specialized giant insect cells known as bacteriocytes and are transmitted maternally, in the eggs, from one generation of aphids to the next. Neither partner is capable of living independently.

The *Buchnera* genome has been sequenced and consists of a chromosome (641 kb) plus two small plasmids. The pLeu plasmid carries genes for leucine synthesis and the pTrp plasmid for tryptophan. Each *Buchnera* cell contains approximately 100 copies of its genome. *Buchnera* retains just fewer than 600 of the 4,000 genes typical of its free-living relatives. There are no insertion sequences, repetitive elements or phage-derived sequences in the *Buchnera* genome.

Genes for the biosynthesis of amino acids that the insect can make are missing from the *Buchnera* genome. Conversely, genes for all the "essential" amino acids are still present. Thus the insect and the endosymbiont each make a different selection of amino acids and exchange them. Note that some "essential" amino acids are made from other amino acids. For example, the insect provides aspartate, which *Buchnera* uses as a precursor to synthesize threonine and lysine, some of which are returned to the insect cells (Fig. 19.08).

In exchange for providing the "essential" amino acids, *Buchnera* receives nutrients and a protected, constant environment. Consequently many regulatory genes for adjusting to environmental changes are missing. *Buchnera* also lacks genes for phospholipid synthesis, although it does have a cell membrane. It is not known whether *Buchnera* imports pre-made phospholipids from the insect or whether it imports the enzymes for their synthesis. Overall, *Buchnera* clearly exhibits its bacterial origin yet is no longer independent and may be regarded as a biosynthetic organelle devoted to the biosynthesis of amino acids and other metabolites that are missing from the specialized diet of the host insect.

Ciliates have Two Types of Nucleus

Many single-celled eukaryotes have strangely arranged genomes, and in addition, may carry out peculiar forms of DNA rearrangement or RNA processing. Most of these primitive eukaryotes are still poorly characterized at the molecular level. For practical reasons, the best-investigated protozoans are mostly disease-causing parasites, and it is possible that many of their genetic peculiarities are due to the relaxed constraints of a parasitic lifestyle. However, the **ciliates** are free-living protozoans, so we will consider their peculiarities first.

Protozoans of the ciliate class are unique in possessing two different kinds of nucleus in the same cell.

The ciliates are a large and widespread group of protozoans that are named after their locomotory organelles, the cilia that cover the cell surface. Perhaps the best known of the ciliates are *Paramecium*, *Tetrahymena* and *Euplotes*. These single-celled

Buchnera Genus of gram-negative bacterial symbionts found in insects that supply their host insect with essential amino acids
ciliates Group of free-living protozoans that move by means of cilia attached to the cell surface
essential amino acids Those amino acids that animals are unable to synthesize for themselves (Arg, Val, Ile, Leu, Phe, Trp, Thr, Met, Lys, His)

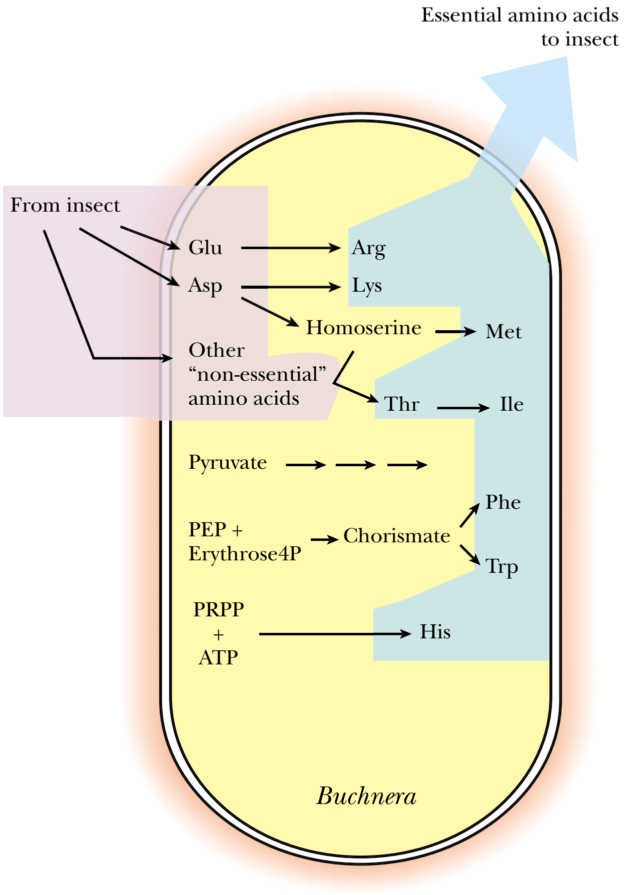


FIGURE 19.08 ***Amino Acid Synthesis in Aphids and Buchnera***

Buchnera live inside aphids and manufacture the amino acids (blue) that aphids cannot manufacture themselves. In return, the aphid provides amino acids that it does manufacture (purple), thus establishing a mutualistic symbiotic relationship.

eukaryotes all contain two types of nucleus, a **macronucleus** and a **micronucleus** within the same cell. The numbers of these nuclei vary from one ciliate to another. *Paramecium* has a single macronucleus and two or more micronuclei, depending on the species. These micronuclei are all diploid and identical because they were derived by mitotic division of the single original micronucleus.

The micronucleus may be regarded as the germline and is used for the sexual exchange of DNA. The genes in the micronucleus are silent and not expressed, whereas, the genes of the macronucleus are transcribed to give mRNA. The macronucleus is thus the somatic component and is polyploid. It may contain from 50 to 1000 copies of the genome (depending on the species) and the copy number may vary depending on environmental conditions. When two cells mate, they exchange haploid micronuclei generated by meiosis. After mating, the haploid micronuclei fuse to give a diploid micronucleus, which then divides. Finally, one of the micronuclei is converted into a new macronucleus and the old macronucleus is disassembled (Fig. 19.09).

Genes capable of being expressed are assembled by deleting internal sequences during formation of the macronucleus.

Micronucleus DNA must be processed in order for the micronucleus to convert into the macronucleus. The DNA of the micronucleus contains many extra sequences that interrupt the germline genes, called **internal eliminated segments (IESs)**. Most IES are less than 100 bp although they may range from 5 to 900 bp in length, and they are very AT rich (75–100% AT). After removal of the IESs the remaining **macronucleus-destined segments** (MDSs) are spliced to form uninterrupted genes that can be expressed (Fig. 19.10). There may be from 5 to 50 IESs interrupting a single gene. Each

internal eliminated segment (IES) Extra sequences in the DNA of the ciliate micronucleus that are eliminated during conversion of a micronucleus to a macronucleus
macronucleus Large somatic nucleus of ciliates that contains multiple copies of genes that are expressed
macronucleus-destined segment (MDS) Segments of DNA that remain during conversion of a ciliate micronucleus to macronucleus and are spliced to form uninterrupted genes that can be expressed
micronucleus Small germline nucleus of ciliates whose genes are not expressed

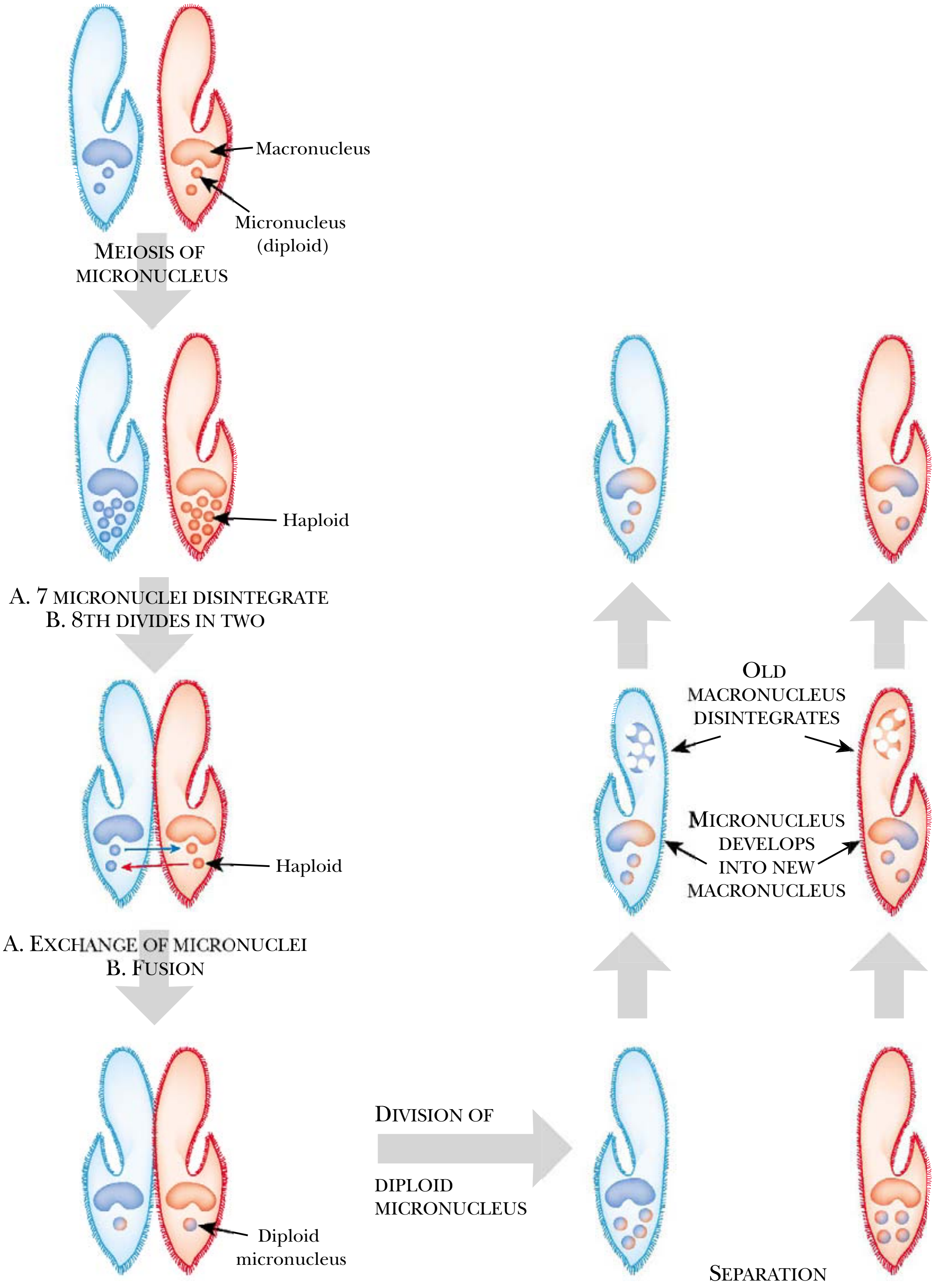


FIGURE 19.09 ***Micronucleus and Macronucleus of* Paramecium**

The micronuclei of *Paramecium* are considered the germline. A typical *Paramecium* contains two micronuclei which go through meiosis when the cells start mating. Meiosis produces eight haploid micronuclei, seven of which disintegrate. The remaining haploid micronucleus performs one mitotic division to form two haploid micronuclei in each cell. Reciprocal exchange of the micronuclei occurs through a bridge between the two mating partners. Each cell then has one original micronuclei and one micronuclei from its mate, which fuse to give a diploid micronucleus. After the mating pair separate, the diploid micronucleus divides by mitosis. This triggers the original macronucleus to disintegrate. One of the mitotic descendents of the micronucleus then develops into a new macronucleus.

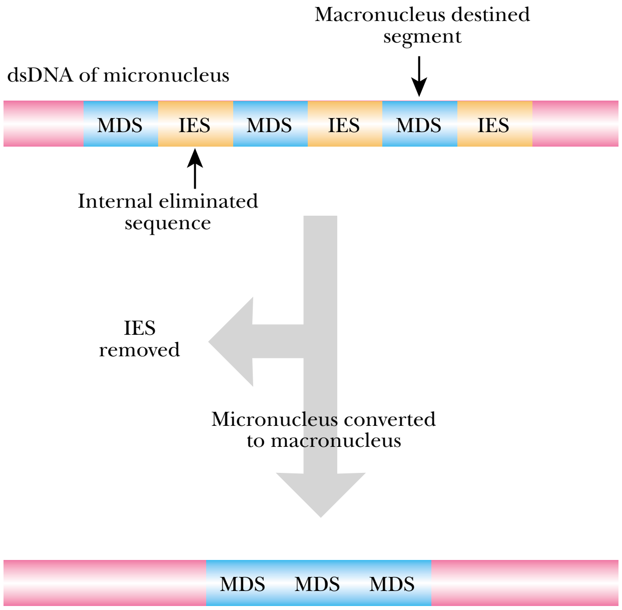

Macronucleus destined segment
dsDNA of micronucleus
MDS
IES
MDS
IES
MDS
IES
Internal eliminated sequence
IES removed
Micronucleus converted to macronucleus
MDS
MDS
MDS

FIGURE 19.10 ***Internal Eliminated Segments of Ciliate DNA***

Internal eliminated segments must be removed from the micronucleus DNA before the micronucleus converts into a macronucleus. Once these segments are spliced from the DNA, a fully functional gene is reconstructed.

IES is flanked by repeats of 2 to 20 bp that lie within the MDS sequences on either side. After splicing, the IES and one of the two repeats are removed. The number of IES sequences varies among the different ciliates but overall, there may be as many as 100,000 IESs per haploid genome, all of which must be removed to allow expression of the genes they interrupt. Note that the IES sequences are not introns because they are removed from the DNA, not from RNA. In addition, IESs are found within both coding and non-coding DNA. As discussed before, introns are only found in coding regions of DNA. Ciliates do have genuine introns, although these relatively rare and are found in only around 20% of the genes.

After joining of the MDSs, the DNA carrying the reconstituted genes is cut into shorter segments and most of the intergenic DNA is eliminated. These segments are then duplicated to give many copies. Thus the macronucleus contains multiple copies of many chromosomal fragments, rather than full-length copies of the original chromosomes. In *Paramecium* and *Tetrahymena*, the DNA of the macronucleus is processed into “mini-chromosomes” of several hundred kilobases. In *Oxytricha* and its relatives (the hypotrichous ciliates), the DNA is processed into molecules of 400–15,000 bp each carrying only one gene. Telomere repeat sequences (5′-CCCCAA-3′ in *Paramecium* and 5′-CCCCAAAA-3′ in *Oxytricha*) are added to the ends of these fragments to promote stability during replication.

At least in some ciliates, the germline DNA is not only interrupted, but the order of the MDS sequences is scrambled relative to the coding sequence of the gene. In *Oxytricha* and its relatives some 25–30% of genes are scrambled in the germline (Fig. 19.11). When the IES sequences are removed, the germline MDSs are unscrambled and reorganized into the correct order.

Trypanosomes Vary Surface Proteins to Outwit the Immune System

Trypanosomes avoid being recognized by the immune system by constantly changing the proteins exposed on their cell surface.

Parasitic eukaryotes live in hostile environments and must elude the defenses of their host. Many adaptations have occurred in parasitic eukaryotes to overcome host defense systems and allow the parasite to thrive in these one-sided relationships. Many parasitic microorganisms attempt to elude the immune system of their host by changing their surface proteins (Fig. 19.12). The idea behind this strategy is straightforward. The immune memory cells recognize the proteins on the surface of the original generation of invading germs. However, if each successive wave of invaders changes its

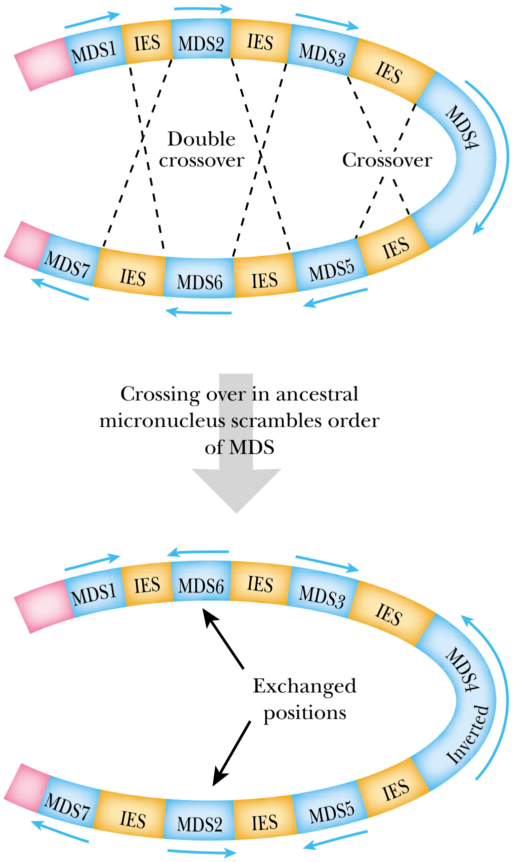


FIGURE 19.11 ***Scrambled Germline Genes of Oxytricha***

The original *Oxytricha* ancestor had all its MDS segments in the same order as the final gene coding sequence. Multiple crossover events during evolution have scrambled the order of the MDS. Modern *Oxytricha* must remove the IES and then unscramble the MDS in order for the gene to be expressed properly.

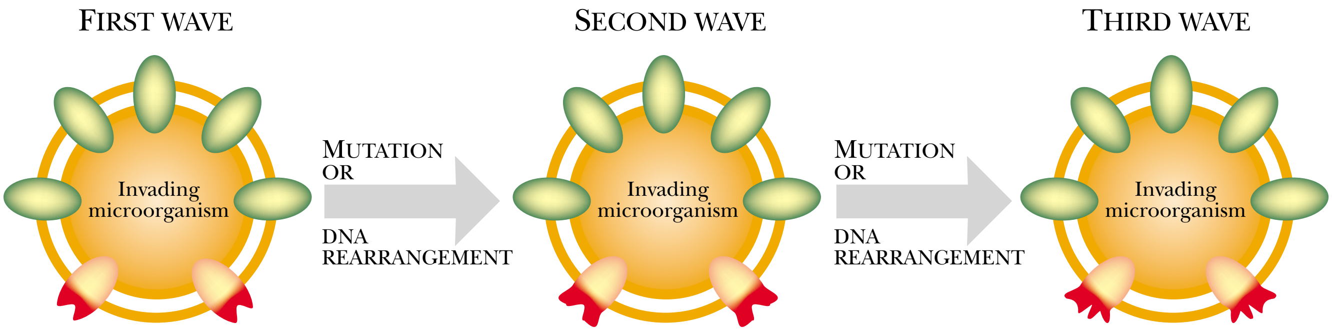


FIGURE 19.12 ***Eluding the Immune System by Changing Surface Proteins***

Invading microorganisms are killed because the host immune system recognizes the surface proteins expressed by the microorganism. To trick the immune system, the microorganism will change its surface proteins by mutating or rearranging the DNA of the respective gene. Each time the surface marker changes, a new set of immune cells must be manufactured to squelch the population of microorganisms.

surface proteins, they will not be recognized. One sophisticated way to achieve this is for the disease-causing microorganism to shuffle its surface proteins by genetic rearrangements.

Although bacteria or viruses cause most diseases of temperate climates, many tropical or subtropical diseases are due to single-celled eukaryotes. Perhaps the best-known diseases are amoebic dysentery, sleeping sickness and malaria. Sleeping sickness in Africa, and the rare Chagas' disease of South America, are both caused by **trypanosomes**. These microorganisms are about 20 microns long (about 20 times longer

trypanosomes Group of parasitic single-celled eukaryotes that cause sleeping sickness and other tropical diseases

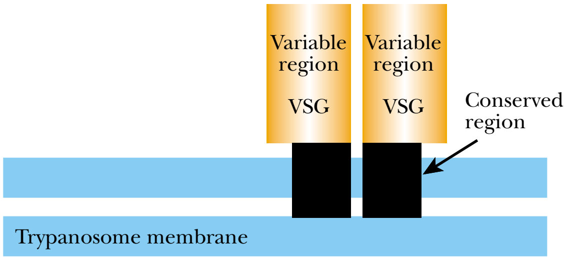


FIGURE 19.13 ***Structure and Location of VSG Protein***

The VSG protein is expressed on the cell surface of the trypanosome. The protein has two domains, the conserved region anchors the protein in the membrane and does not change. The second domain or variable region extends to the outside of the cell and changes shape to evade the immune system.

than average bacteria) and swim around in the blood by means of a flagellum. Insects harbor these parasitic eukaryotes in their saliva, and when the insect bites a person or animal, the eukaryotic parasite invades the bloodstream of the new host. For example, tsetse flies carry the trypanosomes that cause sleeping sickness, whereas malarial parasites are carried by mosquitoes.

While growing and dividing in the insect gut, the trypanosomes are covered with a layer of a protein called procyclin that protects them from digestion by the insect. The trypanosomes then move to the salivary glands where they stop dividing and wait for the insect to bite someone. While waiting, they convert their surface layer to the **variant surface glycoprotein (VSG)**, designed to protect against animal immune systems. The VSG protein has a variable region that is displayed on the trypanosome surface and a conserved portion that anchors it to the membrane. It is found as a dimer, as shown in Figure 19.13.

Trypanosomes cycle between insects and mammals and change their surface proteins when they shift hosts.

After transfer to a human, the trypanosomes grow and divide in the blood until the immune system kills most of them. However, a few of the trypanosomes switch their VSG shape and escape recognition by the immune system. Eventually, the immune system learns about the new surface protein and kills off most of the second wave of trypanosomes. Meanwhile, some of the invaders have switched their VSG type again. This continues and the infection therefore goes in waves, each spreading the invaders further inside each human victim. The immune system never catches up with the constantly changing outer layer of the trypanosome, and the normal result is death of the victim. When tsetse flies suck blood from humans or animals with the disease, they become re-infected. And so the cycle continues (Fig. 19.14).

To understand how the VSG protein is constantly altered, we must first discuss the genomic structure of the trypanosomes. Both the nuclear and mitochondrial genomes of trypanosomes are divided up in a peculiar manner. In addition, trypanosomes indulge in the trans-splicing of many genes at the RNA level as well as RNA-editing as already discussed in Chapter 12.

Each trypanosome cell contains one giant mitochondrion, the so-called "**kinetoplast**". This contains about 50 copies of a large circular DNA molecule that ranges from 20 to 80 kbp depending on the species. These "maxi-circles" encode the normal mitochondrial genes. In addition there are approximately 10,000 mini-circles that encode only the guide RNAs used in splicing (see Ch12). The mini-circles range from 1.5 to 10 kbp and are often catenated (i.e. interlocked).

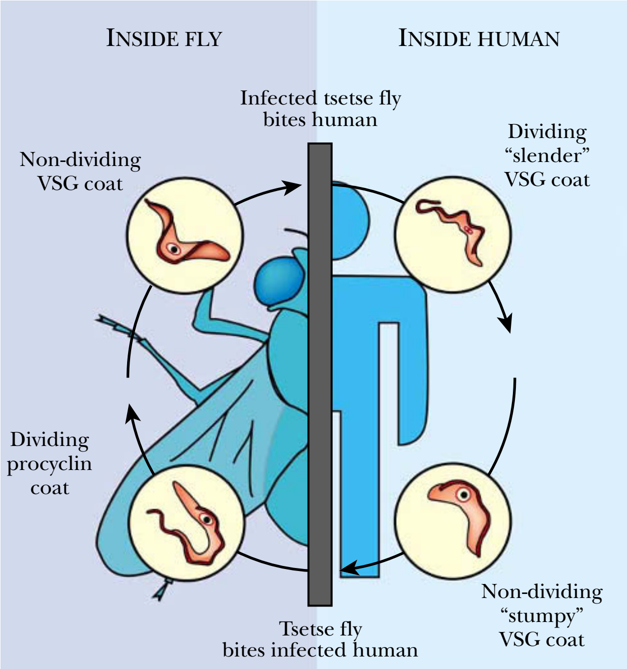
The trypanosome nucleus contains a mixture of normal sized and miniature chromosomes.

The nuclear genome is divided into 11 pairs of large "normal" or "megabase" chromosomes plus about 100 **mini-chromosomes** of 50–150 kbp. The only protein-coding genes found on the mini-chromosomes are silent copies of the *VSG* gene, located close to one or both ends (Fig. 19.15). The rest mostly comprises a

kinetoplast Single giant mitochondrion found inside the cells of protozoans such as trypanosomes
minichromosomes Miniature chromosomes of 50–150 kbp found in trypanosomes that carry silent copies of the *VSG* gene
variant surface glycoprotein (VSG) Glycoprotein found on surface of trypanosomes that is encoded by multiple gene copies and varied to avoid recognition by the animal immune system

FIGURE 19.14
Trypanosome Life Cycle

Trypanosomes alternate between two hosts, humans or other mammals and the tsetse fly. While residing in the fly, trypanosomes acquire a procyclin coat to protect against the fly's digestive enzymes. After moving to the salivary glands of the fly, trypanosomes start to express VSG proteins on the cell surface. When the tsetse fly bites a human, the trypanosome enters the mammalian bloodstream and starts to divide. Changes in VSG surface expression allow a few of the trypanosomes to evade the immune system. If a tsetse fly bites an infected human, the cycle starts all over again.


INSIDE FLY
INSIDE HUMAN
Infected tsetse fly bites human
Non-dividing VSG coat
Dividing "slender" VSG coat
Dividing procyclin coat
Non-dividing "stumpy" VSG coat
Tsetse fly bites infected human

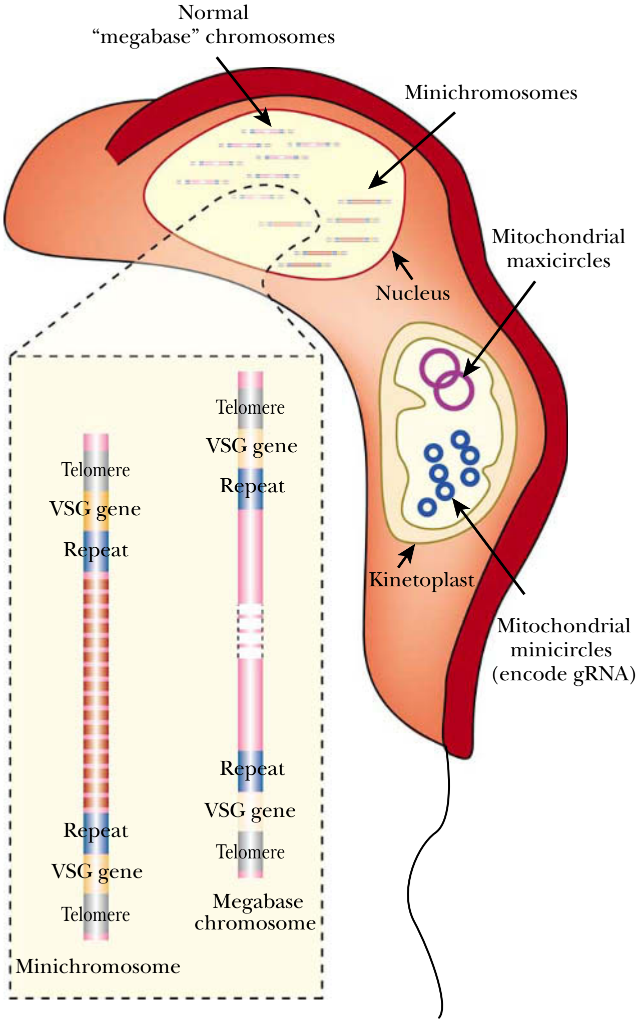


FIGURE 19.15 ***Genome Components of the Trypanosome***

Trypanosome genomes consist of four different genetic elements. The mitochondria or kinetoplast has maxi-circles and mini-circles. The nucleus has eleven pairs of megabase chromosomes, and about 100 mini-chromosomes. The VSG genes are located at the ends of both the megabase and mini-chromosomes.

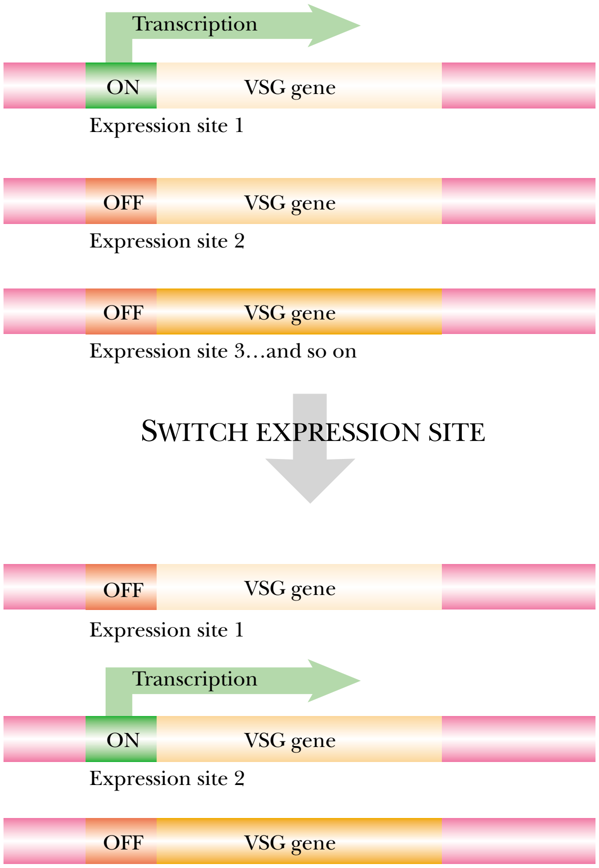


FIGURE 19.16 ***Variation of VSG by Switching Expression Sites***

VSG genes are found at the ends of the megabase chromosome. Only one of these genes is active at any one time, and only that VSG protein is manufactured. Every so often, the trypanosome turns off this expression site and turns on a different one. The new VSG gene produces a VSG protein with a different shape that can evade the immune system.

177 bp repeated sequence of unknown function that makes up 90% of the length of the mini-chromosomes.

Despite being parasites, trypanosomes have about twice as many genes as free-living yeast cells. These genes are carried on the large chromosomes, which are each several megabases in length and (hence the term megabase chromosomes). Confusingly, the two members of a pair of homologous chromosomes often vary in size due to the frequent crossing over between blocks of repeated sequences that occurs during mating.

Trypanosomes possess many copies of the gene for the surface protein expressed inside mammals. Only one copy is expressed at any given time.

The secret to trypanosome success is that it possesses over 1,000 slightly different copies of the gene for the variant surface glycoprotein. At any given time, only one of these genes is expressed. There are three mechanisms for switching: **expression site** activation, chromosome end-swapping, and gene conversion. Among the 1,000 *VSG* genes, there are only about 20 that can actually be expressed. These privileged genes are in special expression sites located just inside the telomeres at the ends of the megabase chromosomes. Only one expression site is active at any given time, and the others are in standby mode. Every so often, the trypanosome switches from one expression site to another, which results in a different VSG being expressed (Fig. 19.16).

The expression site promoters are unusual in being recognized by RNA polymerase I, which normally only transcribes ribosomal RNA genes. The mechanism that ensures that only one of the expression sites is active at any given time is still obscure. However, an unusual modified base is found in regions of trypanosome DNA that are silenced. This is referred to as **J-base (β-D-glucosyl hydroxymethyluracil)**, and is made by modification of thymine. Inactive expression sites have multiple J-bases whereas the active expression site lacks J-base. What controls removal of J-base when an expression site is activated is unknown.

β-D-glucosyl hydroxymethyluracil See J-base
expression site Special location on chromosome where the chosen copy of a gene present in multiple copies may be expressed
J-base β-D-glucosyl hydroxymethyluracil, an unusual base found in regions of trypanosome DNA that are silenced and which is made by modification of thymine

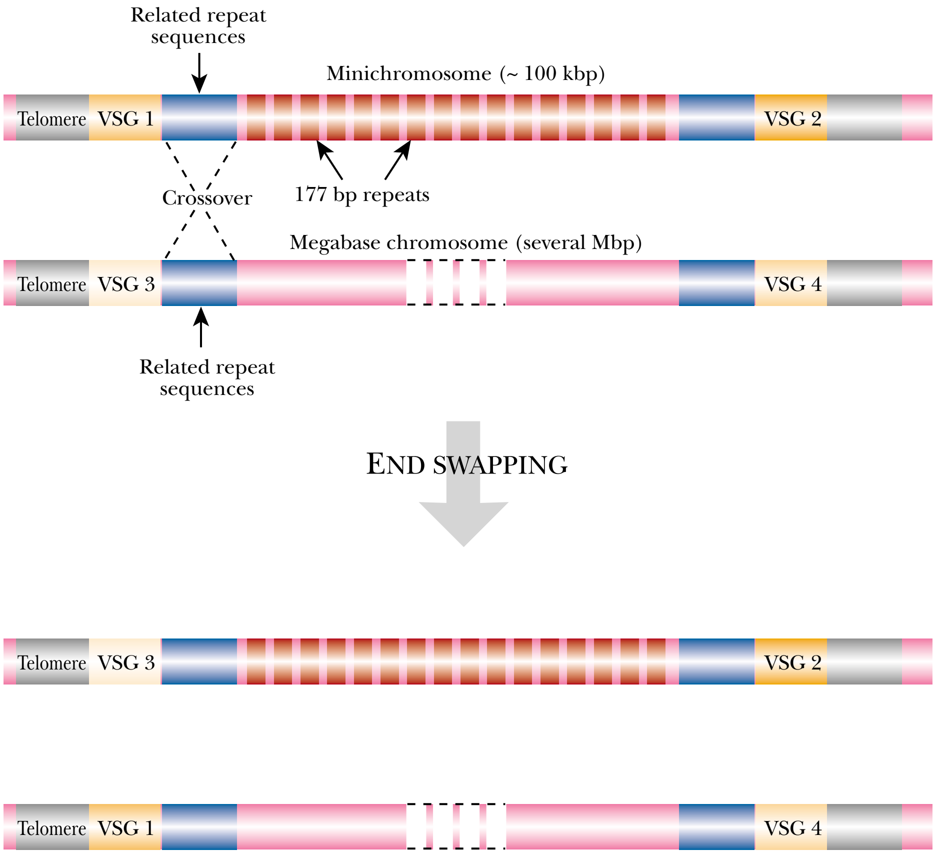


FIGURE 19.17 ***VSG Genes are Transferred between Mini- and Megabase Chromosomes***

Minichromosomes of trypanosomes only contain silent copies of the *VSG* genes that are located at close to their ends just behind the telomeres. These silent copies can be exchanged with the *VSG* genes located on the megabase chromosomes through crossover events between repeated DNA sequences (blue).

Another method to switch expression is called end swapping. The megabase chromosomes and mini-chromosomes can exchange ends by recombination between blocks of repeated sequences just to the inside of the *VSG* genes (Fig. 19.17). This allows about 200 alternative *VSG* genes to be exchanged into the telomeric expression sites of the normal megabase chromosomes.

Not all *VSG* genes are found at the ends of the chromosomes, some are found as tandem arrays scattered throughout the megabase chromosomes. These are not accessible by chromosomal end-swapping and may only be used by **gene conversion**. (See Ch. 14 for the mechanism of gene conversion). All unexpressed copies of the *VSG* gene may be used to supply sequences for splicing into the *VSG* genes in the expression sites. Usually, the complete variable region of the *VSG* gene in the expression site is replaced with the complete variable region from one of the 1,000 extra copies (Fig. 19.18A). The constant region stays unchanged, as its name indicates. Later in infection, segments of various sizes from the spare *VSG* genes are used for replacement; anywhere from just a few base pairs to the whole gene may be used (Fig. 19.18B). Furthermore, just as with the genes encoding mammalian antibodies, point mutations occur in the *VSG* genes at higher than normal frequency. However, in the case of the *VSG* genes, the mutations occur during the segment-swapping process, not afterwards.

More variations result from reshuffling segments of the genes for surface proteins.

Mating Type Determination in Yeast

The single-celled eukaryote *Saccharomyces cerevisiae*, is widely used as a model organism in molecular biology and its major characteristics have been summarized in Chapter 2. Here we will consider its sex determination system. Yeast can live either in the haploid or the diploid state. Both types of cell are structurally and metabolically

gene conversion Recombination and repair of DNA during meiosis that leads to replacement of one allele by another. This may result in a non-Mendelian ratio among the progeny of a genetic cross

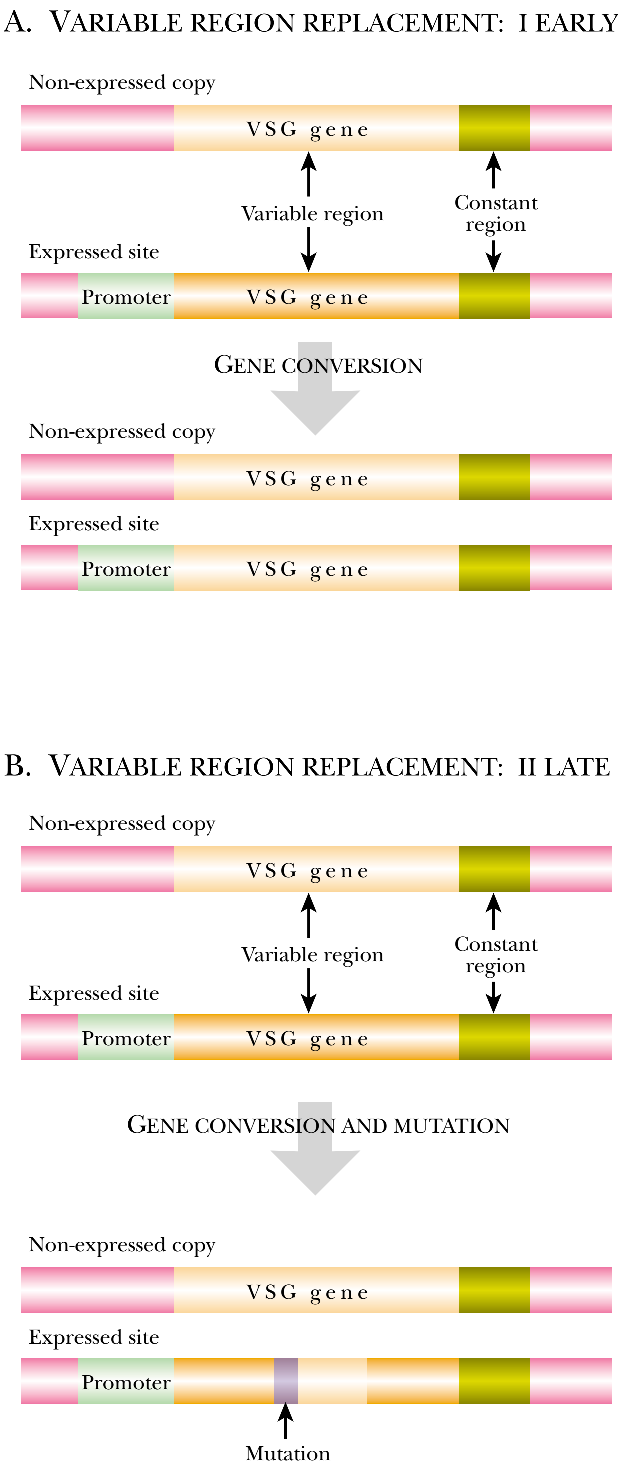


FIGURE 19.18 ***Variation of VSG by Gene Conversion***

(A) Early in infection, the entire VSG gene undergoes gene conversion. The gene at the expressed site (dark orange) is entirely replaced by a non-expressed copy of a VSG gene (light orange). The two VSG genes encode the same protein, but the shape of the variable domain is altered slightly between the two. (B) Later in infection, the expressed gene undergoes mutations and partial gene conversion events. These events amplify the amount of genetic diversity available for this one protein.

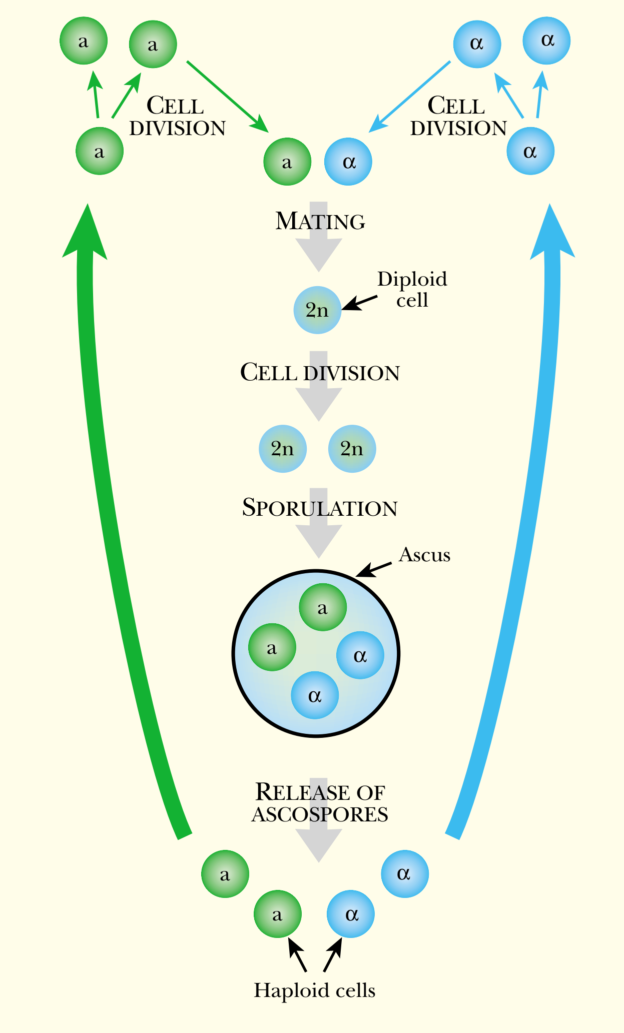


FIGURE 19.19 ***Alternating Haploid and Diploid Phases of Yeast***

Haploid cells come in two mating types, **a** and α (top of figure). When an **a** haploid cell and an α yeast come in contact, they fuse to form a diploid yeast cell. During specific conditions, the diploid yeast cell sporulates and the nucleus undergoes meiosis, forming four haploid cells surrounded by an ascus. The ascus ruptures to release four haploid cells, starting the cycle over again.

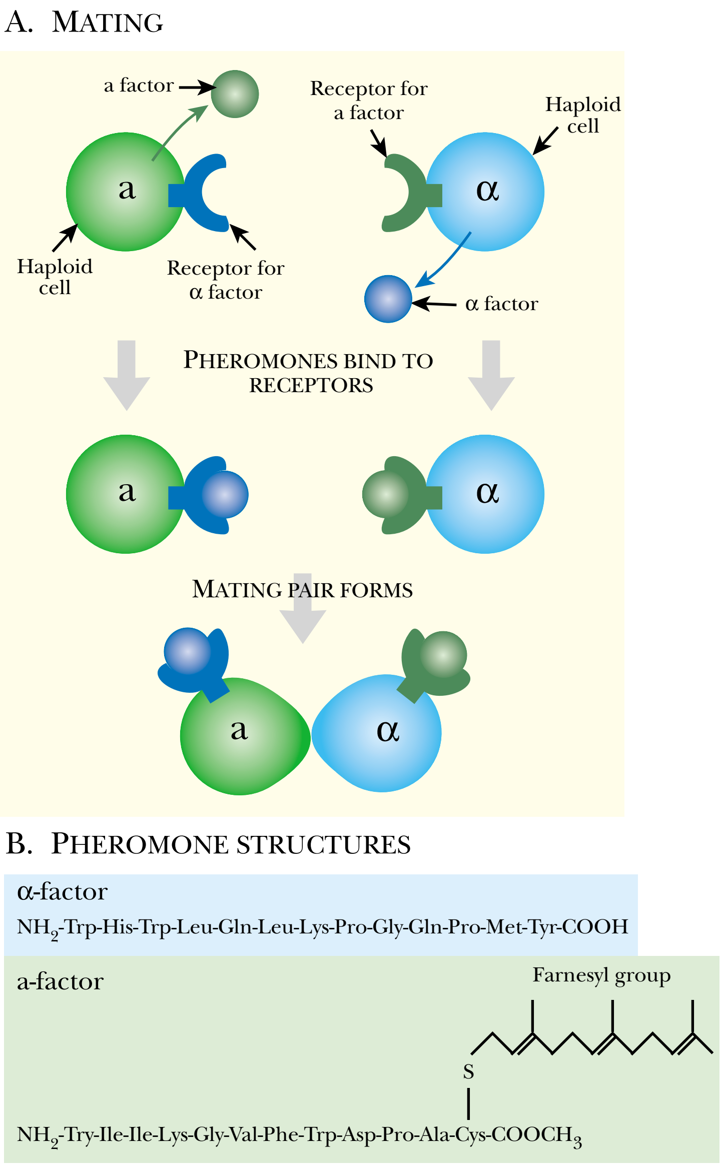
Yeasts alternate between haploid and diploid cells. Formation of diploids involves fusion between haploid cells of different mating types.

similar and grow by budding. In yeast and some other primitive eukaryotes, the cells of the haploid phase may continue to divide or may act as gametes, depending on the circumstances. Although haploid yeast cells are all structurally identical they belong to one of two **mating types**, known as **a** or α. During mating, two haploid cells of different mating types fuse to give a diploid cell (Fig. 19.19). Such diploid cells grow and divide until poor environmental conditions trigger spore formation. Then the diploid cell is converted into an **ascus**, inside which meosis occurs. Four haploid spores (**ascospores**) are produced inside each ascus. After release from the ascus, the ascospores germinate so re-establishing the haploid phase of the life cycle. Although diploid cells alternate with haploid cells, in practice most natural cultures of yeast are diploid. This is because the haploid cells released by meiosis are normally in close proximity and soon mate again. In the laboratory, asci with the spores still inside can be isolated from yeast cultures and the spores may be examined individually. This allows genetic analysis of the inheritance patterns of specific genes.

Yeast cells of different mating types signal each other by releasing distinct pheromones from the cell.

Mating type is controlled by the ***MAT* locus**, which may exist in two alternative states, *MAT**a*** or *MAT*α. Both states contain two genes that are transcribed divergently (either *MAT**a1*** *MAT**a2*** or *MAT**α1*** *MAT**α2***). These gene products (designated MATa1p, MATa2p, MATα1p and MATα2p) activate synthesis of specific peptide

ascospore Type of spore made inside an ascus by fungi of the ascomycete group, including yeasts and molds
ascus Specialized spore forming structure of ascomycete fungus
mating types Equivalent of different sexes found in lower eukaryotes. They are structurally identical but biochemically distinct
MAT* locus** Chromosomal locus in yeast that controls the mating type and exists as two alternative forms, *MAT**a or *MAT**α***



A. MATING

a factor
Receptor for a factor
Haploid cell
a
α
Haploid cell
Receptor for α factor
α factor
PHEROMONES BIND TO RECEPTORS
a
α
MATING PAIR FORMS
a
α

B. PHEROMONE STRUCTURES

α-factor
NH_2-Trp-His-Trp-Leu-Gln-Leu-Lys-Pro-Gly-Gln-Pro-Met-Tyr-COOH
a-factor
Farnesyl group
S
NH_2-Try-Ile-Ile-Lys-Gly-Val-Phe-Trp-Asp-Pro-Ala-Cys-$COOCH_3$

FIGURE 19.20
Pheromones and Receptors in Mating Yeast

(A) Haploid yeast cells of the α type express the genes for the α-factor and **a**-factor receptor. Haploid yeast cells of the a type express the genes for the **a**-factor and α-factor receptor. When **a**-factor binds to the **a**-factor receptor of the α cell, and the α-factor binds the α-factor receptor on the a cell, the two haploid yeast become competent and form a mating pair. (B) The two mating factors are small peptides of 12 to 13 amino acids. The a-factor has a farnesyl group attached to the last amino acid, cysteine, via its sulfhydryl group.

pheromones or mating factors and their corresponding receptors (Fig. 19.20). Thus a *MATa* cell produces **a**-specific pheromone or **a**-factor and the receptor for α-factor, whereas a *MAT*α cell produces α-factor and receptor for a-factor. The pheromones bind to the receptors on the cells of opposite mating type.

Although both versions of the MAT locus express two genes, the two mating types do not behave in a strictly symmetrical manner. The two proteins from the *MAT*α locus control expression of both the **a**-specific and α-specific genes (Fig. 19.21). The MATα1p protein binds to Mcm1p protein forming an activator that switches on α-specific genes (including those for α-factor and **a**-factor receptor). MATα2p also binds to Mcm1p but forms a repressor that switches off **a**-specific genes (including those for **a**-factor and α-factor receptor). The ***MATa*** locus has a rather different role. The MATa1p protein binds to MATα2p forming a repressor that switches off haploid specific genes in diploid cells. The MATa2p protein has no known function.

Haploid cells of yeast frequently and spontaneously change their mating type. This is not due to mutation but to DNA rearrangement at the *MAT* locus (Fig. 19.22).

Haploid yeast cells frequently change their mating type due to rearrangemnets of DNA at the *MAT* locus.

mating factor Chemical messenger or pheromone that indicates the mating type and promotes sexual conjugation
pheromone Chemical messenger that moves between separate individual organisms

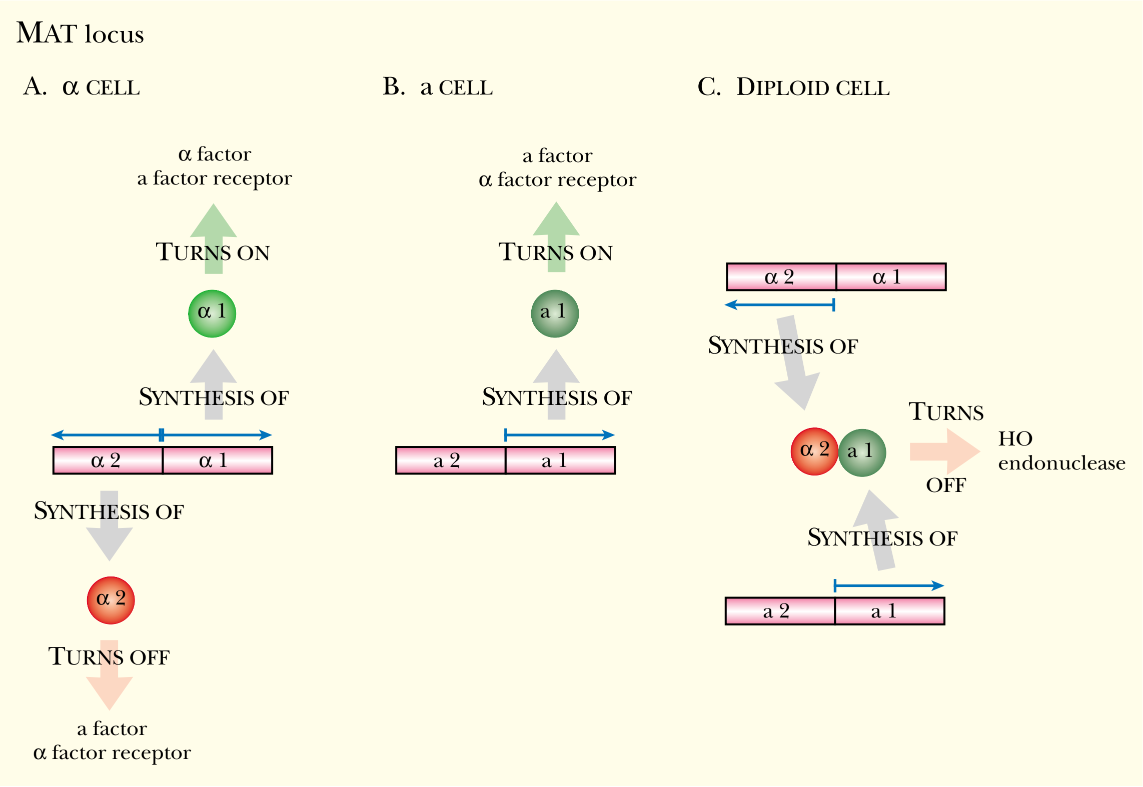
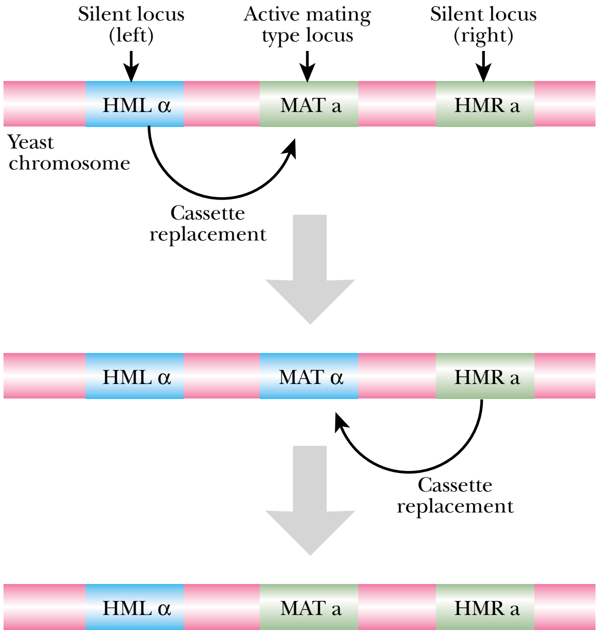


FIGURE 19.21 ***MATa and MATα Proteins are Transcriptional Regulators***

A) A haploid α cell produces two proteins from the MAT locus, α**1** and α**2**. The α**1** protein activates transcription of the α-factor gene and the gene for the **a**-factor receptor. α**2** protein represses transcription of the a-factor and the α-factor receptor. Both α**1** and α**2** act as homodimers. B) A haploid **a** cell also produces two MAT proteins, **a1** and **a2**. The **a1** protein turns on the gene for **a**-factor and the α-factor receptor. The **a1** protein works as a homodimer. C) In a diploid cell, a heterodimer of α**2**/**a1** forms and represses the expression of the HO locus that controls mating type switching.

FIGURE 19.22 ***Switching the Mating Type in Yeast***

The active mating type locus is flanked by two loci called *HML* and *HMR*. *HML* contains a silent copy of the *MATα* gene, and *HMR* contains a silent copy of the *MAT**a*** gene. Yeast cells replace the *MAT* gene at the active mating type locus (center) with the genes located in *HML* and *HMR*. In this example, the *MAT**a*** gene at the active locus is replaced with the *MATα* gene from the HML site, thus switching the phenotype of the yeast cell from an **a** to an α. After a while, the *MATα* gene at the active locus is replaced with the *MAT**a*** gene from *HMR*, thus changing the yeast back to the **a** type.



Located at a considerable distance (over 100 Kb away) and on either side of MAT are two storage loci (*HML* and *HMR*), which contain silent versions of the *MAT*α and *MAT**a*** genes respectively. Switching involves removal of the DNA at the active *MAT* locus and its replacement with a copy of the DNA from either the *HML* or *HMR* locus.

The HO-endonuclease controls switching of the mating type by cutting the DNA at the *MAT* locus. In diploid cells, mating type switching does not occur because transcription of the *HO* gene is repressed. The MATa1 protein product binds to the MATα2 protein, forming a mixed dimer that blocks transcription at the *HO* locus.

Studies of haploid yeast have shown that first generation yeast cells cannot switch mating type because the HO-endonuclease is not expressed. Although HO-endonuclease is only active in the G1 phase of the cell cycle, the *HO* gene requires an activator protein called SWI5, which is only expressed in the G2 phase. Consequently, the *HO* gene is only activated in the G1 phase of the next cell cycle following a G2 phase. Thus haploid yeast must complete at least one cell division before it can switch mating types.

Multi-Cellular Organisms and Homeobox Genes

Even single-celled microorganisms communicate with each other by sending a variety of signals. For example, yeast cells (see above) and gram-positive bacteria (see Ch 18) both secrete pheromones into the external medium when they are ready to mate. Multi-cellular organisms depend on coordinating the activities of many different cells and this requires constant communication. In addition to internal signals between cells, multi-cellular organisms send signals from one organism to another.

Control of body architecture in multi-celled organisms is due to an ordered array of homeobox genes.

Higher animals and plants form permanent multi-cellular structures. As a consequence, their cells differentiate forming different tissues in which different sets of genes are expressed. Development and differentiation involve complex patterns of gene regulation. Despite the great diversity of structures, the genetic mechanisms for laying out the overall body plan have some common themes. Specifying overall body layout, including such things as the number of limbs and segmentation patterns, is the role of the **homeobox genes**. These encode transcription factors containing a **homeodomain** of about 60 amino acid residues forming a helix-turn-helix region that binds DNA and is highly conserved. The regions of the transcription factors outside the homeodomain vary greatly from protein to protein. Homeobox proteins are found in all multicellular animals and also in higher plants.

The best-known family of homeobox genes is the ***Hox* genes**. The Hox proteins are transcription factors that control the expression of many other regulatory proteins, including other transcription factors. They were first identified in *Drosophila* as a result of bizarre developmental mutations. For example, mutations in the *antennapedia* subgroup of *Hox* genes may result in production of a leg where an antenna is normally supposed to grow (hence the name antenna-pedia) and mutations in the *bithorax* cluster can produce four-winged flies.

The *Hox* genes of animals are found in clusters and the order of Hox genes along the chromosome corresponds to their expression in the developing *Drosophila* fruit fly embryo (Fig. 19.23). *Hox* genes at the 3′-end of the cluster are expressed in the head region whereas *Hox* genes at the 5′-end are expressed in the tail region. Consequently, these are known as the anterior and posterior *Hox* genes, respectively. Between these lie the group 3 and central genes. Furthermore, *Hox* genes at the 3′-end are expressed earlier in development than those at the 5′-end. Thus the *Hox* genes

homeobox genes Genes encoding transcription factors containing a homeodomain that help specify the body plan of multicellular organisms
homeodomain Conserved region of about 60 amino acid residues found in homeobox proteins that that binds DNA by a helix-turn-helix motif
***Hox* genes** Family of homeobox genes that control overall body layout by regulating the expression of many other regulatory genes, including those for other transcription factors

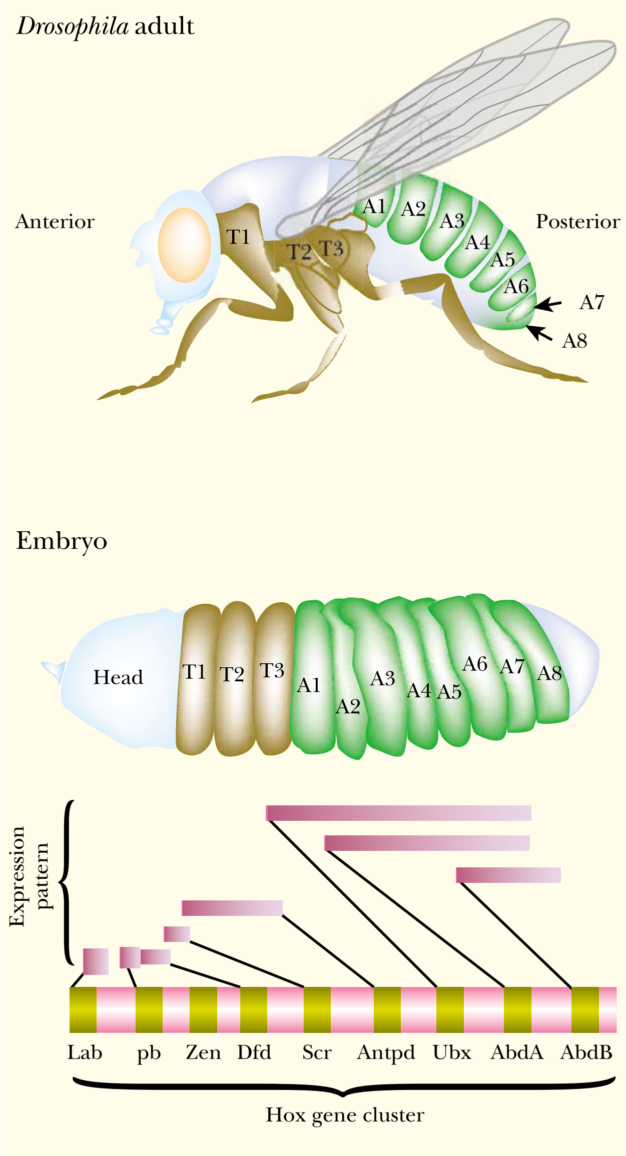


FIGURE 19.23 ***Cluster of Hox Genes in Drosophila***

Different *Hox* genes control development of each segment of the *Drosophila* fruit fly embryo. For example, the embryo segments T1, T2, and T3 express the *antennapedia* gene, encoding a transcription factor of the Hox family. This protein controls the development of the adult legs from the T1, T2 and T3 embryonic segments. The Antennapedia transcription factor is expressed as a gradient in these three segments with the highest concentration in T1. If this protein is expressed in a different segment of the embryo, a leg will grow in the wrong place. When this transcription factor was expressed in the head region, the fly's antenna turned into a leg.

control the expression of groups of other genes both spatially and temporally during development.

More advanced organisms have several sets of homeobox genes due to ancient DNA duplications.

The *Hox* cluster has evolved by duplication of one or several genes and by loss of individual members in some lineages (Fig. 19.24). Jellyfish and other cnidarians have only anterior and posterior *Hox* genes. The ancestral *Hox* cluster of bilateral animals is thought to have contained two anterior, one group 3, four central and one posterior *Hox* gene. The **protostomes**, which includes arthropods, annelids, molluscs, flatworms

protostomes Grouping of animal phyla that includes arthropods, annelids, molluscs, flatworms etc

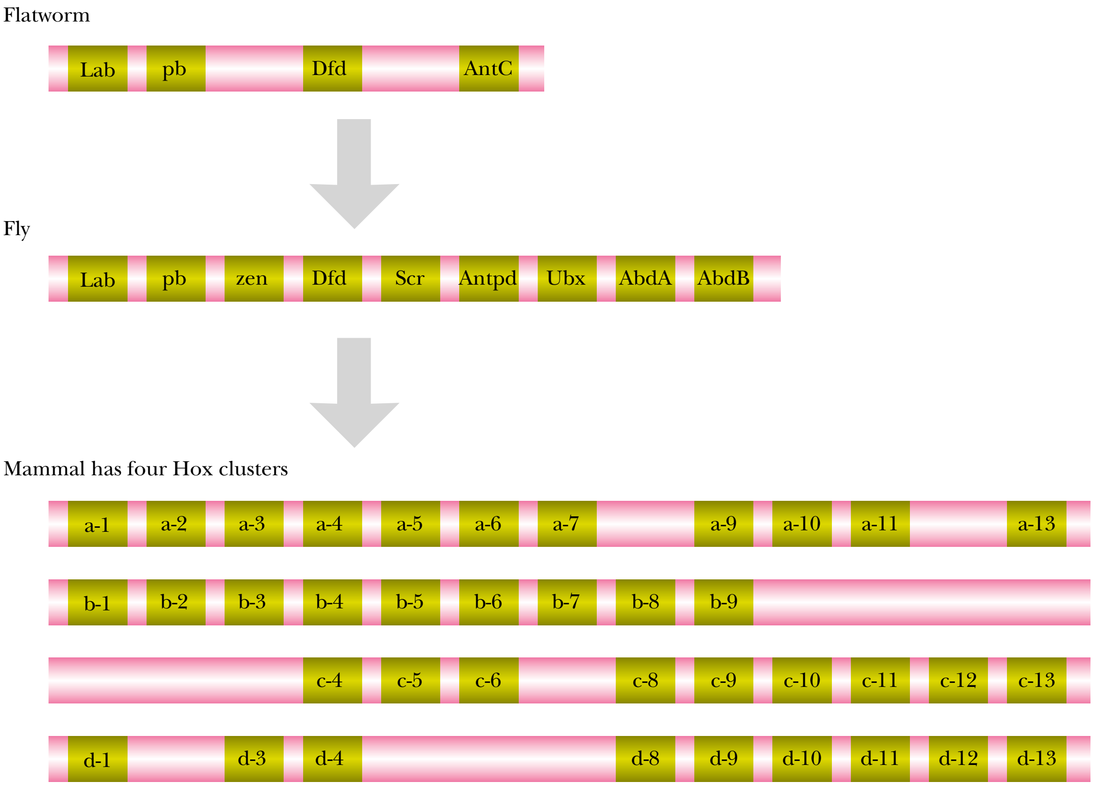


FIGURE 19.24 ***Evolution of Hox Gene Clusters***

The number and complexity of *Hox* genes increase with the complexity of the organism. A simple roundworm that has very little body complexity only has four different Hox genes. In mammals, there are four different Hox gene clusters that each contains from nine to eleven separate Hox genes. The additional genes are required to manage the increasingly complex body patterns.

etc., added two extra central *Hox* genes giving a set of ten. The **deuterostomes**, which includes echinoderms and vertebrates, added another central gene and four more anterior genes, giving a total of 13. In addition, vertebrates have four separate *Hox* clusters (*Hoxa, Hoxb, Hoxc, Hoxd*) as a result of duplicating the whole array.

Although many *Hox* defects are obviously lethal, some inherited conditions are due to mutations in the less vital members of the *Hox* system. Mutation of the human *Hoxd13* gene leads to abnormal hands and feet, in which there are both extra fingers and toes and webbing between them, a condition known as synpolydactyly.

deuterostomes Group of animal phyla including echinoderms and vertebrates

CHAPTER TWENTY

Molecular Evolution

Getting Started—Formation of the Earth

The big bang is estimated to have happened about 20,000,000,000 years ago. About 15,000,000,000 years later, a cloud of interstellar dust and gas coalesced and condensed due to gravity into a large ball of gas that we call the sun, orbited by smaller spherical bodies of variable composition, called the planets. The universe consists mostly of the light molecular weight gases hydrogen and helium, which account for most of the material of the stars. The heavier elements together comprise only about 0.1 percent of the total and form the planets (Table 20.01).

As the earth formed, heat was released by the collapse due to gravity and also by the radioactivity of elements present in the original dust. During its first few hundred million years the Earth was too hot for water to liquefy and so H_2O was only present as steam. Later, as the Earth cooled off, the steam condensed to form oceans and lakes. Life is thought to have originated by means of chemical reactions occurring in the atmosphere followed by further reactions in the primeval oceans and lakes (the hydrosphere).

The Early Atmosphere

Life evolved under a reducing atmosphere lacking any free oxygen.

The Earth's original atmosphere, the **primary atmosphere**, consisted mostly of hydrogen and helium, but the Earth is too small a planet to hold such light gases and they floated away into space. The Earth then accumulated a **secondary atmosphere**, mostly by volcanic out-gassing. Volcanic activity was much greater on the hotter primitive Earth. Volcanic gas consists mostly of steam (95%) and variable amounts of CO_2, N_2, SO_2, H_2S, HCl, B_2O_3, elemental sulfur, and smaller quantities of H_2, CH_4, SO_3, NH_3 and HF but no O_2. Of all these, the concentration of CO_2 was in the second highest amount (about 4%). In addition, water vapor reacted with primeval minerals such as nitrides to give ammonia, with carbides to give methane and with sulfides to give hydrogen sulfide. There was no free oxygen.

Our present atmosphere, the **tertiary atmosphere**, is of biological origin. The methane, ammonia, and other reduced gases have been consumed and the inert components (nitrogen, traces of argon, xenon etc.) have remained largely unchanged. Substantial amounts of oxygen have been produced by photosynthesis. This could not

TABLE 20.01 Elemental Compositions in Atoms per 100,000

Element	Universe	Earth	Crust	Life
H	92,700	120	2,900	60,600
He	7,200	<0.1	<0.1	0
O	50	48,900	60,400	26,700
Ne	20	<0.1	<0.1	0
N	15	0.3	7	2,400
C	8	99	55	10,700
Si	2.3	14,000	20,500	<1
Mg	2.1	12,500	1,800	11
Fe	1.4	18,900	1,900	<1

primary atmosphere The original atmosphere of the earth consisting mostly of hydrogen and helium
secondary atmosphere The atmosphere of the earth after the light gases were lost and resulting mostly from volcanic out-gassing. It contained reduced gases but no oxygen
tertiary atmosphere The present atmosphere of the earth resulting from biological activity

TABLE 20.02 Approximate Evolutionary Time Scale

Millions of Years Ago	Major Events
20,000	Big Bang
5,000	Origin of planets and sun
3,500	Origin of life
3,000	Primitive bacteria start using solar energy
2,500	Advanced photosynthesis releases oxygen
1,500	First eukaryotic cells
1,000	Multicellular organisms
600	First skeletons give nice fossils
1.8	First true humans—*Homo erectus*
0.15	African Eve gives birth to modern Man
0.01	Domestication of crops and animals begins
0.002	Roman Empire
0.0001	Darwin's theory of evolution
0.00005	Double helix discovered by Watson and Crick
0.000003	Human genome sequenced

Oxygen in today's atmosphere is the result of photosynthesis.

occur until the cyanobacteria, the first true photosynthetic organisms, had evolved about 2.5 thousand million years ago (see Table 20.02 for timescale). As more and more photosynthetic organisms evolved, the oxygen content of the atmosphere increased. The oxygen content of the atmosphere reached 1 percent about 800 million years ago and 10 percent about 400 million years ago. Today it is about 20 percent.

Evidence for the increase in O_2 content of the atmosphere comes partly from the finding that rocks of different ages are oxidized to different extents. Thus rocks of age 1,800–2,500 million years sometimes contain UO_2, FeS, ZnS and PbS and FeO, all of which are unstable in the presence of even small amounts of gaseous O_2. Later rocks contain mostly Fe^{3+} rather than Fe^{2+}; and more oxidized ores of U, Zn and Pb.

Oparin's Theory of the Origin of Life

Reactions in the primeval atmosphere led to the accumulation of organic material in oceans and lakes—the primitive soup.

Ultraviolet radiation from the sun, together with lightning discharges, caused the gases in the primeval atmosphere to react, forming simple organic compounds. These dissolved in the primeval oceans and continued to react, forming what is sometimes referred to as the "**primitive soup**". The primitive soup contained amino acids, sugars, and nucleic acid bases among other randomly synthesized molecules (Fig. 20.01). Further reactions formed polymers and these associated, eventually forming globules. Ultimately, these evolved into the first primitive cells. This theory of the origin of life was put forward by the Russian biochemist Alexander Oparin in the 1920s. Charles Darwin himself had actually proposed that life might have started in a warm little pond provided with ammonia and other necessary chemicals. However, it was Oparin who outlined all the necessary steps and realized the critical point: life evolved before there was any oxygen in the air. Oxygen is highly reactive and would have reacted with the organic precursor molecules formed in the atmosphere, oxidizing them back to water and carbon dioxide.

primitive soup Mixture of random molecules, including amino acids, sugars, and nucleic acid bases, found in solution on the primeval earth

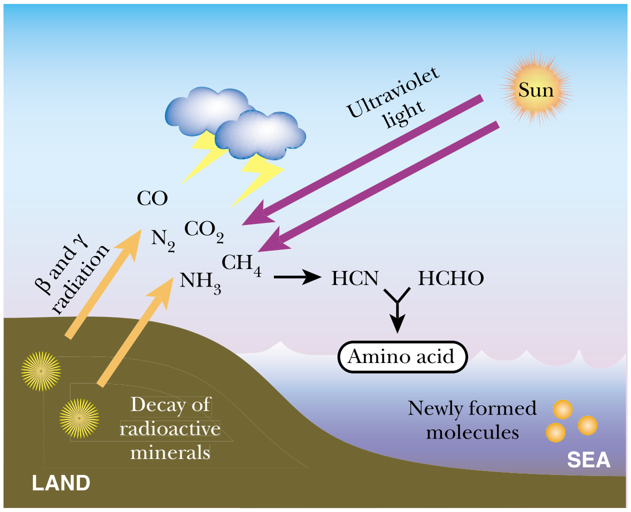


FIGURE 20.01 ***Formation of Primitive Soup***

According to Oparin's theory of the origin of life, conditions on planet earth were sufficient for forming early biological molecules. The atmosphere at this point contained CO, CO_2, CH_4, N_2, and NH_3. When energy was supplied by electrical discharge from lightning, ultraviolet radiation from the sun, and/or β and γ radiation from the earth, early organic compounds such as HCN and HCHO would form. These compounds would combine in the vapor phase and in the water to form amino acids. Dissolving in water protects the precursor molecules from being degraded again by the energy sources that triggered their formation.

The Miller Experiment

In the early 1950s, the biochemist Stanley Miller mimicked the reactions proposed to occur in the primitive atmosphere. An imitation atmosphere containing methane, ammonia and water vapor was subjected to a high voltage discharge (to simulate lightning) or to ultraviolet light (Fig. 20.02). The gases were circulated around the apparatus so that any organic compounds formed in the artificial atmosphere could dissolve in a flask of water, intended to represent the primeval ocean. These compounds could continue to react with each other in the water.

Experiments to mimic reactions in the primeval atmosphere have generated many of the metabolites and monomers found in modern cells.

There are many variants of this experiment (different gas mixtures, different energy sources, etc.). As long as oxygen is excluded, the results are similar. About 10 to 20 percent of the gas mixture is converted to soluble organic molecules and significantly more is converted to a non-analyzable organic tar. First, aldehydes and cyanides are formed, and then a large variety of other organic compounds (Table 20.03). Most of the naturally occurring amino acids, hydroxy-acids, purines, pyrimidines, and sugars have been produced in variants of the Miller experiment. These experiments produce all different isomers of the compounds; only a portion is biologically significant. For example, sarcosine and beta-alanine are both isomers of alanine and all three are generated in the Miller experiment. Many organic molecules, in particular sugars and amino acids, exist as two possible optical isomers, only one of which is normally found in biological macromolecules. In such cases the molecules formed in the Miller experiment consist of an equal mixture of the D- and L-isomers.

Since many chemical reactions are reversible, the same energy sources that produce organic molecules are also very effective at destroying them. The long-term build-up of organic material requires its protection from the energy sources that created it. This is the function of the imitation primeval ocean in the Miller experiment. Water shields molecules from ultraviolet radiation and from electric discharges. The survival of organic molecules on the primitive earth would have depended on

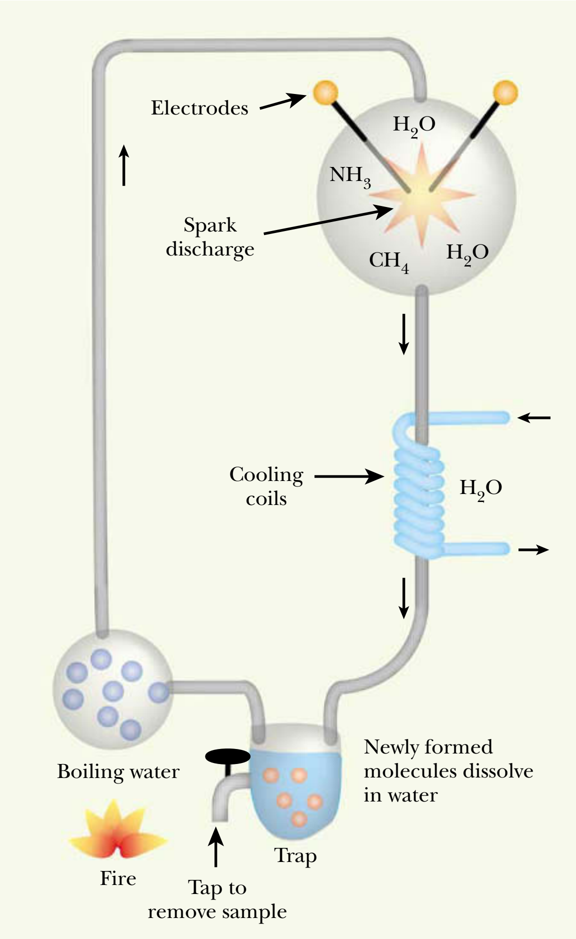


FIGURE 20.02 ***Miller's Experiment***

Miller's experiment used a closed system to simulate the primeval conditions on early planet earth. Water was boiled to make steam (lower left), which then mixed with NH_3 and CH_4 in another chamber. This upper chamber simulates earth's early atmosphere and was subjected to either electrical discharge (shown) or ultraviolet radiation (not shown). The resulting products were cooled and condensed as they passed coils filled with cold water. The condensed products dissolved in a flask of water, which simulated the early oceans. The newly formed molecules were analyzed at various times during the experiment.

TABLE 20.03 Typical Products from Miller's Experiment

Molecule	Name	Relative Yield
H-COOH	formic acid	1000
H_2N-CH_2-COOH	glycine	275
HO-CH_2-COOH	glycolic acid	240
H_2N-CH(CH_3)-COOH	alanine	150
HO-CH(CH_3)-COOH	lactic acid	135
H_2N-CH_2CH_2-COOH	beta-alanine	65
CH_3-COOH	acetic acid	65
CH_3-CH_2-COOH	propionic acid	55
CH_3-NH-CH_2-COOH	sarcosine	20
HOOC-CH_2CH_2-COOH	succinic acid	17
H_2N-CO-NH_2	urea	9
HOOC-CH_2CH_2CH(NH_2)-COOH	glutamic acid	2.5
HOOC-CH_2CH(NH_2)-COOH	aspartic acid	1.7

Water protects organic molecules from destruction by UV radiation or lightning.

their escape from UV radiation and lightning either by dissolving in seas or lakes or by sticking to minerals. Most organic molecules formed too far up in the sky would have been destroyed again very quickly, while those that reached the sea and dissolved would have survived. Note that organic acids, in particular amino acids, are water soluble and non-volatile. Once they are safely dissolved in water, there is little tendency for such molecules to return to the atmosphere. Their precursors, the aldehydes and cyanides, are not only highly reactive but also volatile. Consequently, these molecules do not survive for long. Thus, even at this early stage, there was a form of natural selection between molecules.

Originally it was thought that the primitive secondary atmosphere contained mostly NH_3 and CH_4. However, it seems more likely that most of the atmospheric carbon was CO_2 with perhaps some CO and the nitrogen mostly as N_2. The two reasons for this are: volcanic gas has more CO_2, CO and N_2 than CH_4 and NH_3, and that UV radiation destroys NH_3 and CH_4, therefore, these molecules would have been short-lived. The UV destruction of CH_4 occurs in two steps. First, UV light photolyses H_2O to $H^\bullet$ and $^\bullet OH$ radicals. These then attack methane, giving eventually CO_2 and H_2, which would be lost into space. In Miller's experiment, gas mixtures containing CO, CO_2, N_2, etc. give much the same products as those containing CH_4 and NH_3 so long as there is no O_2. Since CO, CO_2 and N_2 do not supply H atoms; these come mostly from water vapor photolysis. In fact, in order to generate aromatic amino acids under primitive earth conditions it is necessary to use less hydrogen-rich gaseous mixtures. Most of the natural amino acids, hydroxy-acids, purines, pyrimidines, and sugars have been produced in variants of the Miller experiment.

Polymerization of Monomers to Give Macromolecules

Formation of biological polymers requires removal of water.

Polymerization of monomers to give biological macromolecules usually requires the removal of H_2O. Clearly, water is in excess in the oceans and removal of H_2O from dissolved molecules is therefore unfavorable. Consequently, the assembly of macromolecules such as proteins and nucleic acids needs energy to form the linkages and/or remove the water. Before the high-energy phosphates used in modern cells were available, some other form of energy was needed.

Water can be removed by moderate heating, by certain clay minerals or by chemical condensing agents.

Imitation protein polymers, containing randomly linked amino acids, are known as "**proteinoids**." They can be formed by heating dry amino acid mixtures at around 150°C for a few hours (Fig. 20.03A). Whereas biological proteins are bonded using only the α-NH_2 and α-COOH groups of amino acids these "primeval polypeptides" contain substantial numbers of bonds involving side chain residues. They contain up to 250 amino acids and can sometimes perform primitive enzymatic activities. Such dry heat could have occurred near volcanoes or when pools left behind by a changing coastline evaporated. Much of the early work on proteinoids was done by Sydney Fox who proposed their thermal origin. However, another way to randomly polymerize amino acids is by using clay minerals with special binding properties (Fig. 20.03B). Binding of small molecules to the surface of catalytic minerals can promote many reactions. For example, certain clays, such as Montmorillonite, will condense amino acids to form polypeptides up to 200 residues long.

Polymerization of amino acids may have also occurred in solution, but another component, a condensing agent, is required to withdraw water. Several possible primeval condensing agents have been proposed, including reactive cyanide derivatives and, more biologically relevant, **polyphosphates**. Inorganic polyphosphates would have been present in primeval times (formed by volcanic heat from phosphates for example). Polyphosphates can react with many organic molecules to give organic phosphates. Amino acids give two possible products (Fig. 20.04). **Acyl phosphates** have the phosphate group attached to the carboxyl group of the amino acid (NH_2CHRCO-OPO_3H_2) and **phosphoramidates** have the phosphate attached to the amino group of the amino acid (H_2O_3P-NH-CHR-COOH). Gentle heating or irradiation such derivatives will give polypeptides. Modern life uses acyl phosphate derivatives during protein

acyl phosphate Phosphate derivative in which the phosphate is attached to a carboxyl group
phosphoramidate Phosphate derivative in which the phosphate group is attached to an amino group
polyphosphate Compound consisting of multiple phosphate groups linked by high energy phosphate bonds
proteinoid Artificially synthesized polypeptide containing randomly linked amino acids

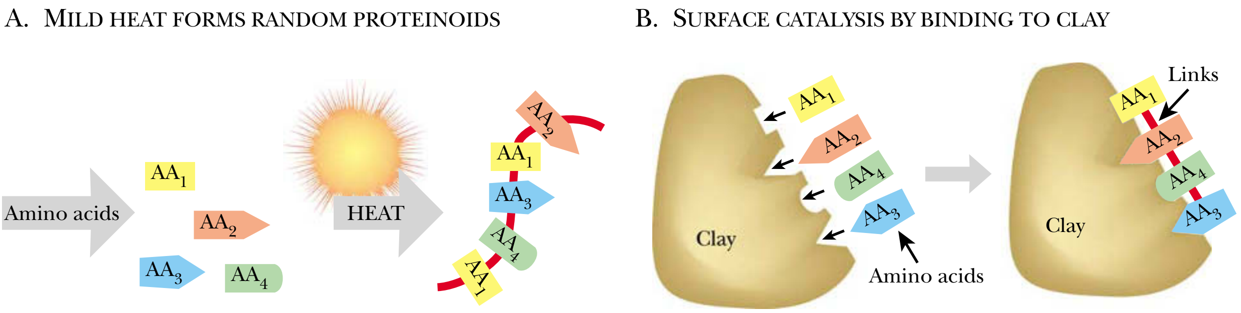
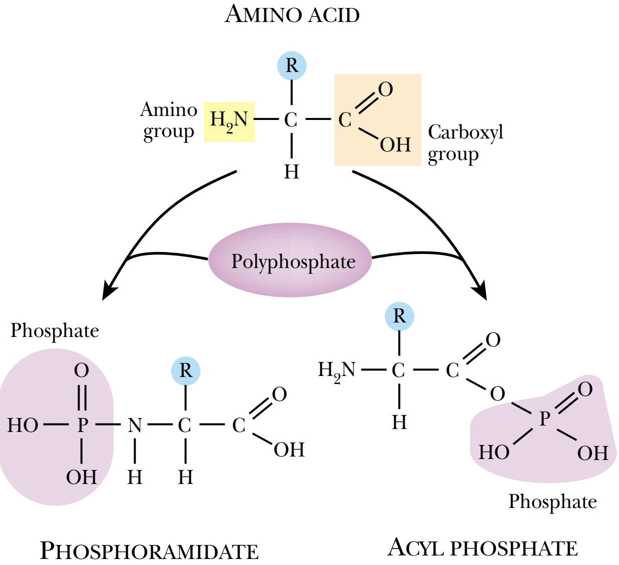

A. MILD HEAT FORMS RANDOM PROTEINOIDS
Amino acids
AA_1
AA_2
AA_3
AA_4
HEAT
B. SURFACE CATALYSIS BY BINDING TO CLAY
Clay
Amino acids
Links
Clay

FIGURE 20.03 ***Formation of Proteinoids by Mild Heat or Clay Catalysis***

(A) A mixture of separate amino acids will form artificial polypeptide chains or "proteinoids" when subjected to heat in the absence of water for a few hours. (B) Amino acids can also form bonds when they bind to certain types of clay. The clay has binding sites for amino acids in close proximity, therefore, once bound the amino acids condense into a proteinoid.

FIGURE 20.04 ***Formation of Acyl Phosphates and Phosphoramidates***

Condensing single amino acids into polypeptide chains could have occurred in solution as long as a condensing agent was present. One possible condensing agent was a polyphosphate that would react with the amino group or carboxyl group of individual amino acids. The two possible products, phosphoramidates and acyl phosphates, can form polypeptide chains by heating in solution.



synthesis, although from a chemical viewpoint, either derivative would work. In fact, laboratory synthesis of DNA does use phosphoramidates. Analogous reactions can produce AMP from adenine plus polyphosphate and polynucleotides can then form by polymerization (see below).

Enzyme Activities of Random Proteinoids

Artificial random proteinoids show inefficient but detectable enzyme activities.

Interestingly, random proteinoids stewed up in modern laboratories under fake primeval Earth conditions will carry out some simple enzyme reactions (Table 20.04). They are far slower and less accurate than enzymes made by real cells, but nonetheless they can perform recognizable enzymatic reactions. For example, random proteinoids can often remove carbon dioxide from molecules like pyruvate or oxaloacetate and split organic esters. About 50% of all modern enzymes contain metal ions as cofactors and the addition of metal ions greatly extends the enzyme activities of random proteinoids. The presence of traces of copper promotes reactions involving amino groups and iron mediates oxidation-reduction reactions. Incorporation of zinc allows the breakdown of ATP, which is used by modern cells, both as a precursor of nucleic acids and as an energy carrier. Most modern enzymes that process nucleic acids possess a zinc atom as cofactor.

TABLE 20.04 Enzyme Activities of Random Proteinoids

Reaction	Requirements	Substrate
esterase	histidine	p-nitrophenyl-phosphate
ATPase	Zn^{2+}	ATP
amination and deamination	Cu^{2+}	α-ketoglutarate ↔ glutamate
peroxidase and catalase	heme, basic proteinoids	H_2O_2 and H-donors e.g. hydroquinone, NADH
decarboxylation	basic proteinoids, or acidic proteinoids	oxaloacetate, pyruvate

Origin of Informational Macromolecules

Primeval synthesis of proteins or nucleic acids gives polymers with a mixture of natural and unnatural linkages.

Biological information is passed on by template-specific polymerization of nucleotides. A mixture of polyphosphate, purines and pyrimidines will produce random nucleic acid chains if ribose or deoxyribose is included. One problem, not yet solved, is that life uses 3′, 5′ linked nucleic acids whereas primeval syntheses give RNA molecules with a mixture of linkages, but mostly 2′, 5′. In contrast, deoxyribose has no 2′-OH and so cannot give 2′, 5′ links. However, it is generally thought that RNA probably provided the first informational molecule and that DNA is a later invention designed to store information in a more stable and accurate form.

Zn or Pb ions can catalyse non-enzymatic synthesis of RNA.

When an RNA template is incubated with a mixture of nucleotides, plus a primeval condensing agent, a complementary piece of RNA is synthesized. This non-enzymatic reaction is catalyzed by lead ions, with an error rate of about 1 wrong base in 10. With zinc ions, a great improvement is seen and lengths of up to 40 bases are produced, with an error rate of about 1 in 200. All modern day RNA and DNA polymerases contain zinc. If a 3′, 5′ linked RNA template is used about 75% of the newly formed RNA is 3′, 5′ linked. However, this does not surmount the problem that the original formation of random RNA type polymers favors the non-biological 2′, 5′ linkage very heavily.

The first strands of RNA to form act as templates for the assembly of later molecules.

If a mixture of nucleoside triphosphates (or nucleotides plus polyphosphate) is incubated under primeval conditions, using Zn as a catalyst, a molecule of single-stranded RNA with a random sequence will form. This original polymerization step is very slow. However, once an initial RNA polymer is present, it will act as a template for the assembly of a complementary strand. Template-directed synthesis is much more rapid, even in the absence of any enzyme. The complementary strand will in turn act as a template to generate more of the original RNA molecule. The net result is that once the first random sequence emerges, it will multiply rapidly and take over the incubation mixture (Fig. 20.05). The result will be a set of sequences with frequent mistakes but which are nonetheless clearly related (a molecular "**quasi-species**"). If a series of similar incubations are carried out, each individual sample will yield a "quasi-species" of related sequences. However, the sequence that takes over will be different for each incubation mixture.

Ribozymes and the RNA World

The ability of RNA to act as an enzyme as well as encode genetic information suggests that RNA came first.

Which came first the chicken or the egg?—Protein or nucleic acid? Since it is possible for random RNA molecules to be assembled and duplicate themselves under primeval conditions, it is generally held that nucleic acids appeared first. Furthermore, although most modern-day enzymes are indeed proteins, examples of RNA acting as enzyme and catalyzing reactions without help from proteins do exist. This scenario suggests

quasi-species A set of closely related sequences whose individual members vary from consensus by frequent errors or mutations

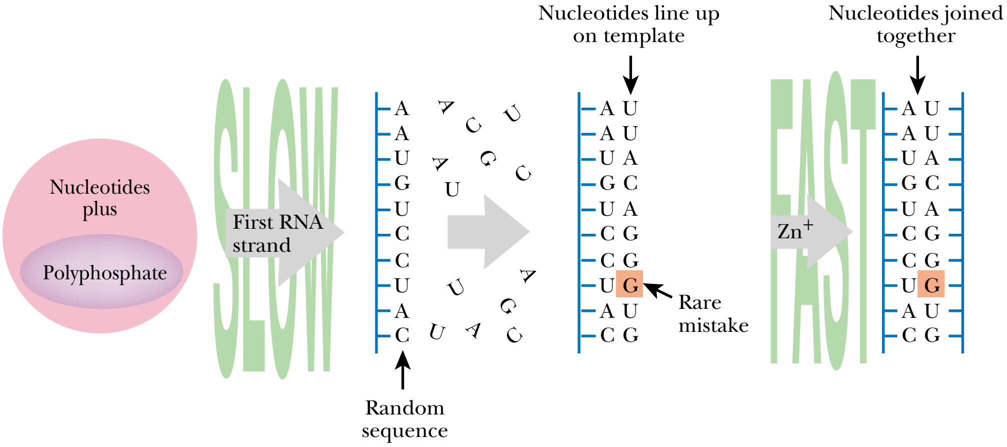


FIGURE 20.05 ***Assembly and Duplication of Random RNA***

A mixture of nucleosides and polyphosphates can form random stretches of RNA in the presence of zinc ions. The first strand of RNA forms very slowly. However, once the first strand is assembled, it may be used as a template to assemble complementary strands of RNA. Such template driven assembly of nucleotides is much faster than formation of the original random RNA strand. However, the non-enzymatic synthesis of RNA incorporates many wrongly paired bases.

that the primitive nucleic acid replicated alone and a protective protein coat was added later.

One rather extreme viewpoint is the idea that the earliest organisms had both genes and enzymes made of RNA and formed a so-called "**RNA world**." This idea was proposed by Walter Gilbert in 1986 and seeks to avoid the paradoxical problem that nucleic acids are needed to encode proteins, but that enzymes made of protein are needed to replicate nucleic acids. During the RNA world stage, RNA supposedly carried out both functions. Later, proteins infiltrated and took over the role of enzymes and DNA appeared to store the genetic information, leaving RNA as a mere intermediate between genes and enzymes.

Several examples illustrate the ability of RNA to perform enzymatic reactions as well as encode genetic information. These cases favor the primacy of RNA:

1. **Ribozymes** are single-use RNA molecules that are enzymatically active. A genuine enzyme processes large numbers of other molecules, and does not become altered during the reaction. Therefore, self-splicing RNA is not a true enzyme because it works only once. There is a growing list of known and suspected ribozymes. The most important is the ribosomal RNA of the large subunit, which is directly involved in the reactions of protein synthesis (see Ch. 8). One of the best-known ribozymes is **ribonuclease P**. This enzyme has both RNA and protein components and processes certain transfer RNA molecules. It is the RNA part of ribonuclease P that carries out the reaction. The protein serves only to hold together the ribozyme and the transfer RNA it operates on. In concentrated solution, the protein is not even necessary, and the RNA component will work on its own.

RNA molecules with enzyme activity include ribosomal RNA, ribonuclease P, self-splicing introns and viroids.

2. Self-splicing introns ("group I" introns) are an example of catalytic RNA. The genes of eukaryotic cells are often interrupted by non-coding regions (the introns), which must be removed from the messenger RNA before translation into protein. Normally, this is done by a spliceosome made up of several pro-

ribonuclease P A ribozyme found in many bacteria that processes certain transfer RNA molecules
ribozyme RNA molecule that is enzymatically active
RNA world The hypothetical stage of early life in which RNA encoded genetic information and carried out enzyme reactions without the need for either DNA or protein

teins and small RNA molecules. Occasionally, the intron RNA splices itself out without help from any protein. Such self-splicing is found in a few nuclear genes of some protozoans, in the mitochondria of fungal cells, and the chloroplasts of plant cells (see Ch. 12 for details).

3. Viroids are infectious RNA molecules that infect plants. As noted in Chapter 17, viroid RNA carries out a self-cleavage reaction during replication, i.e. viroids act as ribozymes.

Newly made nucleic acid molecules all start with a stretch of RNA.

4. DNA polymerase cannot initiate new strands but can only elongate pre-existing strands (see Ch. 5). Primers made of RNA must be used whenever new strands of DNA are started. RNA polymerase is capable both of initiation and elongation. This suggests that RNA polymerase may have evolved before DNA polymerase.
5. Small guide molecules of RNA are used in a variety of processes. These include the removal of introns, the modification and editing of messenger RNA (see Ch. 12 for details) and the extension of the ends of eukaryotic chromosomes by telomerase.
6. Riboswitches are recently discovered binding motifs in RNA that directly bind small molecules and so control gene expression in the absence of regulatory proteins (see Ch. 11 for details).

Artificial ribozymes can be isolated by screening pools of random RNA sequences for particular enzymatic reactions.

In a way, the critical question is whether RNA can copy itself without the involvement of DNA or help from protein enzymes. Although no RNA polymerases that are ribozymes still exist, it has proven possible to generate them artificially. Altered RNA molecules can be selected by a form of Darwinian evolution at the molecular level. Pre-existing ribozymes can be used as starting materials. Alternately, some experimenters have used random pools of artificially generated RNA sequences. In one experiment, RNA molecules showing primitive RNA ligase activity were selected from a pool of random RNA sequences. Such artificial ribozymes can link together two chains of RNA by a typical ligase reaction just like protein enzymes in modern cells. The best of these RNA ligase ribozymes was then subjected to further rounds of mutation and selection. The result was a ribozyme of 189 bases that uses an RNA template to synthesize a complementary strand of RNA with about 96–99% accuracy. This ribozyme adds single nucleotides, one at a time, to an RNA primer using nucleoside triphosphates as substrates (Fig. 20.06). However, it is very slow and can only extend chains by around 14 nucleotides because it is not "processive". In other words, the ribozyme dissociates from the template after adding each nucleotide, whereas true polymerases remain attached and proceed along the template adding nucleotides in quick succession.

One problem with the "RNA world" concept is that RNA is more reactive than DNA. Although RNA would form more easily than DNA under primeval conditions, it would also be less stable. Thus DNA, though slower to form initially, might tend to accumulate under such conditions. Moreover, the primeval soup would contain a mixture of the sub-components of both types of nucleic acid as well as proteins, lipids and carbohydrates. So it seems perhaps more likely that an ill-defined mixture, perhaps even hybrid nucleic acid molecules with both RNA and DNA components, emerged first.

The First Cells

The first cells probably used RNA in multiple roles some of which were later taken over by proteins or DNA.

Forming biologically relevant molecules in the primitive earth would have been the first step on the road to forming the first primitive cells. Possibly random proteins and greasy lipid molecules collected around the primeval RNA (or DNA), so forming a microscopic membrane-covered organic blob. Eventually this proto-cell learned how to use RNA to code for its protein sequences. The lipids formed a membrane around the outside to keep the other components together. Early on, protein and RNA

FIGURE 20.06 ***Artificial Evolution of Ribozyme RNA Polymerase***

A random pool of RNA was generated and screened for the ability to seal a nick in one strand of a broken piece of double-stranded RNA. Occasional random molecules of RNA that had this enzymatic activity were found and isolated. Through successive rounds of mutation and selection, this primitive ribozyme was altered just enough to actually catalyze the elongation of the RNA using a single-stranded template with a primer. Unlike true protein RNA polymerases, the RNA polymerizing ribozyme only added one nucleotide at a time, dissociating after each addition.

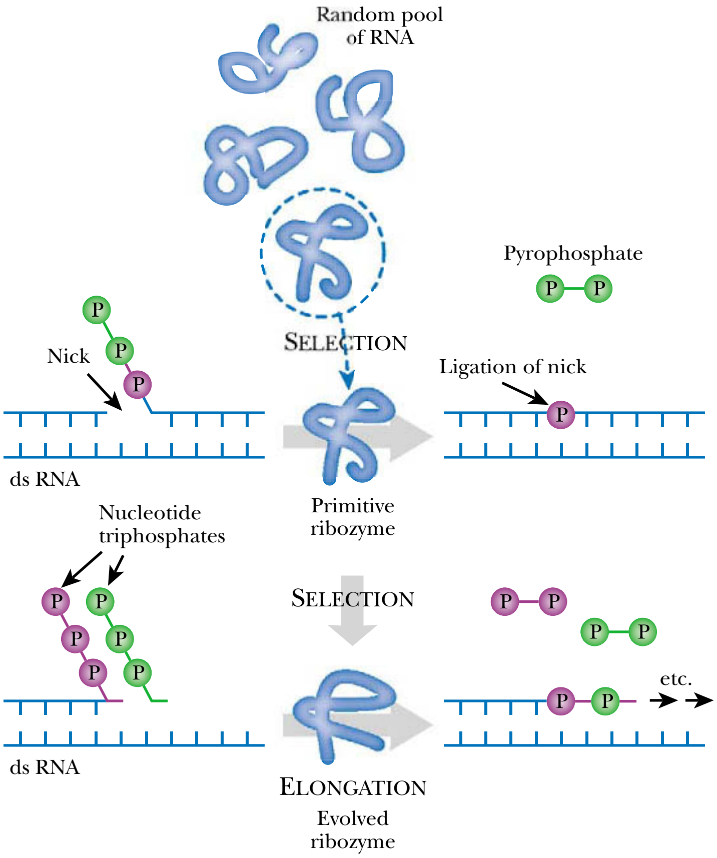
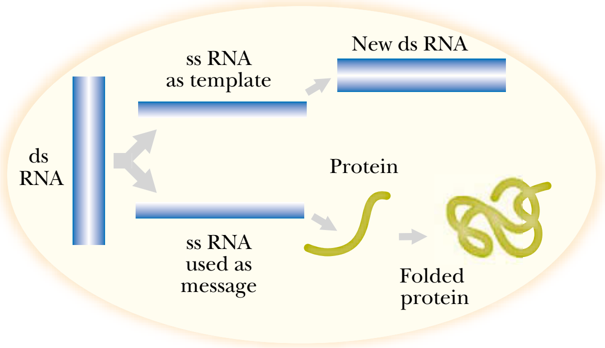


FIGURE 20.07 ***Emergence of the Proto-Cell***

An early primitive cell may have contained RNA molecules as the information coding material. RNA molecules would have acted as ribozymes to make copies of the RNA genetic material. Over time, the RNA would have evolved the ability to synthesize proteins to do much of the enzymatic work.

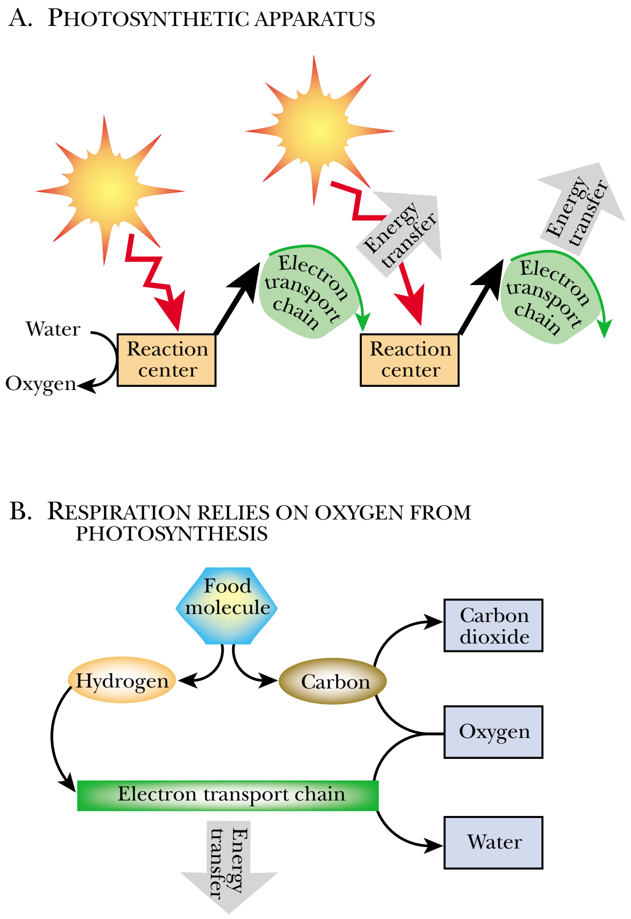


shared the enzymatic functions. Later, RNA lost most of its enzymatic roles as the more versatile proteins took these over (Fig. 20.07). It is generally thought that RNA was the first information storage molecule and that DNA was a later invention. Because DNA is more stable than RNA, it would store and transmit information with fewer errors.

Primitive photosynthesis probably used sulfur compounds as a source of electrons, more advanced photosynthesis resulted in the release of oxygen from water.

This primitive cell vaguely resembles primitive bacteria, and lived off the organic compounds in the primitive soup. Eventually the supply of pre-made organic molecules was consumed. The proto-cell was forced to find a new source of energy and it turned to the sun (Fig. 20.08A). The earliest forms of photosynthesis probably used solar energy coupled to the use of sulfur compounds to provide reducing power. Later, more advanced photosynthesis used water instead of sulfur compounds. The water was split, releasing oxygen into the atmosphere.

Before this, the atmosphere had been void of oxygen. The addition of oxygen completely altered the primitive earth. Once oxygen became available, respiration could


A. PHOTOSYNTHETIC APPARATUS

B. RESPIRATION RELIES ON OXYGEN FROM PHOTOSYNTHESIS

FIGURE 20.08
Development of photosynthesis sets the stage for respiration

(A) When early primitive cells ran out of pre-made organic molecules to supply energy, the cells started using the energy of the sun. The light energy was harvested by reaction centers that used water to provide the reducing power and released molecular oxygen into the atmosphere. Electrons carrying energy derived from the sunlight passed down electron transport chains. This allowed the conversion of the energy to a form used by the cell. (B) The accumulation of oxygen in the atmosphere allowed some of the proto-cells to change their biochemistry. Instead of using water and sunlight, these cells started to use oxygen to oxidize organic matter. This provided the cell with energy and released carbon dioxide and water into the environment.

be developed. Cells reorganized components from the photosynthetic machinery to release energy by oxidizing food molecules with oxygen (Fig. 20.08B). Photosynthesis emits oxygen and consumes carbon dioxide, whereas respiration does the reverse. The overall result is an ecosystem where plants and animals complement each other biochemically.

The Autotrophic Theory of the Origin of Metabolism

There is an alternative theory for the chemical origin of life. According to this view, the first proto-cells were not heterotrophic scavengers of organic molecules but were autotrophic and fixed carbon dioxide into organic matter themselves. An autotroph is defined as any organism that uses an inorganic source of carbon and makes its own organic matter as opposed to a heterotroph, which uses pre-made organic matter. The most familiar autotrophs are plants that use energy from sunlight to convert carbon dioxide into sugar derivatives. However, a variety of bacteria exist that fix carbon dioxide without light but instead rely on other sources of energy. Furthermore, the pathways of carbon dioxide fixation vary. In particular, some autotrophic bacteria incorporate carbon dioxide into carboxylic acids rather than generating sugar derivatives like plants.

The autotrophic theory of the origin of life postulates the chemical oxidation of readily available iron compounds as the primeval energy source. In particular, the conversion of ferrous sulfide (FeS) to pyrite (FeS_2) by hydrogen sulfide (H_2S) releases energy and provides H atoms to reduce carbon dioxide to organic matter. [Anaerobic bacteria are found today that generate energy by the oxidation of iron Fe^{2+} compounds to Fe^{3+}, as well as others that generate energy by oxidizing sulfur compounds. Thus a primeval metabolism based on iron and sulfur seems reasonable.]

An alternative theory suggests that energy released by the reaction of sulfur compounds powered the earliest life forms.

Several possible schemes have been suggested for the first carbon dioxide fixation reactions. One scheme involves iron catalyzed insertion of CO_2 into sulfur derivatives of those carboxylic acids still found today as metabolic intermediates (e.g., acetic acid, pyruvic acid, succinic acid, etc.). These early reactions would have occurred on the surface of iron sulfide minerals buried underground, rather than in a primeval soup. This leaves open the question of where such organic acids came from originally. One possibility is that they resulted from a Miller type synthesis, as described above. More radical is the suggestion that the first organic molecules were derived directly from carbon monoxide plus hydrogen sulfide. It has been demonstrated that a mixed FeS/NiS catalyst can convert carbon monoxide (CO) plus methane thiol (CH_3SH) into a thioester (CH_3-CO-SCH_3), which then hydrolyses into acetic acid. Inclusion of catalytic amounts of selenium allows conversion of CO plus H_2S alone to CH_3SH (and then to the thioester and acetic acid).

Organic matter can be generated from simple gas molecules using metallic catalysts.

Recently it was shown that carbon monoxide (CO) activated by the same mixed FeS/NiS catalyst can also drive the formation of peptide bonds between alpha-amino acids in hot aqueous solution. Not surprisingly, this system will also hydrolyze polypeptides.

Evolution of DNA, RNA and Protein Sequences

Consider the genes of an ancient ancestral organism. Over millions of years, mutations will occur in the DNA sequences of its genes at a slow but steady rate (see Ch. 13). Most mutations will be selected against because they are detrimental, but some will survive. Most mutations that are incorporated permanently into the genes will be neutral mutations with no harmful or beneficial effects on the organism. Occasionally, mutations that improve the function of a gene and/or the protein encoded by it will occur, although these are relatively rare. Sometimes a mutation that was originally harmful may turn out to be beneficial under new environmental conditions.

Due to mutation, the sequences of DNA and the encoded proteins will gradually change over long periods of time.

The actual function of the protein matters the most, not the exact sequence of the gene. If the protein can still operate normally, a mutation in the gene may be acceptable. Many of the amino acids making up a protein chain can be varied, within reasonable limits, without damaging the function of the protein too much. Replacement of one amino acid by a similar one (i.e. a conservative substitution—see Chapter 13) will rarely abolish the function of a protein. If we compare the sequences of the same protein taken from many different modern-day organisms we will find that the sequences can be lined up and are very similar. For example, the α chain of hemoglobin is identical in humans and chimpanzees. Yet, 13% of the hemoglobin amino acids are different in pigs compared with humans, 25% are different in chickens, and 50% are different in fish. This divergence in sequence is much what is expected from other estimates of evolutionary relatedness. Fig. 20.09 has an example of one of these alignments. It shows a highly conserved iron binding site found in a related family of enzymes—a group of alcohol dehydrogenases found in microorganisms—that use iron in their active site mechanism.

Related organisms contain genes and proteins with related sequences. These may be used to construct evolutionary trees.

It is possible, then, to construct an evolutionary tree using a set of sequences for a protein as long as it is found in all the creatures being compared. The α chain of hemoglobin is only found in our blood relatives. In contrast, cytochrome c is a protein involved in energy generation in all higher organisms, including plants and fungi. It even has recognizable relatives in many bacteria. A cytochrome c tree is shown in Figure 20.10. Humans and fish differ in amino acid sequence by only 18% for cytochrome c. From humans to either plants or fungi gives about 45% divergence. However plants and fungi also differ by 45%, which tells us that, by this measure, plants have diverged as far from fungi as animals have from plants.

Individual mutations may revert and restore the ancestral sequence of a gene or protein at a particular location. However, genes almost never mutate backwards to resemble the ancestors they diverged from many mutations ago. This is essentially a matter of probability. There is nothing forbidding any particular mutation to revert to

FIGURE 20.09 ***Alignment of Related Sequences***

Amino acid sequences of related polypeptides are given in one letter code. The conserved iron binding site is shown in bold. In particular, the iron atom is bound by the two conserved histidine (H) residues. The glycerol dyhydrogenase from *Bacillus* is related to the other members of this protein family but no longer uses iron and, as can be seen, the iron binding sequence has diverged and both histidines have been replaced by other amino acids.

```
ECO   NPVARERVHS  AATIAGIAFA  NAFLGVCHSM  AHKLGSQFHI  PHGLANALLI  CHNVIRYNAND
SALM  NPVARERLHS  AAYIAPIAFA  NAFLGVCHWM  AHKLPAQLHI  PHGPFNARYR  HSVRR   AQS
CLOE  NEKAREKMAH  ASTMAGMASA  NAFLGLCHSM  AIKLSSEHNI  PSGIANALLI  EEVIRKFNAVD
CLOB  E  AREQMHY  AQCLAGMAFS  NALLGICHSM  AHKTGAVFHI  PHGCANAIYL  PYVIKFNSKT
ZYMMO DMPAREAMAY  AQFLAGMAFN  NASLGYVHAM  AHQLGGYYNL  PHGVCNAVLL  PHVLAYNASV
ENTH  DLEAREKMHN  AATIAGMAFA  SAFLGMDHSM  AHKVGAAFHL  PHGRCVAVLL  PYHVIRYNGQ
YEAST DKKARTDMCY  AEYLAGMAFN  NASLGYVHAL  AHQLGGFYHL  PHGVCNAVLL  PHVQ  EANM
PRD   D  AGEEMAL  GQYVAGMGFS  NVGLGLVHGM  AHPLGAFYNT  PHGVANAILL  PHVMRYNADF
GLD   EKCEQTLFKY  GK             LAYESVK  AKVVTPALE     AVVEANTL  LSGLGFESGG

                               IRON BINDING SITE
```

ECO = *E. coli* alcohol dehydrogenase
SALM = *Salmonella typhimurium* alcohol dehydrogenase
CLOB = *Clostridium acetobutylicum* butanol dehydrogenase
CLOE = *C. acetobutylicum* alcohol dehydrogenase
ZYMMO = *Zymomonas mobilis* alcohol dehydrogenase II
ENTH = *Entamoeba histolytica* alcohol dehydrogenase
YEAST = Yeast alcohol dehydrogenase IV
PRD = *E. coli* propanediol dehydrogenase
GLD = *Bacillus* glycerol dehydrogenase (a zinc enzyme)

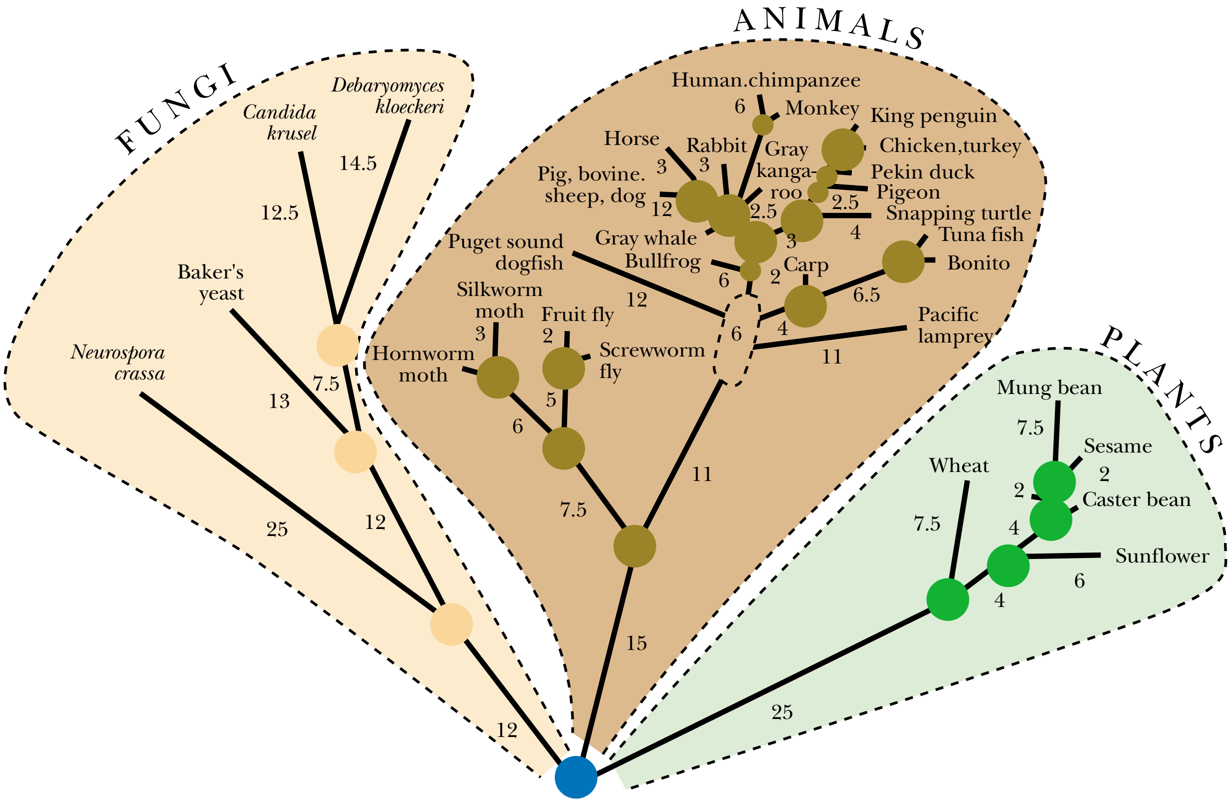
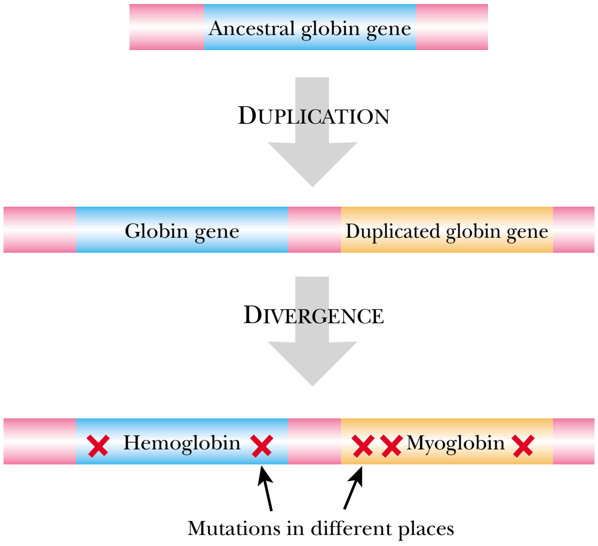


FIGURE 20.10 ***Evolutionary Tree based on Cytochrome c***

All three kingdoms, fungi, animals, and plants, have the gene for cytochrome c, which is involved in production of energy. This phylogenetic tree was constructed by comparing the amino acid sequence of cytochrome c from each of these organisms. The closer the branches are together, the more similar the sequences. The numbers along the branches represent the number of amino acid differences.

FIGURE 20.11 ***Duplication Creates New Genes***

During evolution, the entire ancestral globin gene was duplicated. The two copies were then able to diverge independently of each other. The first gene developed into hemoglobin, which is only found in red blood cells. The second copy developed into myoglobin, which is found in muscle tissue. Both proteins still carry oxygen, but they have tissue specificity.



the original sequence, but the likelihood of precisely reversing each of many mutations is infinitesimally small.

Creating New Genes by Duplication

Duplication of segments of DNA creates a supply of new genes.

How do new genes arise? The standard way is via gene duplication. As discussed in chapter 13, mutations may cause the duplication of a segment of DNA that carries a whole gene or several genes. The original copy must be kept for its original function but the extra copy is free to mutate and may be extensively altered. In most cases the mutations that accumulate will inactivate the duplicate copy. Less often, the extra copy will remain active and be altered so as to perform a related but different function from the original copy.

The steady accumulation of mutations in duplicated genes results in families of genes with related sequences.

Multiple duplication followed by sequence divergence may result in a family of related genes that carry out related functions. One of the best examples is the **globin** family of genes. Hemoglobin carries oxygen in the blood, whereas myoglobin carries it in muscle. These two proteins have much the same function, have similar 3-D shapes, and their sequences are related. After the ancestral globin gene duplicated, the two genes for hemoglobin and myoglobin slowly diverged as they specialized to operate in different tissues (Fig. 20.11).

The actual hemoglobin of mammalian blood has two alpha (α)-globin and two beta (β)-globin chains forming an α2/β2 tetramer, unlike myoglobin, which is a monomer of a single polypeptide chain. The α-globin and β-globin were derived by further duplication of the ancestral hemoglobin gene. In addition, the ancestral α-globin gene split again, to give modern α-globin and zeta (ζ)-globin. The ancestral β-globin gene split again, twice, to give modern β-globin and the gamma (γ)-, delta (δ)-, and epsilon (ε)-globins (Fig. 20.12).

These globin variants are used during different stages of development. At each stage, the hemoglobin tetramer consists of two α-type and two β-type chains. The ζ-globin and ε-globin chains appear in early embryos, which possess ζ2/ε2 hemoglobin. In the fetus, the ε-chain is replaced by the γ-chain and the ζ-chain is replaced by the α-chain, so giving α2/γ2 hemoglobin. A fetus needs to attract oxygen away from the mother's blood, so the α2/γ2 hemoglobin binds oxygen better than the adult α2/β2 hemoglobin (Fig. 20.12).

globins Family of related proteins, including hemoglobin and myoglobin, that carry oxygen in the blood and tissues of animals

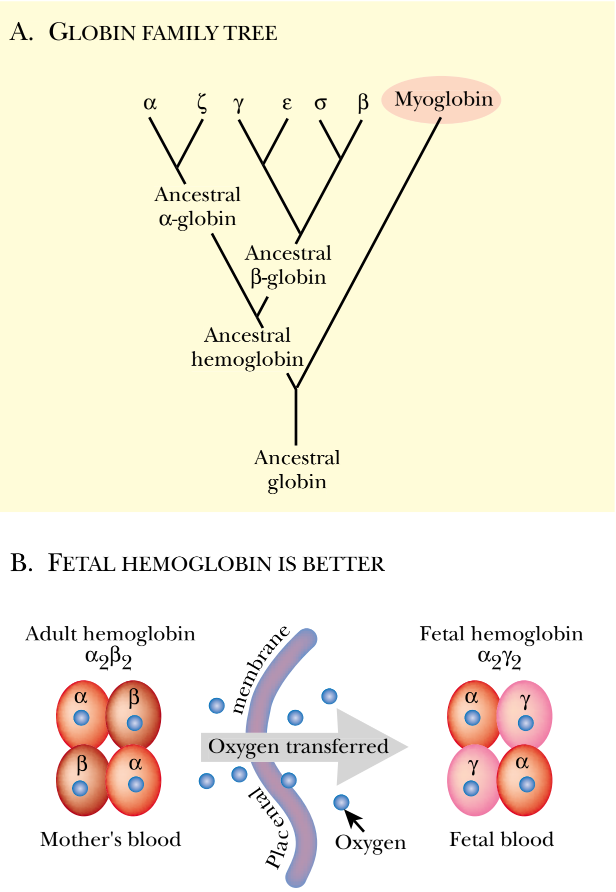


FIGURE 20.12 ***Origin of Globin Family of Genes***

(A) Over the course of evolution, a variety of gene duplication and divergence events gave rise to a family of closely related genes. The first ancestral globin gene was duplicated giving hemoglobin and myoglobin. After another duplication, the hemoglobin gene diverged into the ancestral α-globin and ancestral β-globin genes. Continued duplication and divergence created the entire family of globin genes. (B) The different members of the hemoglobin family are adapted for specific functions during development. Thus the fetus uses two α chains and two γ chains to form its hemoglobin tetramer. This form is able to extract the oxygen from the mother's blood because it has a higher affinity for oxygen than the adult form of hemoglobin.

The globin genes are an example of a **gene family**, a group of closely related genes that arose by successive duplication. The individual members are obviously related in their sequences and carry out similar roles. During evolution, continued gene duplication may give rise to multiple new genes whose functions steadily diverge until their ancestry may be difficult to recognize; this gives a **gene superfamily**. The genes of the immune system provide good examples of gene families and superfamilies.

Reverse transcriptase may generate duplicate genes that lack introns and promoters and are located far away from the original copy.

In eukaryotes, retro-elements that encode reverse transcriptase are relatively common (see Ch. 15). Consequently, occasional reverse transcription of cellular mRNA molecules may occur. This gives a complementary DNA copy that may be integrated into the genome. This results in a duplicate copy of the gene, although this lacks the introns and promoter of the original gene. Such inactive copies are known as pseudogenes and usually accumulate mutations that inactivate the coding sequence. Rarely, a pseudogene may end up next to a functional promoter and be expressed. This gives a duplicate functional copy of the original gene that may be altered by mutation as already discussed.

Occasionally whole genomes may be duplicated.

Rare mistakes during cell division may result in the whole genome being duplicated. In particular, errors in meiosis may give diploid gametes. Fusion of two diploid gametes would give a tetraploid zygote and hence a tetraploid individual. More often, a triploid individual forms by fusion of one mutant diploid gamete plus one normal haploid gamete. Most triploids are sterile, as they give gametes with incorrect numbers of chromosomes. But occasionally triploids will be able to generate tetraploid progeny. Aberrant ploidy levels are fairly common in plants. Around 5 in 1000 plant gametes

gene family Group of closely related genes that arose by successive duplication and perform similar roles
gene superfamily Group of related genes that arose by several stages of successive duplication. Members of a superfamily have often diverged so far that their ancestry may be difficult to recognize

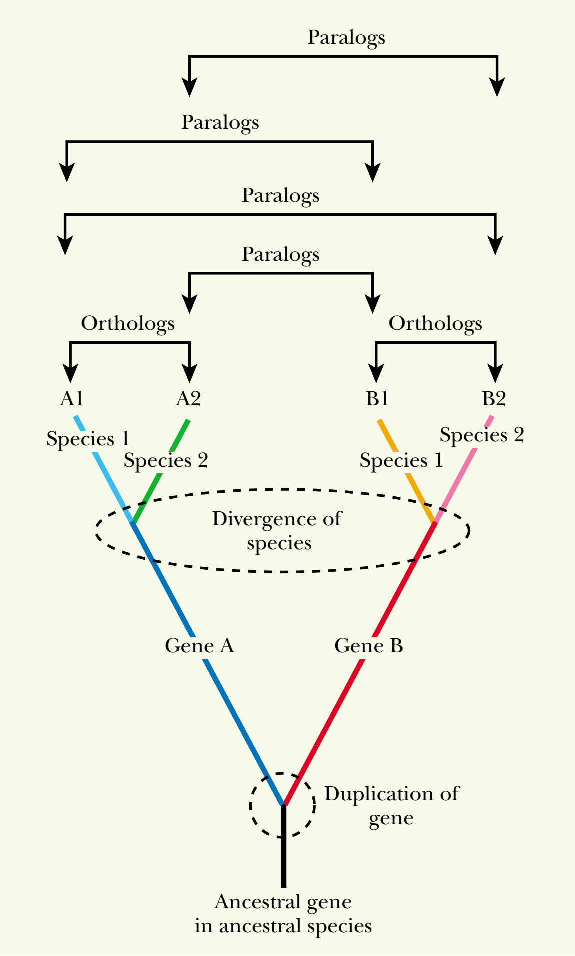


FIGURE 20.13 ***Paralogous and Orthologous Sequences***

In this example, an ancestral gene duplicated and diverged into genes A and B, which by definition are paralogs. These two genes were both present when the ancestral species diverged into species 1 and species 2. Thus both species 1 and species 2 have genes for A and B (referred to as A1 & B1 and A2 & B2 respectively). Each such pair are still paralogs. However, since species 1 and 2 are now two separate species, the A1 and A2 genes are orthologs, and the B1 and B2 genes are also orthologs.

are diploid. Therefore, in a cross between two different parents, approximately 2.5 in 10^{-5} zygotes will be tetraploid. Over time, the duplicate copies of genes in a tetraploid organism will gradually diverge. Eventually, once the duplicate copies diverged far enough to be distinct and have assumed new functions, the organism will effectively become "diploid" again.

Paralogous and Orthologous Sequences

A gene may be related to similar genes in other organisms and also to other genes of similar sequence within the same organism.

Sequences are said to be **homologous** when they share a common ancestral sequence. If several organisms each contain single copies of a particular gene that all derive from the same common ancestor then sequence comparison should give an accurate evolutionary tree. However, gene duplication may result in multiple copies of the same gene within a single organism. These alternatives are illustrated in Fig. 20.13. **Orthologous** genes are those found in separate species and which diverged when the organisms containing them diverged. **Paralogous** genes are multiple copies located within the same organism due to gene duplication.

To generate an accurate evolutionary tree orthologous genes must be compared. For example, we must compare the sequences of α-globin from one animal with the orthologous α-globin from another and not with the paralogous β-globin. Since paralogous sets of genes have similar sequences, these may cause confusion unless their

homologous Related in sequence to an extent that implies common genetic ancestry
orthologous genes Homologous genes that are found in separate species and which diverged when the organisms containing them diverged
paralogous genes Homologous genes that are located within the same organism due to gene duplication

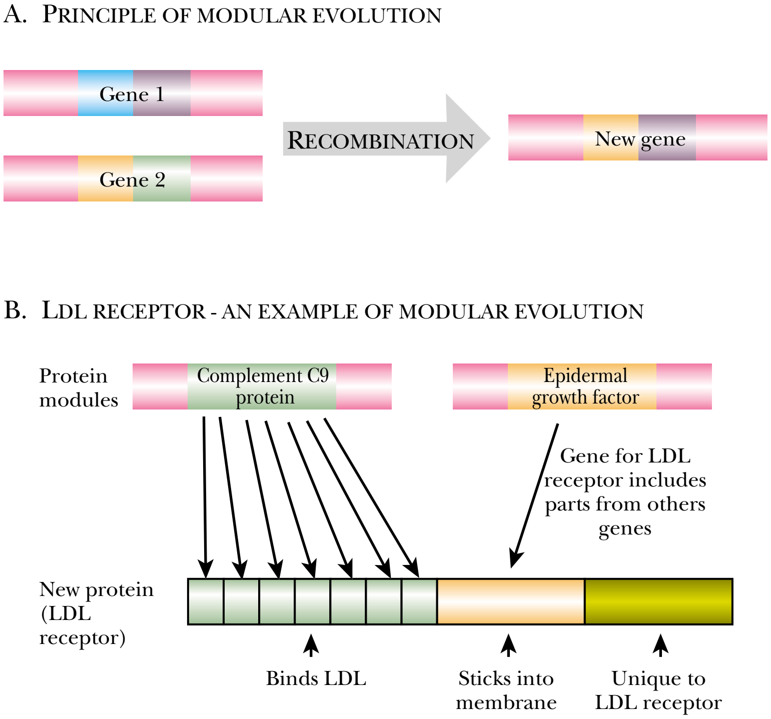

A. PRINCIPLE OF MODULAR EVOLUTION
Gene 1
Gene 2
RECOMBINATION
New gene
B. LDL RECEPTOR - AN EXAMPLE OF MODULAR EVOLUTION
Protein modules
Complement C9 protein
Epidermal growth factor
Gene for LDL receptor includes parts from others genes
New protein (LDL receptor)
Binds LDL
Sticks into membrane
Unique to LDL receptor

FIGURE 20.14 ***Novel Gene Derived from Pre-Existing Modules***

(A) Modular evolution of a new gene may involve the fusion of separate gene modules, or functional units. For example, the orange module of gene 1 may provide a function that complements the purple module of gene 2. If these two domains fuse they might then form a novel but functional gene. (B) The LDL receptor has domains found in several other proteins. The first part of the LDL receptor gene has seven repeated modules or functional units also found in the C9 complement factor of the immune system. The next module allows the protein to bind the cell membrane. This module is very similar to a portion of the epidermal growth factor receptor. Finally, the LDL receptor has a module that is unique.

correct ancestry is known. In particular, we need to know whether or not an organism contains multiple sequences derived from the same ancestor. Suppose that, due to partial information, we only knew of α-globin from pigs and β-globin from dogs. If we were unaware of the presence of other members of the globin family in these organisms we might compare these two sequences as if they were orthologous. This would lead to an incorrect relationship.

Creating New Genes by Shuffling

New genes may be made by shuffling of modular segments of DNA.

Another way to create new genes is by using pre-made modules. Segments from two or more genes may be fused together by DNA rearrangements so generating a novel gene consisting of regions derived from several sources (Fig. 20.14A). An example of the formation of a new gene from several diverse components is the LDL receptor (Fig. 20.14B). LDL is low-density lipoprotein that carries cholesterol around in the blood. The LDL receptor is found on the surface of cells that take up LDL. The gene for the LDL receptor consists of several regions, two of which are derived from other genes. Towards the front are seven repeats of a sequence also appearing in the C9 factor of complement, an immune system protein. Farther along is a segment related to part of epidermal growth factor (a hormone). When such a gene mosaic is transcribed and translated, we get a patchwork protein consisting of several different domains.

Different Proteins Evolve at Very Different Rates

The sequences of different genes and proteins evolve at different rates. Less critical sequences are free to evolve faster.

Obviously, we should not rely on a single protein to build an evolutionary tree. If we make trees for several proteins, we often get rather similar evolutionary relationships. However, different proteins evolve at different speeds. As noted above, humans and fish differ by 50% in the α chain of hemoglobin but by less than 20% in their cytochrome c. If we plot the number of amino acid changes versus the evolutionary time scale (Fig. 20.15), we can see this easily for cytochrome c (slow), hemoglobin (both α and β chains evolve at medium speed), and fibrinopeptides A and B (rapid evolution).

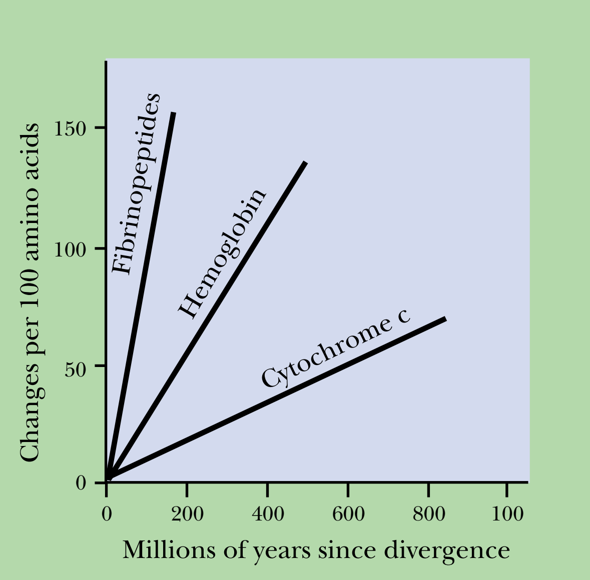


FIGURE 20.15 ***Rates of Protein Evolution***

During the course of evolution, some proteins accumulate more mutations than others. The cytochrome c gene is very stable, and only 50 changes/100 amino acids have occurred in 800 million years. Fibrinopeptides A and B, on the other hand, have accumulated 50 changes/100 amino acids in less than 100 million years.

TABLE 20.05 Rates of Evolution for Different Proteins

Protein	Rate of Evolution
Neurotoxins	110–125
Immunoglobulins	100–140
Fibrinopeptide B	91
Fibrinopeptide A	59
Insulin C peptide	53
Lysozyme	40
Hemoglobin α chain	27
Hemoglobin β chain	30
Somatotropin	25
Insulin	7.1
Cytochrome c	6.7
Histone H2	1.7
Histone H4	0.25

The rate of evolution is given as the number of mutations per 100 amino acid residues per 100 million years.

Table 20.05 gives the evolutionary rates for an assortment of proteins. Fibrinopeptides are involved in the blood clotting process. They need an arginine at the end and must be mildly acidic overall. Apart from this they can vary widely as there are so few constraints on what is needed. In contrast, histones bind to DNA and are responsible for its correct folding. Almost all changes to a histone would be lethal for the cell, so they evolve extremely slowly.

Cytochrome c is an enzyme whose function depends most critically on a few amino acid residues at the active site, which bind to its heme cofactor. Consequently, these active site residues rarely vary, even though amino acids around them change. Of 104 residues, only Cys-17, His-18 and Met-80 are totally invariant. In other places variation is low; large, nonpolar, amino acid residues always fill positions 35 and 36. Several cytochrome c molecules have been examined by X-ray crystallography and all have the same 3-D structure. Although cytochrome c molecules may vary by as many as 88% of their residues, they retain the same 3-D conformation. Thus, little variation is seen with the amino acids that are essential to the function or structure of cytochrome c.

Insulin is a hormone that evolves at much the same rate as cytochrome c. Insulin consists of two protein chains (A and B) encoded by a single insulin gene. During protein synthesis, a long pro-insulin molecule is made. This has the middle, the C-peptide, cut out and discarded. Disulfide bonds hold the A and B chains together. Since the C chain is not part of the final hormone, it is free to evolve much faster and it changes at almost 10 times the rate of the A and B chains. Notice that all these proteins maintain the critical residues throughout evolution. It is important to note that mutations are random. A mutation is just as likely to occur in the A, B, and C portion of the insulin gene. The mutations that do occur in the A and B portions may be detrimental to the organism, therefore, these mutations never get passed to a new generation. On the other hand mutations in the C portion occur, do not harm the function of the protein, and get passed onto the progeny.

Molecular Clocks to Track Evolution

A rapidly evolving protein will eventually become so altered in sequence between diverging organisms that the relationship will no longer be recognizable. Conversely, a protein that evolves very slowly will show little or no difference between two different organisms. Therefore, we need to use slowly changing sequences to work out distant evolutionary relationships and fast-evolving sequences for closely related organisms.

Rapidly evolving sequences reveal the relationships between closely related organisms. Slowly changing sequences are needed to compare distantly related organisms.

Most human proteins have identical sequences to those of the closely related chimpanzee. Even if we examine the rapidly evolving fibrinopeptides, humans and chimpanzees end up on the same branch of the evolutionary tree. So how can we tell people apart from chimps? Mutations that do not affect the sequence of proteins accumulate much faster during evolution, since they have little or no detrimental effect. So if instead of protein sequences we look at the DNA sequences of very closely related organisms we find many more differences. These are found mainly in non-coding sequences and in the third codon position. As discussed in Chapter 8, changing the third base of most codons does not alter the encoded amino acid. So changing the DNA sequence at the third base of most codons leaves the encoded protein unaltered (Fig. 20.16).

Introns are non-coding sequences that are spliced out of the primary transcript and so do not appear in the messenger RNA (see Ch. 12). Intron sequences are therefore not represented in the final protein. Apart from the intron boundaries and splice recognition sites, the DNA sequence of an intron is free to mutate extensively. Other non-coding sequences exist between genes and, if not involved in regulation, they are also relatively free to mutate.

The early data on cytochrome c, hemoglobin, etc., were obtained by direct sequencing of proteins. Since DNA sequencing is much easier and more accurate, most of the recently discovered protein sequences are deduced from the DNA sequence. Hence, we have a lot of DNA information on closely related animals. Using this data will help fine-tune the evolutionary relationship between animals such as humans and chimpanzees.

Ribosomal RNA—A Slowly Ticking Clock

Ribosomal RNA sequences have been used to construct a global evolutionary tree.

A major problem is how to construct an evolutionary tree that includes all living organisms and shows the relationships between all the main groups of organisms. To achieve this, first we need a molecule that is present in all organisms. Second, the chosen molecule must evolve slowly so as to still be recognizable in all the major groups of living things.

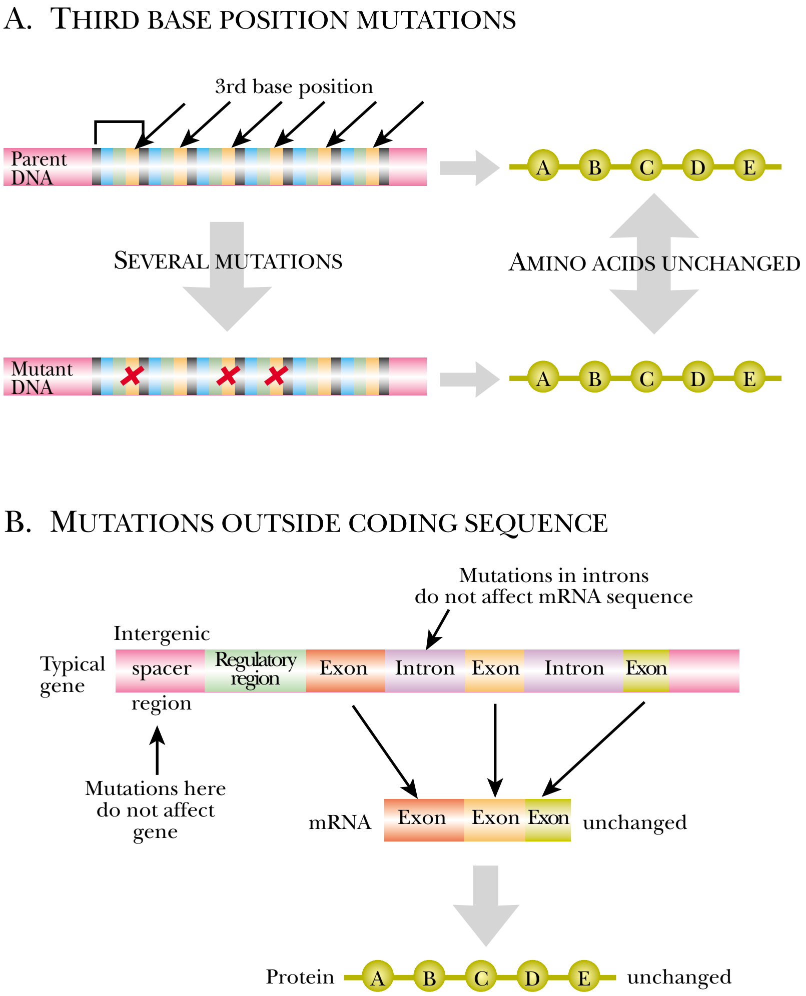


FIGURE 20.16 ***Non-Coding DNA Evolves Faster***

(A) During evolution mutations in the third position of the triplet codon rarely alter the amino acid sequence of the protein; therefore they are rarely deleterious. (B) Non-coding DNA regions often have no obvious function and so many mutations accumulate within these regions. Mutations in intragenic spacer regions or in introns have no effect on protein sequence or function.

Although histones evolve very slowly, only eukaryotic cells possess them; they are missing from bacteria. The solution is to use ribosomal RNA. In practice, the DNA of the genes that encode the RNA of the small ribosomal subunit (16S or 18S rRNA) is sequenced and the rRNA sequence is then deduced. All living organisms have to make proteins and they all have ribosomes. Furthermore, since protein synthesis is so vital, ribosomal components are highly constrained and evolve slowly. The only group excluded is the viruses, which have no ribosomes. (Whether viruses are truly alive is debatable and their evolutionary origins are still controversial; see Ch. 17.)

Use of relationships based on ribosomal RNA has allowed the creation of large-scale evolutionary trees encompassing all the major groups of organisms. Higher organisms consist of three main groups—plants, animals and fungi (Fig. 20.17). Analysis of rRNA indicates that the ancestral fungus was never photosynthetic but split off from the plant ancestor before the capture of the chloroplast. Despite traditionally being studied by botanists, fungi are actually more closely related to animals than plants. A variety of single celled organisms sprout off the eukaryotic tree near the bottom and do not fall into any of the three major kingdoms.

Sequence analysis indicates that fungi are closer to being immobile animals than non-photosynthetic plants.

As discussed in Chapter 19, most eukaryotic cells contain mitochondria and, in addition, plant cells contain chloroplasts. These organelles are derived from symbiotic bacteria and contain their own ribosomes. The rRNA sequences of mitochondria and chloroplasts reveal their relationship to the bacteria. Relationships among eukaryotes, like that shown in Fig. 20.17, are therefore made by using the rRNA of the ribosomes found in the cytoplasm of eukaryotic cells. These ribosomes have their ribosomal RNA coded for by genes in the cell nucleus.

Sequencing of rRNA reveals that chloroplasts and mitochondria are related to the bacteria.

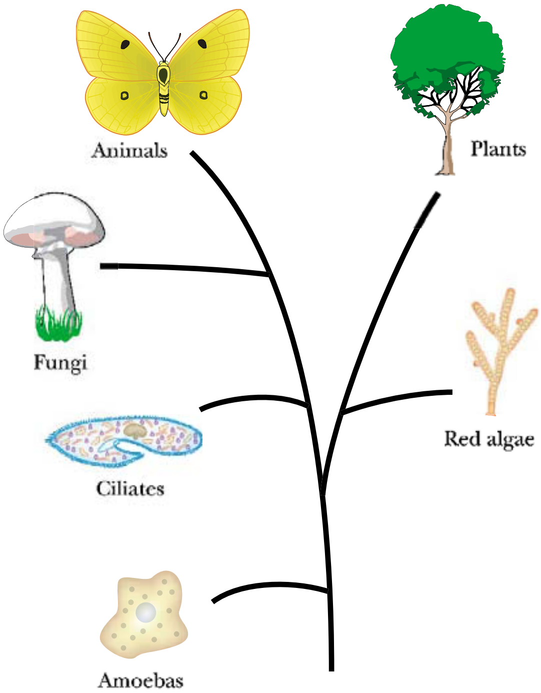


FIGURE 20.17 ***Eukaryote Kingdoms Based on Ribosomal RNA Sequences***

Comparing the ribosomal RNA sequences has allowed scientists to deduce how closely the major divisions of organisms are related. Protozoans such as amoebas were the earliest groups to branch from the ancestral eukaryotic lineage. A division between photosynthetic and non-photosynthetic organisms occurred next. The two branches evolved separately, the photosynthetic branch formed algae and higher plants, whereas the non-photosynthetic branch developed into ciliates, fungi and animals.

Life consists of three domains—the eubacteria, the archaebacteria and the eukaryotes.

When an rRNA-based tree is made that includes both prokaryotes and eukaryotes it turns out that life on Earth consists of three lineages (Fig. 20.18). These three domains of life are the **eubacteria** ("true" bacteria, including the organelles), the **archaea** or **archaebacteria** ("ancient" bacteria) and the eukaryotes. There is as much difference between the two genetically distinct types of prokaryote as between prokaryotes and eukaryotes. Sequencing of organelle rRNA indicates that mitochondria and chloroplasts belong to the eubacterial lineage.

One bizarre aspect of classifying life forms by ribosomal RNA is that the organism itself is not needed. A sample of DNA containing the genes for 16S rRNA is sufficient. Although many microorganisms present in the sea or in soil have never been successfully cultured, DNA can be extracted from the soil or seawater directly. Using PCR (see Ch. 23) it is possible to amplify the DNA from a single cell and get enough of the 16S rRNA gene to obtain a sequence. Several new groups of bacteria that branched off very early from the archaebacterial lineage have been discovered by this method, despite not being successfully cultured.

The Archaebacteria versus the Eubacteria

Although most common bacteria are eubacteria, there is another group, the archaebacteria or archaea as they have recently been renamed. Both types of bacteria have microscopic cells without a nucleus. They both have single circular chromosomes and divide in two by simple binary fission. In short, they both conform to the definition of a prokaryotic cell and there was no obvious reason to suspect from their superficial structure that they were so radically different. However, sequence analysis of ribosomal RNA indicates that there is about as much genetic difference between the eubacteria and archaea as between either of these two groups and eukaryotic cells.

The archaebacteria are somewhat closer to the eukaryotes than to the eubacteria.

Of the two groups of prokaryotes, the archaea are probably slightly more closely related to the **urkaryote**, the primeval ancestor of the eukaryotic nucleus. Some

archaebacteria One of the three domains of life comprising the "ancient" bacteria
archaea New name for archaebacteria, one of the three domains of life
eubacteria One of the three domains of life comprising the "true" bacteria, including the organelles
urkaryote The hypothetical primeval ancestor of the eukaryotic nucleus

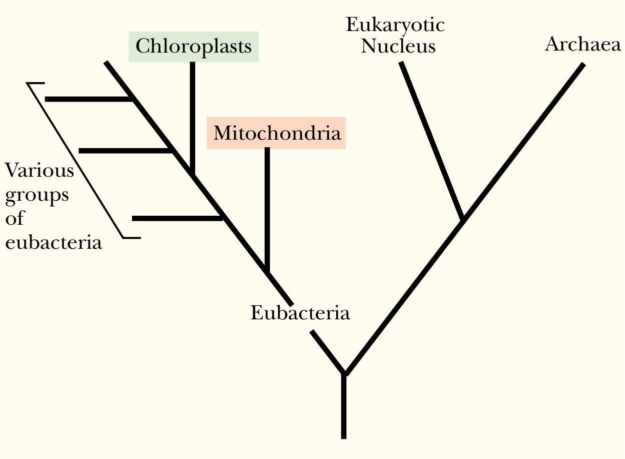


FIGURE 20.18 ***The Three Domains of Life***

All of today's organisms belong to one of three main divisions based on relationships among ribosomal RNA: the eubacteria, the archaebacteria (or archaea) and the eukaryotes. Mitochondria and chloroplasts have rRNA that most closely resembles eubacteria.

archaea have their DNA packaged by histone like proteins that show some sequence homology to the true histones of higher organisms. In addition, the details of protein synthesis and the translation factors of archaea resemble those of eukaryotes, rather than eubacteria. These similarities have led to the suggestion that the primeval eukaryote evolved from an archaeal ancestor.

Archaea differ biochemically from eubacteria in several other major respects. Archaea have no peptidoglycan and their cytoplasmic membrane contains unusual lipids, which are made up from C5 isoprenoid units rather than C2 units as with normal fatty acids (Fig. 20.21). Moreover, the isoprenoid chains are attached to glycerol by ether linkages instead of esters. Some double-length isoprenoid hydrocarbon chains stretch across the whole membrane.

Archaea tend to be found in bizarre environments and many of them are adapted to extreme conditions. They are found in hot sulfur springs, thermal vents in the ocean floor, in the super salty Dead Sea and Great Salt Lake and also in the intestines of cows (and other animals) where they make methane. Some examples of archaea are:

Halobacteria: These are extremely salt-tolerant and grow in up to 5M NaCl but will not grow below 2.5M NaCl (sea water is only 0.6M). They trap energy from sunlight by using bacterio-rhodopsin, a molecule related to the rhodopsin pigment used as a photo-detector in animal eyes.

Methanogens (methane producing bacteria): These are obligate anaerobes and are very sensitive to oxygen. They convert H_2 plus CO_2 to CH_4 (methane). Their metabolism is unique—they contain coenzymes found in no other living organisms but have no typical flavins or quinones.

Sulfolobus: Lives in geothermal springs and grows best at a pH optimum of 2–3 and a temperature of 70–80°C. These archaea oxidize sulfur to sulfuric acid. Many other sulfur-metabolizing archaea are also found living under a variety of extreme conditions.

DNA Sequencing and Biological Classification

Before the sequencing of DNA became routine, animals and plants were classified reasonably well, fungi and other primitive eukaryotes were classified poorly, and bacterial classification was a lost cause due to lack of observable characters. Using gene sequences for classification was developed for bacteria and has since spread to other types of organism. Nowadays ancestries may be traced by comparing the sequences of DNA, RNA or proteins that are more representative of fundamental genetic relationships than are many superficial characteristics. Furthermore, in situations where division into species, genera, families, etc., is arbitrary, sequence data can provide

Instant Evolution of Ribosomal RNA

Consider an essential molecule that evolves slowly, such as histones or ribosomal RNA. It is possible that certain combinations of two mutations might yield a functional molecule, but that either alone would be lethal. For example, a mutation from G to C that destroyed a critical GC base pair in a stem loop structure might be fatal in 16S rRNA. However, replacing the GC with a CG base pair might well allow function (Fig. 20.19). During normal evolution, this replacement is highly unlikely since either single mutation is lethal and the likelihood of simultaneous mutations in just these two bases is very low.

Consequently, a CG base pair in this particular position will probably be very rare among the 16S rRNA sequences of existing life forms. To fully analyze the relationship of structure and function in a molecule such as rRNA, several artificial mutations must be introduced simultaneously. This may be done by a procedure known as "instant evolution" that was developed in the laboratory of Dr. Philip R. Cunningham at Wayne State University. In this approach, 16S rRNA is mutated and those mutations that prevent protein synthesis are isolated. Next, suppressor mutations that restore protein synthesis are selected. Alternatively, multiple random mutations may be simultaneously introduced into a short region of the rRNA that is suspected to play an important role in protein synthesis. In either case, most of these mutations in rRNA would be lethal under normal circumstances; to avoid killing the bacteria they must be manipulated so that the mutant form of the 16S rRNA does not interfere with normal cellular protein synthesis.

The following technology was developed to prevent the mutated form of rRNA from affecting the normal bacterial functions (Fig. 20.20):

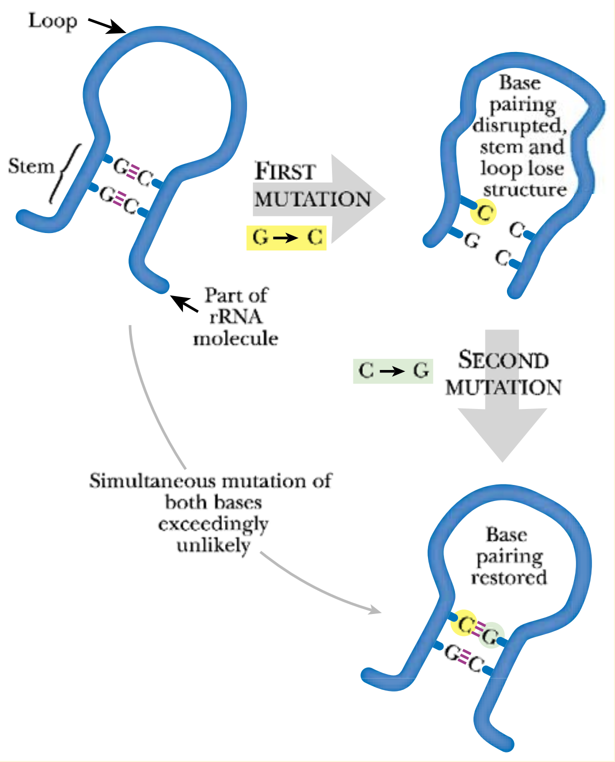

Loop
Stem
G≡C
G≡C
Part of rRNA molecule
FIRST MUTATION
G → C
Base pairing disrupted, stem and loop lose structure
C
G
C
C
C → G
SECOND MUTATION
Simultaneous mutation of both bases exceedingly unlikely
Base pairing restored
C≡G
G≡C

FIGURE 20.19 ***Single Change Lethal, Two Changes Functional***

Mutations in ribosomal RNA are lethal since they alter the stem-loop structure of the molecule. In this example, mutating the guanine to cytosine prevents a critical base pair from completely the stem portion. If a second mutation were to change the cytosine to a guanine, the base pair would reform and the ribosomal RNA would be functional again. Although changing both positions simultaneously is extremely unlikely, this would not be detrimental to the organism and the mutation would be passed onto successive generations.

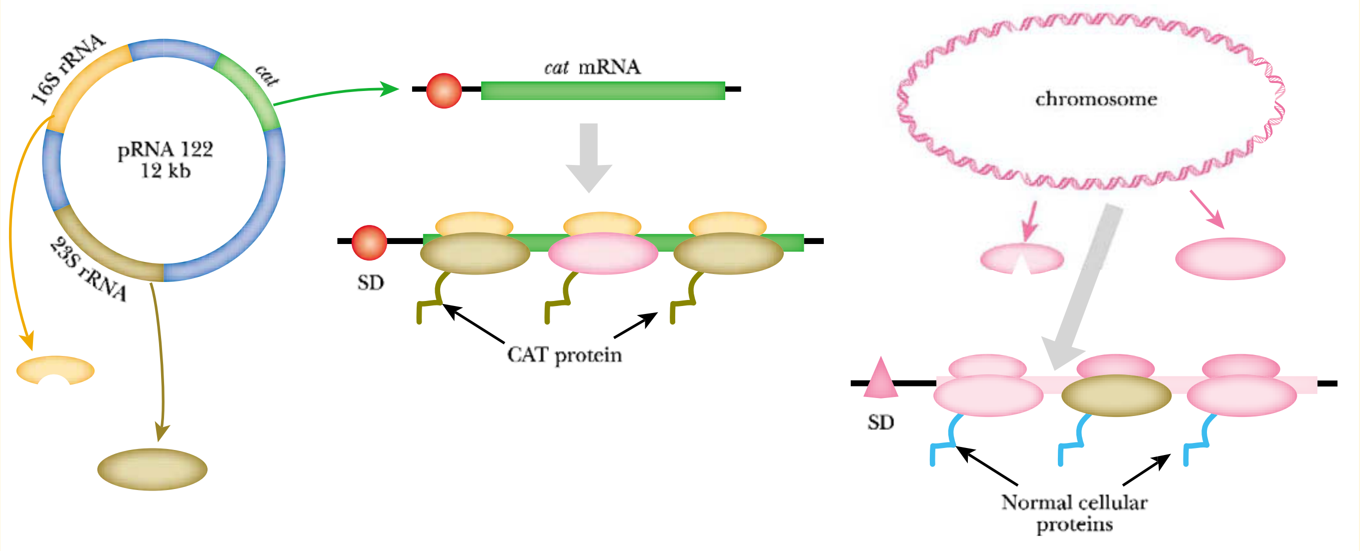


FIGURE 20.20 ***Instant Evolution of Ribosomal RNA***

The plasmid pRNA122 carries the sequence for altered 16S rRNA and the reporter gene (in this example, CAT). Separation of chromosomal and plasmid-directed protein synthesis occurs primarily due to the changes in the Shine-Dalgarno (SD) sequence. The 16S rRNA encoded by the plasmid cannot recognize the SD sequence of normal cellular mRNA, but does recognize the SD sequence upstream of the reporter gene (*cat*). If a mutation in the 16S rRNA prevents it from functioning, the reporter gene is not translated into CAT protein, and therefore the bacteria are not chloramphenicol resistant. Translation of normal cellular proteins occurs without interruption because the chromosomal copy of the 16S rRNA (small pink oval) is used. The chromosomal copy of the 16S rRNA does not recognize the SD sequence of the *cat* mRNA, and so does not allow synthesis of CAT protein.

a. A copy of the 16S rRNA gene was put on a plasmid and mutated. Since the genomic copy of the 16S rRNA is still functional, most of the cell's ribosomes will be normal. Only a fraction of the ribosomes will have the mutant 16S rRNA.

b. The anti-Shine-Dalgarno sequence of the plasmid 16S rRNA is altered so that it no longer recognizes normal cellular mRNA, thus lethal mutations in this 16S rRNA copy will not interfere with normal protein synthesis.

c. A reporter gene is designed with an altered Shine-Dalgarno sequence that matches the plasmid or mutated form of the 16S rRNA. Thus only the translation of the mRNA from the reporter gene responds to the mutations in the plasmid-borne copy of the 16S rRNA. Two different reporter genes were used, chloramphenicol acetyl transferase (CAT), which gives chloramphenicol resistance to the bacteria, and green fluorescent protein (GFP), which causes the bacteria to turn green under fluorescent light. The mutant 16S rRNA is therefore functionally isolated from the rest of the cell and can be analyzed by monitoring the expression of the two proteins CAT and GFP. Lethal mutations in 16S rRNA merely prevent expression of CAT and GFP without affecting the normal protein synthesis of the bacteria.

In these experiments nearly 60,000 different SD-anti-SD combinations were tried but only 13 were found to be functional without killing the cell. Researchers in the Cunningham laboratory showed that almost any change in the nucleotide sequence of the anti-SD was lethal to bacteria-probably because it disrupted the balance of protein synthesis. Since its development, instant evolution has been used by several researchers throughout the world to study the role of ribosomal RNA in protein synthesis. This technology may also allow the development of new antibiotics targeted against critical regions of the ribosome and that are not susceptible to the development of drug resistance.

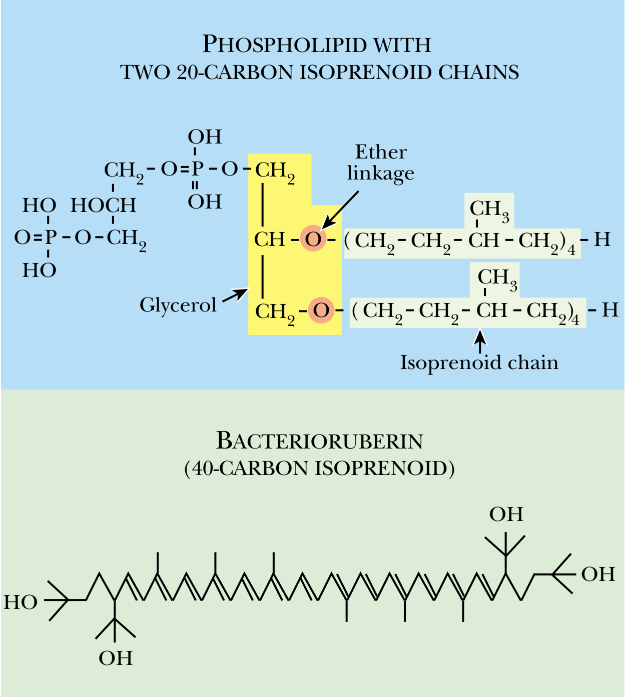


FIGURE 20.21 ***Unusual Lipids of Archaea***

The archaebacteria have lipid chains made of five carbon isoprenoid units rather than two carbon units as seen in eubacteria. The isoprenoid chains are linked to a glycerol via an ether link rather than an ester link. In some instances the isoprenoid lipid chains may contain 40 carbons (bacterioruberin, for example). These longer lipids span the whole membrane of the archaea.

quantitative measurements of genetic relatedness. Even if we cannot unambiguously define a species, we can be consistent in how much sequence divergence is needed to allocate organisms to different species or families.

Evolutionary trees are often constructed that show the number of sequence differences in rRNA.

Originally ribosomal RNA sequences were used for classification. However as ever more sequence data is obtained, including whole genomes, it is possible to take an increasing number of other genes into account. Computer programs exist for calculating the relative divergence of the sequences and can generate trees such as that in Figure 20.22. Here we have four bacteria, all in different genera but belonging to the same family, the Enterobacteria. To root such a tree correctly we also need the sequence from an organism in an "out-group;" in this case we have used the bacterium *Pseudomonas*, which is only distantly related to enteric bacteria. The nodes in Figure 20.22 represent the deduced common ancestors. The branch lengths are often scaled to represent the number of mutations needed and the numbers indicate how many base changes are needed to convert the sequence at each branch point into the next. (The total length of the 16s rRNA of Enteric bacteria is 1542 bases.)

Sequences are especially useful in classifying aberrant organisms that have lost structural characters normally used for comparison.

As discussed in Ch. 19, parasites have many adaptations and quirks that developed due to the unusual environment they inhabit. Establishing phylogenetic relationships among parasites is very difficult based on simple trait analysis. Fortunately, gene sequences can often be used to trace the ancestry of parasitic or aberrant life forms. Aberrant development due to unusual environments is not only seen in parasites. Animals such as moles have adapted to living underground or in caves, and have lost their eyes in the process since these organs are not useful. Sometimes vestigial remnants of structures remain even though the animal has no use for the structure. Whales have the atrophied remains of hind limbs, which implies that whales are not real fish, but mammals that have become fish-like in general form as they have adapted to life in the ocean. Until gene sequencing emerged, it remained unknown which mammals are the whale's closest relatives. It now appears that whales are related to the artiodactyls, hoofed mammals such as hippos, giraffes, pigs and camels.

The ancestries of sequences can be confirmed if they share major insertions or deletions.

One major problem with sequence comparison is that base changes can revert. Although statistical comparison of multiple sequences with many altered sites is often sufficient to establish a lineage, ambiguity sometimes remains. A useful way to help resolve ambiguities is by using conserved insertions or deletions—known as signature sequences or "indels". Although a single base insertion or deletion might possible revert, the likelihood that an insertion or deletion of several bases might revert so as

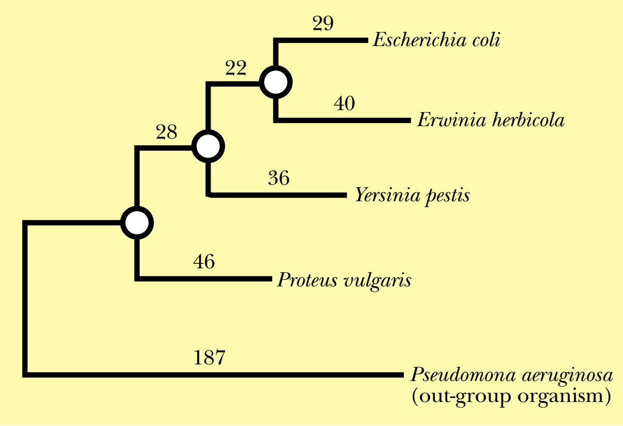


FIGURE 20.22 ***Phylogenetic Tree for Enteric Bacteria***

The phylogenetic relationship between bacteria can be deduced by comparing ribosomal RNA sequences. Here the sequences of the 16S rRNA genes are compared for the four enteric bacteria, *E. coli*, *Erwinia herbicola*, *Yersinia pestis*, and *Proteus vulgaris*. The relatively unrelated bacterium *Pseudomona aeruginaosa*, is used as an outgroup organism to provide the base or root of the tree. From these comparisons, it can be deduced that *P. vulgaris* was the first to branch from the primitive ancestor and that *E. coli* and *E. herbicola* were the latest.

to exactly restore the original length and sequence is vanishingly small. Consequently, if a subgroup of a family of related sequences all contain an indel of defined length and sequence at the same location, they must all have been derived from the same ancestral sequence.

Mitochondrial DNA—A Rapidly Ticking Clock

Although mitochondria contain circular molecules of DNA reminiscent of bacterial chromosomes, the mitochondrial genome is much smaller. The mitochondrial DNA codes for a few of the proteins and ribosomal RNA of the mitochondrion, but most components are now encoded by the eukaryotic nucleus as already discussed in Chapter 19. What concerns us here is that the mitochondrial DNA of animals accumulates mutations much faster than the nuclear genes. In particular, mutations accumulate rapidly in the third codon position of structural genes and even faster in the intergenic regulatory regions. This means that mitochondrial DNA can be used to study the relationships of closely related species or of races within the same species. Most of the variability in human mitochondrial DNA occurs within the D-loop segment of the regulatory region. Sequencing this segment allows us to distinguish between people of different racial groups.

Mitochondrial DNA changes fast enough to be used to classify subgroups within the same species.

One apparent drawback to using mitochondrial DNA is that mitochondria are all inherited from the mother. Although sperm cells do contain mitochondria, these are not released during fertilization of the egg cell and are not passed on to the descendants. On the other hand, analysis of mitochondria gives an unambiguous female ancestry, as complications due to recombination may be ignored. Furthermore, a eukaryotic cell contains only one nucleus but has many mitochondria so there are often thousands of copies of the mitochondrial DNA. This makes extraction and sequencing of mitochondrial DNA easier from a technical viewpoint.

Mitochondrial DNA can sometimes be obtained from museum samples and extinct animals. Mitochondrial DNA extracted from frozen mammoths found in Siberia differed in four to five bases out of 350 from both Indian elephants and African elephants. The DNA analysis supports the three-way split proposed based on anatomical relationship. The quagga is an extinct animal, similar to the zebra. It grazed the plains of Southern Africa only a little over a hundred years ago. A pelt preserved in a

German museum has yielded muscle fragments from which DNA has been extracted and sequenced. The two gene fragments used were from the quagga mitochondrial DNA. The DNA from the quagga differed in about 5 percent of its bases from the modern zebra. The quagga and mountain zebra are estimated from this to have had a common ancestor about three million years ago.

DNA has also been successfully extracted from Egyptian mummies. Although the amounts of DNA obtained are only 5 percent or so of those from fresh, modern, human tissue, DNA sequences have been obtained from a mummy 2,400 years old. Although several thousand base pairs were sequenced, no actual human genes were identified. Since the DNA of higher animals consists mostly of non-coding sequences, this is hardly surprising. Nonetheless, the mummy DNA did contain Alu elements that are characteristic of human DNA.

The African Eve Hypothesis

Attempts to sort out human evolution from skulls and other bones led to two alternative schemes. The multi-regional model proposes that *Homo erectus* evolved gradually into *Homo sapiens* simultaneously throughout Africa, Asia and Europe. The Noah's Ark model proposes that most branches of the human family became extinct and were replaced, relatively recently, by descendants from only one local sub-group (Fig. 20.23). Although anthropologists take both theories seriously, few geneticists regard the multi-regional model as plausible. This model implies continuous genetic exchange between widespread and relatively isolated tribes over a long period of prehistory. Not surprisingly, recent molecular analysis has tended to support the Noah's Ark model.

Mitochondrial sequence analysis suggests all modern humans are derived from a small group of ancestors who lived in Africa around 100,000 years ago.

Although mitochondria evolve fast, the overall variation among people of different races is surprisingly small. Calculations based on the observed divergence and the estimated rates suggest that our common ancestor lived in Africa between 100,000 and 200,000 years ago. Since mitochondria are inherited maternally, this ancestor has been named "**African Eve**". This African origin is supported by the deeper "genetic roots" of modern-day African populations. In other words, different sub-groups of Africans branched off from each other before the other races branched off from the Africans as a whole (Fig. 20.24).

The ancestors of today's Europeans split off from their Euro-Asian forebears and wandered into Europe via the Middle East around 40,000 to 50,000 years ago (Fig. 20.25). American Indians appear to derive from two major migrations originating from mainland Asian populations. The earlier Paleo-Indians (around 30,000 years ago) populated the whole American continent, while the more recent migration (less than 10,000 years ago) produced the Na-Dene peoples who are mostly North American Indians.

Besides using mitochondrial DNA, sequences of microsatellite regions of the chromosomes have been compared among the different races. The phylogenetic results are very similar. They also give a primary African—non-African split, and if anything, they suggest an even more recent date for the common ancestor, nearer 100,000 years ago.

But what about Adam, or "**Y-guy**" as he is sometimes called by molecular biologists? The shorter human Y chromosome does not recombine with its longer partner, the X chromosome over most of its length. This allows us to follow the male lineage without complications due to recombination. For example, the *ZFY* gene on the Y chromosome is handed on from father to son and is involved in sperm maturation. The sequence data for *ZFY* suggest a split between humans and chimps about 5 million years ago and a common male ancestor for modern mankind about 250,000 years ago. However, recent data from a much larger number of genetic markers on the Y chro-

African Eve Hypothetical female human ancestor thought to have lived in Africa around 100,000–200,000 years ago
Y-guy Hypothetical male human ancestor thought to have lived in Africa around 100,000–200,000 years ago

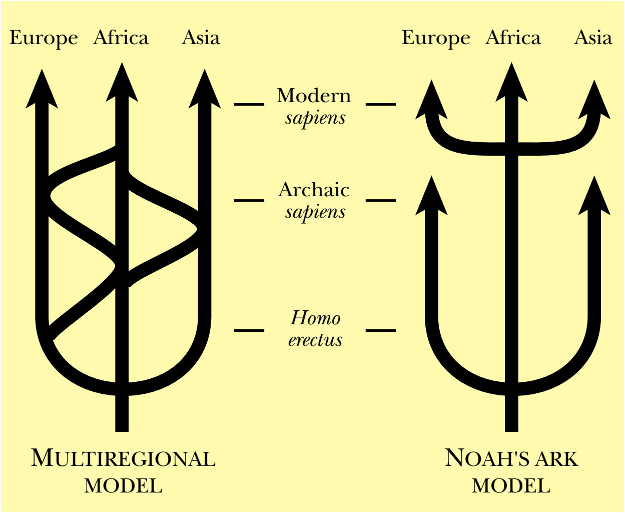


FIGURE 20.23 ***Multi-Regional and Noah's Ark Models of Human Evolution***

The multiregional model of human evolution (left) suggests that *Homo sapiens* developed from multiple interactions between several ancestral lines. The early *Homo erectus* ancestor branched and migrated from Africa to Asia and Europe. Traits developing in each branch were transmitted to the other branches implying genetic exchanges between the three branches, even though many thousands of miles separated the early ancestral groups. The Noah's Ark model (right) seems more plausible based on genetic analysis. The model suggests that modern *Homo sapiens* developed from one ancestral group in Africa. Other branches of archaic *sapiens* did develop and inhabited different regions in Europe and Asia for a while before dying out. The modern *sapiens* branch has then evolved into several branches from a relatively recent African ancestor.

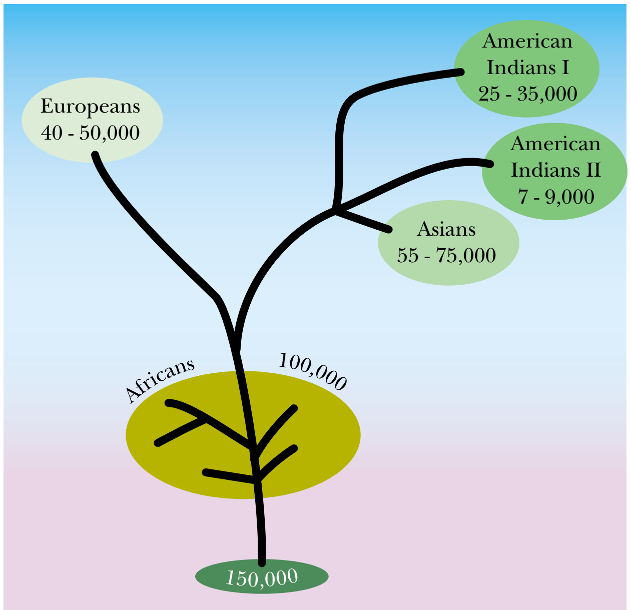


FIGURE 20.24 ***African Eve Hypothesis I—DNA***

This phylogenetic relationship was deduced by comparing mitochondrial DNA sequences from living humans. Numbers shown are estimated years before the present (BP). According to the African Eve theory, early humans developed in Africa about 150,000 years BP and diverged into many different tribal groups, most of which remained in Africa. The European and Asian races are derived from those relatively few groups of African ancestors who emigrated into Eurasia via the Middle East.

Y chromosome sequences confirm the recent African origin of humans.

mosome dates Y-guy to somewhat less than 100,000 years ago. Recent analyses of clusters of mutations on the Y chromosome in are incompatible with the multi-regional model and confirm the recent African origin of modern humans.

A final sad note concerns Neanderthal Man. Although Neanderthal Man survived to live alongside the modern races of *Homo sapiens* in Europe and the Middle East until relatively recently (approximately 30,000 years ago), sequence analysis suggests that the Neanderthals came to a dead end. Comparison of DNA sequences suggests

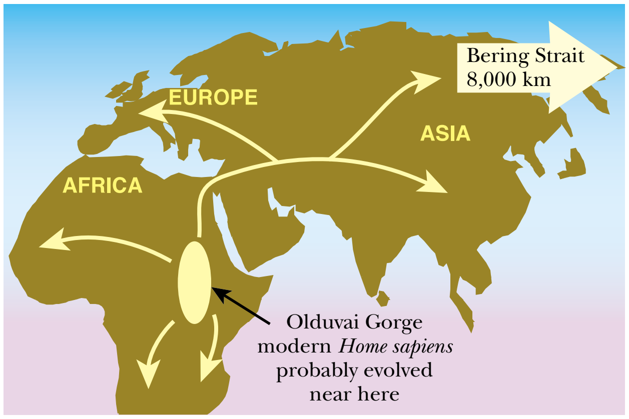


FIGURE 20.25 ***African Eve Hypothesis II—Migrations***

The divergence of the African ancestor into the modern African, European and Asian races included migration into different parts of the world. Scientists believe that modern *Homo sapiens* evolved in eastern Africa, around the Olduvai Gorge. Descendents of these early ancestors migrated to Europe and Asia as well as other areas in Africa. Descendents of some Asian groups crossed the Bering Strait to inhabit the American continent. Once isolated, these various groups evolved independently.

that the Neanderthals did not interbreed with modern Man or contribute significantly to the present-day human gene pool.

Ancient DNA from Extinct Animals

Apart from the occasional mummy or mammoth, DNA sequences from still living creatures are normally used to construct evolutionary schemes. However, ancient DNA extracted from the fossilized remains of extinct creatures can provide a valuable check on estimated evolutionary rates. The oldest available DNA so far successfully analyzed comes from amber. Amber is a polymerized, hardened resin produced by extinct trees that has gradually solidified to a glassy consistency over millions of years. Sometimes small animals were stuck in the resin when it oozed out of the trees and have been preserved there ever since (Fig. 20.27). Most of the trapped animals are insects, but occasionally worms, snails, and even small lizards have also been found. Amber acts as a preservative and the internal structure of individual cells from trapped insects can still be seen with an electron microscope. It has proven possible to recover DNA that is 25 to 125 million years old from some insects and some of this has been amplified by PCR and sequenced.

Small stretches of DNA sequence have been rescued from extinct life forms.

The largest chunks of amber are no more than 6 inches across, so larger animals such as dinosaurs cannot be preserved. Nonetheless, a few blood cells preserved in the gut of a blood-sucking insect could, in theory, provide the complete DNA sequence of a large animal. This was the scenario for Michael Crichton's high-tech thriller, Jurassic Park, where dinosaurs were resurrected by having their DNA inserted into amphibian eggs. In real life, such ancient dinosaur DNA would be severely damaged and only short segments would be readable. Nonetheless, the possibility of someday obtaining a few short fragments of some *Tyrannosaurus rex* genes is no longer a total fantasy.

Although DNA has indeed been isolated from samples that are several million years old it is so severely degraded that identification has not been possible. To date, the oldest identified animal DNA is approximately 50,000 years old and comes from mammoths preserved in the permafrost of Siberia. The permafrost has also yielded identifiable plant DNA from grasses and shrubs around 300,000–400,000 years old.

Microorganisms may also be trapped in amber and in such cases it may be possible to revive the whole organism, not merely obtain DNA samples. In particular,

Genghis Khan's Y Chromosome

Large-scale surveys have shown that about 1 in 12 men in Asia carry a variant of the Y chromosome that originated in Mongolia roughly 1,000 years ago. Around 30 natural genetic markers were surveyed in several thousand men. The markers included deletions and insertions, sequence polymorphisms and repetitive sequences. Most men carry Y chromosomes with more or less unique combinations of such DNA markers. However, about 8% of Asian males carry Y chromosomes with the same (or almost the same) combination of genetic markers. This phenomenon was not seen among men from other continents. Furthermore, the Asian men with the special "Mongol cluster" of genetic markers were found only among those populations who formed part of the Mongol Empire of Genghis Khan. For example, the "Mongol cluster" was absent from Japan and southern China, which were not incorporated into the Mongolian Empire, but was present in 15 different populations throughout the area of Mongolian domination (Fig. 20.26). In addition, although very few Pakistanis have the "Mongol cluster", about 30% of a small tribal group known as the Hazara do possess it. The Hazara are known to be of Mongolian origin and claim to be direct descendents of Genghis Khan. Since present-day Pakistan is outside Genghis Khan's area of conquest they presumably migrated to their present location later.

This particular variant has therefore been proposed to be the Y chromosome of Genghis Khan the great Mongolian conqueror. About 800 years ago the warlord Temujin united the Mongols and in 1206 assumed the title of Genghis Khan ("Lord of Lords"). The Mongols massacred many of the males and impregnated many of the women in areas they conquered. The present day distribution of Y chromosomes apparently reflects these practices. Whether this special variant of the Y chromosome was present in Genghis Khan himself or just frequent among his Mongol warriors cannot be known for certain. Nonetheless, it is more likely than not that Genghis Khan himself had this Y chromosome, as all the warriors in such tribes were usually closely related.

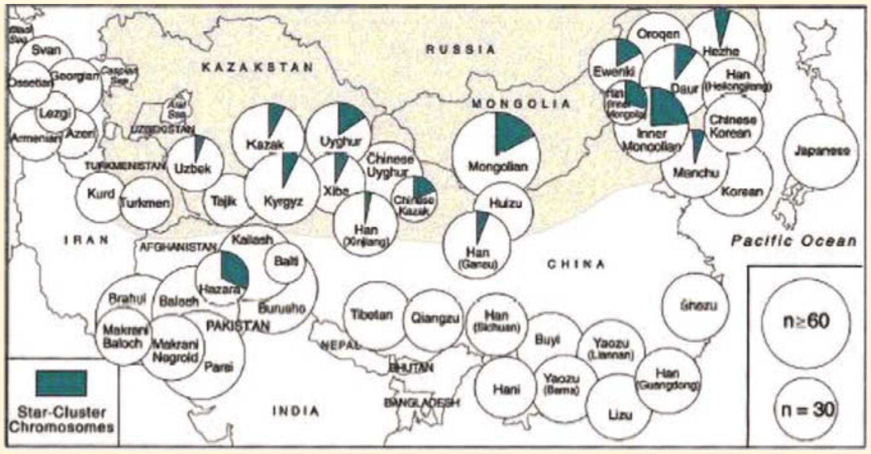


FIGURE 20.26 ***Genghis Khan's Empire and Y Chromosome***

The relative proportion of Mongol cluster chromosomes at various geographical locations is represented by the green segments in the circles. The size of the circles indicates sample size. From Zerjal and Tyler-Smith, The genetic legacy of the Mongols, *American Journal of Human Genetics* (2003) 72:717–721.

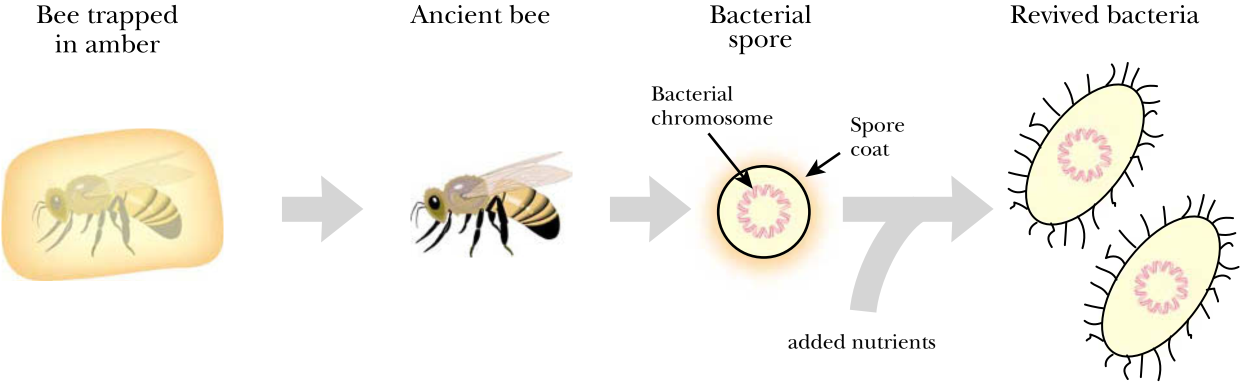


FIGURE 20.27 ***Ancient DNA Preserved in Amber***

At some point millions of years ago, a bee was trapped in sap from a tree. The sap gradually hardened and solidified into solid clear yellowish material—amber. The bee was completely preserved together with the bacterial spores that it carried. After extraction from the amber resin, occasional spores are capable of growth given the right nutrients and environmental conditions.

spores, covered by a protective coat are formed by some bacteria to survive bad conditions and may survive for extremely long periods. Some 30 million-year-old bacterial spores have been found inside bees trapped in pieces of amber. When provided with nutrients, the spores grew into bacterial colonies. These reawakened bacteria were identified as *Bacillus sphaericus*, which is found today in association with bees. DNA from the ancient *Bacillus sphaericus* was similar in sequence to its modern relative, but not identical, as it would have been if the ancient bacteria were just contaminants. More recently, spores of another bacillus species were isolated and revived from a brine inclusion within a 250-million year old salt crystal.

Evolving Sideways: Horizontal Gene Transfer

Genetic information may be passed "vertically" from an organism to its direct descendents or "horizontally" to other organisms that are not descendents.

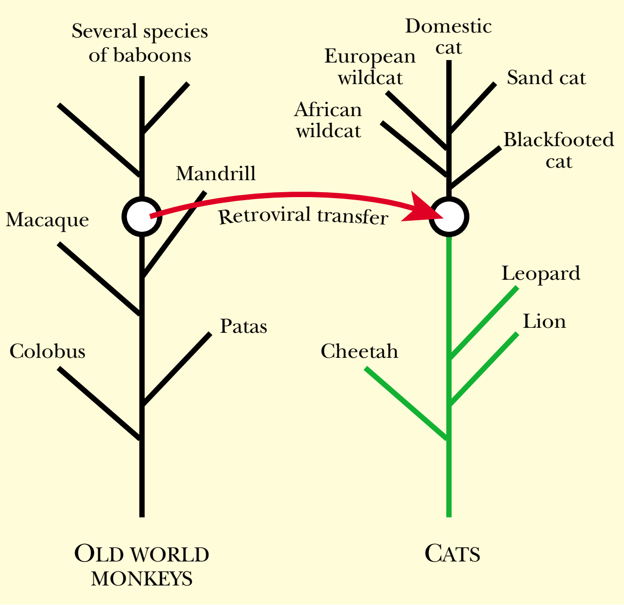
Standard Darwinian evolution involves alterations in genetic information passed on from one generation to its descendants. However, it is also possible for genetic information to be passed "sideways" from one organism to another that is not one of its descendents or even a near relative. The term **vertical gene transfer** refers to gene transmission from the parental generation to its direct descendants. Vertical transmission thus includes gene transmission by all forms of cell division and reproduction that create a new copy of the genome, whether sexual or not. This contrasts with "**horizontal gene transfer**" (also known as "**lateral gene transfer**") in which genetic information is passed sideways, from a donor organism to another that is not its direct descendent.

For example, when antibiotic resistance genes are carried on plasmids they can be passed between unrelated types of bacteria (see Ch. 16). Since genes carried on plasmids are sometimes incorporated into the chromosome, a gene can move from the genome of one organism to an unrelated one in a couple of steps. The complete genomes of many bacteria have now been fully sequenced. Estimates using this data suggest that about 5–6% of the genes in an average prokaryotic genome have been acquired by horizontal transfer. The effects of horizontal transfer are especially noticeable in a clinical context. Both virulence factors and antibiotic resistance are commonly carried on transmissible bacterial plasmids.

horizontal gene transfer Transfer of genetic information "sideways" from one organism to another that is not directly related
lateral gene transfer Movement of genes sideways between unrelated organisms. Same as horizontal gene transfer
vertical gene transfer Transfer of genetic information from an organism to its descendents

FIGURE 20.28 ***Horizontal Transfer of Type-C Virogene in Mammals***

The type-C virogene was present during evolution of old world monkeys from their common ancestor. Surprisingly, a version of this gene closely related to the one in baboons was identified in North African and European cats. Since baboons and cats are not closely related, the gene must have moved from one group to another via horizontal transfer. Further supporting the idea of horizontal transfer, the gene is not found in cats like the lion or cheetah, which developed before the North African and European cats branched off.



Horizontal gene transfer usually involves viruses, plasmids or transposons.

Such horizontal transfer may occur between members of the same species (e.g. the transfer of a plasmid between two closely related strains of *Escherichia coli*) or over major taxonomic distances (e.g. the transfer of a Ti-plasmid from bacteria to plant cells). Horizontal gene transfer over long distances depends on carriers that cross the boundaries from one species to another. Viruses, plasmids and transposons are all involved in such sideways movement of genes and have been discussed in their own chapters (see Chs. 15–17). Retroviruses, in particular, are capable of inserting themselves into the chromosomes of animals, picking up genes and moving them into another animal species.

One well-described example of horizontal transfer in animals concerns the type-C virogene shared by baboons and all other Old World monkeys. The type-C virogene was present in the common ancestor of these monkeys, about 30 million years ago, and since then has diverged in sequence just like any other normal monkey gene. Related sequences are also found in a few species of cats. Only the smaller cats of North Africa and Europe possess the baboon type-C virogene. American, Asian and Sub-Saharan African cats all lack this sequence. Therefore, the original cat ancestor did not have this type-C virogene. Furthermore, the sequence found in North African cats resembles that of baboons more closely than the sequences in monkeys closer to the ancestral stem (see Fig. 20.28). This suggests that about 5–10 million years ago a retrovirus carried the type-C virogene horizontally from the ancestor of modern baboons to the ancestor of small North African cats. The domestic European pussycat originally came from Egypt, so it also carries the type-C virogene. However, other cats that diverged more than 10 million years ago lack these sequences.

Problems in Estimating Horizontal Gene Transfer

When the human genome was sequenced, several hundred human genes were at first attributed to horizontal transfer from bacteria. However, later analyses indicated that very few of these were genuine cases of horizontal transfer (see Ch. 24). Several factors have contributed to such over-estimates of horizontal transfer, both for the human genome and in other cases.

a. Sampling Bias. Relatively few eukaryotic genomes have been fully sequenced whereas hundreds of bacterial genomes have been sequenced. Thus the absence of sequences homologous to a human gene from a handful of other eukaryotes is insufficient evidence for an external (bacterial) origin. As more eukaryotic sequence data has become available many genes supposedly of "bacterial" origin have in fact been found in other eukaryotes.

b. The loss of homologs in related lineages may suggest that a gene originated externally to the group of organisms that retain it. As in the related case (a) above, the solution to this artifact is the collection of more sequence data from many related lineages.

c. Gene duplication followed by rapid divergence may give rise to apparently novel genes that are missing from the direct vertical ancestor of a group of organisms.

d. Intense evolutionary selection for a particular gene may result in a greatly increased rate of sequence alteration. Those genes that evolve faster than normal will tend to be misplaced when sequence comparison is used to construct evolutionary trees.

e. The ease of horizontal transfer of genetic information by plasmids, viruses, transposons under laboratory conditions is misleading. Under natural conditions there are major barriers to such movements. Furthermore, the results of horizontal transfer are often only temporary. Newly acquired genes, especially those on plasmids, transposons, etc., are easily lost. Such genes tend to be acquired in response to selection such as antibiotic resistance and, conversely, they will be lost when the original selective conditions disappear.

f. Experimental problems such as DNA contamination. Bacterial and viral parasites are associated with esentially all higher organisms and completely purifying the eukaryotic DNA is not always easy.

Many of the originally proposed examples of widespread horizontal gene transfer have been severely compromised by the above factors. However, some examples do seem to be valid. One of the most interesting is the recent finding of relatively frequent horizontal gene transfer between the mitochondrial genomes of flowering plants. The genes for certain mitochondrial ribosomal proteins have apparently been transferred from an early monocotyledonous lineage to several different dicotyledonous lineages. Examples include transfer of the *rps2* gene to kiwifruit (*Actinidia*) and the *rps11* gene to bloodroot (*Sanguinaria*).

CHAPTER TWENTY-ONE

Nucleic Acids: Isolation, Purification, Detection, and Hybridization

Isolation of DNA

Almost every molecular biologist has collected DNA from the organism they are studying. Initially, isolating DNA was a long and arduous process with large amounts of DNA collected. Advancing technology has resulted in the amount of DNA needed for either analysis or cloning of genes to steadily decrease. Nowadays, for example, enough DNA can be collected for genetic manipulations in laboratory mice or rats from a small piece of the tail. Human DNA may be analyzed using small blood samples or a few cells scraped from the inside of the cheek. The decrease in the amount of DNA required for analysis has allowed scientists to streamline the process so that DNA can be isolated in a few hours instead of a few days.

Modern technology has led to a steady decrease in the amount of DNA needed for analysis.

Extracting DNA from plants, animals, and bacteria, all require that the cellular contents be liberated into a solution. Since the bacteria are single cells and contain no bone, fat, gristle, etc., the DNA is relatively easy to extract. In contrast, samples from animals and plants must often be ground into tiny fragments before proceeding. Since plant cells have very rigid cell walls, the scientist must mechanically break the cells open in a blender, or add special degradative enzymes to digest the cell wall components. Similarly, to extract DNA from a mouse's tail, enzymes are added to degrade the connective tissue and disperse the cells. By far the easiest way to get DNA is to extract it from bacteria. A few drops of a bacterial culture will give plenty of DNA for most purposes. First, the bacterial cell wall is easily digested by **lysozyme**, an enzyme that degrades the **peptidoglycan** layer of the cell wall. A successive treatment with **detergent** dissolves the lipids of the cell membrane. Chelating agents, such as **EDTA (ethylene diamine tetraacetate)**, are also used, especially with gram-negative bacteria, to remove the metal ions that bind components of the outer membrane together. In all these samples, the cellular contents, including the DNA, are then liberated into solution and are purified by a further series of steps.

Cell walls and membranes must be broken down to liberate the DNA from the cell.

Purification of DNA

Two general types of procedure are used for purification of DNA, **centrifugation** and chemical extraction. The principle of centrifugation is as follows. The sample is spun at high speed and the centrifugal force causes the larger or heavier components to sediment to the bottom of the tube. For example, destroying the cell wall of bacteria by lysozyme and detergents leaves a solution containing the fragments of the cell wall, which are small, and the DNA, which is a gigantic molecule. When the sample is centrifuged, DNA and some other large components are sedimented to the bottom of the tube. The fragments of cell wall, together with many other soluble components, remain in solution and are discarded.

Macromolecules such as DNA may be separated from smaller molecules by centrifugation.

The sedimented DNA is then re-dissolved in an appropriate buffer solution. However, it still has a lot of protein and RNA mixed in with it. These are generally removed by chemical means. One step used in many DNA purifications is **phenol extraction**. Phenol, also known as carbolic acid, is very corrosive and extremely dangerous because it dissolves and denatures the proteins that make up 60 to 70 percent of all living matter. Consequently, phenol may be used to dissolve and remove all of the proteins from a sample of DNA.

Phenol dissolves proteins and is used to remove them from DNA.

centrifugation Process in which samples are spun at high speed and the centrifugal force causes the larger or heavier components to sediment to the bottom

detergent Molecule that is hydrophobic at one end and highly hydrophilic at the other and which is used to dissolve lipids or grease

EDTA (ethylene diamine tetraacetate) A widely used chelating agent that binds di-positive ions such as Ca^{2+} and Mg^{2+}

lysozyme An enzyme found in many bodily fluids that degrades the peptidoglycan of bacterial cell walls

peptidoglycan Mixed polymer of carbohydrate and amino acids that comprises the structural layer of bacterial cell walls

phenol extraction Technique for removing protein from nucleic acids by dissolving the protein in phenol

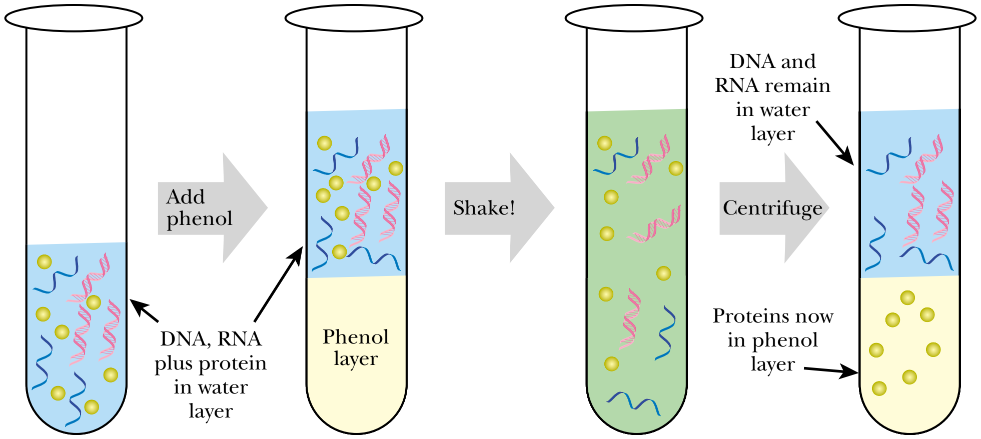

Add phenol
DNA, RNA plus protein in water layer
Phenol layer
Shake!
DNA and RNA remain in water layer
Centrifuge
Proteins now in phenol layer

FIGURE 21.01 ***Phenol Extraction Removes Proteins from Nucleic Acids***

Proteins can be removed from a solution of DNA or RNA by adding an equal volume of phenol. The phenol dissolves the proteins without disrupting the DNA or RNA. Since phenol is very dense, it forms a separate layer at the bottom of the tube. When the two solutions are shaken, the proteins dissolve into the phenol. The two layers separate again after a brief spin in the centrifuge. The top phase, which now contains just DNA and RNA, can be isolated.

When phenol is added to water, the two liquids do not mix to form a single solution; instead, the denser phenol forms a separate layer below the water. When shaken, the two layers mix temporarily, and the proteins dissolve in the phenol. When the shaking stops, the DNA solution and phenol containing the proteins separate into two layers (Fig. 21.01). To ensure that no phenol is trapped with the DNA, the sample is centrifuged briefly. Then the water containing the DNA and RNA is sucked off and kept. Generally, several successive phenol extractions are performed to purify away the proteins from DNA.

Nucleic acids may be purified on columns containing resins that bind DNA and RNA.

A variety of newer techniques have been developed that avoid phenol extraction. Most of these involve purifying DNA by passing it through a column containing a resin that binds DNA but not other cell components. The two main choices are silica and anion exchange resins. Silica resins bind nucleic acids rapidly and specifically at low pH and high salt concentrations. The nucleic acids are released at higher pH and low salt concentration. Anion exchange resins, such as diethylaminoethyl-cellulose, are positively charged and bind DNA via its negatively charged phosphate groups. In this case binding occurs at low salt concentrations and the nucleic acids are eluted by high concentrations of salt, which disrupt the ionic bonding.

Removal of Unwanted RNA

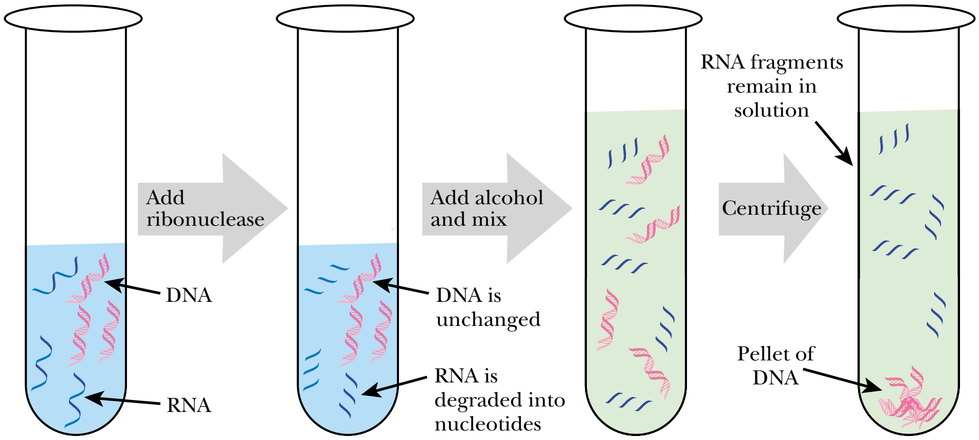
Unwanted RNA is often removed by degrading it enzymatically.

Special enzymes remove contaminating RNA from a DNA sample. The enzyme **ribonuclease** degrades RNA into short oligonucleotides but leaves the giant DNA macromolecule unchanged. A mixture of DNA and RNA is first incubated with the ribonuclease at the optimal temperature for enzyme activity. Next, an equal volume of alcohol is added. The alcohol precipitates large macromolecules, including long chains of DNA, out of solution. However, the small RNA fragments remain dissolved. Note that alcohol treatment is not very specific and will precipitate most large carbohydrates and many proteins as well as intact macromolecules of both DNA and RNA. Thus, alcohol precipitation can only be used after these components have been removed from the DNA by centrifugation and phenol extraction. Next the DNA is sedimented to the bottom of the tube by centrifugation and the supernatant solution containing

ribonuclease An enzyme that degrades RNA

FIGURE 21.02 ***Removal of RNA by Ribonuclease***

A mixture of RNA and DNA is incubated with ribonuclease, which digests all the RNA into small fragments and leaves the DNA unaltered. An equal volume of alcohol is added, and the larger pieces of DNA are precipitated out of solution. The solution is centrifuged, and the large insoluble pieces of DNA form a small pellet at the bottom of the tube. The RNA fragments remain in solution.



the RNA fragments is discarded (Fig. 21.02). The tiny pellet of DNA left at the bottom of the tube is often scarcely visible. Nonetheless, it contains billions of DNA molecules, sufficient for most investigations. This DNA is dissolved into buffered water and is now ready for use in genetic engineering.

Gel Electrophoresis of DNA

Electrophoresis separates DNA and RNA on the basis of charge.

Perhaps the most widely used physical method in all of molecular biology is **gel electrophoresis**. This technique separates and purifies fragments of DNA or RNA as well as proteins. The basic idea of **electrophoresis** is to separate the molecules based on their intrinsic electrical charge. Electrically positive charges attract negative charges and repel other positive charges. Conversely, negative charges attract positive charges and repel other negative charges. Two electrodes, one positive and the other negative, are connected up to a high voltage source. Positively charged molecules move towards the negative electrode and negatively charged molecules move towards the positive electrode (Fig. 21.03).

Since DNA carries a negative charge on each of the many phosphate groups making up its backbone, it will move towards the positive electrode during electrophoresis. The bigger a molecule, the more force required to move it. However, the longer a DNA molecule, the more negative charges it has. In practice, these two factors cancel out because all fragments of DNA have the same number of charges per unit length. Consequently, DNA molecules in free solution will all move toward the positive electrode at the same speed, irrespective of their molecular weights.

Gel electrophoresis separates nucleic acid molecules according to their molecular weight.

Electrophoresis of DNA is usually used to separate the DNA into different sizes. For example, scientists will often isolate DNA from bacteria. In addition to the chromosome, bacteria often contain plasmids; however, gel electrophoresis will separate the two different sized molecules of DNA. The gel that separates the fragments consists of a matrix of cross-linked polymer chains. Most DNA is separated using **agarose gel electrophoresis**. **Agarose** is a polysaccharide extracted from seaweed. When agarose and water are mixed and boiled, the agarose melts into a homogeneous solution. As the solution cools, it gels to form a meshwork, which has small pores or openings filled with water. The cooled gel looks much like a very concentrated mixture of gelatin without the food coloring. The pore size of agarose is suitable for separating nucleic acid polymers consisting of several hundred nucleotides or longer. Shorter fragments of DNA as well as proteins are usually separated on gels made of **polyacry-**

agarose A polysaccharide from seaweed that is used to form gels for separating nucleic acids by electrophoresis
agarose gel electrophoresis Technique for separation of nucleic acid molecules by passing an electric current through a gel made of agarose
electrophoresis Movement of charged molecules due to an electric field. Used to separate and purify nucleic acids and proteins
gel electrophoresis Electrophoresis of charged molecules through a gel meshwork in order to sort them by size
polyacrylamide Polymer used in separation of proteins or very small nucleic acid molecules by gel electrophoresis

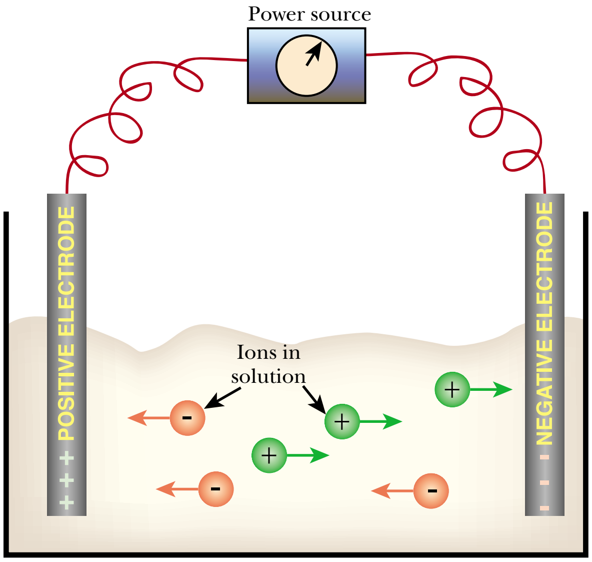


FIGURE 21.03 ***Principle of Electrophoresis***

Creating an electrical field in a solution of positive and negatively charged ions allows the isolation of the ions with different charges. Since DNA has a negative charge due to its phosphate backbone, electrophoresis will isolate the negatively charged DNA from other components.

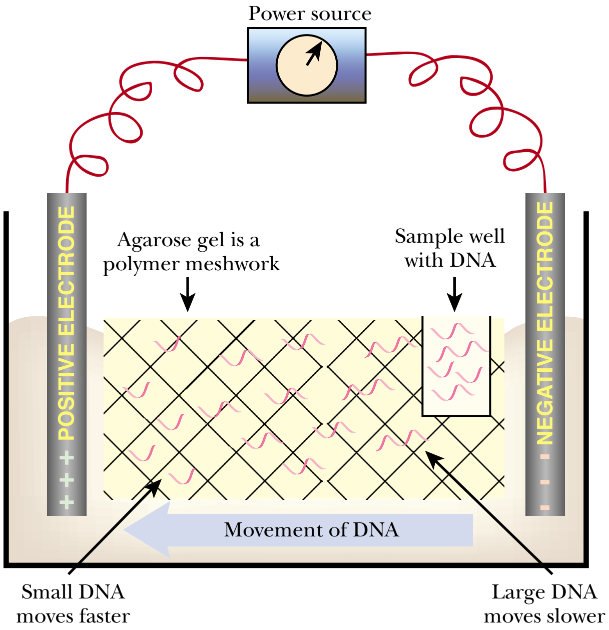


FIGURE 21.04 ***Agarose Gel Electrophoresis of DNA***

Agarose gel electrophoresis separates fragments of DNA by size. Negatively charged DNA molecules are attracted to the positive electrode. As the DNA migrates, the fragments of DNA are hindered by the cross-linked agarose meshwork. The smaller the piece of DNA, the less likely it will be slowed down. Therefore, smaller fragments of DNA migrate faster.

lamide. The meshwork formed by this polymer has smaller pores than agarose polymers. The samples of protein or DNA are loaded into a slot or sample well at the end of the gel closest to the negative electrode. DNA molecules move through the gel away from the negative electrode and towards the positive electrode. As the DNA molecules move through the gel they are hindered by the meshwork of fibers that make up the gel. The larger molecules find it more difficult to squeeze through the gaps but the smaller ones are slowed down much less. The result is that the DNA fragments separate in order of size (Fig. 21.04). In our example, the rings of plasmid DNA will move farther in the gel than the chromosome.

DNA can be detected by radioactive labeling or by staining with a dye such as ethidium bromide.

Agarose gels are normally square slabs that allow multiple samples to be run side by side. Because DNA is naturally colorless, some way of visualizing the DNA after running the gel is needed. The DNA may be radioactively labeled and detected by autoradiography, as explained below. Alternatively, the gel can be stained with **ethidium bromide**, which binds tightly and specifically to DNA or RNA. Ethidium bromide

ethidium bromide A stain that specifically binds to DNA or RNA and appears orange if viewed under ultraviolet light

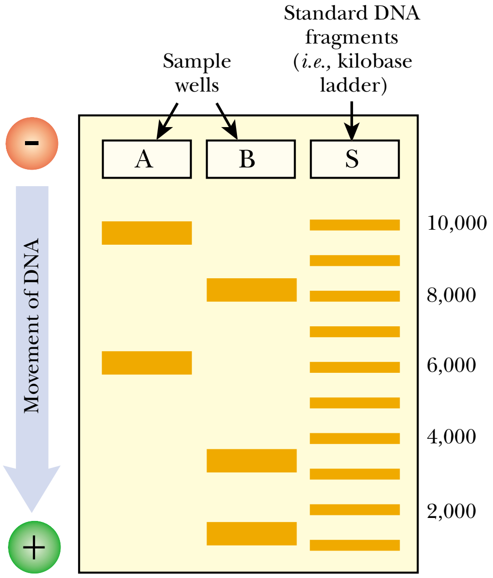


FIGURE 21.05 ***Agarose Gel Separation of DNA: Staining and Standards***

To visualize DNA, the agarose gel containing the separated DNA fragments is soaked in a solution of ethidium bromide, which intercalates between the base pairs of the DNA. Excess ethidium bromide is removed by rinsing in water, and the gel is placed under a UV light source. The UV light excites the ethidium bromide and causes it to fluoresce orange. In sample A, there are two fragments of DNA each of a different size and each forming a separate band in the gel. To determine the size of fragments, a standard set of DNA fragments of known sizes is run alongside the sample to be analyzed.

intercalates between the base pairs of DNA and RNA, therefore, DNA binds much more molecules of ethidium bromide than RNA. The gel must then be examined under UV light, where ethidium bromide bound to DNA fluoresces bright orange. RNA also fluoresces under UV light, but not as intense. After the bands are located, they may be cut out of the agarose slab and the DNA extracted to yield a pure fragment.

Agarose gel electrophoresis can be used to purify DNA for use in genetic engineering or it can be used to measure the sizes of different fragments. To find the size of an unknown piece of DNA, a set of standard DNA fragments of known sizes is run alongside, on the same gel (Fig. 21.05). A set of DNA fragments that are exact multiples of 1000 bp are often used. This is known as a **kilobase ladder**.

Pulsed Field Gel Electrophoresis

Very large DNA molecules may be separated by pulsed field electrophoresis.

Standard agarose gels can be used to separate nucleic acids that range from around 200 bp to perhaps 50 kb. Specialized variants of gel electrophoresis are needed to separate DNA molecules that are extremely large or, conversely, very short. Polyacrylamide gels are capable of separating DNA or RNA fragments differing by only one base pair in size, but their range is from 5 to 1,000 bp. Such gels are widely used in DNA sequencing (see Ch. 24). **Pulsed field gel electrophoresis (PFGE)** is used for analysis of very large DNA molecules from 10 kb to 10 megabases. It can even be used to separate whole chromosomes. For example, the protozoan parasite *Leishmania* has chromosomes that are too small to be seen using a light microscope but which have been separated by PFGE.

kilobase ladder A set of DNA fragments that are exact multiples of 1000 bp and are often used as standards in gel electrophoresis
pulsed field gel electrophoresis (PFGE) Type of gel electrophoresis used for analysis of very large DNA molecules and which uses an electric field of "pulses" delivered from a hexagonal array of electrodes

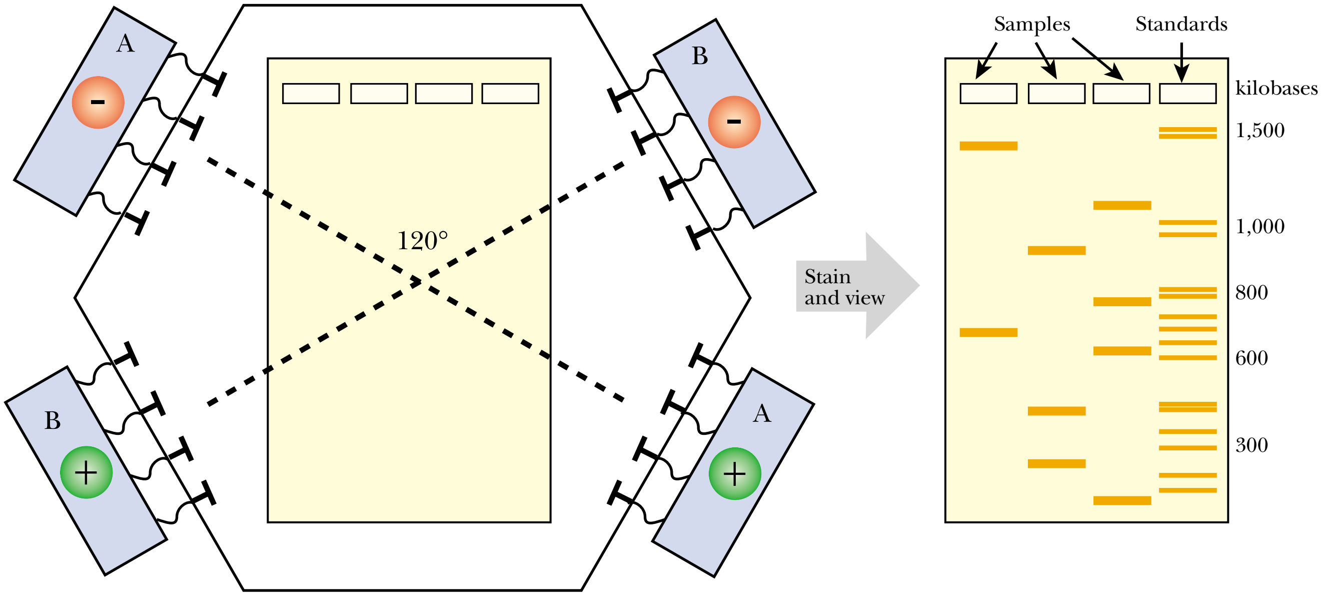


FIGURE 21.06 ***Pulsed Field Gel Electrophoresis***

The principle behind pulse field gel electrophoresis (PFGE) is the same as standard agarose gel electrophoresis, that is, negatively charged DNA is attracted to the positive electrode as it passes through a gel. However, instead of one electric field that propels the DNA vertically, PFGE uses two electric fields situated at 120° with respect to each other. The electric field alternates between the two sets of electrodes, A and B. As DNA leaves the sample well, it first migrates towards the positive A electrode, then the electric field switches and the DNA migrates toward the positive B electrode. And so on. By constantly switching directions, the larger pieces of DNA do not get caught in the gel meshwork and migrate faster in the gel. As usual, the gel is stained with ethidium bromide and viewed under UV light to see the bands. Notice that the bands shown here range from 1,500 kilobases (1,500,000 bases) to 100 kilobases (100,000 bases).

During standard electrophoresis, DNA molecules align with the electric field and so travel sideways on through the gel meshwork. Consequently, large DNA molecules are trapped in the gel matrix and scarcely move. In PFGE, the electric field consists of "pulses" which are delivered from a hexagonal array of electrodes (Fig. 21.06). This creates essentially two fields that intersect at 120 degrees. Large molecules will travel in the direction of the first field for a short time and eventually get caught in the agarose matrix. When the field is changed, the DNA realigns by 120 degrees and then moves in this direction for a while. Assuming the two pulse times are equal, overall movement is in the forward direction.

Denaturing Gradient Gel Electrophoresis

Denaturing gradient gel electrophoresis (DGGE) separates DNA molecules differing in sequence by only a single base. As its name indicates, this technique combines gel electrophoresis with DNA denaturation. It is applicable to double-stranded DNA fragments of a few hundred bases, which are often generated by PCR (see Ch. 23). During DGGE, double-stranded DNA is subjected to a gradient of increasing denaturation to separate the strands. The dsDNA melts in stages in which discrete zones known as "melting domains", become unpaired along the piece of DNA. Partially melted DNA migrates more slowly through an agarose gel during electrophoresis because the small streamlined DNA fragments open into larger structures that get

Denaturing DNA during electrophoresis allows separation of molecules which differ only very slightly.

denaturing gradient gel electrophoresis (DGGE) Combination of gel electrophoresis with DNA denaturation that allows separation of DNA molecules differing in sequence by only a single base

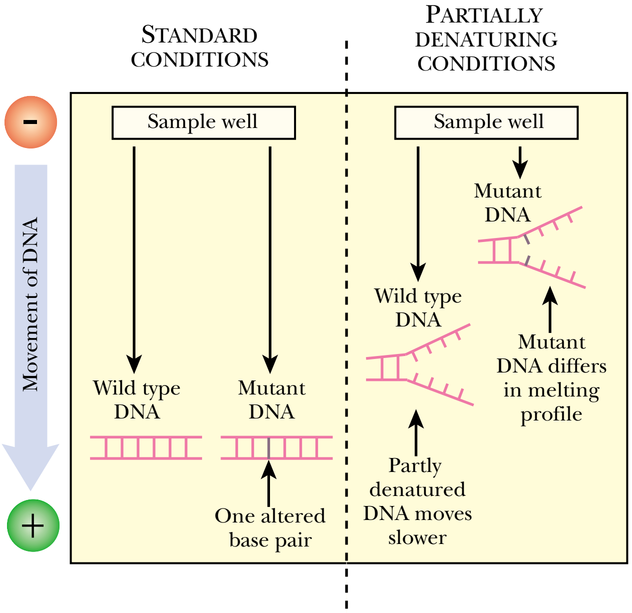


FIGURE 21.07 ***Denaturing Gradient Gel Electrophoresis***

Under standard conditions, short pieces of double-stranded DNA migrate fairly quickly and are not hindered by the gel matrix. During DGGE, the agarose gel contains a gradient of a denaturing agent that melts the DNA into single strands. Close to the sample well, the concentration of denaturing agent is low, and the DNA does not melt. As the DNA migrates into higher concentrations of denaturing agent, it melts at one end, forming a Y-shaped molecule. This is much more likely to get caught in the agarose meshwork than the undenatured version. How easily DNA denatures depends upon its base sequence. Therefore, wild type and mutant versions of the same DNA sequence will denature at different places in the denaturation gradient. Small pieces of DNA that differ by as little as one base pair may be separated by this technique.

caught in the agarose meshwork. Since the melting temperature of each "melting domain" is sequence specific, the melting profile of DNA with a single base mutation will differ significantly from the wild type (Fig. 21.07).

In practice, denaturation is due to a combination of moderately high but constant temperature (usually between 50 and 65°C) plus chemical denaturation with a mixture of urea and formamide. A concentration gradient of urea plus formamide is set up across the gel when it is formed. The chemical denaturation gradient may be arranged parallel or perpendicular to the direction of migration of the DNA. Parallel DGGE gives bands in unique positions, and the results resemble those of normal agarose gel electrophoresis. In perpendicular DGGE, a mixture of DNA molecules differing by one or a few base pairs are loaded across the whole gel slab, and a series of sigmoid bands are obtained (Fig. 21.08).

DGGE is used in the analysis of mutations, especially those such as base substitutions, which do not change the length of the DNA. For example, DGGE has been used to screen mutations in the *BRCA1* and *BRCA2* genes that are involved in causing breast cancer. In addition, DGGE is widely used in screening natural populations for genetic variability and/or relatedness. In particular, PCR of DNA extracted from soil or other natural habitat followed by DGGE has been used to analyze the phylogenetic relationships of microbial populations without the need to culture living microorganisms.

Chemical Synthesis of DNA

An alternative to isolating DNA from natural sources is to synthesize it artificially. Molecular biologists routinely use artificially manufactured lengths of DNA for a

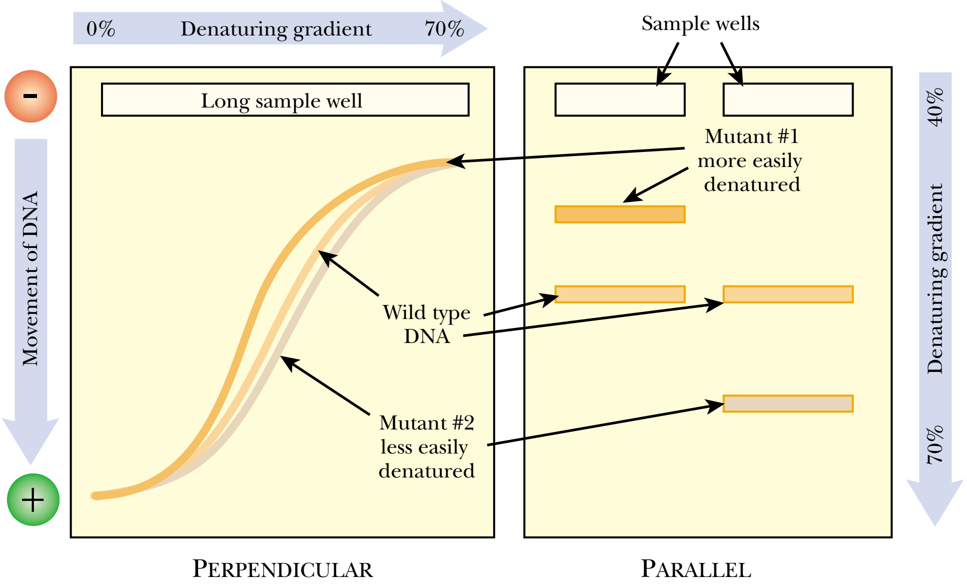


FIGURE 21.08 ***Parallel and Perpendicular DGGE***

In perpendicular DGGE, the gradient of denaturing agent is at right angles to the electric field. A mixture of DNA fragments (wild-type, mutant 1 and mutant 2 in this example) is loaded into a long well that spans the entire gel. On the left where the amount of denaturant is low, none of the three DNA fragments melt and they all migrate together. In the middle of the denaturation gradient, the three DNA fragments have melted to differing extents. Mutant 1 denatures most easily and therefore migrates slower than the other fragments. Mutant 2 DNA melts the least and so migrates the farthest. On the far right of the gel, where the concentration of denaturant is greatest, all three DNA fragments are fully denatured and run together as one band of single-stranded DNA.

In parallel DGGE, the denaturing gradient runs from top to bottom, in the same direction as the electric field. Here, a mixture of mutant 1 and wild-type were loaded in the first well, and a mixture of wild-type and mutant 2 in the second. Mutant 1 denatures most easily and migrates the least. Mutant 2 is most resistant to denaturation and so migrates the fastest.

Short to medium lengths of DNA are routinely made by chemical synthesis nowadays.

variety of purposes. Short stretches of single-stranded DNA are used as probes for hybridization (see below), primers in PCR (see Ch. 23) and primers for DNA sequencing (see Ch. 24). Short lengths of double-stranded DNA are made by synthesizing two complementary single-strands and allowing them to anneal. Such pieces may be used as linkers or adaptors in genetic engineering (see Ch. 22). It is also possible to synthesize whole genes, although this is more complicated (see below).

The first step in chemical synthesis of DNA is to anchor the first nucleotide to a solid support, most often a porous glass bead. **Controlled pore glass (CPG)** beads that have pores of uniform sizes are most commonly used. The beads are packed into a column and the reagents are poured down the column one after another. Nucleotides are added one by one and the growing strand of DNA remains attached to the glass beads until the synthesis is complete (Fig. 21.09). Chemical synthesis of DNA is performed in practice by an automated machine (Fig. 21.10). After loading the machine with chemicals, the required sequence is typed into the control panel. Gene machines take a few minutes to add each nucleotide and can make pieces of DNA 100 nucleotides or more long. Modern DNA synthesizers are also usually capable of adding fluorescent dyes, biotin or other groups used in labeling and detection of DNA (see below).

Chemical synthesis of DNA is carried out with the growing chain of DNA attached to a solid support.

Since the DNA is made only with chemical reagents, the process requires some special modifications not necessary if biological enzymes were to manufacture the

controlled pore glass (CPG) Glass with pores of uniform sizes that is used as a solid support for chemical reactions such as artificial DNA synthesis

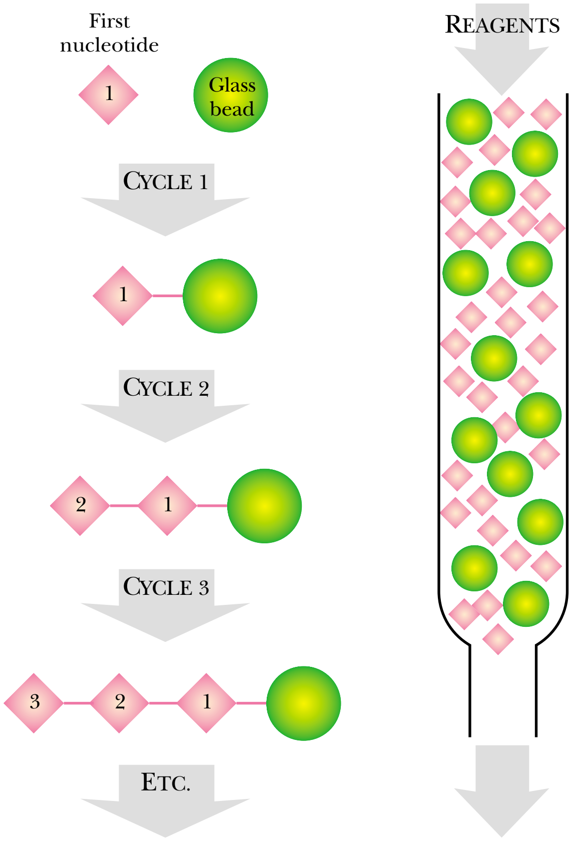


FIGURE 21.09 ***Chemical Synthesis of DNA on Glass Beads—Principle***

DNA is synthesized attached to porous glass beads in a column. Chemical reagents are trickled through the column one after the other. The first nucleotide is linked to the beads and each successive nucleotide is linked to the one before. After the entire sequence has been assembled, the DNA is chemically detached from the beads and eluted from the column.

In the most common method of making DNA, phosphoramidite groups react to link nucleotides together.

DNA. The first problem is that each deoxynucleotide has two hydroxyl groups, one used for bonding to the nucleotide ahead of it and the other for bonding to the nucleotide behind it in the nucleic acid chain. Chemical reagents cannot distinguish between these two hydroxyls. Therefore, each time a nucleotide is added, one of its hydroxyl groups must first be chemically blocked and the other must be activated. The standard **phosphoramidite method** for artificial chemical synthesis of DNA proceeds in the 3′ to 5′ direction. Consequently, before adding any new nucleotide to the chain, the 5′-hydroxyl of the previous nucleotide is blocked with a **dimethoxytrityl (DMT) group** and the 3′-hydroxyl is activated with a phosphoramidite. Note that the reagents for chemical synthesis are phosphoramidite nucleotides (with a single phosphorus), not nucleoside triphosphates as in biosynthesis (Fig. 21.11). Also, chemical synthesis occurs in the 3′ to 5′ direction, the opposite of biological DNA synthesis, which always occurs 5′ to 3′.

The first nucleotide is anchored to a glass bead via its 3′-OH group. The first base is actually added as a nucleoside without any phosphate group. It is bound to the glass bead via a spacer molecule (Fig. 21.12). The spacer helps prevent the bases in the growing nucleotide chain from reacting with the glass bead surface.

Next, acid (often trichloroacetic acid, TCA) is poured through the column to remove the DMT blocking group and expose the 5′-hydroxyl group on the first nucleotide. The second phosphoramidite nucleotide is added to the column. This nucleotide links to the first one via the single phosphate in the phosphoramidite moiety (Fig. 22.13). After each reaction step, the column is washed by acetonitrile to remove unreacted reagents and then flushed with argon to remove any traces of acetonitrile. The cycle continues until a full-length sequence is manufactured.

dimethoxytrityl (DMT) group Group used for blocking the 5′-hydroxyl of nucleotides during artificial DNA synthesis
phosphoramidite method Method for artificial synthesis of DNA that utilizes the reactive phosphoramidite group to make linkages between nucleotides

FIGURE 21.10 ***DNA Synthesizer***

Biologist programs an automated DNA synthesizer to produce a specific oligonucleotide for her research. Credit: Hank Morgan, Photo Researchers, Inc.

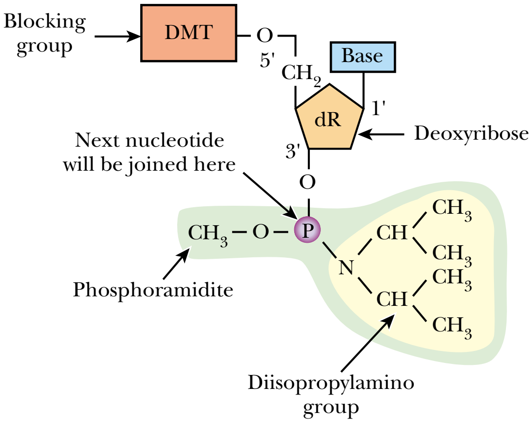


FIGURE 21.11 ***Phosphoramidite Nucleotides are used for Chemical Synthesis of DNA***

During chemical synthesis of DNA, modifications must be added to each nucleotide to ensure that the correct group reacts with the next chemical reagent. Each nucleotide has a blocking DMT group attached to the 5′-OH. The 3′-OH is activated by attaching a phosphoramidite group.

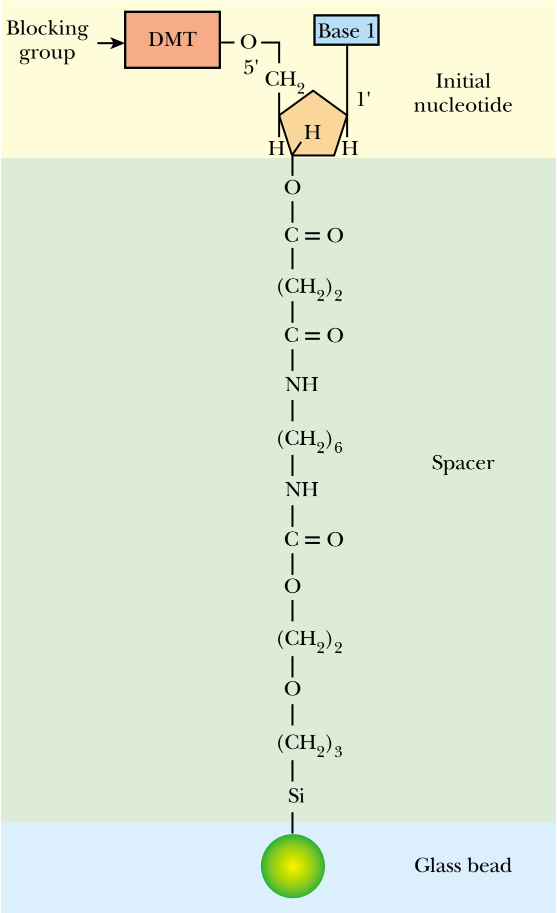


FIGURE 21.12 ***Addition of Spacer Molecule and First Base to Glass Bead***

The first nucleotide is linked to a glass bead via a spacer molecule attached to its 3′-OH group.

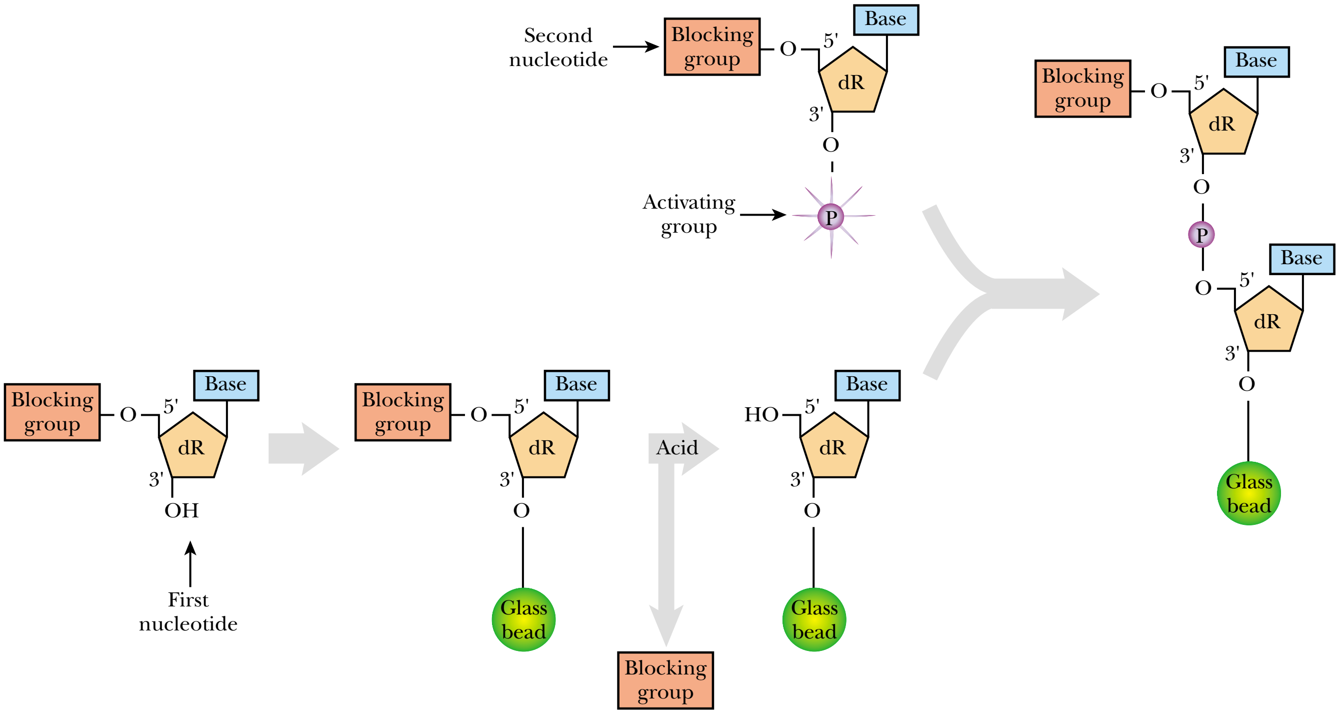


FIGURE 21.13 ***Chemical Synthesis of DNA—Nucleotide Addition***

During chemical synthesis of DNA, the DNA is added in a 3′ to 5′ direction. In order to add the successive bases correctly, the 3′-OH of incoming bases must be activated by phosphoramidite (purple) but the 5′-OH must be blocked with a dimethoxytrityl (DMT) group. After the first nucleotide is anchored to the glass bead its DMT blocking group is removed by acid. The next nucleotide can then link to the exposed 5′-OH by a phosphate linkage. Notice that the second nucleotide still has its 5′-OH blocked with DMT. The process continues by using acid to remove DMT from the second nucleotide, adding the third nucleotide and so on (not shown here).

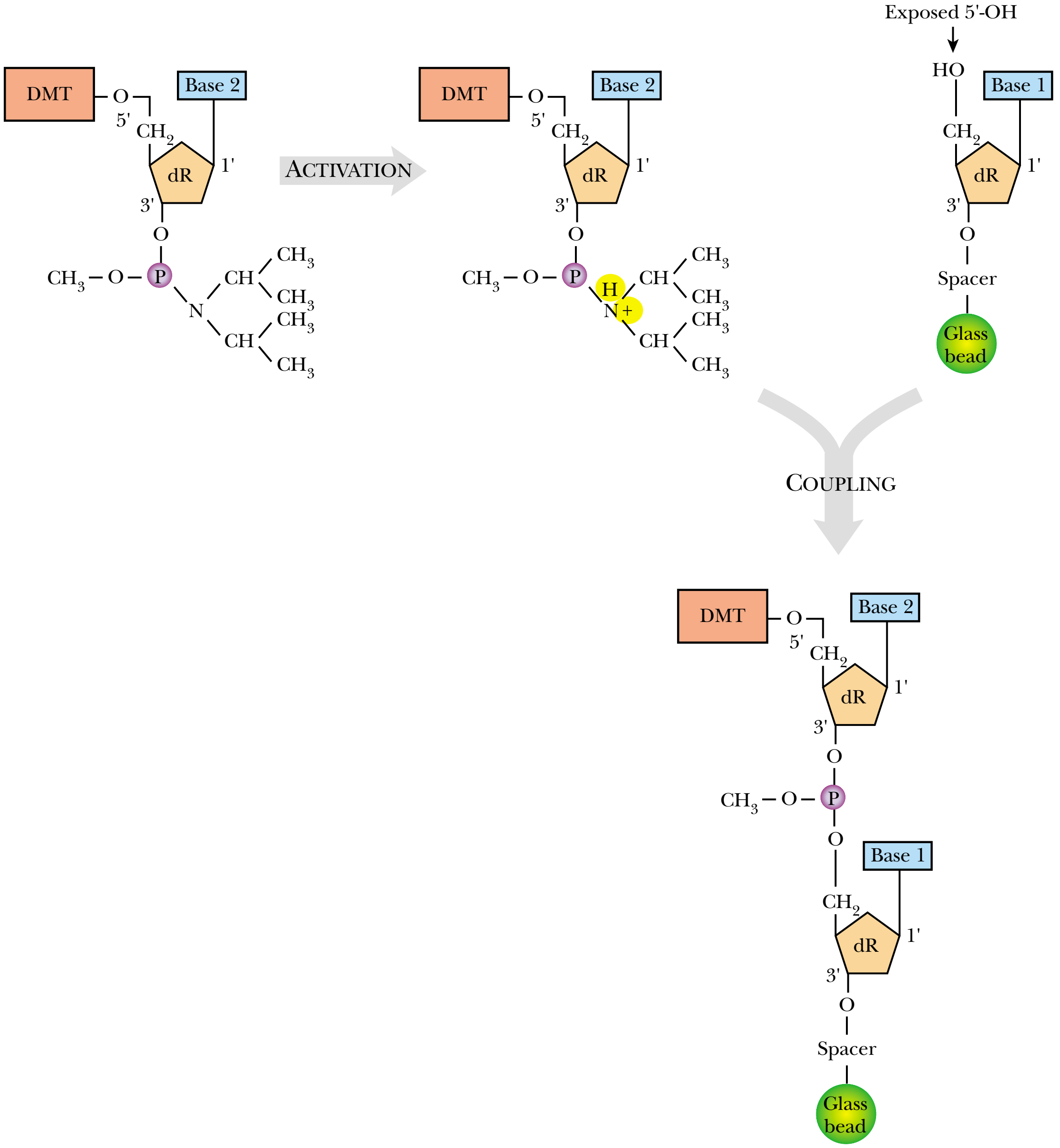


FIGURE 21.14 ***Chemical Synthesis of DNA—Coupling***

In order to couple a phosphoramidite nucleotide to the growing chain of DNA, the phosphoramidite moiety must be activated. Tetrazole activates the N of the diisopropylamino group by adding a proton. The diisopropylamino group is then displaced by the exposed 5′-OH of the acceptor nucleotide. The coupling reaction results in two nucleotides linked by a phosphite triester. Further reaction with iodine oxidizes this to a phosphate triester, which is then hydolyzed to a phosphodiester link by removal of the methyl group in the third position.

Blocking agents are used throughout the procedure to prevent the wrong groups from reacting.

The specific coupling of one nucleotide to another requires the addition of tetrazole, which activates the phosphoramidite group on the new nucleotide. The phosphoramidite then forms a bond with the exposed 5′-hydroxyl group of the previous nucleotide (Fig. 21.14). Since this coupling reaction is not 100% efficient, any remaining unreacted 5′-hydroxyl groups must be blocked by acetylation before adding the next nucleotide. This is done using acetic anhydride plus dimethylaminopyridine. Unused but unblocked 5′-hydroxyl groups would react in the next synthetic cycle and give rise to incorrect DNA sequences. After coupling, nucleotides are linked by a relatively unstable phosphite triester (Fig. 21.14). This is oxidized by iodine to a phosphate triester, or phosphotriester. The methyl group that occupies the third position of the phosphotriester is removed to give a phosphodiester linkage only after the whole DNA strand has been synthesized. After the phosphite oxidation step, the column is washed and treated with acid to expose the 5′-hydroxyl of the nucleotide that was just added. It is then ready for another nucleotide.

Since the amino groups of the bases are reactive they must also be protected throughout the whole reaction sequence. Benzoyl groups are normally added to protect the amino groups of adenine and cytosine, whereas, guanine is protected by isobutyryl group (thymine needs no protection as it has no free amino group). These groups are not shown in any of figures 21.11 through 21.14 and are only removed after synthesis of the whole strand of DNA. When all nucleotides have been added, the various protective groups are removed and the 5′-end of the DNA is phosphorylated, either chemically or by ATP plus polynucleotide kinase. After detaching from the column, the final product is purified by HPLC or gel electrophoresis to separate it from the defective shorter molecules that are due to imperfect coupling.

Chemical Synthesis of Complete Genes

The coupling efficiency limits the length of DNA that can be chemically synthesized. For example, a coupling efficiency of 98% would give a yield of around 50% for a 40-mer oligonucleotide and 10% for a 100-mer. Very short complete genes with sequences of up to 80 or 100 nucleotides can therefore be synthesized as single pieces of DNA. Longer sequences must be made in segments and assembled.

Whole genes have been made by synthesizing smaller fragments and then assembling them.

To chemically synthesize a complete gene, a series of overlapping fragments, representing both strands of the gene to be assembled are manufactured. These are purified and annealed together as shown in Fig. 21.15. Two alternatives are possible. In the first case, the fragments comprise the whole gene with only nicks remaining between the assembled fragments. DNA ligase is then used to seal the nicks between the segments. In the second case, only part of each strand is made and large single-stranded gaps remain after annealing the fragments. In this case, DNA polymerase I is used to fill in the gaps before joining the segments with ligase. In either variant, regulatory sequences may be included along with the coding sequence of the gene. In addition, artificial restriction sites (see Ch. 22) are often added flanking the artificial gene to allow its insertion into a cloning vector.

Peptide Nucleic Acid

Peptide nucleic acid (**PNA**) is a totally artificial molecule that is used as a DNA analog in genetic engineering. PNA is just what its name indicates, consisting of a polypeptide backbone with nucleic acid bases attached as side chains. Note that the polypeptide backbone of PNA is not identical to that of natural proteins (Fig. 21.16). PNA is designed to space out the bases that it carries at the same distances as found in the genuine nucleic acids. This enables a strand of PNA to base pair with a complementary strand of DNA or RNA.

PNA was deliberately designed with an uncharged backbone. The objective was that single strands of complementary PNA should form a triple helix with double stranded DNA. What actually happened was that one of the DNA strands was displaced and a triple helix was formed from two single strands of PNA plus one strand of DNA (Fig. 21.17). Not only is the repulsion due to phosphate in the backbone absent in PNA, but extra H-bonds form between the N atoms of the PNA peptide backbone and the phosphate of the single DNA strand. If two identical PNA strands are joined by a flexible linker (a "**PNA clamp**"), this binds even more avidly to its target DNA and forms a virtually irreversible triple helix.

Peptide nucleic acid is useful in many practical applications as it is not digested by either nucleases or proteases.

peptide nucleic acid (PNA) Artificial analog of nucleic acids with a polypeptide backbone
PNA clamp Two identical PNA strands that are joined by a flexible linker and are intended to form a triple helix with a complementary strand of DNA or RNA

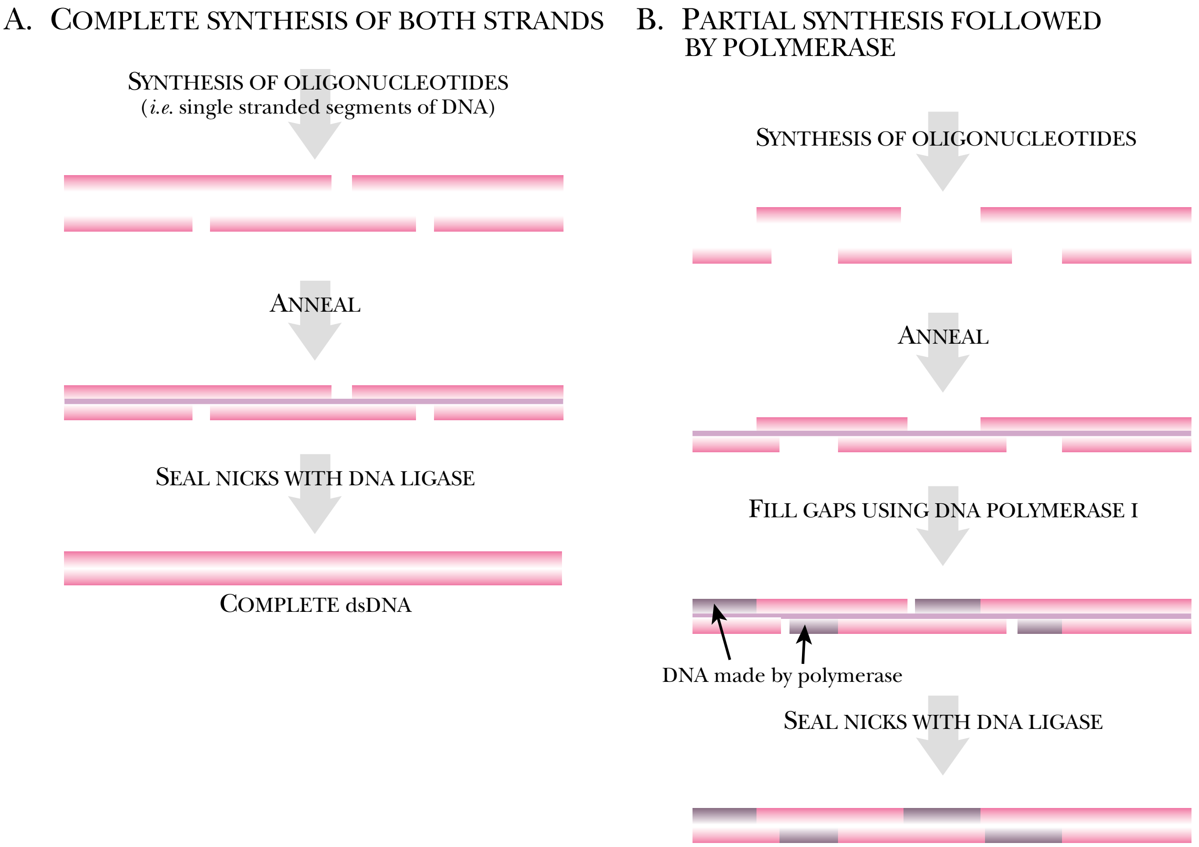

A. COMPLETE SYNTHESIS OF BOTH STRANDS
SYNTHESIS OF OLIGONUCLEOTIDES
(*i.e.* single stranded segments of DNA)
ANNEAL
SEAL NICKS WITH DNA LIGASE
COMPLETE dsDNA
B. PARTIAL SYNTHESIS FOLLOWED BY POLYMERASE
SYNTHESIS OF OLIGONUCLEOTIDES
ANNEAL
FILL GAPS USING DNA POLYMERASE I
DNA made by polymerase
SEAL NICKS WITH DNA LIGASE

FIGURE 21.15 ***Synthesis and Assembly of a Gene***

(A) Complete synthesis of both strands. Small genes can be chemically synthesized by making overlapping oligonucleotides. The complete sequence of the gene, both coding and non-coding strands, is made from small oligonucleotides that anneal to each other forming a double-stranded piece of DNA with nicks at intervals along the backbone. The nicks are then sealed using DNA ligase. (B) Partial synthesis followed by polymerase. To manufacture larger stretches of DNA, oligonucleotides are synthesized so that a small portion of each oligonucleotide overlaps with the next. The entire sequence is manufactured, but gaps exist in both the coding and non-coding strands. These gaps are filled using DNA polymerase I and the remaining nicks are sealed with DNA ligase.

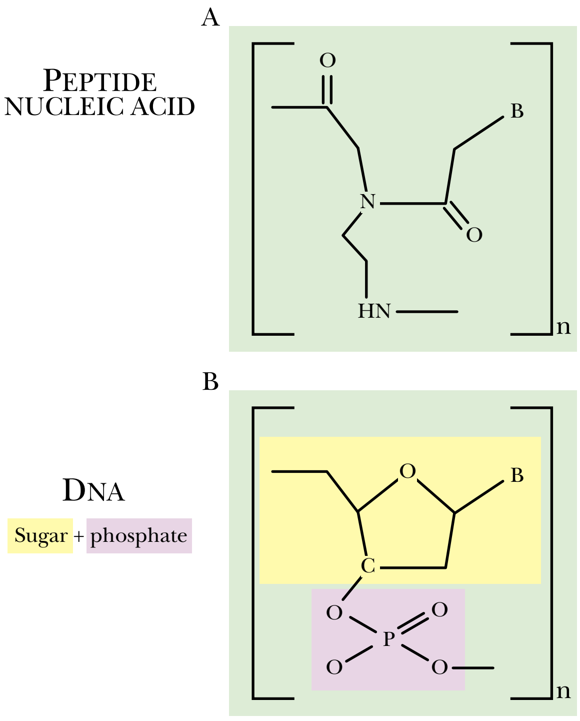


FIGURE 21.16 ***Structure of Peptide Nucleic Acid Compared to DNA***

A) The PNA polypeptide backbone. B) The sugar and phosphate backbone of normal DNA. B = nucleic acid base. The brackets with "n" indicate a polymer with multiple repeats.

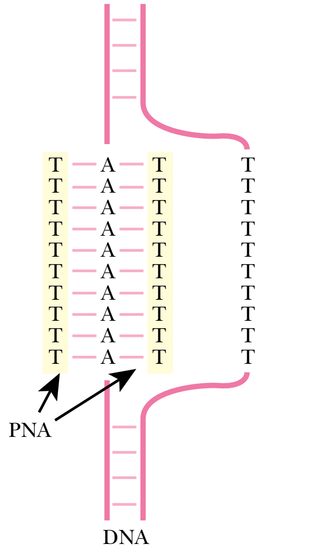


FIGURE 21.17 ***Triple Helix of Peptide Nucleic Acid with DNA***

PNA displaces one of the strands of a DNA double helix. Two strands of PNA pair with the adenine-rich strand of DNA to form a triple helix. Regions of DNA that are bound by PNA cannot be transcribed.

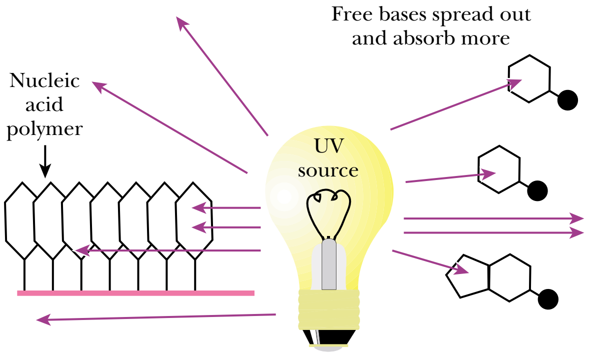


FIGURE 21.18 ***Absorption of UV Radiation by Nucleic Acids***

All nucleic acids absorb UV light by the aromatic rings of the bases. The phosphate backbone (pink line or black dots) is not involved in UV absorption. The structure of the nucleic acid dictates how much light the aromatic rings absorb. On the right side of the light bulb, free nucleotides are shown spread out such that each ring can absorb the UV light. Overall, the free nucleotides absorb more UV. In contrast, as shown on the left, the aromatic rings are stacked along the phosphate backbone in a nucleic acid polymer. In this configuration the rings shield each other and absorb less UV light.

The PNA backbone is very stable and is not degraded by any naturally occurring nucleases or proteases. PNA can be used to bind to and block target sequences of DNA in purine-rich regions and prevent transcription of DNA to give mRNA. It can also bind to RNA and prevent translation of mRNA into protein; i.e., it acts like antisense RNA. PNA is useful for laboratory applications and may also be used clinically in the near future. For example, antisense PNA that inhibits the translation of *gag-pol* mRNA of HIV-1 has been shown to reduce virus production by 99 percent in tissue culture. Similarly, antisense PNA can stop the *in vitro* translation of *Ha-ras* and *bcl-2* mRNA (both from cancer cells). In the case of *bcl-2*, PNA was also shown to inhibit gene expression by binding to the DNA at a purine-rich tract.

The major problem with using PNA clinically is that it penetrates cells poorly—much worse than natural nucleic acids. Recent developments suggest that PNA can enter cells effectively if it is coupled to other molecules that are taken up readily or if it is carried by positively charged liposomes.

Measuring the Concentration of DNA and RNA with Ultraviolet Light

The aromatic rings of the bases found in DNA and RNA absorb ultraviolet light with an absorption maximum at 260 nm. If a beam of UV light is shone through a solution containing nucleic acids, the proportion of the UV absorbed depends on the amount of DNA or RNA. This approach is widely used to measure the concentration of a solution of DNA or RNA. The amount of UV light absorbed by a series of standard DNA concentrations is measured to calibrate the technique. The amount of UV light absorbed by the unknown DNA is then determined and plotted on the standard curve to deduce the concentration.

The concentration of DNA or RNA is usually measured by absorption of ultraviolet light.

Proteins absorb UV at 280 nm, largely due to the aromatic ring of tryptophan. The relative purity of a preparation of DNA may be assessed by measuring its absorbance at both 260 and 280 nm and computing the ratio. Pure DNA has a A260/A280 ratio of 1.8. If protein is present, the ratio will be less than 1.8 and if RNA is present, the ratio will be greater than 1.8. Pure RNA has a A260/A280 ratio of approximately 2.0.

In a solution of unlinked nucleotides, the bases are more spread out and absorb more radiation. In a DNA double helix, the bases are stacked on top of each other and relatively less UV light is absorbed (Fig. 21.18). In single-stranded RNA (or in single-stranded DNA), the situation is intermediate. UV light is preferentially absorbed by the purine and pyrimidine rings, not the sugar or phosphate components of nucleic acids. The stacking of bases in a DNA double helix tends to shield them from UV radiation, as compared to free nucleotides in solution. In single stranded nucleic acids (whether DNA or RNA), the bases are only partially shielded and they absorb an intermediate amount of light. [UV absorption by nucleic acids is due to energy transitions of the delocalized electrons of the aromatic rings of the bases. When the bases stack, these electrons interact and no longer absorb UV radiation so readily. Hence the shielding effect.] So for a given amount of nucleotides, dsDNA absorbs less than ssRNA, which absorbs less than free nucleotides.

Radioactive Labeling of Nucleic Acids

Various sources of DNA can be distinguished by adding a specific label to one source. For example, when a bacteriophage attacks bacteria, the bacteriophage DNA enters the bacteria. A scientist who is studying this process will have to distinguish the bacteriophage DNA from the bacterial DNA by labeling one of the two. Originally incorporating **radioisotopes** into the DNA of the virus or bacteria would distinguish the two. For example, radioactive nucleotides would be provided to the bacteria so that during the next round of DNA synthesis, some of the radioactive molecules would be incorporated into the bacteria's chromosome. The bacterial DNA would be "**hot**," or radioactive, and the viral DNA would be "**cold**," or non-radioactive.

Radioactive isotopes of sulfur or phosphorus are used to label DNA or RNA.

Radioisotopes are the radioactive forms of an element. In molecular biology, two are especially important: the radioactive isotopes of phosphorus, ^{32}P, and sulfur, ^{35}S. Nucleic acids consist of nucleotides linked together by phosphate groups, each containing a central phosphorus atom. If ^{32}P is inserted at this position we have radioactive DNA or RNA (Fig. 21.19). The half life of ^{32}P is 14 days, which means that half of the radioactive phosphorus atoms will have disintegrated during this time period, so experiments using ^{32}P need to be done fast!

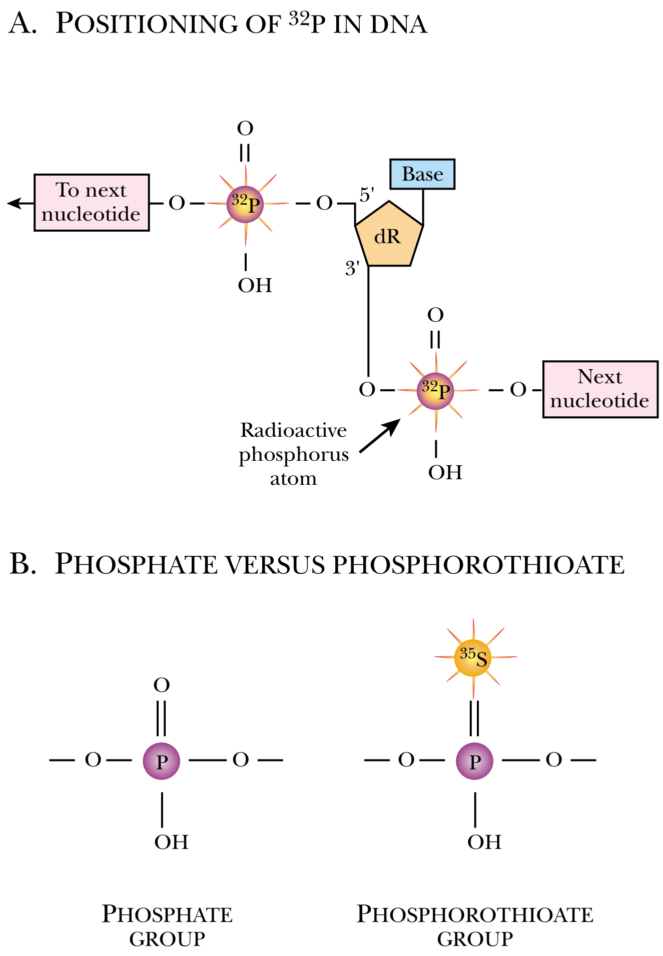
The sulfur isotope, ^{35}S, is also widely used. Since sulfur is not a normal component of DNA or RNA, we use **phosphorothioate** derivatives. A normal phosphate group has four oxygen atoms around the central phosphorus. In a phosphorothioate one of these is replaced by sulfur (Fig. 21.19). To introduce ^{35}S into DNA or RNA, phosphorothioate groups containing radioactive sulfur atoms are used to link together the nucleotides. Despite these complications, ^{35}S is usually preferable to ^{32}P in most molecular biology applications. There are two reasons: first, the half life of ^{35}S is 88 days, so it doesn't disappear so fast; second, the radiation emitted by ^{35}S is of lower energy than for ^{32}P. Therefore the radiation doesn't travel so far and the radioactive bands are more precisely located and not so fuzzy. In short, ^{35}S is more accurate.

Detection of Radio-Labeled DNA

The two methods most widely used in molecular biology to measure radioactivity are **scintillation counting** and **autoradiography**. If a sample is in liquid or on a strip of filter paper, the amount of radioactivity is measured using scintillation counting. If the sample is flat, such as an agarose gel, scientists use autoradiography to pinpoint the location of radioactive bands or spots.

Scintillation counting relies on special chemicals called **scintillants**. These emit a flash of light when high energy electrons known as beta-particles are released by the radioactive isotopes in DNA. The light pulses from the scintillant are detected by a

autoradiography Allowing radioactive materials to take pictures of themselves by laying them flat on photographic film
"cold" Slang for non-radioactive
"hot" Slang for radioactive
phosphorothioate A phosphate group in which one of the four oxygen atoms around the central phosphorus is replaced by sulfur
radioisotope Radioactive form of an element
scintillant Molecule that emits pulses of light when hit by a particle of radioactivity
scintillation counting Detection and counting of individual microscopic pulses of light



A. POSITIONING OF ^{32}P IN DNA

B. PHOSPHATE VERSUS PHOSPHOROTHIOATE

FIGURE 21.19 ***Use of ^{32}P and ^{35}S to Label Nucleic Acids***

(A) Positioning of ^{32}P in DNA. To make radioactive DNA, the phosphorus atom in the phosphate backbone is replaced with its radioactive isotope, ^{32}P. (B) Phosphate versus phosphorothioate. Instead of replacing the phosphorus atom of the phosphate group, one of the oxygen atoms can be replaced by the radioactive isotope of sulfur, ^{35}S, giving a phosphorothioate.

Scintillation counters measure radioactivity in liquid samples whereas autoradiography is used to locate radioactive molecules on gels or membranes.

photocell (Fig. 21.20). A **scintillation counter** is simply a very sensitive device for recording faint light pulses. To use the scintillation counter, radioactive samples to be measured are added to a vial containing scintillant fluid and loaded into the counter. The counter prints out the number of light flashes it detects within a designated time. The researcher can then compare the number of flashes with a series of standards to determine the relative amount of radioactive DNA in the sample.

Scintillation counters can be used to measure light generated by chemical reactions, also. In this case, the light is emitted directly so no scintillant fluid is needed and the luminescent sample is merely inserted directly. The detection of light emitted by luciferase and lumi-phos in genetic analysis is described in Ch. 25.

Autoradiography is used for detecting radioactively labeled DNA or RNA in a gel after separation by electrophoresis. Very often DNA or RNA bands in a gel are transferred onto membranes by blotting to allow more convenient handling during autoradiography. In addition, autoradiography can detect radioactive DNA bound to filter paper during hybridization experiments. Whether the radioactive DNA is in a gel or on a membrane or piece of filter paper, the radioactive isotope emits beta-particles. The particles will turn regular photographic film black, exactly the same way that light turns photographic film black. To do autoradiography, the gel or filter containing radioactive DNA is dried so the photographic film does not stick. Next, a sheet of photographic film is laid on top of the gel or filter and left for several hours or sometimes even for days. The film darkens where the radioactive DNA bands or spots are found (Fig. 21.21). Exposing the film must be carried out in a dark room to avoid visible light.

scintillation counter Machine that detects and counts pulses of light

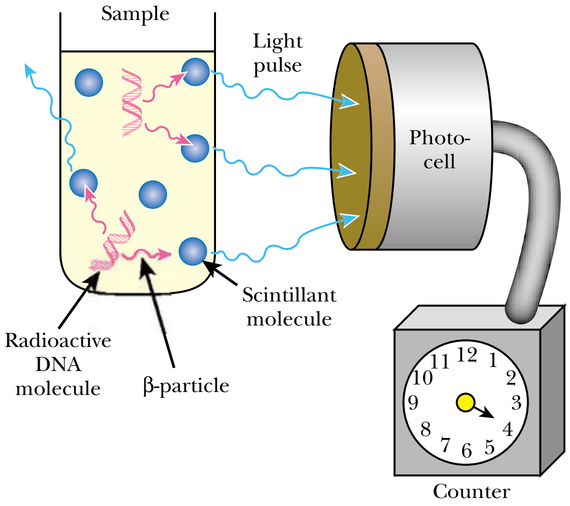
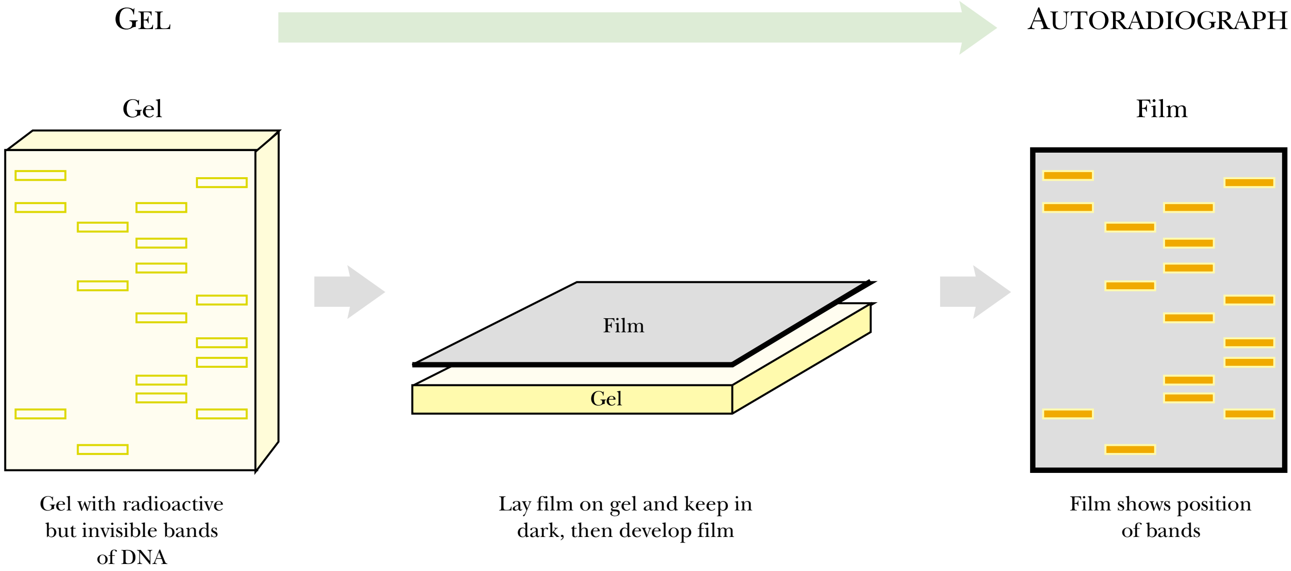


FIGURE 21.20 ***Scintillation Counter is Used to Measure Radioactivity***

Radioactive DNA is mixed with a liquid scintillant. The scintillant molecules absorb the β-particles emitted by the ^{32}P in the DNA, and in turn emit a flash of light. The photocell counts the number of light pulses in a specific time period.



Gel with radioactive but invisible bands of DNA

Lay film on gel and keep in dark, then develop film

Film shows position of bands

FIGURE 21.21 ***Autoradiography to Detect Radio-Labeled DNA or RNA***

A gel containing radioactive DNA or RNA is dried and a piece of photographic film is laid over the top. The two are loaded into a cassette case that prevents light from entering. After some time (hours to days), the film is developed and dark lines appear where the radioactive DNA was present.

Fluorescence in the Detection of DNA and RNA

DNA or RNA may be labeled with fluorescent dyes.

Most classic work in biochemistry and molecular biology was done with radioactive tracers and probes. Although the low levels of radioactivity used in laboratory analyses are little real hazard, the burden of storage and proper disposal of the waste has made it relatively cheaper and quicker to use other detection methods. Some of the newer DNA detection methods use **fluorescence**, chemical tagging and hybridization.

Fluorescence occurs when a molecule absorbs light of one wavelength and emits light of lower energy at a longer wavelength (Fig. 21.22). Detection of fluorescence requires both a beam of light to excite the dye and a photo-detector to detect the fluorescent emission. Fluorescent dyes can be attached to DNA molecules and modern automated methods for DNA sequencing make use of such fluorescent tagging (see

Fluorescence Process in which a molecule absorbs light of one wavelength and then emits light of another, longer, lower energy wavelength

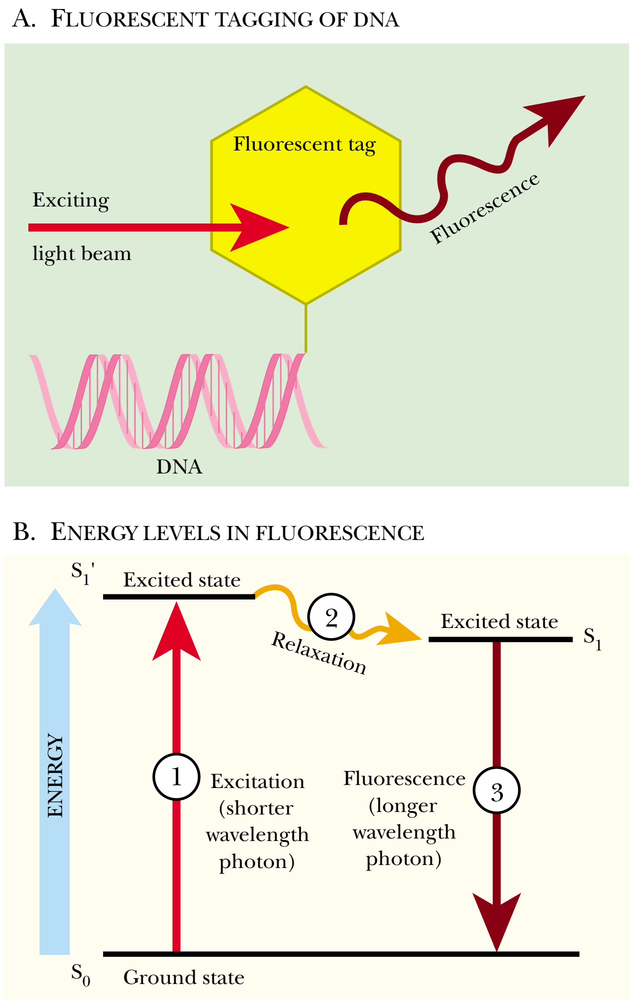

A. FLUORESCENT TAGGING OF DNA
Fluorescent tag
Exciting
light beam
Fluorescence
DNA
B. ENERGY LEVELS IN FLUORESCENCE
S_1'
Excited state
2
Relaxation
Excited state
S_1
ENERGY
1
Excitation (shorter wavelength photon)
Fluorescence (longer wavelength photon)
3
S_0
Ground state

FIGURE 21.22
Fluorescence Detection

(A) Fluorescent tagging of DNA. During DNA synthesis, a nucleotide linked to a fluorescent tag is incorporated at the 3′ end of the DNA. A beam of light excites the fluorescent tag, which in turn releases light of a longer wavelength (fluorescence).
(B) Energy levels in fluorescence. The fluorescent molecule attached to the DNA has three different energy levels, S_0, S_1', and S_1. The S_0 or ground state is the state before exposure to light. When the fluorescent molecule is exposed to a light photon of sufficiently short wavelength, the fluorescent tag absorbs the energy and enters the first excited state, S_1'. Between S_1' and S_1, the fluorescent tag relaxes slightly, but doesn't emit any light. Eventually the high-energy state releases its excess energy by emitting a longer wavelength photon. This release of fluorescence returns the molecule back to the ground state.

Ch. 24). In this case, each of the nucleotide bases is labeled with a different fluorescent dye. As each of the bases pass through the sequencing machine, a laser activates the dye which fluoresces by emitting light of a lower wavelength. A detector records the wavelength of emitted light, and translates the data to give the identity of each passing base. The nucleotide sequence is printed for the researcher.

Another instrument that uses fluorescence is the **Fluorescence Activated Cell Sorter or FACS**. Its job, originally, was sorting cells labeled with a fluorescent tag from those that were untagged. The new generation of more sensitive FACS machines is capable of sorting chromosomes labeled by hybridization with fluorescent tagged DNA probes (Fig. 21.23). Another novel use for a FACS machine is in sorting whole small organisms such as the nematode worm *C. elegans*, which is widely used in genetic analysis. Gene expression may be monitored by making gene fusions to green fluorescent protein (see Ch. 25 for gene fusion analysis). Consequently, organisms with different levels of fluorescence are produced and may need to be screened by automatic sorting if generated in large numbers.

Another recent and growing use for FACS technology is in fluorescent bead sorting. Many reactions in modern high-throughput screening involve anchoring one or more reagents to microscopic polystyrene beads. In some reaction schemes, fluorescently labeled molecules may bind to colorless reagents, such as DNA or proteins, previously attached to the beads. In other cases, the beads may be color coded by fluorescent dyes before reaction occurs. In either case, the beads are sorted after

fluorescence activated cell sorter (FACS) Instrument that sorts cells (or chromosomes) based on fluorescent labeling

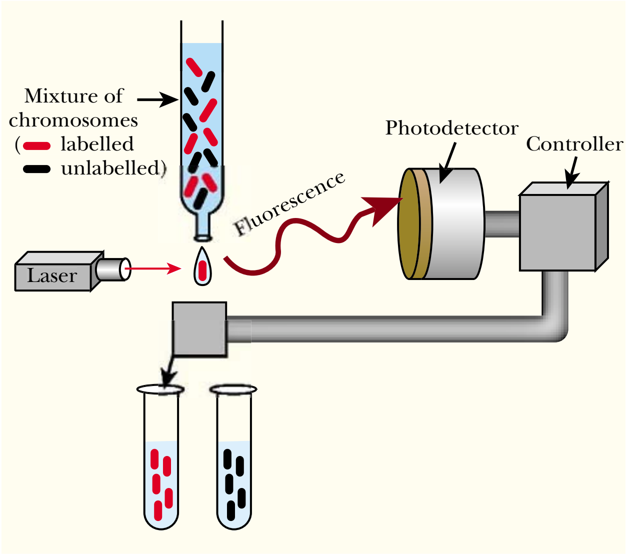


FIGURE 21.23 ***FACS Machine Can Sort Chromosomes***

FACS machines can separate fluorescently labeled chromosomes from unlabeled ones. Liquid carrying a mixture of labeled and unlabeled chromosomes passes by a laser, which excites the fluorescent tags. Whenever the photo-detector detects fluorescence, the controller module directs that drop into the test tube on the left. When no fluorescence is emitted, the controller module directs the drop into the test tube on the right. This sorting procedure allows the separation of fluorescently labeled particles from unlabeled ones.

reaction. Modern equipment can distinguish between beads with different colored fluorescent dyes, not merely separate fluorescent beads from unlabeled ones based on brightness.

Chemical Tagging with Biotin or Digoxigenin

Chemical tagging of DNA allows detection by a two stage approach that generates a colored product.

Biotin (a vitamin) and **digoxigenin** (a steroid from foxglove plants) are two molecular tags widely used for labeling DNA. Both biotin and digoxigenin are linked to uracil, which is normally a component of RNA not DNA. Therefore, to label DNA, uracil must be incorporated into the DNA instead of thymine. Although DNA polymerase will not incorporate uridine triphosphate (UTP) it will use deoxyuridine triphosphate (deoxyUTP), the deoxyribose containing version of this nucleotide. If deoxyUTP labeled with biotin or digoxigenin is added to the synthesis reaction, DNA polymerase will incorporate the labeled uridine where thymidine would normally be inserted. The biotin or digoxigenin tags stick out from the DNA without disrupting its structure (Fig. 21.24).

Biotin and digoxigenin are not colored or fluorescent molecules, but they can be detected in a two-stage process. The first step is to bind a molecule that can be detected. Biotin is a vitamin required both by animals and many bacteria. Scientists learned the biology of this vitamin to find a molecule that binds to biotin. Chickens lay highly nutritious eggs that would be a paradise for invading bacteria. One of the defense mechanisms to protect the egg from bacterial attack deploys a protein known as **avidin**. This protein, found in egg white, binds biotin so avidly that invading bacteria become vitamin deficient. Molecular biologists use avidin to bind the biotin tag. Attached to other side of the avidin is another molecule that provides the actual detection system. Digoxigenin also requires a second, detectable molecule. In this case, an antibody that recognizes and binds to the digoxigenin is used. (Antibodies are proteins made by the immune system, which specifically recognize foreign molecules.) As before, the detection system is attached to the other side of the antibody.

There are different options for the second step in detecting biotin/avidin or digoxigenin/antibody complexes. The first option is to attach an enzyme that generates a colored product to the avidin or the antibody. For example, avidin can be conjugated to

avidin A protein from egg white that binds biotin very tightly
biotin One of the B family of vitamins that is also widely used for chemical labeling of DNA molecules
digoxigenin A steroid from foxglove plant widely used for chemical labeling of DNA molecules

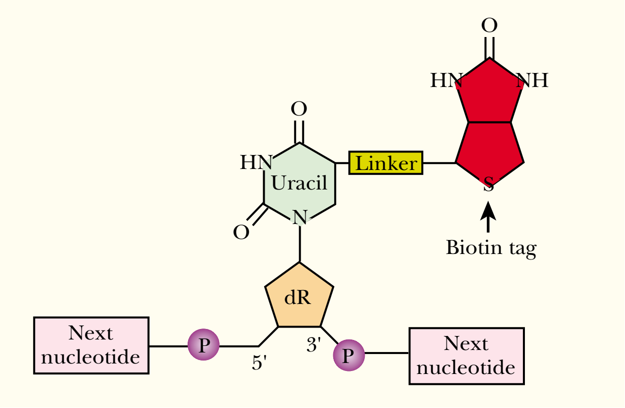


FIGURE 21.24 ***Labeling DNA with Biotin***

Uracil can be incorporated into a strand of DNA if the nucleotide has a deoxyribose sugar. Prior to incorporation, the uracil is tagged with a biotin molecule attached via a linker. This mode of attachment allows the biotin to stick out from the DNA helix without disrupting its structure.

Colored or luminescent products may be released when enzymes split specially designed artificial substrates.

alkaline phosphatase, an enzyme that snips phosphate groups from a wide range of molecules. One substrate for alkaline phosphatase, an artificial **chromogenic substrate** known as "**X-phos**", produces a blue dye when cleaved. X-phos consists of a dye precursor bound to a phosphate group, and when alkaline phosphatase cleaves the phosphate off, the dye precursor is converted to a blue dye by oxygen in the air (Fig. 21.25).

Another option to detect biotin/avidin or digoxigenin/antibody is an enzyme that produces light by a chemical reaction, known as **chemiluminescence**. Alkaline phosphatase is still conjugated to the avidin or antibody, but a different substrate, called "**lumi-phos**", is added. Lumi-phos consists of a light-emitting group bound to the phosphate group. When alkaline phosphatase splits off the phosphate, the unstable luminescent group emits light. Detecting and recording the light is accomplished by using photographic film if the DNA is on a filter or in a gel, or scanned by an instrument capable of detecting light emissions if the DNA is in a solution.

The Electron Microscope

Bacteria are just visible under a light microscope.

With an ordinary light microscope, objects down to approximately a micron (a millionth of a meter) in size can be seen. Typical bacteria are a micron or two long by about half a micron wide. Although bacteria are visible under a light microscope, their internal details are too small to see. The resolving power of a microscope depends on the wavelength of the light. In a light microscope, if two dots are less than about half a wavelength apart, they cannot be distinguished. Visible light has wavelengths in the range of 0.3 (blue) to 1.0 (red) micron, so bacteria are just at the limits of visible detection and most viruses cannot be seen.

Electron microscopes allow viruses, subcellular components and even single macromolecules to be visualized.

A beam of electrons has a much smaller wavelength than visible light and so can distinguish detail far beyond the limits of resolution by light. Electron beams may be focused like visible light except that the lenses used for electron beams are not physical (glass absorbs electrons) but electromagnetic fields that alter the direction in which the electrons move. Using an electron microscope allows visualization of the layers of the bacterial cell wall and of the folded-up bacterial chromosome, which appears as a light patch against a dark background. When an electron beam is fired through a sample, materials that absorb electrons more efficiently appear darker. Because electrons are easily absorbed, even by air, an electron beam must be used inside a vacuum chamber and the sample must be sliced extremely thin (Fig. 21.26).

To improve contrast, cell components are usually stained with compounds of heavy metals such as uranium, osmium or lead, all of which strongly absorb electrons.

alkaline phosphatase An enzyme that cleaves phosphate groups from a wide range of molecules
chemiluminescence Production of light by a chemical reaction
chromogenic substrate Substrate that yields a colored product when processed by an enzyme
lumi-phos Substrate for alkaline phosphatase that releases light upon cleavage
X-phos Substrate for alkaline phosphatase that is cleaved to release a blue dye

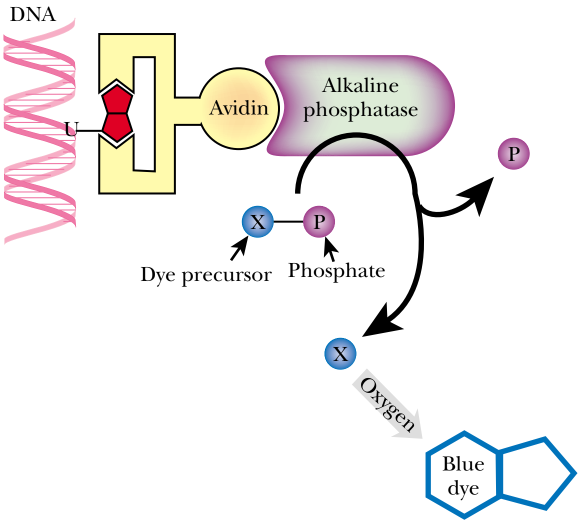
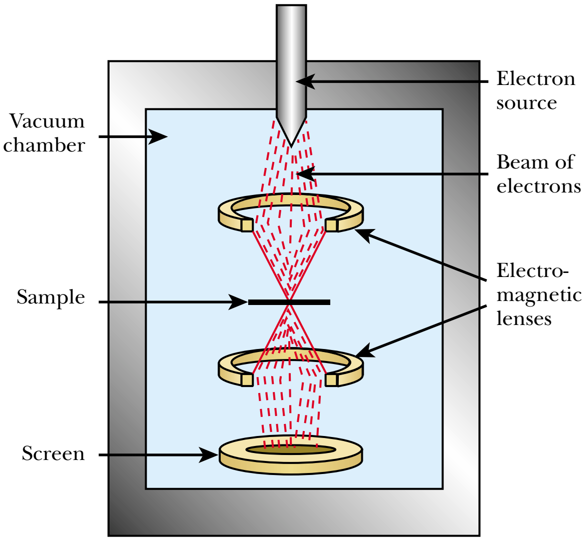


FIGURE 21.25 ***Detection Systems for Biotin***

DNA that has biotin attached via uracil can be detected with a two-step process. First, avidin is bound to the biotin. The avidin is conjugated to an enzyme called alkaline phosphatase, which cleaves phosphate groups from various substrates. Second, a substrate such as X-phos (shown) or lumi-phos (not shown) is added. Alkaline phosphatase removes the phosphate group from either substrate. In the case of X-phos, cleavage releases a precursor that reacts with oxygen to form a blue dye. If the substrate is lumi-phos, cleavage allows the unstable luminescent group to emit light.

FIGURE 21.26 ***Principle of the Electron Microscope***

Electron microscopy can reveal the substructures of animal cells, viruses, and bacteria. A beam of electrons is emitted from a source and is focused on the sample using electromagnetic lenses. When the electrons hit the sample, components such as cell walls, membranes, etc, absorb electrons and appear dark. The image is viewed on a screen or may be transferred to film for a permanent record. Since air molecules also absorb electrons, the entire process must be done in a vacuum chamber.



Individual, uncoiled DNA molecules can be seen if they are shadowed with metal atoms to increase electron absorption (Fig. 21.27). Shadowing is done by spreading the DNA out on a grid and then rotating it in front of a hot metal filament. Metal atoms evaporate and cover the DNA. Gold, platinum, or tungsten are typically used for shadowing.

Replicating plasmid DNA, showing the replication forks has been visualized under the electron microscope, as have a variety of other DNA and RNA molecules. A more recent example of this approach was the direct visualization of the introns found in eukaryotic genes (see Ch. 12). The messenger RNA and the DNA both contain the exons that comprise the coding sequence, but the final mRNA lacks the introns (noncoding regions). If mRNA is hybridized to single stranded DNA from the corresponding gene, the results are regions of base pairing (the exons) interrupted by loops due

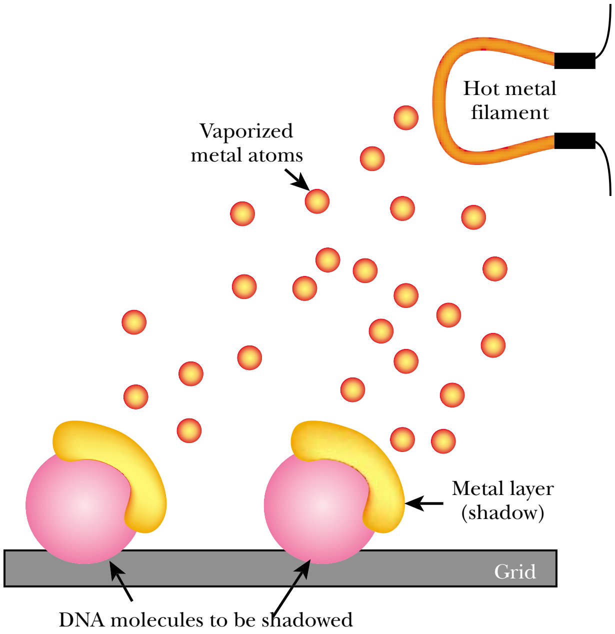


FIGURE 21.27 ***Metal Shadowed DNA Molecules are Visible under an Electron Microscope***

A hot metal filament releases vaporized metal atoms into the chamber containing a sample of DNA. The sample is rotated around the filament and metal ions attach to the exposed surface of the DNA. Once the DNA has a coat of metal atoms, it can be visualized by electron microscopy.

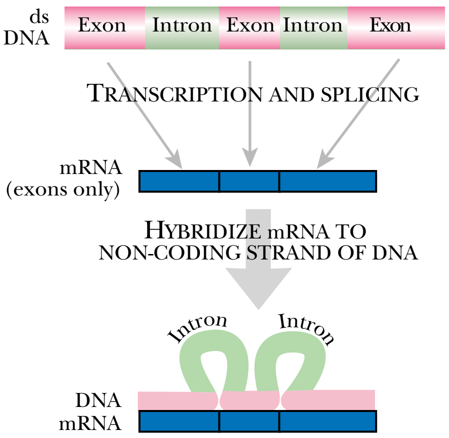

ds DNA
Exon
Intron
Exon
Intron
Exon
TRANSCRIPTION AND SPLICING
mRNA (exons only)
HYBRIDIZE mRNA TO NON-CODING STRAND OF DNA
Intron
Intron
DNA
mRNA

FIGURE 21.28 ***R-Loop Analysis to Visualize Introns***

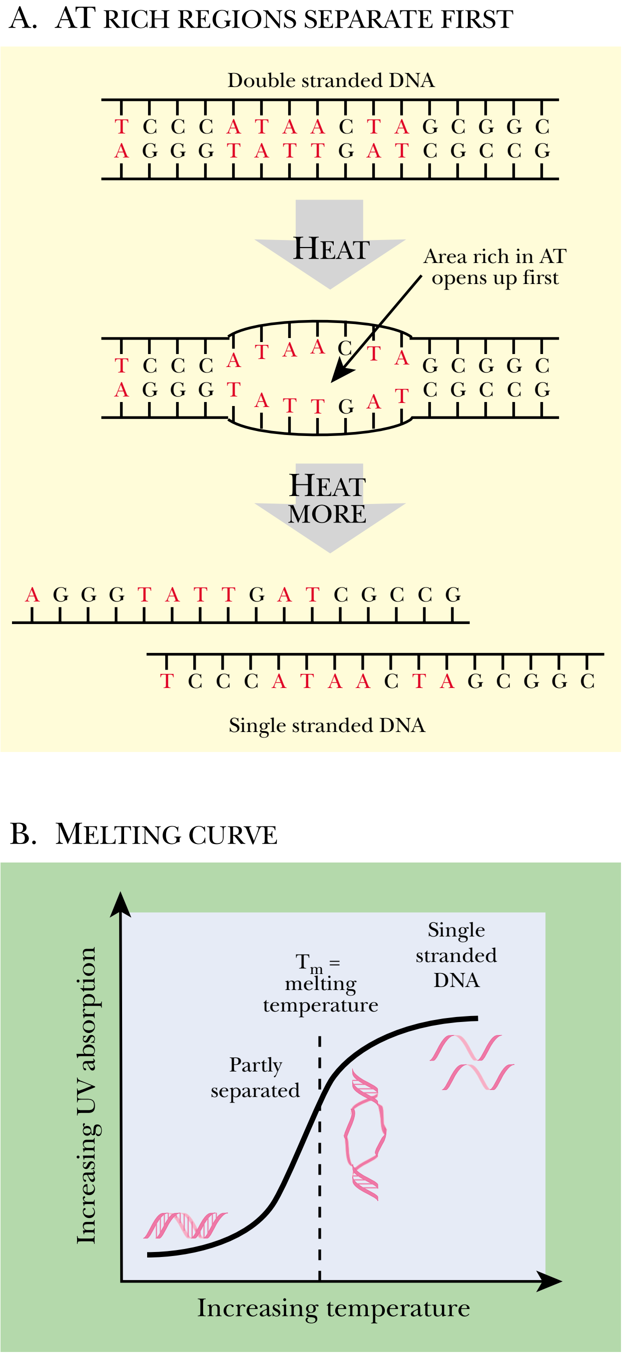
R-loop analysis can be used to visualize the introns of eukaryotic genes. First, a double-stranded DNA molecule is denatured into two single strands. One strand is annealed to the corresponding mRNA. Because RNA lacks introns the DNA and RNA anneal only in the coding regions. The introns, which are only found in the DNA, remain single-stranded and loop out from the heteroduplex. The entire complex can be visualized by electron microscopy after shadowing with metal ions.

to the extra intron sequences in the DNA. This is called **R-loop analysis** (Fig. 21.28). The resulting loops can be directly seen under the electron microscope.

Hybridization of DNA and RNA

The DNA double helix consists of two strands of nucleotides twisted around each other and held together by hydrogen bonding between the bases. If a solution of DNA is heated, the input of energy makes the molecules vibrate and the hydrogen bonds start coming apart. If the temperature is high enough, the DNA comes completely apart

R-loop analysis Hybridization of the DNA copy of a gene to the corresponding mRNA that results in the appearance of loops, which represent the intervening sequences in the DNA that have no partners in the mRNA



A. AT RICH REGIONS SEPARATE FIRST

Double stranded DNA
T C C C A T A A C T A G C G G C
A G G G T A T T G A T C G C C G
HEAT
Area rich in AT opens up first
T C C C A T A A C T A G C G G C
A G G G T A T T G A T C G C C G
HEAT MORE
A G G G T A T T G A T C G C C G
T C C C A T A A C T A G C G G C
Single stranded DNA

B. MELTING CURVE

Increasing UV absorption
Increasing temperature
T_m = melting temperature
Single stranded DNA
Partly separated

FIGURE 21.29 ***Melting of DNA at High Temperature***

(A) AT rich regions separate first. Molecules of double stranded DNA usually have some regions with relatively more AT than GC base pairs (red area). When the DNA is heated, the two hydrogen bonds of an AT base pair dissolve before the three hydrogen bonds of a GC base pair. As more heat is applied, the GC base pairs also come apart and two single strands of DNA are formed.
(B) Melting curve. The fraction of DNA that is melted or denatured can be determined by measuring the UV absorbance. As double-stranded DNA (bottom left of graph) is more compact it absorbs less UV light. Upon heating, the AT rich regions start to open up. The partially denatured DNA absorbs more UV since some of its base pairs are more exposed. Finally, the DNA is fully denatured into single strands. The bases are more exposed still and show maximum absorbance. The midpoint of the curve is defined as the T_m, or melting temperature for that specific DNA sequence.

Heating of the DNA double helix melts it into single strands.

into two separate strands (Fig. 21.29A). This is known as **denaturation** or "**melting**". Since the GC base pair has three hydrogen bonds compared to the two holding AT together, GC base pairs are stronger than AT base pairs. Therefore, as the temperature rises, AT pairs come apart first and regions of DNA with lots of GC base pairs melt at higher temperatures.

The **melting temperature** of a DNA molecule is defined as the temperature at the halfway point on the melting curve. The halfway point is used because it is more accurate than trying to guess precisely where melting is complete. Melting is followed by measuring the UV absorption, since disordered DNA absorbs more UV light

denaturation When describing proteins or other biological polymers, refers to the loss of correct 3-D structure
melting When used to describe DNA, refers to its separation into two strands as a result of heating
melting temperature The temperature at which the two strands of a DNA molecule are half-way unpaired

(Fig. 21.29B). Overall, the higher the proportion of GC base pairs, the higher the melting temperature of a DNA molecule.

Single strands will base pair to recreate a double helix if slowly cooled together.

If denatured DNA is cooled again, the single DNA strands will recognize their partners by base pairing and double stranded DNA will re-form. This is referred to as **annealing**. For proper annealing, the DNA must be cooled slowly to allow the single strands time to find the correct partners. Consider two completely different DNA molecules. If they are mixed, melted and then cooled to re-anneal the single strands, each single strand will recognize and pair with its original complementary strand (Fig. 21.30). Suppose on the other hand, two closely related DNA molecules are used. Although the sequences may not match perfectly, nonetheless, if they are similar enough, some base pairing will occur. The result will be the formation of **hybrid DNA** molecules.

Hybrid double helices may be formed by annealing single strands that are related in sequence.

The formation of hybrid DNA molecules has a wide variety of uses. For example, how closely two DNA molecules are related may be tested. To do this, a sample of the first DNA molecule is heated to melt it into single-stranded DNA. The single strands are then attached to a suitable filter. Next, the filter is treated chemically to block any remaining sites that would bind DNA. Then, after melting, a solution of the second DNA molecule is poured through the filter (Fig. 21.31). Some of the single strands of DNA molecule No. 2 will base pair with the single strands of DNA molecule No. 1 and will stick to the filter. (As discussed above, DNA molecule No. 2 must be labeled by radioactivity, fluorescence or some other way to enable its detection.) The more closely related the two molecules are, the more hybrid molecules will be formed and the higher the proportion of molecule No. 2 that will be bound by the filter. For example, if the DNA for a human gene, such as hemoglobin, was fully melted and bound to a filter, then DNA for the same gene but from different animals could be tested. We might expect gorilla DNA to bind strongly, frog DNA to bind weakly and mouse DNA to be intermediate. [Several variants of nucleic acid hybridization are in use. This version, involving the binding of DNA to DNA is known as Southern blotting—see below.]

Another use for **hybridization** is in isolating genes for cloning. Suppose we already have the human hemoglobin gene and want to isolate the corresponding gorilla gene. First the human DNA is bound to the filter as before. Then gorilla DNA is cut it into short segments with a suitable restriction enzyme (see Ch. 22 for details). The gorilla DNA is heated to melt it into single strands and poured over the filter. The DNA fragment that carries the gorilla gene for hemoglobin will bind to the human hemoglobin gene and remain stuck to the filter. Other, unrelated genes will not hybridize. This approach allows the isolation of new genes provided a related gene is available for hybridization.

Probes are labeled molecules of DNA (or RNA) that are used to detect complementary sequences by hybridization.

A wide range of methods based on hybridization is used for analysis in molecular biology. The basic idea in each case is that a known DNA sequence acts as a "**probe**." Generally the **probe molecule** is labeled by radioactivity or fluorescence for ease of detection. The probe is used to search for identical or similar sequences in the experimental sample of target molecules. Both the probe and the target DNA must be treated to give single-stranded DNA molecules that can hybridize to each other by base pairing. In the previous example, the probe DNA would be the human hemoglobin DNA since the sequence is already known. The gorilla DNA would be the sample of target molecules.

Southern, Northern, and Western Blotting

Isolating new genes from related species provides a wealth of information for a scientist. Many times, scientists that are studying human genes need to find similar genes

annealing The rejoining of separated single strands of DNA to form a double helix
hybrid DNA Artificial double-stranded DNA molecule made by two single strands from two different sources
hybridization Formation of double-stranded DNA molecule by annealing of two single strands from two different sources
probe Short for probe molecule
probe molecule Molecule that is tagged in some way (usually radioactive or fluorescent) and is used to bind to and detect another molecule

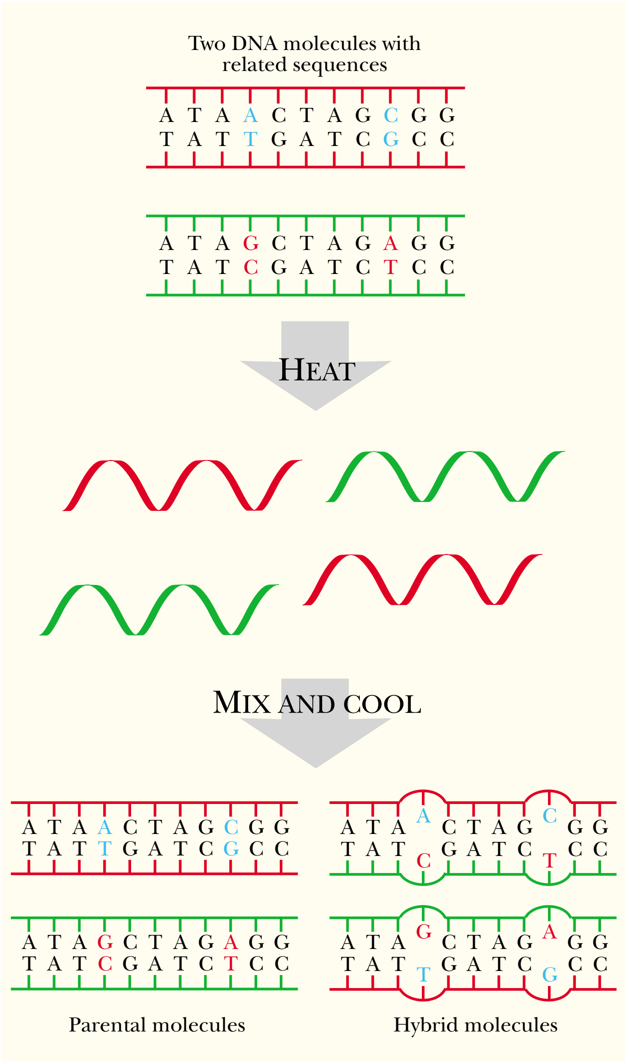


FIGURE 21.30 ***Hybrid DNA Formed by Melting and Re-Annealing***

Two double-stranded DNA molecules with similar but not identical sequences are heated to form single-stranded DNA. The mixture is cooled slowly, allowing the sequences to anneal. Sometimes, the two original sequences will anneal into the parental molecules. Occasionally, the two will mix, forming a hybrid molecule with one green strand of DNA and one red strand. Since the sequences are not identical, there are regions that do not base pair.

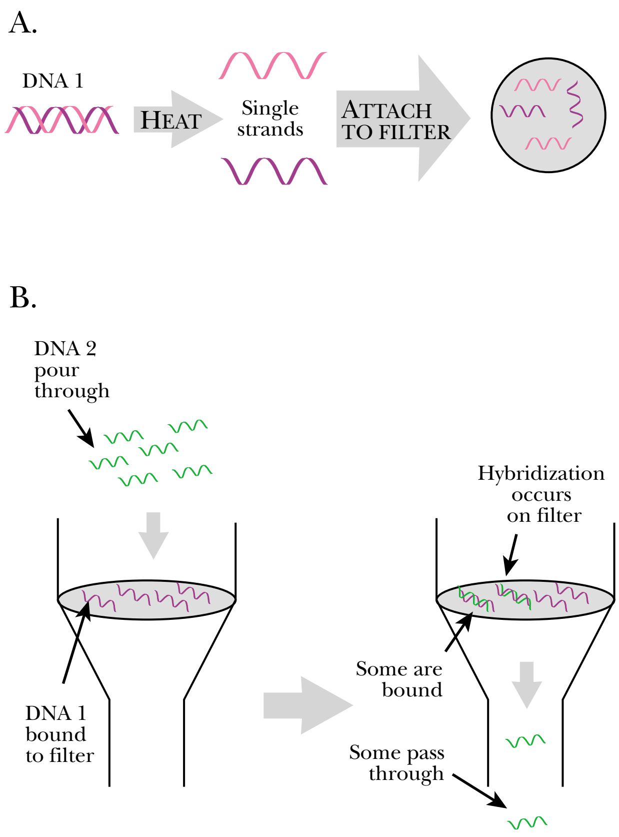


FIGURE 21.31 ***Relatedness of DNA by Filter Hybridization***

(A) DNA No. 1 is denatured and attached to a filter. (B) When DNA No. 2 is added to the filter, some of the DNA strands will hybridize, provided that the sequences are similar enough. If the sequence is identical, all of the single-stranded red DNA should hybridize to strands of green DNA. If the sequences are very different, little or none of the green DNA will hybridize with the red.

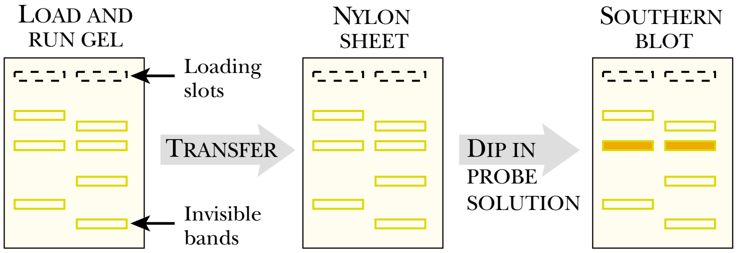


FIGURE 21.32 ***Southern Blotting: DNA:DNA Hybridization***

Southern blotting requires the target DNA to be cut into smaller fragments and run on an agarose gel. The fragments are denatured chemically to give single strands, then transferred to a nylon membrane. Notice that the DNA is invisible both in the gel and on the membrane. A radioactive probe (also single-stranded) is passed over the membrane. When the probe DNA finds a related sequence, a hybrid molecule is formed. Surplus probe that has not bound is washed away. Photographic film is placed on top of the membrane. The location of radioactive hybrid molecules is revealed by black bands on the film.

TABLE 21.01 Different Types of Blotting

Type of Blotting	Molecule on Membrane	Probe Molecule
Southern blotting	DNA	DNA
Northern blotting	RNA	DNA
Western blotting	Protein	Antibody
South-Western blotting	Protein	dsDNA

Using probes to detect DNA sequences by hybridization can be carried out on a membrane and is then referred to as "blotting".

in other organisms such as yeast or *Drosophila*. Many scientists use **Southern blotting** to identify the gene in a different organism. Southern blotting is a technique in which one DNA sample is hybridized to another DNA sample. Suppose we have a large DNA molecule, such as the yeast chromosome, and we wish to locate the particular gene whose sequence is similar to the human gene of interest. First the target or yeast DNA is cut with a restriction enzyme, and the fragments are separated by gel electrophoresis. The double-stranded fragments are melted into single-stranded fragments by soaking the gel in alkaline denaturing solution such as sodium hydroxide. Then the DNA fragments are transferred to a nylon membrane. Finally, the membrane is dipped in a solution of labeled DNA probe molecules, in this example, a radioactively labeled piece of the human gene (Fig. 21.32). The probe binds only to those fragments with sequences similar enough to base pair. When the probe hybridizes to the corresponding DNA, the filter will be "hot" or radioactive in that area, and if a piece of photographic film is placed over the filter, a black spot corresponding to the hybrid molecule will appear. Southern blotting only refers to hybridization of DNA to DNA.

Although Southern blotting was actually named after its inventor, Edward Southern, it set a geographical trend for naming other types of hybridization techniques (Table 21.01). **Northern blotting** refers to hybridization that uses RNA as the target molecule and DNA as a probe. For example, DNA probes may be used to locate messenger RNA molecules that correspond to the same gene. The mixture of RNA is run on the gel and transferred to the filter. The filter is then probed just as above.

Western blotting does not even involve nucleic acid hybridization. Proteins are separated on a gel, transferred to a membrane and detected by antibodies or other methods. Since Western blotting applies to proteins it is described more fully in Ch.

Northern blotting Hybridization technique in which a DNA probe binds to an RNA target molecule
Southern blotting A method to detect single stranded DNA that has been transferred to nylon paper by using a probe that binds DNA
Western blotting Detection technique in which a probe, usually an antibody, binds to a protein target molecule

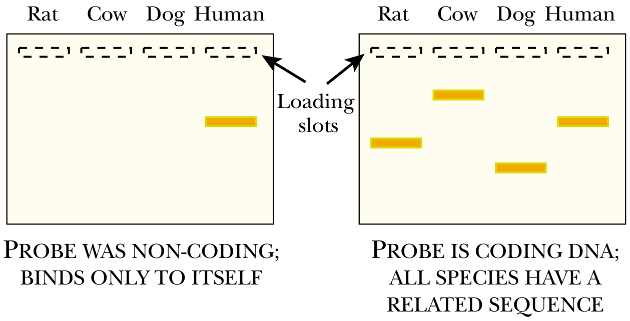

Rat
Cow
Dog
Human
Loading slots
Rat
Cow
Dog
Human
PROBE WAS NON-CODING; BINDS ONLY TO ITSELF
PROBE IS CODING DNA; ALL SPECIES HAVE A RELATED SEQUENCE

FIGURE 21.33 ***Zoo Blotting Reveals Coding DNA***

A specialized form of Southern blotting, called zoo blotting, is used to distinguish coding DNA from non-coding regions. The target DNA includes several samples of genomic DNA from different animals, hence the term "zoo". The probe is a segment of human DNA that may or may not be from a coding region. Blotting is carried out as usual. On the left, the only hybrid seen was between the probe and the human DNA. Therefore, related sequences were not found in other species and the probe is probably non-coding DNA. In the example on the right, the probe binds to related sequences in other animals; therefore, this piece of DNA is probably from a coding region.

26, Proteomics. In **South-Western blotting**, proteins suspected of binding to DNA are stuck to the membrane for testing. The probe is a DNA fragment, in this case double-stranded, since DNA binding proteins normally bind to the DNA double helix.

Zoo Blotting

Zoo blotting is not a distinct method but a neat trick using Southern blotting. One problem encountered when cloning human genes is that most DNA in higher animals is non-coding DNA, yet most scientists want the coding regions. The question is how to identify coding regions in the large amount of non-coding DNA. During evolution, the base sequence of non-coding DNA mutates and changes rapidly, whereas coding sequences change much more slowly and can still be recognized after millions of years of divergence between two species (see Ch. 20).

Therefore DNA is extracted from a series of related animals, such as a human, monkey, mouse, hamster, cow, etc. Samples in this DNA "zoo" are each cut up with a suitable restriction enzyme and the fragments are run on a gel and transferred to a nylon membrane. They are probed using DNA that is suspected of being human coding DNA. If a DNA sample really does include a coding sequence, it will probably hybridize with some fragment of DNA from most other closely related animals (Fig. 21.33). If the DNA is non-coding DNA, it will probably hybridize only to the human DNA.

Fluorescence in Situ Hybridization (FISH)

All the previous techniques require the scientist to isolate the DNA, RNA or protein from its cellular environment. In contrast, **Fluorescence in Situ Hybridization**, better known as **FISH**, is used to detect the presence of a gene, or the corresponding messenger RNA, within the actual cell (Fig. 21.34). DNA sequences from the gene of interest must first be generated for use as a probe. These may be obtained by cloning the gene or, more usually nowadays, amplified by PCR (see Ch. 23 for details). In

FISH See Fluorescence in Situ Hybridization
Fluorescence in Situ Hybridization (FISH) Using a fluorescent probe to visualize a molecule of DNA or RNA in its natural location
South-Western blotting Detection technique in which a DNA probe binds to a protein target molecule
Zoo blotting Comparative Southern blotting using DNA target molecules from several different animals to test whether the probe DNA is from a coding region

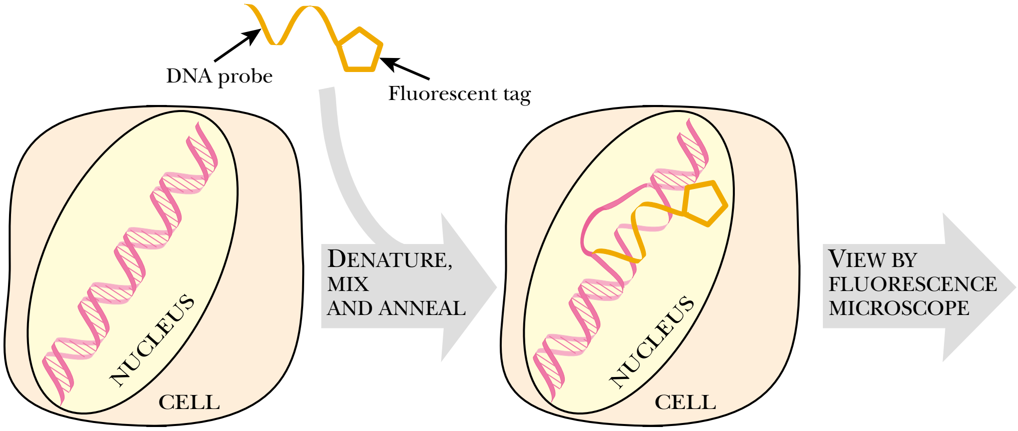


FIGURE 21.34 ***Fluorescence in Situ Hybridization—Principle***

A cell with intact DNA in its nucleus is treated to denature the DNA, forming single-stranded regions. The fluorescently labeled DNA probe is added, and the single-stranded probe can anneal with the corresponding sequence inside the nucleus. The hybrid molecule will fluoresce when the light from a fluorescence microscope excites the tag on the probe. This technique can localize the gene of interest to different areas of the nucleus or to individual chromosomes.

DNA or RNA sequences may be detected in their natural location inside the cell by hybridization to fluorescent probes.

practice, it is rarely necessary to use the whole gene sequence as a probe, unless distinguishing between closely related genes is essential. As the name indicates, the DNA probe is labeled with a fluorescent dye whose localization will eventually be observed under a fluorescence microscope. The tissue or cell must also be treated to denature the chromosomal DNA, but this is done on the actual tissue section, leaving the DNA within the nuclei.

A thin section of tissue from a particular animal, a mouse for example, may be treated with a DNA probe for a known mouse gene. In this case, the mouse probe will hybridize to the mouse DNA in the nucleus of all the cells. This tells us that the genes are in the nucleus, which we knew anyway. Some more useful applications are as follows:

1. Using a virus gene as a probe reveals which cells contain virus genes, and whether the virus genes are in the cytoplasm or have penetrated the nucleus.
2. Besides DNA within the nucleus, FISH can be used to identify where a particular gene is in the metaphase chromosome. First, a chromosome smear can be made on a microscope slide, and then probed with a fluorescently labeled gene of interest. The place where the probe binds reveals which chromosome carries the gene corresponding to the probe. With sufficiently sophisticated equipment, the gene may be localized to a specific region on the chromosome (Fig. 21.35).
3. A DNA probe can be used to detect mRNA within the target tissue since one of the two strands of the denatured DNA will bind to the RNA. Since mRNA is already single-stranded, the cells do not have to be treated with high heat or chemical denaturants. Cells actively transcribing the gene of interest will have high levels of the corresponding mRNA, which will bind the probe and light up (Fig. 21.36). The greater the gene expression, the brighter the cell will fluoresce. Identifying the location of a particular mRNA can be real helpful when comparing tissues. For example, comparing the amount of mRNA for a particular gene in liver cells versus heart cells can help determine the function of the gene of interest. Levels of many mRNA molecules are low and hence would give only a weak signal by FISH. In practice, such mRNA is often amplified by RT-PCR before detection (see Ch. 23). Alternatively, more modern and more sensitive techniques, such as microarrays are used for mRNA detection (see Ch. 25).

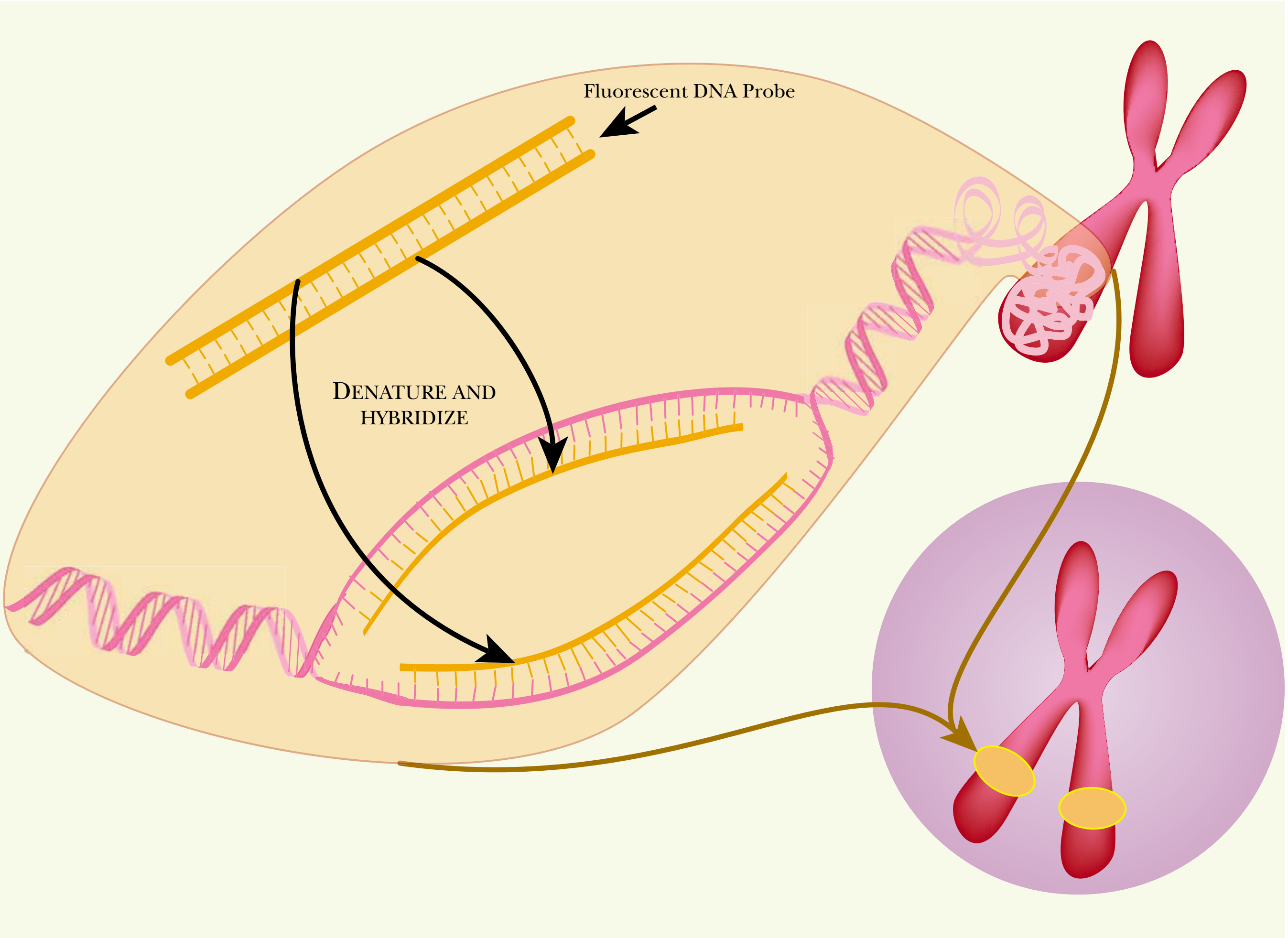


FIGURE 21.35 ***FISH To Localize Genes on Chromosomes***

FISH can localize a gene to a specific place on a chromosome. First, metaphase chromosomes are isolated and attached to a microscope slide. The chromosomal DNA is denatured into single-stranded pieces that remain attached to the slide. The fluorescent probe hybridizes to the corresponding gene. When the slide is illuminated, the hybrid molecules fluoresce and reveal the location of the gene of interest.

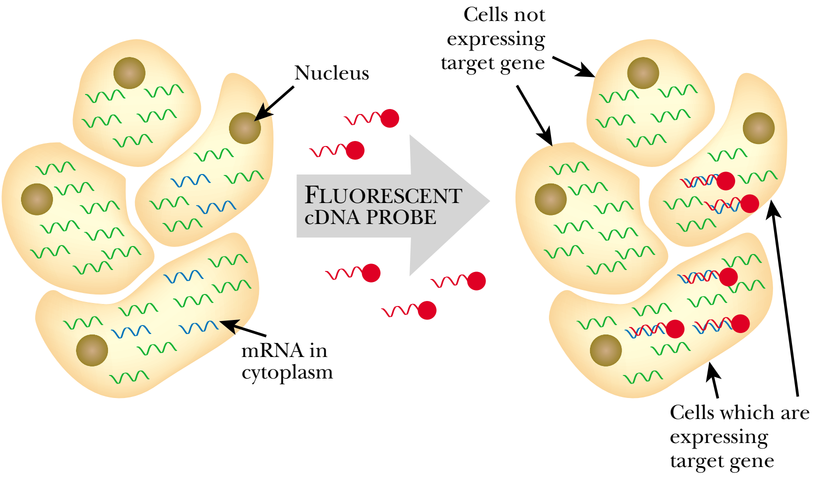

Nucleus
mRNA in cytoplasm
FLUORESCENT cDNA PROBE
Cells not expressing target gene
Cells which are expressing target gene

FIGURE 21.36 ***FISH To Detect and Measure mRNA Levels***

A variety of different mRNA species are expressed at any one time within a tissue. Not all cells express the same genes or express them at the same level. If the target cells are probed with fluorescently labeled DNA (red), this will bind to the corresponding mRNA (blue). Notice that the probe DNA does not bind to the nuclear DNA because in this procedure the cells were not treated to denature the chromosomal DNA. The target gene in this example was only expressed in two of the cells.

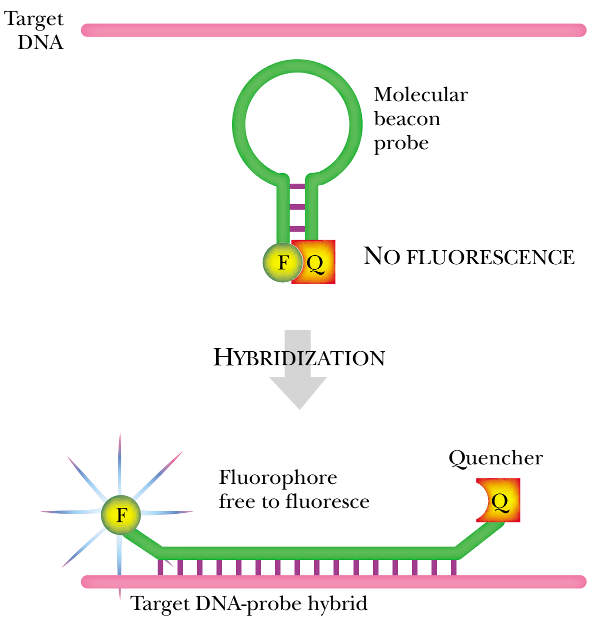


FIGURE 21.37 ***Molecular Beacon***

A molecular beacon is a probe that has two engineered regions at the ends of the probe sequence. On the 5′ side, a fluorescent tag is added (F), and on the 3′ side, a quenching group is added (Q). Just inside the two tags are six base pairs that can form a stem-loop structure. In this state the probe cannot fluoresce. When the probe binds to the target sequence, the stem-loop structure is lost. Since the quenching group is no longer next to the fluorescent tag, the probe can now fluoresce.

Molecular Beacons

Fluorescent probes may be designed that only fluoresce after binding to the target DNA sequence.

A **molecular beacon** is a fluorescent probe molecule that is designed to fluoresce only when it binds to a specific DNA target sequence. The probe contains both a fluorescent group, or **fluorophore**, and a quenching group at opposite ends of a DNA sequence of 20 to 30 bases. The central region of the probe is complementary to the target sequence. The terminal half dozen bases at each end of the probe are complementary and form a short double-stranded region as shown in Fig. 21.37. In the stem and loop conformation the quenching group is next to the fluorophore and so prevents fluorescence. When the molecular beacon binds to the target sequence it is linearized. This separates the quenching group from the fluorophore, which is now free to fluoresce. Care is needed to avoid disrupting the short stem structure. For example, high temperatures will cause unpairing and give a false positive response.

fluorophore A fluorescent chemical group
molecular beacon A fluorescent probe molecule that contains both a fluorophore and a quenching group and that fluoresces only when it binds to a specific DNA target sequence

CHAPTER TWENTY-TWO

Recombinant DNA Technology

Introduction

Molecular cloning involves cutting out the gene of interest from its original location and placing it on a vector for subsequent manipulation.

This chapter discusses the basic techniques involved in molecular cloning. This involves two general stages. First DNA from some particular source is cut to liberate a gene or other fragment of interest. This fragment is then "cloned" by inserting it into another DNA macromolecule, known as a vector. After cloning, the chimeric DNA is normally inserted into an appropriate host cell. A **chimera** is any hybrid molecule of DNA, such as a vector plus a cloned gene, which has been engineered from two different sources of DNA. We will first consider the enzymes used to cut and join fragments of DNA. Then we will discuss the vectors used in cloning and how DNA fragments are inserted into them. Ultimately, cloned genes may be used in the manufacture of high levels of recombinant protein or may be applied in gene therapy to cure inherited defects.

Nucleases Cut Nucleic Acids

Assorted nucleases are known that cut DNA or RNA in the middle or remove single nucleotides from the ends.

Nucleases are enzymes that degrade nucleic acids. **Ribonucleases** (RNases) attack RNA and **deoxyribonucleases** (DNases) attack DNA. Most nucleases are specific, though the degree of specificity varies greatly. Some nucleases will only attack single-stranded nucleic acids, others will only attack double-stranded nucleic acids and a few will attack either kind. **Exonucleases** attack at the end of nucleic acid molecules and usually remove just a single nucleotide, or sometimes a short oligonucleotide. Any particular exonuclease attacks either the 3′-end or the 5′-end but not both. **Endonucleases** cleave the nucleic acid chain in the middle. Some endonucleases are non-specific, others, in particular the restriction enzymes, are extremely specific and will only cut DNA after binding to specific recognition sequences. All these enzymes have proved extremely useful both in genetic analysis and genetic engineering.

Restriction and Modification of DNA

Restriction enzymes cut DNA at specific sequences.

Methylation of DNA by bacteria is used to distinguish the cell's own DNA from DNA of foreign origin. In nature, foreign DNA entering a bacterial cell would most likely be due to virus infection. When viruses attack bacteria, the virus coat is left outside and only the virus DNA enters the target cell (see Ch. 17). The virus DNA will take over the victim's cellular machinery and use it to manufacture more virus particles unless the bacterial cell fights back. The key is to degrade only foreign DNA without endangering the bacterial cell's own DNA. Bacteria make restriction and modification enzymes that respectively cut and methylate DNA, ensuring than the foreign DNA is recognized and destroyed. Whenever a bacterial cell makes a restriction enzyme, it also makes the corresponding modification enzyme that modifies and protects its own DNA. The result is that the DNA in a bacterial cell is immune to that cell's own restriction enzymes. In contrast, incoming, unmodified DNA will be degraded by the restriction enzyme.

Modification enzymes protect DNA from the corresponding restriction enzymes by adding methyl groups at the recognition site.

Restriction enzymes are endonucleases that cut double-stranded DNA. They bind to DNA at a specific sequence of bases, called the recognition site and then proceed to cut the DNA. **Modification enzymes** bind to the DNA at the same recognition site as the corresponding restriction enzymes and methylate the DNA. Modification enzymes usually add the methyl group to adenine or cytosine within the recognition

chimera Hybrid molecule that includes DNA from more than one source
deoxyribonuclease (DNase) Enzyme that cuts or degrades DNA
endonuclease Enzyme that cleaves nucleic acid molecule in the middle
exonuclease Enzyme that cleaves nucleic acid molecule at the end and usually removes just a single nucleotide
modification enzyme Enzyme that binds to the DNA at the same recognition site as the corresponding restriction enzyme but methylates the DNA
nuclease Enzyme that cuts or degrades nucleic acids
restriction enzyme Type of endonuclease that cuts double stranded DNA at a specific sequence of bases, the recognition site
ribonuclease (RNase) Enzyme that cuts or degrades RNA

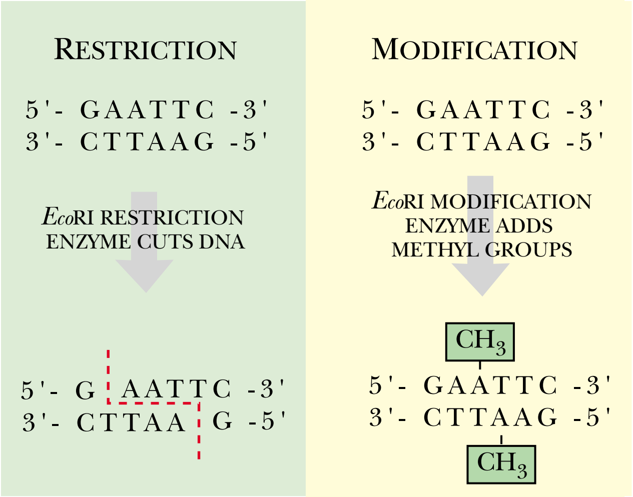


FIGURE 22.01 ***Restriction and Modification Systems***

Restriction enzymes recognize non-methylated double-stranded DNA and cut it at specific recognition sites. For example, *Eco*RI recognizes the sequence, 5′-GAATTC-3′, and cuts after the G. Since this sequence is an inverted repeat, the enzyme also cuts the other strand after the corresponding G, giving a zig-zag cut. Modification enzymes are paired with restriction enzymes and recognize the same sequence. Modification enzymes methylate the recognition sequence, which prevents the restriction enzyme from cutting it.

site (Fig. 22.01). Once methylated, DNA is protected from the restriction endonuclease. Only non-methylated DNA will be cut and destroyed by the restriction enzyme. All the DNA in a bacterial cell, including the chromosome and any plasmids, is normally protected by modification.

Recognition of DNA by Restriction Endonucleases

Most recognition sites for restriction enzymes are inverted repeats of 4, 6 or 8 bases.

Due to their ability to recognize specific sites in DNA, restriction endonucleases have become one of the most widely used tools in genetic engineering. Restriction enzyme recognition sites are usually four, six or eight bases long and the sequence forms an inverted repeat. Thus the sequence on the top strand of the DNA is the same as the sequence of the bottom strand read in the reverse direction, as shown in above.

Several hundred different restriction enzymes are now known and each has its own specific recognition site. Some recognition sites require a specific base at each position. Others are less specific and may require only a purine or a pyrimidine at a particular position. Some examples are shown in Table 22.01.

Since any random series of four bases will be found quite frequently, four base-recognizing enzymes cut DNA into many short fragments. Conversely, since any particular eight-base sequence is less likely, the eight-base-recognizing enzymes cut DNA only at longer intervals and generate fewer larger pieces. The six-base enzymes are the most convenient in practice, as they give an intermediate result.

Naming of Restriction Enzymes

Restriction enzymes have names derived from the initials of the bacteria they come from. The first letter of the genus name is capitalized and followed by the first two letters of the species name (consequently, these three letters are in italics). The strain is sometimes represented e.g. the R in *Eco*RI refers to *Escherichia coli* strain RY13. The roman letter indicates the number of restriction enzymes found in the same species. For example, *Moraxella bovis* has two different restriction enzymes called *Mbo*I and *Mbo*II. Some examples are shown in Table 22.01.

TABLE 22.01 Examples of Restriction Endonucleases

Enzyme	Source Organism	Recognition Sequence
*Hpa*II	*Haemophilus parainfluenzae*	C/CGG GGC/C
*Mbo*I	*Moraxella bovis*	/GATC CTAG/
*Nde*II	*Neisseria denitrificans*	/GATC CTAG/
*Eco*RI	*Escherichia coli* RY13	G/AATTC CTTAA/G
*Eco*RII	*Escherichia coli* RY13	/CC(A or T)GG GG(T or A)CC/
*Eco*RV	*Escherichia coli* J62/pGL74	GAT/ATC CTA/TAG
*Bam*HI	*Bacillus amyloliquefaciens*	G/GATCC CCTAG/G
*Sau*I	*Staphylococcus aureus*	CC/TNAGG GGANT/CC
*Bgl*I	*Bacillus globigii*	GCCNNNN/NGGC CGGN/NNNNCCG
*Not*I	*Nocardia otitidis-caviarum*	GC/GGCCGC CGCCGG/CG
*Dra*II	*Deinococcus radiophilus*	RG/GNCCY YCCNG/GR

/ = position where enzyme cuts
N = any base, R = any purine, Y = any pyrimidine

Different restriction enzymes may share the same recognition sequence although they do not necessarily cut at precisely the same place.

If two restriction enzymes from different species share the same recognition sequence they are known as **isoschizomers**. Note that isoschizomers may not always cut in the same place even though they bind the same base sequence. For example, the sequence GGCGCC is recognized by four enzymes, each of which cuts in different places: *Nar*I (GG/CGCC), *Bbe*I (GGCGC/C), *Ehe*I (GGC/GCC), and *Kas*I (G/GCGCC).

Cutting of DNA by Restriction Enzymes

It might seem logical for the DNA to be cut at the recognition site where the restriction enzyme binds. This is often true, but not always. There are two major classes of restriction enzyme that differ in where they cut the DNA, relative to the recognition site.

Type I restriction enzymes cut the DNA a long way from the recognition sequence.

Type I restriction enzymes cut the DNA a thousand or more base pairs away from the recognition site. This is done by looping the DNA around so that the enzyme binds both at the recognition site and the cutting site (Fig. 22.02). Since the exact length of the loop is not constant, and since the base sequence at the cut site is not fixed, these enzymes are of little practical use to molecular biologists. Even more bizarre is that type I restriction enzymes are suicidal. Most enzymes carry out the same reaction over and over again on a continual stream of target molecules. Each molecule of a type I restriction enzyme can cut DNA only a single time and then it is inactivated!

Type I restriction systems consist of a single protein with three different subunits. One subunit recognizes the DNA, another methylates the recognition sequence and the third cuts the DNA at a distance from the recognition sequence. In type II restric-

isoschizomers Restriction enzymes from different species that share the same recognition sequence
type I restriction enzyme Type of restriction enzyme that cuts the DNA a thousand or more base pairs away from the recognition site

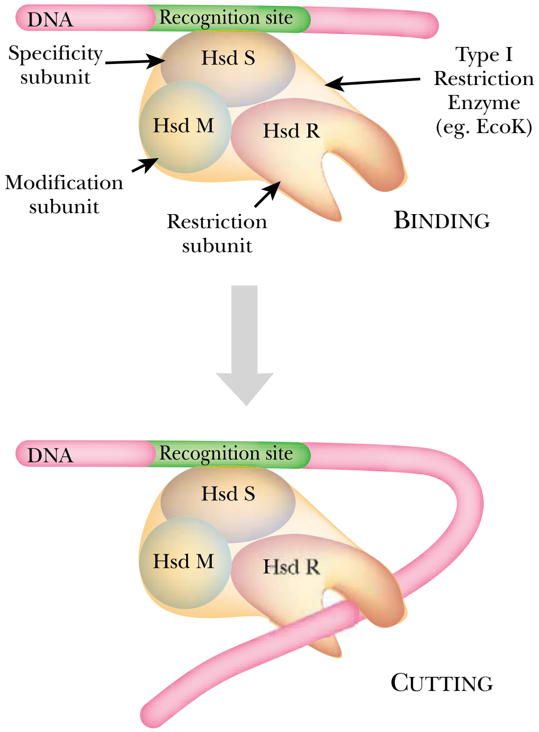


FIGURE 22.02 ***Type I Restriction Enzyme***

Type I restriction enzymes have three different subunits. The specificity subunit recognizes a specific sequence in the DNA molecule. The modification subunit adds a methyl group to the recognition site. If the DNA is non-methylated, the restriction subunit cuts the DNA, but at a different site, usually over 1000 base pairs away. In the *Eco*K restriction enzyme, the subunits are HsdS, HsdM, and HsdR.

tion systems the restriction endonuclease and the methylase are two separate proteins that operate independently but recognize the same DNA sequence.

Type II restriction enzymes cut the DNA within the recognition sequence. Some generate blunt ends, others give sticky ends.

Type II restriction enzymes cut the DNA in the middle of the recognition site. Since the exact position of the cut is known, these are the restriction enzymes that are normally used in genetic engineering. There are two different ways of cutting the recognition site in half. One way is to cut both strands of the double stranded DNA at the same point. This leaves **blunt ends** as shown in Figure 22.03. The alternative is to cut the two strands in different places, which generates overhanging ends. The ends made by such a staggered cut will base pair with each other and consequently they are known as **sticky ends**.

Enzymes that generate sticky ends are the most useful. If two different pieces of DNA are cut with the same restriction enzyme or enzymes that generate the same overhang, the same sticky ends are generated. This allows fragments of DNA from two different original DNA molecules to be bound together by matching the sticky ends (Fig. 22.04). Such pairing is temporary since the pieces of DNA are only held together by hydrogen bonding between the base pairs, not by permanent covalent bonds. Nonetheless, this assists the permanent bonding of the sugar-phosphate backbone by DNA ligase. When two sticky ends made by the same enzyme are ligated, the junction may be cut apart later by using the same enzyme again. However, if two sticky ends made by two different enzymes are ligated together, a hybrid site is formed that cannot be cut by either enzyme (as would happen with *Bam*HI and *Bgl*II in Fig. 22.04).

Sticky ends are more convenient than blunt ends when joining together fragments of DNA using DNA ligase.

DNA Fragments are Joined by DNA Ligase

The enzyme **DNA ligase** is used to join DNA fragments covalently. DNA ligase operates during DNA replication where it joins up the fragments of the lagging strand (Ch.

blunt ends Ends of a double-stranded DNA molecule that are fully base paired and have no unpaired single-stranded overhang
DNA ligase Enzyme that joins DNA fragments covalently, end to end
sticky ends Ends of a double-stranded DNA molecule that have unpaired single-stranded overhangs, generated by a staggered cut
type II restriction enzyme Type of restriction enzyme that cuts the DNA in the middle of the recognition site

FIGURE 22.03 ***Type II Restriction Enzymes—Blunt Versus Sticky Ends***

*Hpa*I is an blunt end restriction enzyme, that is, it cuts both strands of DNA in exactly the same position. *Eco*RI is a sticky end restriction enzyme. The enzyme cuts between the G and A on both strands, which generates an a four base pair overhang on the ends of the DNA. Since these bases are free to base pair with any complementary sequence, they are considered "sticky".

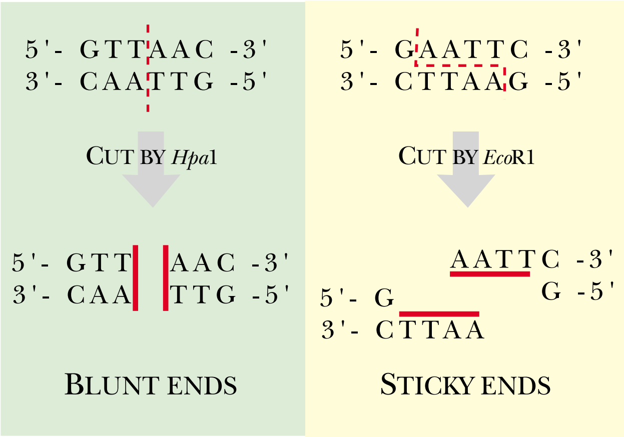
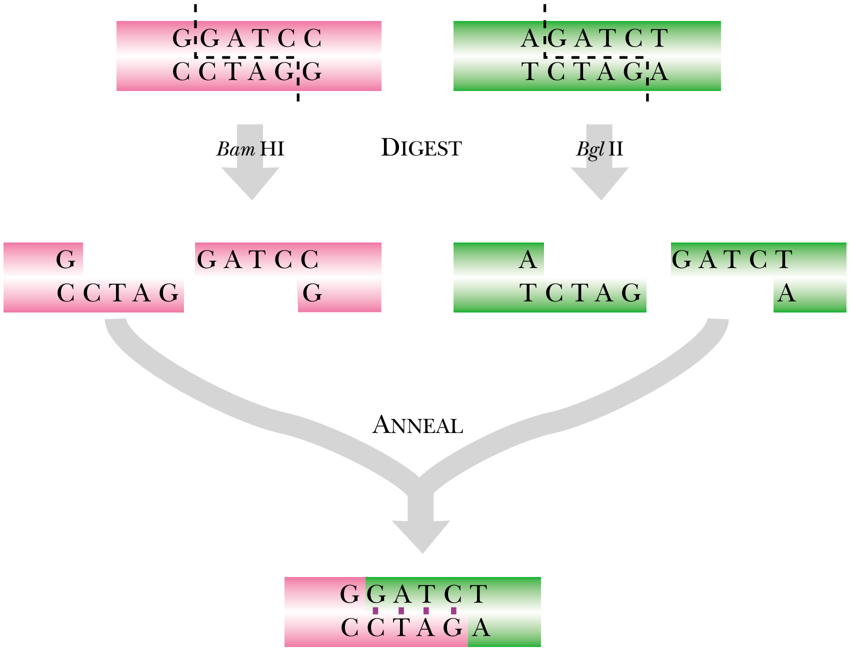


FIGURE 22.04 ***Matching of Compatible Sticky Ends***

*Bam*HI and *Bgl*II generate the same overhanging or sticky ends. *Bam*HI recognizes the sequence 5′-GGATCC-3′ and cuts after the first 5′ G, which generates the 3′-CTAG-5′ overhang on the bottom strand. *Bgl*II recognizes the sequence 5′-AGATCT-3′ and cuts after the first 5′ A, which generates a 5′-GATC-3′ overhang on the top strand. If these two pieces are allowed to anneal, the complementary sequences will hydrogen bond together, allowing the nicks to be sealed more easily by DNA ligase.



5). If DNA ligase finds two DNA fragments touching each other end to end, it will ligate them together (Fig. 22.05). In practice, segments of DNA with matching sticky ends will tend to stay attached much of the time and consequently DNA ligase will join them efficiently. Since DNA fragments with blunt ends have no way to bind each other, they drift apart most of the time. Ligating blunt ends is very slow and requires a high concentration of DNA ligase, as well as, a high concentration of DNA. In fact, bacterial ligase cannot join blunt ends at all. In practice, **T4 ligase** is normally used in genetic engineering as it can join blunt ends if need be. T4 ligase originally came from bacteriophage T4, although nowadays it is manufactured by expressing the gene that encodes it in *E. coli*.

Making a Restriction Map

A diagram that shows the location of restriction enzyme cut sites on a segment of DNA is known as a **restriction map**. The first step in generating such a map is to

restriction map A diagram showing the location of restriction enzyme cut sites on a segment of DNA
T4 ligase Type of DNA ligase from bacteriophage T4 and which is capable of ligating blunt ends

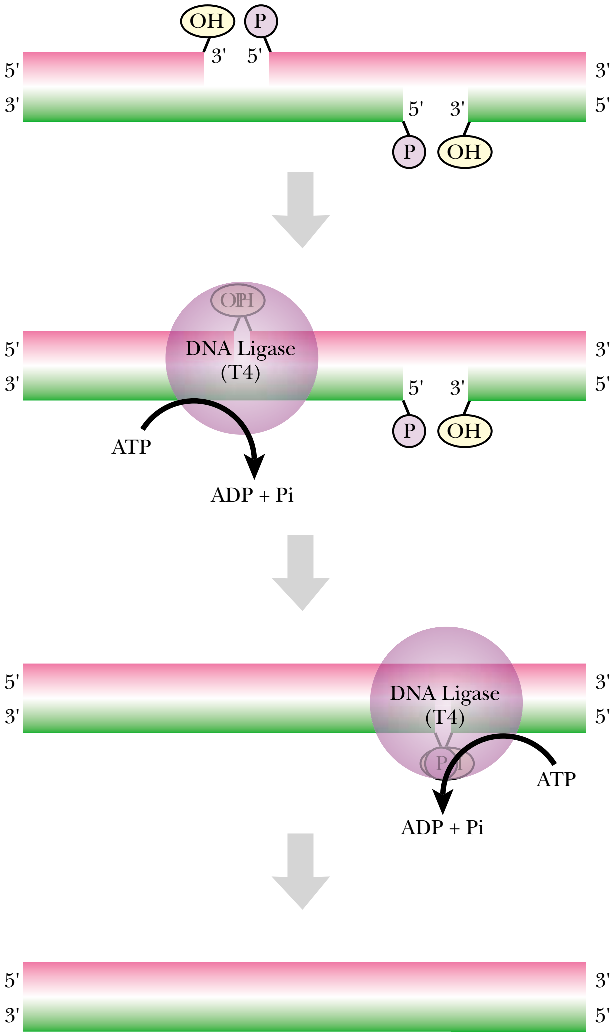


FIGURE 22.05 ***DNA Ligase Joins Fragments of DNA***

T4 DNA ligase connects the sugar-phosphate backbone of two pieces of DNA. In the example, overlapping sticky ends connect a double stranded piece of DNA, but the backbone of each strand has not been connected. T4 DNA ligase recognizes these nicks or breaks in the backbone and uses energy from the hydrolysis of ATP to drive the ligation reaction.

A restriction map is a diagram showing the location of cut sites on DNA for a variety of restriction enzymes.

Restriction maps are deduced by cutting the target DNA with a selection of restriction enzymes, both alone and in pairs.

digest the DNA with a series of restriction enzymes, one at a time. The fragments of digested DNA are separated by agarose gel electrophoresis (as described in Ch. 21). Comparison with appropriate standards allows the sizes of the fragments to be estimated. This reveals how many recognition sites each enzyme has in the DNA and their distances apart. What remains unknown is the relative order of the fragments.

For example, suppose we start with a 5,000 base pair (bp) piece of DNA that is cut twice by the restriction enzyme *Bam*HI giving three fragments of 3,000 bp, 1,500 bp, and 500 bp. There are three alternative arrangements for three fragments (Fig. 22.07A). You might think there should be six possible arrangements, but the other three theoretical arrangements are merely the first three, turned back to front, they are not genuinely different physical molecules. Fig. 22.07A shows the backward arrangement just for fragment number III.

To decide which of the three possible arrangements are correct, double digests using two restriction enzymes must be performed. The DNA is cut with each enzyme alone and with both simultaneously (Fig. 22.07B). The results of gel electrophoresis for the two single digests and the double digest are shown. Suppose that the second restriction enzyme is *Eco*RI and that alone it cuts just once to give two fragments of 4,000 bp and 1,000 bp. Thus, for *Eco*RI alone there is only one possible arrangement. In the double digest, the largest fragment seen in the *Bam*HI lane has disappeared. This means that there is an *Eco*RI cut site within this 3,000 bp *Bam*HI fragment. Since, in this example, there is only one *Eco*RI cut site, only one of the *Bam*HI fragments disappears in the double digest. This allows us to reduce the possibilities to the restriction maps shown in Figure 22.07B.

Gene Disruption by Engineered Insertional Mutagenesis

A variety of techniques have been used to construct mutations utilizing genetic engineering technology. These techniques are usually known as site-directed mutagenesis (see Ch. 13). In particular, mutations that serve to completely inactivate a gene are useful in genetic analysis. So, genes may be deliberately disrupted by the insertion of foreign DNA. To do this, it is first necessary to clone the gene onto some convenient vector such as a bacterial plasmid (see below). To disrupt the gene, a deliberately designed segment of DNA is used. Known as a **gene cassette** (Fig. 22.06), it usually carries with it a gene for resistance to some antibiotic such as chloramphenicol or kanamycin. This way, the inserted DNA cassette can easily be detected, because cells carrying it will become resistant to the antibiotic. At each end, the cassette has several convenient restriction enzyme sites. The target gene is cut open with one of these restriction enzymes and the cassette is cut from its original location with the same enzyme. The cassette is then ligated into the middle of the target gene (Fig. 22.06). The disrupted gene is then put back into the organism from which it came.

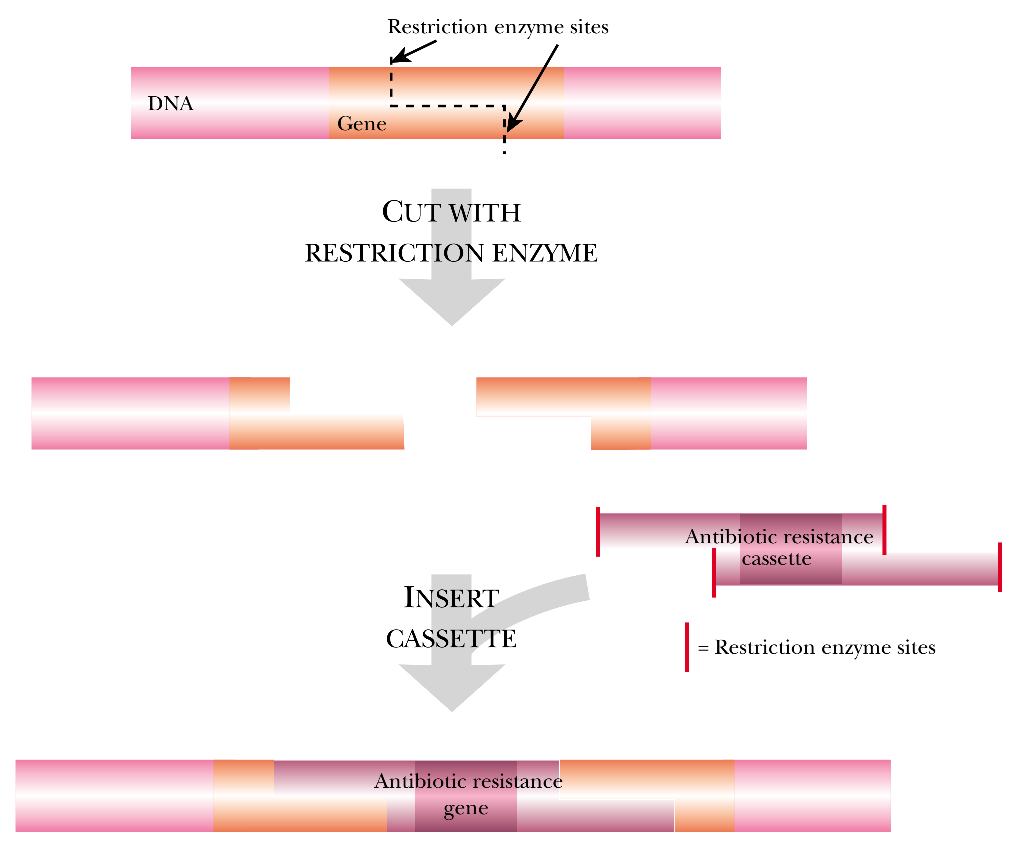


FIGURE 22.06 ***Gene Disruption Using a Cassette***

A gene to be disrupted is cut with a restriction enzyme. An artificially constructed cassette that confers antibiotic resistance is inserted into the cut site and ligated into the gene. The new DNA construct formed can be detected easily since it provides resistance to antibiotics.

Note that two alternatives remain. To decide between these needs a third enzyme. Double digests with BamHI plus enzyme III and of EcoRI plus enzyme III would be analyzed, as above. Eventually, this approach allows the construction of a complete restriction map of any segment of DNA. This may then be used as a guide to further manipulations.

gene cassette Deliberately designed segment of DNA that is flanked by convenient restriction sites and usually carries a gene for resistance to an antibiotic or some other easily observed character

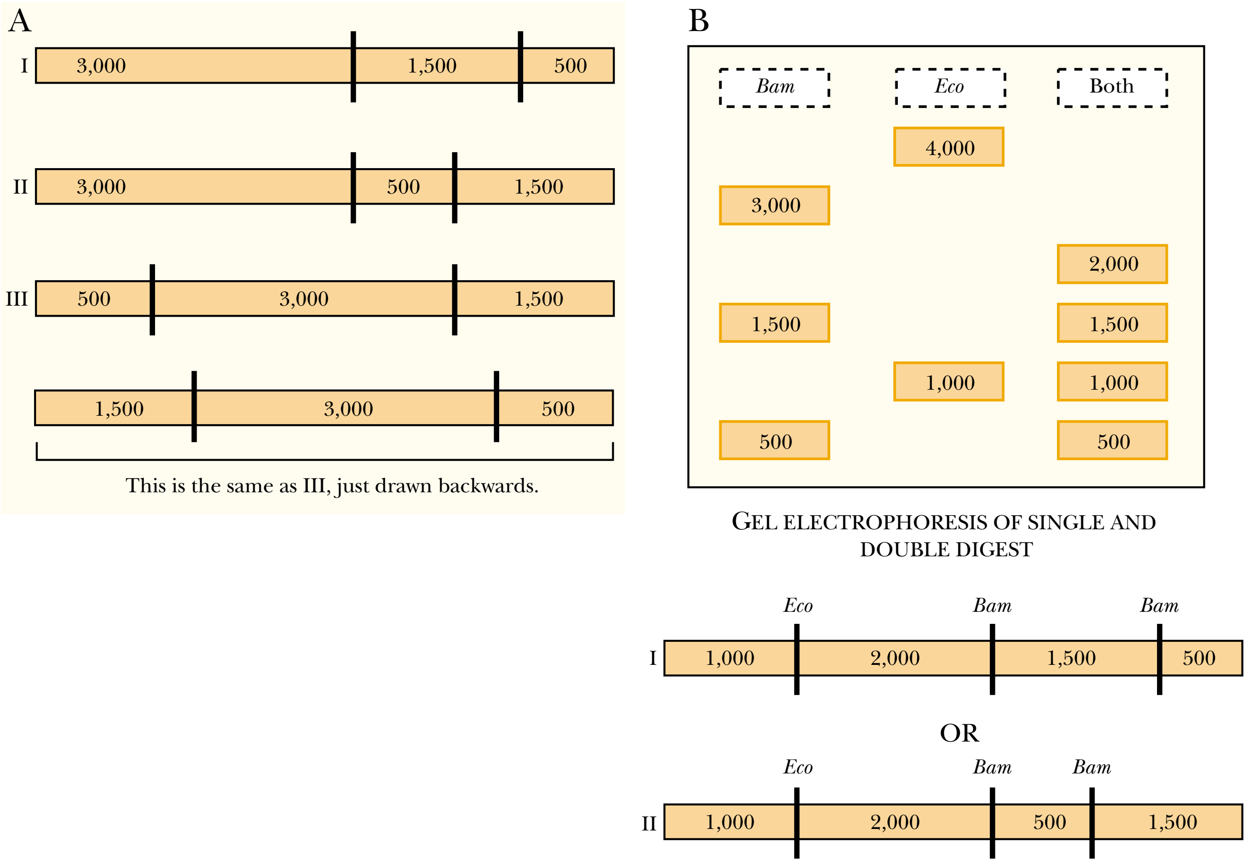


FIGURE 22.07 ***Restriction Mapping***

(A) To determine the location and number of restriction enzyme sites, a segment of DNA is digested with a restriction enzyme. In this example, the piece of DNA is 5,000 base pairs in length. Cutting with *Bam*HI gave three fragments: of 3,000 bp, 1500 bp, and 500 bp. The figure shows the three possible arrangements of these three fragments. The fourth arrangement shown is not really different but is merely the third possible arrangement drawn in the opposite orientation.
(B) Double digestion is the next step in compiling a restriction map. Cutting the DNA with *Eco*RI alone would give two fragments: of 4000 bp and 1000 bp. When the DNA is cut with both *Eco*RI and *Bam*HI simultaneously, four fragments are resolved by gel electrophoresis. Two of these are identical to the 1500 bp and 500 bp fragments from the *Bam*HI single digest; therefore, no *Eco*RI sites are present within these fragments. The remaining two fragments, 2000 bp and 1000 bp, add up to give the 3000 bp *Bam*HI fragment. Therefore, the single *Eco*RI site must be within the 3000 bp *Bam*HI fragment. Of the three possible arrangements shown in part (A), the third arrangement is ruled out (if it was cut with *Eco*RI alone, it could not give two fragments of 4000 bp and 1000 bp).

Restriction Fragment Length Polymorphisms (RFLPs)

Related molecules of DNA, such as different versions of the same gene from two related organisms, normally have very similar sequences. Consequently they will have similar restriction maps. However, occasional differences in base sequence will result in corresponding differences in restriction sites. Each restriction enzyme recognizes a specific sequence (usually of four, six or eight bases). If even a single base within this recognition sequence is altered, the enzyme will no longer cut the DNA (Fig. 22.08). Consequently, restriction sites that are present in one version of a sequence may be missing in its close relatives.

If two such related but different DNA molecules are cut with the same restriction enzyme, segments of different lengths may be produced. Consequently, a difference

If two related DNA molecules differ in sequence *at* a cut site, fragments of different sizes will result if the molecules are digested.

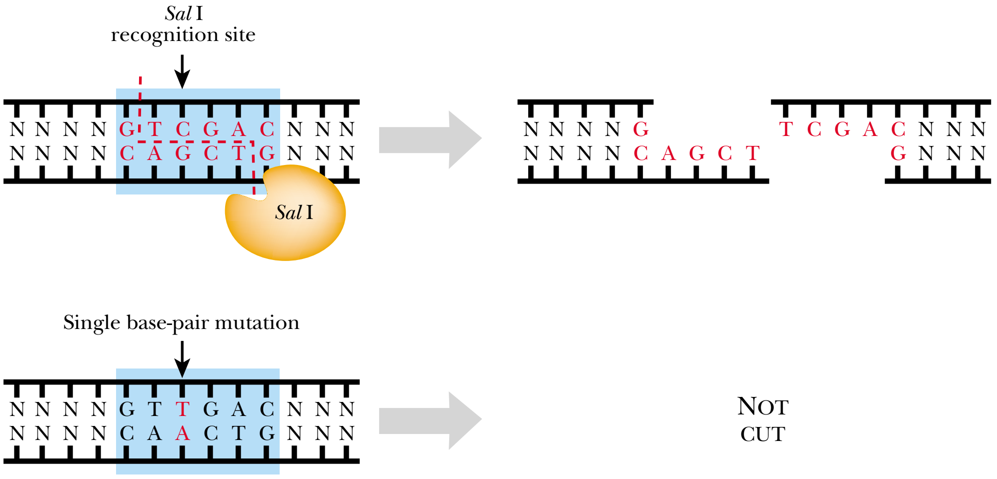

Sal I
recognition site
N N N N G T C G A C N N N
N N N N C A G C T G N N N
Sal I
N N N N G
N N N N C A G C T
T C G A C N N N
G N N N
Single base-pair mutation
N N N N G T T G A C N N N
N N N N C A A C T G N N N
NOT CUT

FIGURE 22.08 ***Single Base Changes Prevent Cutting by Restriction Enzymes***

The recognition sequence for a particular restriction enzyme is extremely specific. Changing a single base will prevent recognition and cutting. The example shown is for *Sal*I, whose recognition sequence is GTCGAC.

between two DNA sequences that affects a restriction site is known as a **restriction fragment length polymorphism (RFLP)**. When these are separated on a gel, bands of different sizes will appear (Fig. 22.09). RFLPs may be used to identify organisms or analyze relationships even though we do not know the function of the altered gene. In fact, since we are examining the DNA directly, the alteration may be in non-coding DNA or an intervening sequence. It does not need to be in the coding region of a gene. RFLPs are widely used in forensic analysis.

Properties of Cloning Vectors

Once the desired fragments of DNA have been isolated, further manipulations require them to be inserted into a **cloning vector**. In principle, any molecule of DNA, which can replicate itself inside a cell, could work as a vector. For convenience in manipulation, the following factors must be considered:

The term "vector" refers to self-replicating DNA molecules used to carry cloned genes (or any other pieces of cloned DNA).

1. The vector should be a reasonably small and manageable DNA molecule.
2. Moving the vector from cell to cell should be relatively easy.
3. Generating and purifying large amounts of vector DNA should be straightforward.

In addition to these basic requirements, most vectors have been designed to provide some convenient means to perform the following:

1. Detect the presence of the vector
2. Directly select for cells that contain the vector
3. Insert genes into the vector
4. Detect the presence of an inserted gene on the vector.

In practice, bacterial plasmids come closest to these requirements and are the most widely used vectors. Many viruses are also used as vectors, especially when attempting the engineering of higher organisms. For some special purposes, where very large fragments of DNA are to be cloned, whole chromosomes are sometimes used as vectors. We will first discuss the requirements for vectors using plasmids as examples.

cloning vector Any molecule of DNA that can replicate itself inside a cell and is used for carrying cloned genes or segments of DNA. Usually a small multicopy plasmid or a modified virus

restriction fragment length polymorphism (RFLP) A difference in restriction sites between two related DNA molecules that results in production of restriction fragments of different lengths

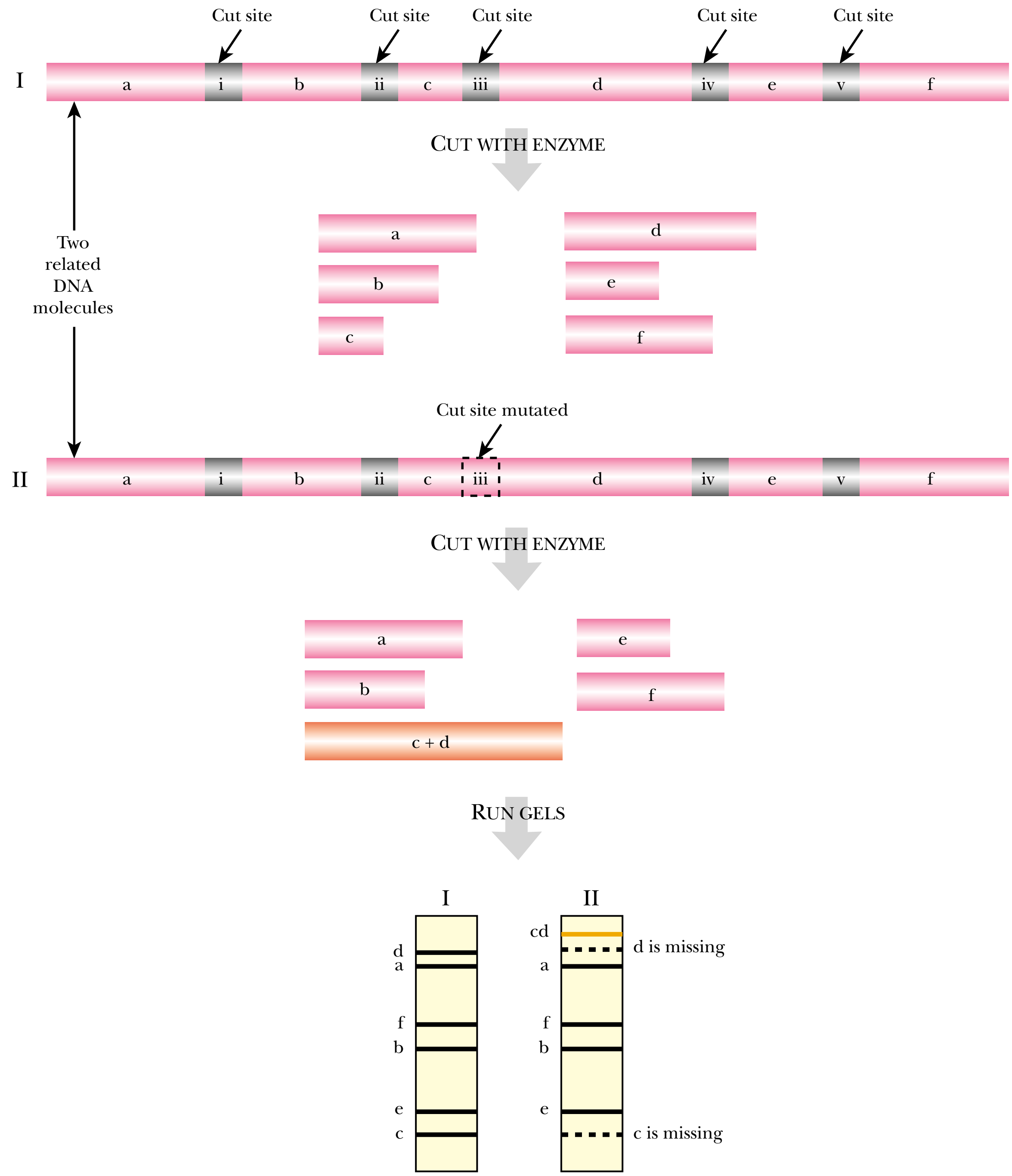


FIGURE 22.09 ***Restriction Fragment Length Polymorphism (RFLP)***

DNA from related organisms shows small differences in sequence that result in changes in restriction map patterns. In the example shown, cutting a segment of DNA from the first organism yields six fragments of different sizes (labeled a–f on the gel). If the equivalent region of DNA from a related organism is digested with the same enzyme we would expect a similar pattern. Here, a single nucleotide difference is present, which eliminates one of the restriction sites. Consequently, digesting this DNA only produces five fragments, since site iii has been mutated and the original fragments c and d are no longer separated. Instead a new fragment, the size of c plus d is seen.

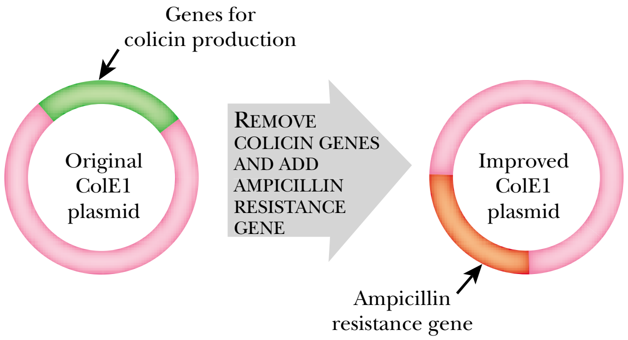

Genes for colicin production
Original ColE1 plasmid
REMOVE COLICIN GENES AND ADD AMPICILLIN RESISTANCE GENE
Improved ColE1 plasmid
Ampicillin resistance gene

FIGURE 22.10 ***Antibiotic Resistant Cloning Plasmid***

The ColE1 plasmids of *E. coli* have been modified for use as cloning vectors. The original colicin genes have been deleted so that the bacteria carrying these plasmids no longer produce these antibacterial toxins. In addition, a gene for antibiotic resistance has been added. This provides an easily identifiable phenotype to the bacteria that carry the altered plasmid.

Multicopy Plasmid Vectors

Multicopy plasmids are convenient as vectors since they make plenty of plasmid DNA and also express cloned genes at high levels.

Vectors derived from the small multicopy plasmids of bacteria (see Ch. 16) were the first to be used and are still the most widespread. The **ColEI plasmid** of *Escherichia coli* is a small circular DNA molecule that forms the basis of many vectors widely used in molecular biology. It exists in up to 40 copies per cell so obtaining plenty of plasmid DNA is relatively easy and it can be moved from cell to cell by transformation as described in Chapter 18.

Although the original ColE1 plasmid was once used directly as a vector, most modern ColE1-based vectors contain a range of artificial improvements. First, the genes for colicin E1, a toxic protein for killing bacteria (see Ch. 16), are removed, since these are obviously not needed. Next, a gene for resistance to an antibiotic was added. The most popular antibiotic for this purpose is **ampicillin**, a widely used penicillin derivative (Fig. 22.10). The ampicillin resistance gene is known as ***amp*** or ***bla***, which refers to **beta-lactamase**. This enzyme degrades penicillins and related antibiotics.

Vectors are often designed to be selected by antibiotic resistance.

When this vector is transformed into bacteria we can directly select those cells that get the plasmid by incubating them in a growth medium containing ampicillin. Those cells containing a plasmid survive, while those that did not get a plasmid are killed. Any time we wish to check that the vector is still present, we test the cells for ampicillin resistance.

Inserting Genes into Vectors

The simplest way to insert a segment of DNA into a vector is to cut both the target DNA and the vector with the same restriction enzyme. If a restriction enzyme that generates sticky ends is used, the vector and the insert will have matching overhangs. A mixture of the two is treated with DNA ligase, which links together DNA strands. The result is the ligation of the target DNA fragment into the vector as shown in Figure 22.11. If a restriction enzyme that generates blunt ends is used, ligation is more difficult and T4 ligase must be used as discussed above.

Vectors are often engineered to contain large numbers of convenient restriction cut sites.

This procedure relies on the vector possessing only one site for the chosen restriction enzyme. If there were more than one cut site in the vector the restriction enzyme

***amp* gene** Gene conveying resistance to ampicillin and related antibiotics and encoding beta-lactamase. Same as *bla* gene
ampicillin A widely used antibiotic of the penicillin family
beta-lactamase (β-lactamase) Enzyme that degrades beta-lactam antibiotics, including penicillins and cephalosporins
***bla* gene** Gene conveying resistance to ampicillin and related antibiotics and encoding beta-lactamase. Same as *amp* gene
ColEI plasmid Small multicopy plasmid of *Escherichia coli* that forms the basis of many cloning vectors widely used in molecular biology

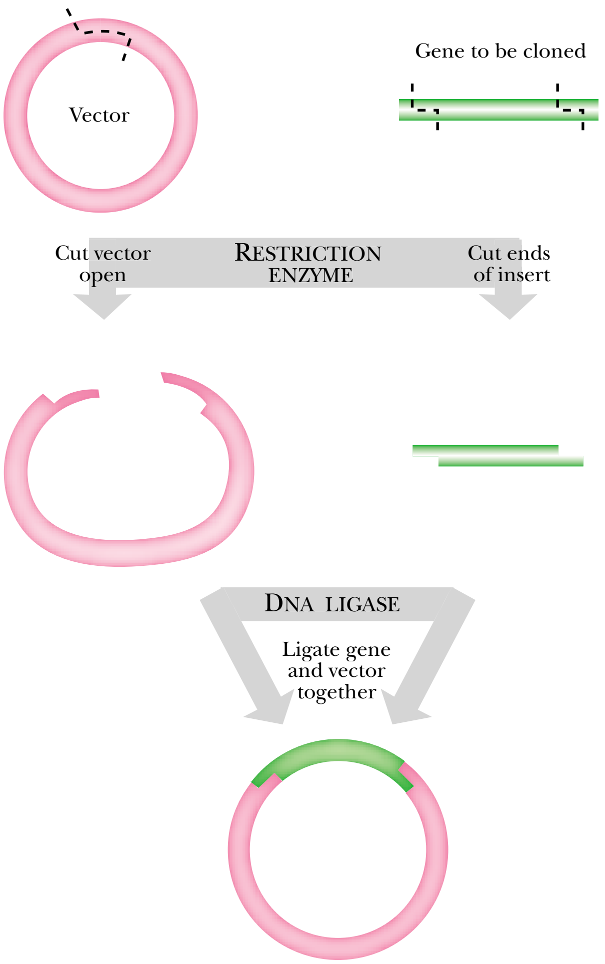


FIGURE 22.11 ***Insertion of DNA into Cloning Vector***

To insert the gene of interest into a plasmid vector, both the vector and the gene of interest should have compatible sticky ends. To achieve this both the gene and the vector must be digested with the same restriction enzyme. The two pieces are mixed together with DNA ligase, which joins the ends yielding a closed double-stranded circular plasmid carrying the gene of interest.

would cut it into multiple fragments. Furthermore, we must avoid inserting the cloned gene into any of the genes needed by the plasmid for its own replication and survival within the cell. Moreover, since there are many different restriction enzymes, it would be convenient to have a wide range of restriction recognition sites in the vector. These issues are all resolved by inserting a **polylinker**, or **multiple cloning site (MCS)**, into the cloning vector. This is a stretch of artificially synthesized DNA, about 50 base pairs long, which contains cut sites for seven or eight widely used restriction enzymes. This not only allows a wide choice of restriction enzymes, but ensures that the insert does not damage the plasmid and goes into more or less the same location each time (Fig. 22.12).

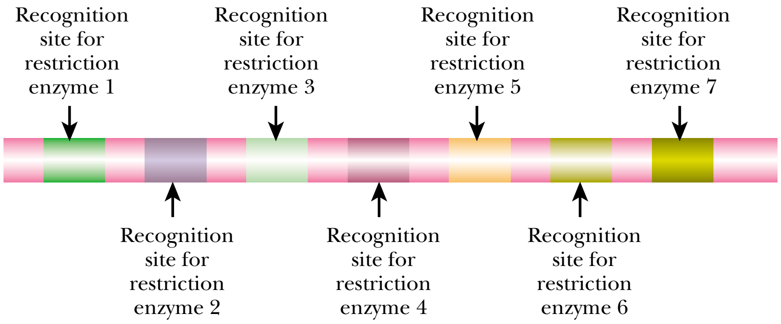
The remainder of the plasmid should not contain any cut sites for any of the enzymes represented in the polylinker. One way to ensure this is to choose only enzymes with zero cut sites in the original plasmid. Alternatively, we can get rid of unwanted cut sites by the approach shown in Figure 22.13. Due to spontaneous mutation (see Ch. 13 for mutations), occasional plasmids will suffer a base change within the cut site that needs to be eliminated. This will abolish recognition of the site by the restriction enzyme. The question is how to find this one rare mutant plasmid. First plasmid DNA is prepared and treated with the restriction enzyme in question. The plasmid DNA is then transformed into fresh bacterial cells without re-ligating the break. Wild type bacteria rapidly degrade incoming linear DNA; therefore, the majority of plasmids will be destroyed by this procedure. Those few that have lost the cut site by mutation will remain circular and survive.

multiple cloning site (MCS) A stretch of artificially synthesized DNA that contains cut sites for seven or eight widely used restriction enzymes. Same as polylinker

polylinker A stretch of artificially synthesized DNA that contains cut sites for seven or eight widely used restriction enzymes. Same as multiple cloning site (MCS)

FIGURE 22.12 ***Polylinker or Multiple Cloning Site***

Many plasmid vectors contain an artificial region of DNA that has many different restriction enzyme sites. Such polylinkers or multiple cloning site are designed so that all the restriction enzyme sites in the polylinker are unique, and the corresponding enzymes only cut the plasmid once.


Recognition site for restriction enzyme 1
Recognition site for restriction enzyme 3
Recognition site for restriction enzyme 5
Recognition site for restriction enzyme 7
Recognition site for restriction enzyme 2
Recognition site for restriction enzyme 4
Recognition site for restriction enzyme 6

Detecting Insertions in Vectors

Once a gene or other fragment of DNA has been cloned into a plasmid vector and transformed into a bacterial cell, we face the problem of detecting its presence. The plasmid itself may be detected by conferring antibiotic resistance on the host cell, but this leaves the question of whether the presumed insert is actually there. If the cloned gene itself codes for a product that is easy to detect, there is no problem. In most cases, however, the presence of the inserted DNA itself must be directly monitored.

Inserts in a vector can be checked by isolating DNA, cutting with a restriction enzyme, and seeing how many fragments are generated.

The least sophisticated and most tedious method is to screen a large number of suspects for the inserted DNA. Many separate bacterial colonies that received the plasmid vector, hopefully with DNA inserted, are grown in separate vials. Plasmid DNA is extracted from each of these bacterial cultures and cut with the restriction enzyme used in the original cloning experiment. If there is no insert in the plasmid, this merely converts the plasmid from a circular to a linear molecule of DNA. If the vector contains inserted DNA, two pieces of DNA are produced, one being the original plasmid and the other the inserted DNA fragment. To see how many fragments of DNA are present, the cut DNA is separated by agarose gel electrophoresis. If enough transformed colonies are tested, sooner or later one carrying a plasmid with the inserted DNA fragment will be found. This approach was of necessity used in the early days of genetic engineering. Today, modified vectors are available that facilitate screening by a variety of approaches.

Inserts are sometimes screened by the change in growth properties due to disrupting a gene on the vector.

Rather less laborious is to use a plasmid with two antibiotic resistance genes. One antibiotic resistance gene is used to select for cells, which have received the plasmid vector itself. The second is used for the insertion and detection of cloned DNA (Fig. 22.14). The cut site for the restriction enzyme used must lie within this second antibiotic resistance gene. When the cloned fragment of DNA is inserted this antibiotic resistance gene will be disrupted. This is referred to as **insertional inactivation**. Consequently, cells that receive a plasmid without an insert will be resistant to both antibiotics. Those receiving a plasmid with an insert will be resistant to only the first antibiotic.

Inserts are often detected by blue and white screening with Xgal. Inserts abolish production of beta-galactosidase and result in white (rather than blue) colonies.

The most convenient and widely used method to screen for inserts uses color screening. The most common procedure uses **β-galactosidase** and **X-gal** to produce bacterial colonies that change color when an insert is present within the vector. The process, called **blue/white screening**, has a unique vector that carries the 5′-end of the *lacZ* gene. This truncated gene encodes the **alpha fragment** of β-galactosidase, which consists of the N-terminal region or first 146 amino acids. A specialized bacterial host strain is required whose chromosome carries a *lacZ* gene missing the front portion but encodes the rest of the β-galactosidase protein. If the plasmid and chromosomal gene segments are active they produce two protein fragments that associate to give an active enzyme. This is referred to as **alpha complementation** (Fig. 22.15). Note that assem-

alpha complementation Assembly of functional β-galactosidase from N-terminal alpha fragment plus rest of protein
alpha fragment N-terminal fragment of β-galactosidase
beta-galactosidase (β-galactosidase) Enzyme that cleaves lactose and other β-galactosides so releasing galactose
blue/white screening Screening procedure based on insertional inactivation of the gene for β-galactosidase
insertional inactivation Inactivation of a gene by inserting a foreign segment of DNA into the middle of the coding sequence
X-gal Chromogenic substrate that is split by β-galactosidase so releasing an insoluble blue dye

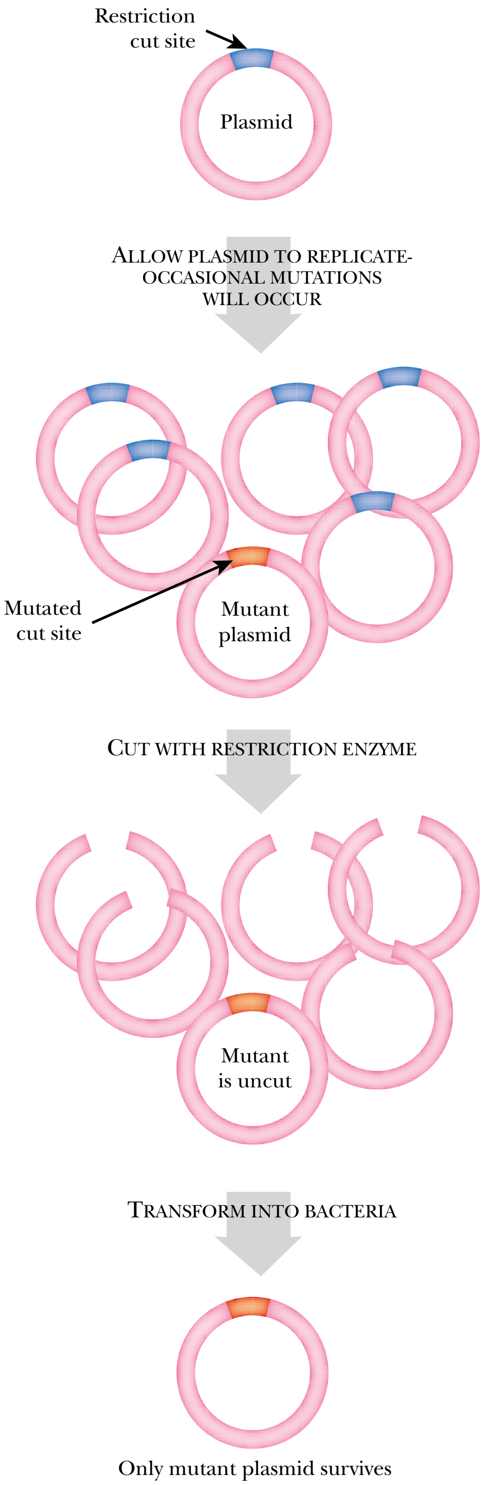


FIGURE 22.13 ***Eliminating Unwanted Restriction Sites***

The restriction site shown in blue is unwanted. During normal DNA replication, occasional mutations occur. Consequently a very small percentage of the plasmids will carry a random mutation (red) that alters this particular restriction recognition sequence. A sample of the plasmid DNA is isolated from a bacterial culture. The plasmids are treated with the appropriate restriction enzyme. All will be cut, except those with mutant restriction sites. The mixture is then transformed back into bacterial cells. Bacteria receiving cut, linearized plasmids will degrade them. Only mutant plasmids which remain circular will survive.

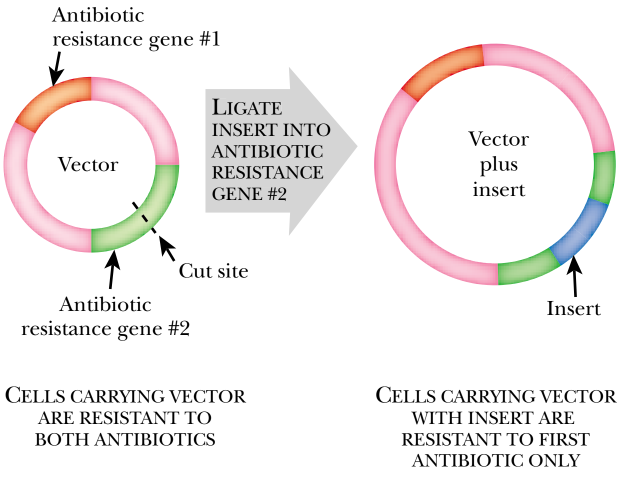

Antibiotic resistance gene #1
Vector
Cut site
Antibiotic resistance gene #2
LIGATE INSERT INTO ANTIBIOTIC RESISTANCE GENE #2
Vector plus insert
Insert
CELLS CARRYING VECTOR ARE RESISTANT TO BOTH ANTIBIOTICS
CELLS CARRYING VECTOR WITH INSERT ARE RESISTANT TO FIRST ANTIBIOTIC ONLY

FIGURE 22.14 ***Screening for Insert by Disruption of Antibiotic Resistance***

A plasmid that has a unique restriction enzyme site within an antibiotic resistance gene can be used to identify those plasmids into which a cloned gene has been inserted successfully. If the gene of interest is ligated into this restriction site, the antibiotic resistance gene will no longer be active. Any bacteria harboring the plasmid with an insert will no longer be resistant to this particular antibiotic.

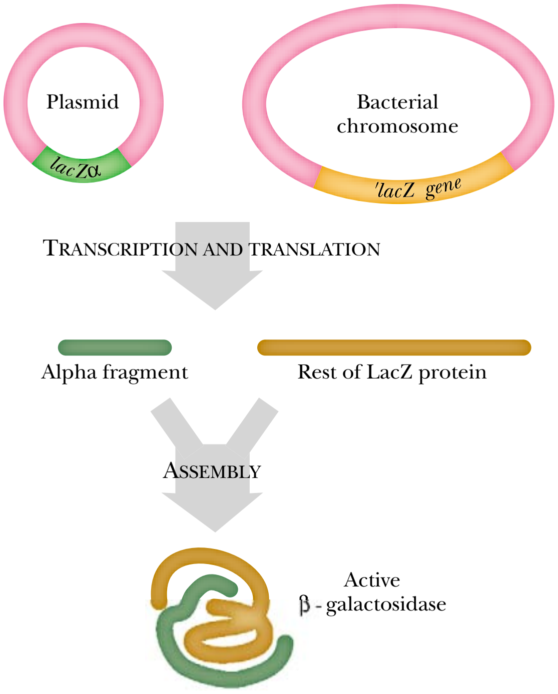


FIGURE 22.15 ***Alpha complementation***

The β-galactosidase protein is unique since it can be expressed as two pieces that come together to form a functional protein. The two protein fragments can be encoded on two different molecules of DNA within the bacterial cell. The alpha fragment can be expressed from a plasmid and the remainder of the β-galactosidase can be expressed from the chromosome.

bling an active protein from fragments made separately is normally not possible. Fortunately, β-galactosidase is exceptional in this respect. The reason for splitting *lacZ* between plasmid and host is that the *lacZ* gene is unusually large (approximately 3000 bp—almost as large as many small plasmids) and it greatly helps if cloning plasmids are small.

In order to utilize this unique protein for cloning, a polylinker is inserted into the *lacZα* coding sequence on the plasmid, very close to the front of the gene. Luckily, the very front most part of the β-galactosidase protein is inessential for enzyme activity. As long as the polylinker is inserted without disrupting the reading frame, the small addition does not affect the enzyme. However, if a foreign segment of DNA is inserted into the polylinker, the alpha fragment of β-galactosidase is disrupted and no active enzyme can form (Fig. 22.16). The active form of β-galactosidase splits X-gal, which produces a blue color (see Ch. 7 for details). Plasmids without a DNA insert will produce β-galactosidase and the bacterial cell that carries them will turn blue. Plasmids with an insert will be unable to make β-galactosidase and the cells will stay white.

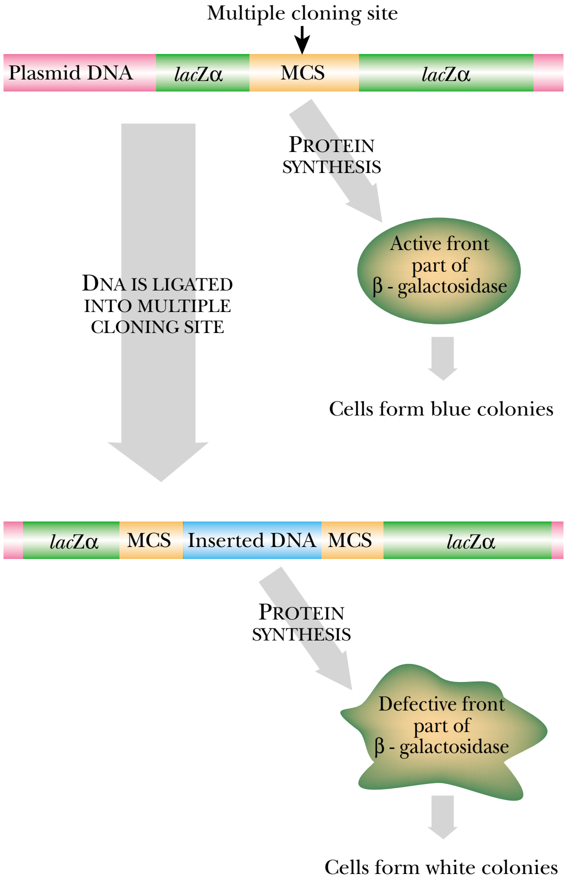

Multiple cloning site
Plasmid DNA
lacZα
MCS
lacZα
PROTEIN SYNTHESIS
Active front part of β - galactosidase
Cells form blue colonies
DNA IS LIGATED INTO MULTIPLE CLONING SITE
lacZα
MCS
Inserted DNA
MCS
lacZα
PROTEIN SYNTHESIS
Defective front part of β - galactosidase
Cells form white colonies

FIGURE 22.16 ***Blue/White Screening for β-Galactosidase***

In order to screen for inserts in a plasmid with the *lacZα* gene, a small polylinker is inserted into the extreme N-terminal portion of *lacZα*. The small insertion is cloned in-frame, therefore, the alpha fragment is still active when complexed with the remainder of β-galactosidase expressed from the chromosomal gene. Bacteria containing this construct will turn blue in the presence of X-gal. However, if a large segment of DNA, such as a cloned gene, is inserted into the multiple cloning site, the alpha fragment is disrupted and β-galactosidase is no longer active. Bacteria harboring this plasmid cannot cleave X-gal, and therefore remain white.

Moving Genes between Organisms: Shuttle Vectors

Specialized vectors have been made that can replicate in more than one organism. This allows the same gene to be expressed in different hosts.

The plasmid vectors we have discussed so far are designed to work in bacteria. Even when investigating genes from animals or plants, they are normally cloned first onto bacterial plasmids. Eventually, however, cloned genes are often moved from one organism to another. This may be done using a **shuttle vector**. As its name implies, this is a vector that can survive in more than one type of host cell. Obviously, the detailed requirements for vectors will vary depending on the host organism, but the general ideas are the same.

The earliest shuttle vectors were designed to shuttle between bacteria, such as *E. coli*, and yeast, a simple eukaryote (Fig. 22.17). Starting with a bacterial plasmid vector, several extra components are needed to create such a shuttle vector:

Shuttle vectors must have separate origins of replication and separate selection mechanisms for each host organism.

A. An origin of replication that works in yeast. Prokaryotic replication origins do not work in eukaryotes or vice versa, since the required DNA sequences differ substantially. However, the sequences of replication origins are rather similar in different eukaryotes, and so the yeast origin will work in many other higher organisms, at least to some extent.

B. A **centromere sequence** to allow correct partition of the plasmid in yeast. When a yeast cell divides, the duplicated chromosomes are pulled apart by micro-

centromere (Cen) sequence Sequence at centromere of eukaryotic chromosome that is needed for correct partition of chromosomes during cell division
shuttle vector A vector that can survive in and be moved between more than one type of host cell

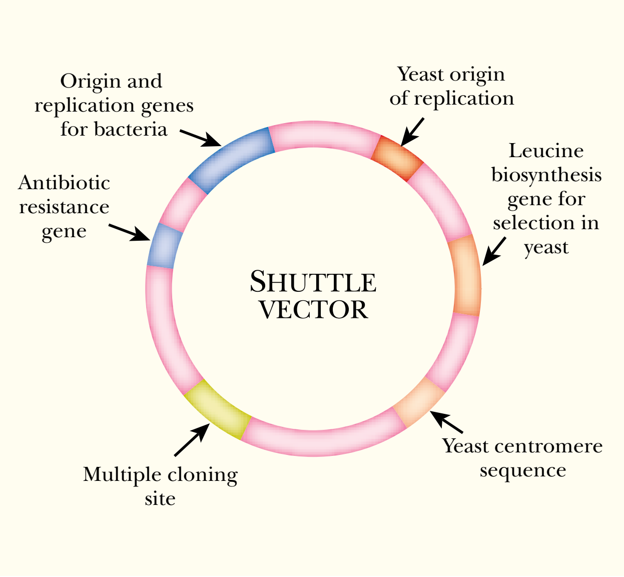

Origin and replication genes for bacteria
Yeast origin of replication
Leucine biosynthesis gene for selection in yeast
Antibiotic resistance gene
SHUTTLE VECTOR
Multiple cloning site
Yeast centromere sequence

FIGURE 22.17 *Shuttle Vector for Yeast*

In order for a shuttle vector to grow in both yeast and *E. coli*, it must have several essential elements: two origins of replication, one for *E. coli* and one for yeast; a yeast centromere sequence so that it is partitioned into the daughter cells during yeast replication ; selectable markers for both yeast and *E. coli*; and a multiple cloning site for inserting the gene of interest.

tubules attached to their centromeres, so that each daughter cell gets a full set. Shuttle vectors must be segregated correctly at cell division also. To achieve this, the shuttle vector must contain a segment of DNA from the centromere of the yeast chromosomes, the **Cen** sequence. This is recognized by the microtubules that drag new chromosomes apart.

C. A gene to select for the plasmid in yeast. The problem here is that yeast is not affected by most of the antibiotics that kill bacteria. In practice, a less satisfactory selection technique is used. A yeast host strain that has a defect in a gene for making an amino acid, say leucine, is used. The corresponding biosynthetic gene is present on the vector. In the absence of leucine the yeast will starve. Only if it obtains the plasmid carrying the *leu*$^+$ gene will it survive.

Bacteriophage Lambda Vectors

Viruses may be used as vectors, especially if they can adopt a lysogenic or latent state where they replicate in step with the host cell.

Bacteriophage lambda, which infects *E. coli*, has been widely used as a cloning vector. As described in Ch. 18, lambda is a well-characterized virus with both lytic and lysogenic alternatives to its life cycle. Although lambda DNA circularizes for replication and insertion into the *E. coli* chromosome, the DNA inside the phage particle is linear (Fig. 22.18). At each end are complementary 12 bp long overhangs known as ***cos* sequences (cohesive ends)**. Once inside the *E. coli* host cell, these pair up and the cohesive ends are ligated together by host enzymes forming the circular version of the lambda genome.

If essential genes are removed from a virus vector, a helper virus may be needed to allow replication.

Only DNA molecules of between 37 and 52 kb can be stably packaged into the head of the lambda particle. Small fragments of extra DNA may be inserted into the lambda genome without preventing packaging. However, to accommodate longer inserts it is necessary to remove some of the lambda genome. The left hand region has essential genes for the structural proteins and the right hand region has genes for replication and lysis. The middle region (~15 kb) of the lambda genome is non-essential and may be replaced with approximately 23 kb of foreign DNA (Fig. 22.19). Since the middle region of lambda has the genes for integration and recombination, such lambda replacement vectors cannot integrate into the host chromosome and form lysogens by

bacteriophage lambda Virus of *E. coli* with both lytic and lysogenic alternatives to its life cycle, which is widely used as a cloning vector
Cen sequence See centromere sequence
***cos* sequences (lambda cohesive ends)** Complementary 12 bp long overhangs found at each end of the linear form of the lambda genome

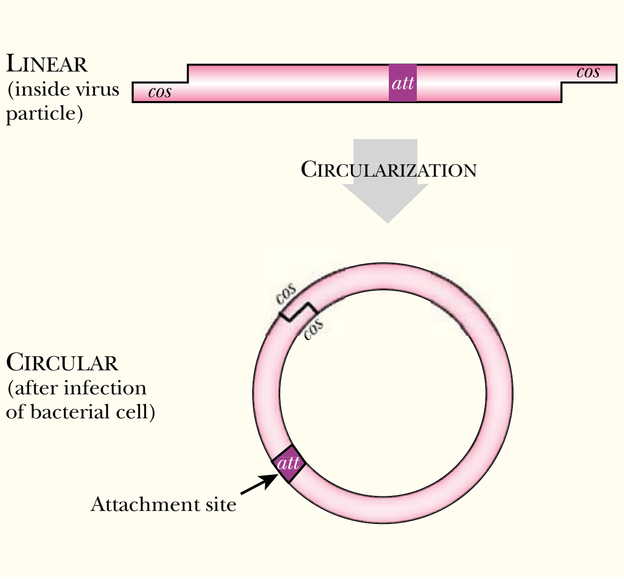


FIGURE 22.18 ***Lambda—Linear and Circular Genomes***

In the lambda phage particle, the genome is a linear DNA molecule with two *cos* sequences at each end. After the phage injects its DNA into the bacterial host, the DNA circularizes. The two cohesive ends base pair and are ligated together by bacterial enzymes so forming a circle.

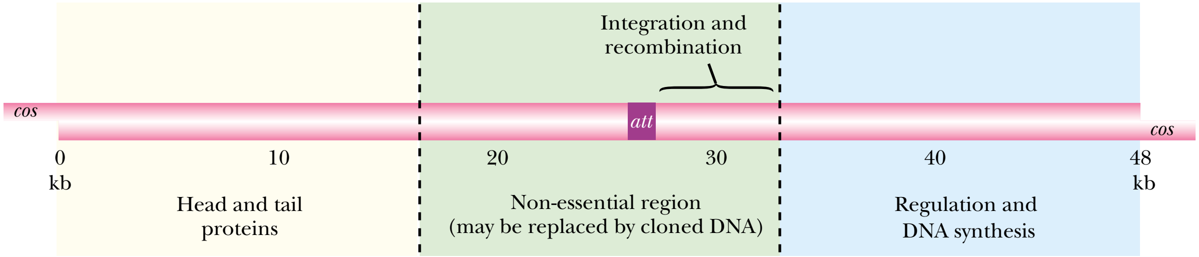

Integration and recombination
cos
att
cos
0
kb
10
20
30
40
48
kb
Head and tail proteins
Non-essential region (may be replaced by cloned DNA)
Regulation and DNA synthesis

FIGURE 22.19 ***Lambda Replacement Vector***

Since lambda phage is easy to grow and manipulate, the genome has been modified to accept foreign DNA inserts. The green region of the genome has genes that are non-essential for lambda growth and packaging. This region can be replaced with large inserts of foreign DNA (up to about 23 kb). When used with a helper phage, such modified lambdas provide useful cloning vectors.

themselves. To generate lysogens it is necessary to use a helper phage to provide the integration and recombination functions.

If foreign DNA is inserted into the middle of lambda, the result is a linear DNA molecule with two cohesive ends. To get such constructs into an *E. coli* host cell efficiently requires ***in vitro* packaging** (Fig. 22.20). In this technique, a mixture of lambda proteins is mixed with the recombinant lambda DNA *in vitro* to form phage particles. Infecting two separate *E. coli* cultures with two different defective lambda mutants generates the necessary lambda proteins. Each of the two mutants lacks an essential head protein and cannot form particles containing its own DNA. A mixture of the two lysates gives a full set of lambda proteins and when mixed with lambda DNA can generate infectious phage particles.

Cosmid Vectors

In vitro packaging using lambda lysates is a powerful technique. Packaging of DNA into a lambda head does not require the lambda genes, if fact, it is possible to fill almost

***in vitro* packaging** Procedure in which virus proteins are mixed with DNA *in vitro* to assemble infectious virus particles. Often used for packaging recombinant DNA into bacteriophage lambda

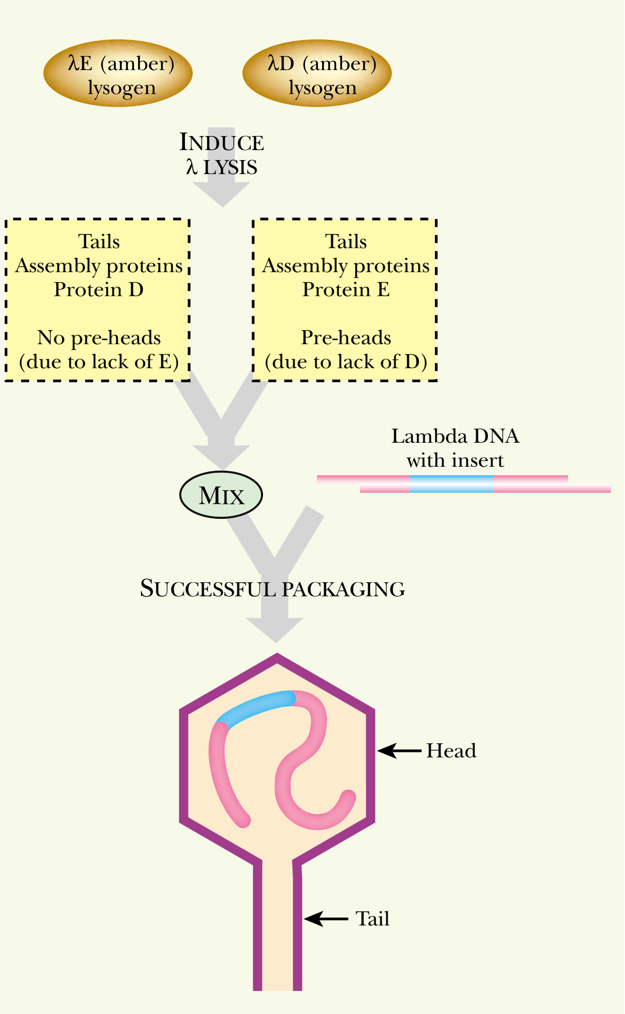


FIGURE 22.20 ***In Vitro Packaging of Lambda Replacement Vector***

A lambda cloning vector containing cloned DNA must be packaged in a phage head before it can infect *E. coli*. Before the DNA can be packaged, the phage head proteins must be isolated. To do this, a culture of *E. coli*, is infected with a mutant lambda which lacks the gene for one of the head proteins called E. A different culture of *E. coli* is infected with a different lambda mutant, which lacks phage head protein D. Both *E. coli* cultures are grown with the mutant lambdas and the viruses are induced to enter the lytic cycle. Although the *E. coli* are lysed by the phage, they cannot form complete heads. Instead a soluble mixture of phage proteins is isolated. Each lysate contains phage tails, assembly proteins, and components of the heads, except either D or E. These two lysates are mixed along with the lambda vector containing the cloned DNA. Although mixing is done *in vitro*, the components can self-assemble into a functional phage that can infect *E. coli*.

the whole of a lambda particle with cloned DNA by using **cosmid** vectors. Cosmids themselves are small multicopy plasmids that carry *cos* sites (Fig. 22.21). The cosmid is first linearized so that each end has a *cos* sequence. In order to clone a gene of interest into the cosmid, both the gene and cosmid are cut either with the same restriction enzyme or with two enzymes that give identical sticky ends (e.g., *Bam*HI and *Mbo*I, as in Fig. 22.21). The target DNA is often only partially digested, i.e., some sites are left uncut. First, this allows large segments of a genome to be isolated. Second, if a cut

cosmid Small multicopy plasmid that carries lambda *cos* sites and can carry around 45 kb of cloned DNA

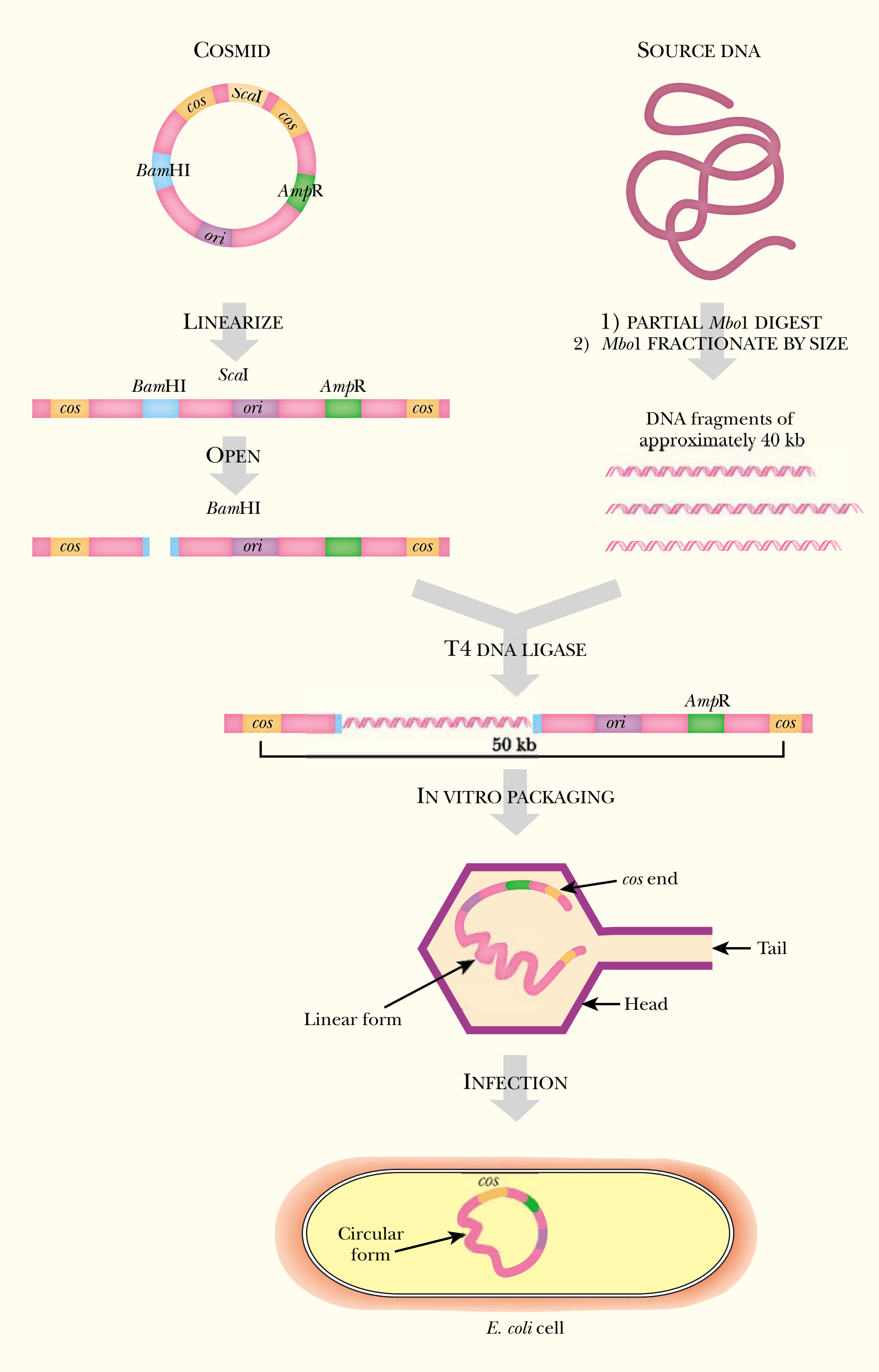


FIGURE 22.21 ***Cosmid Vector***

To clone large pieces of DNA into cosmid vectors, both must have compatible sticky ends. The cosmid vector is first linearized so that each end has a *cos* site. Then the linear cosmid is cut with *Bam*HI, which generates sticky ends with the overhang sequence GATC. The genomic DNA from the source of interest is also digested. Instead of *Bam*HI, this DNA is partially digested with *Mbo*I, which also generates a GATC overhang. Partial digestion leaves some sites uncut and allows large segments of a genome to be isolated. These segments are mixed with the two halves of the cosmid and joined using ligase. The final constructs are packaged into lambda particles *in vitro* and are used to infect *E. coli*.

site lies within a gene of interest, some fragments will still carry the intact gene. Ligation of the two cosmid pieces to either side of the target DNA results in a length of DNA with a *cos* site at each end. This construct can be packaged into lambda particles *in vitro*, and then used to infect *E. coli.* Using a small cosmid, of say 4 kb, allows inserts of up to about 45 kb to be cloned.

TABLE 22.02 Insert Sizes and Cloning Vectors

Vector	Maximum Insert Size
Multicopy plasmid	10 kb
Lambda replacement vector	20 kb
Cosmid	45 kb
P1 plasmid vector	100 kb
PAC (P1 artificial chromosome)	150 kb
BAC (bacterial artificial chromosome)	300 kb
YAC (yeast artificial chromosome)	2,000 kb

Artificial chromosomes are vectors designed to carry huge amounts of DNA from higher organisms.

Yeast Artificial Chromosomes

Analysis of the genomes of higher organisms requires the cloning of much larger fragments than for bacteria. Because eukaryotic genes contain introns they may be hundreds of kilobases in length. Such large DNA fragments require special vectors. The largest capacity vectors derived from bacteriophage can handle at most 100 kb (Table 22.02). Consequently, "artificial chromosomes" have been developed to carry huge lengths of eukaryotic DNA.

Huge segments of DNA, up to 2,000 kb or 2 million basepairs may be carried on **yeast artificial chromosomes or YACs** (Fig. 22.22). For any replicon, whether plasmid or chromosome, to survive, the vector must have a yeast specific origin of replication and a centromere recognition sequence (Cen sequence). The YAC has both of these elements. In addition, as required by all eukaryotic chromosomes, telomere sequences are present on both ends. A yeast cell will treat this structure, although artificial, as a chromosome. Of course, for practical use a selectable marker and a suitable multiple cloning site are also included.

Colossal amounts of cloned DNA can be inserted into a YAC and may thus be replicated inside yeast cells. Because the recognition sequences for replication origins, centromeres and telomeres are so similar among higher organisms, an added bonus is that YACs will survive in mice and are even passed on from parent to offspring. Admittedly, not every baby mouse inherits the YAC, nonetheless, this opens the way for cloning the huge DNA sequences needed for engineering higher animals and for sequencing their genomes.

Bacterial and P1 Artificial Chromosomes

Multicopy vectors, such as ColE1 derived plasmids, are valuable because they give higher yields of DNA than single copy vectors. However, they also have disadvantages. In particular, the inserts may be unstable especially if they are very long and contain repeated sequences. Many times, unstable inserts are deleted from the plasmid by recombination events. Eukaryotic DNA is particularly unstable in plasmids. Therefore, cloning large segments of eukaryotic DNA in bacteria is now done using **bacterial artificial chromosomes (BACs)**. These are single copy vectors based on the F-plasmid of *E. coli*. They can accept inserts of 300 kb or more. Electroporation is necessary to trans-

bacterial artificial chromosome (BAC) Single copy vector based on the F-plasmid of *E. coli* that can carry very long inserts of DNA. Widely used in the human genome project

yeast artificial chromosome (YAC) Single copy vector based on yeast chromosome that can carry very long inserts of DNA. Widely used in the human genome project

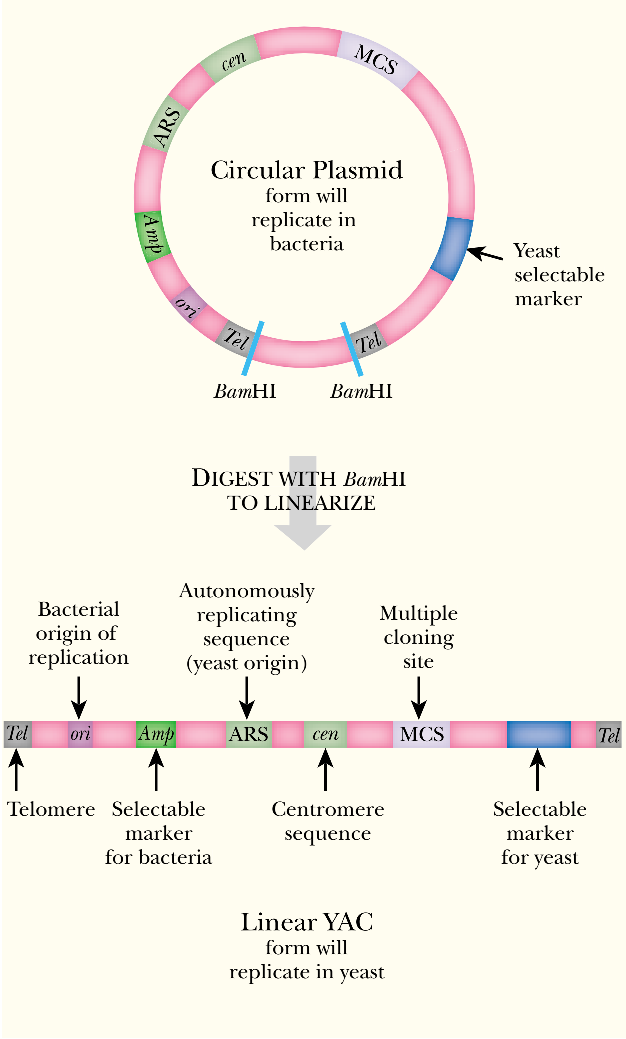


FIGURE 22.22 ***Yeast Artificial Chromosome (YAC)***

The YAC has two forms, a circular form for growing in bacteria, and a linear form for growing in yeast. The circular form can be manipulated and grown like any other plasmid in bacteria since it has a bacterial origin of replication and an antibiotic resistance gene. In order to use this in yeast, the circular form is isolated and linearized such that the yeast telomere sequences are on each end. This form can accommodate up to 2,000 kb of cloned DNA inserted into its multiple cloning site (MCS).

form these large constructs into *E. coli* host cells and the yields of DNA are low. Nonetheless bacterial artificial chromosomes have been widely used in the human genome project and other eukaryotic genome sequencing projects.

Another cloning vector used for larger eukaryotic DNA segments is the **P1 artificial chromosome (PAC)**. This cloning vector is derived from bacteriophage P1, and has been used to carry inserts of up to 150 kb. Just like the lambda derived vectors (see above), these PACs require *in vitro* packaging. Artificial chromosomes based on P1 have also been made for use in *E. coli* host cells.

A DNA Library Is a Collection of Genes from One Organism

Gene libraries or **DNA libraries** are collections of cloned genes that are big enough to contain at least one copy of every gene from a particular organism. The size of the

DNA library Collection of cloned segments of DNA that is big enough to contain at least one copy of every gene from a particular organism. Same as gene library
Gene library Collection of cloned segments of DNA that is big enough to contain at least one copy of every gene from a particular organism. Same as DNA library
P1 artificial chromosome (PAC) Single copy vector based on the P1-phage/plasmid of *E. coli* that can carry very long inserts of DNA

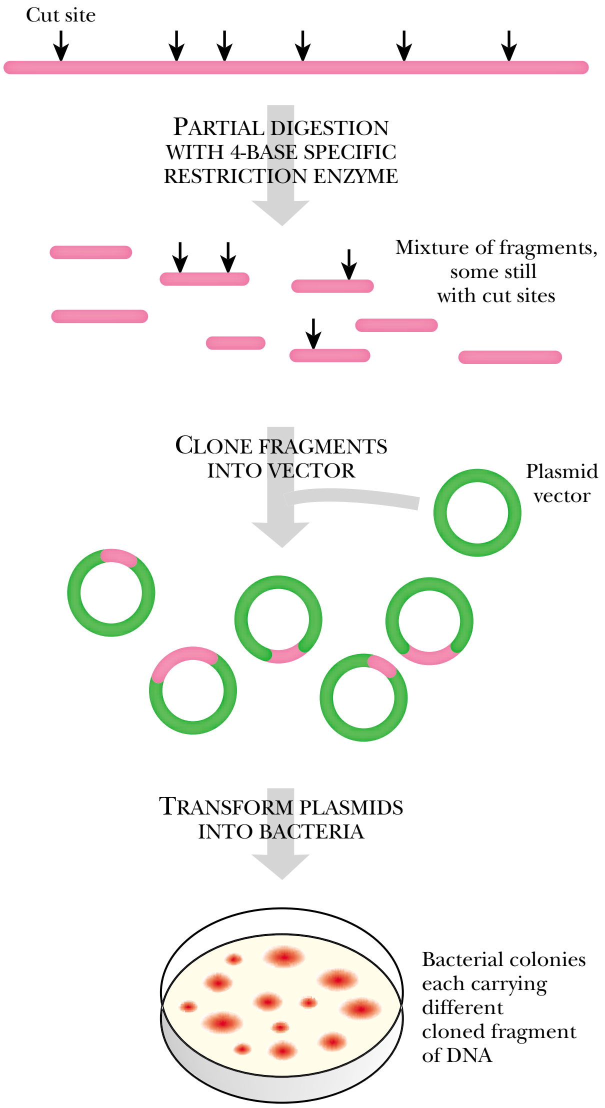

Cut site
PARTIAL DIGESTION WITH 4-BASE SPECIFIC RESTRICTION ENZYME
Mixture of fragments, some still with cut sites
CLONE FRAGMENTS INTO VECTOR
Plasmid vector
TRANSFORM PLASMIDS INTO BACTERIA
Bacterial colonies each carrying different cloned fragment of DNA

FIGURE 22.23 ***Creating a DNA Library***

A DNA library contains as many genes from the organism of interest as possible. The genomic DNA from the organism of interest is isolated and digested with a restriction enzyme. Usually, the restriction enzyme used has a recognition sequence of four base pairs, therefore, the DNA would be cut into fragments much smaller than the average gene. Therefore, the digestion is carried out for a brief period that leaves many of the restriction sites uncut. A suitable vector for the required insert size is chosen and is cut with a restriction enzyme that produces compatible sticky ends. The digested genomic DNA and the vector are ligated together and transformed into bacterial host cells. A large number of transformed bacterial colonies must be isolated and kept to ensure that all possible genes from the genome of interest are represented on at least one vector.

Collections of cloned genes carried on a vector are used to screen for particular genes.

genes and the organism the library will be propagated in, dictate which of the vectors are used for holding the inserts. The genes of prokaryotes are relatively short, averaging about a 1000 bp each. In contrast, eukaryotic genes are much longer, largely due to the presence of introns. Different strategies must therefore be followed for eukaryotic gene libraries as discussed below.

To make a prokaryotic gene library, the complete bacterial chromosomal DNA is cut with a restriction endonuclease and the fragments are inserted into a vector, usually a simple ColE1-derived plasmid (Fig. 22.23). This mixture of cloned fragments is transformed into a suitable bacterial host strain and a large number of colonies containing vector plus insert are kept. These must then be screened for the gene of interest. If the gene has an observable phenotype, this may be used. Otherwise, more general methods such as hybridization or immunological screening are necessary.

Gene libraries are often made using a 4-base specific restriction enzyme to cut the genomic DNA. This cuts DNA every 256 bases on average. Since this is shorter than an average gene, the DNA is only partially digested by only allowing a short amount of time for the restriction enzyme to cut the DNA. This generates a mixture of fragments of various lengths, many of which still have restriction sites for the enzyme used. The hope is that an intact copy of every gene, even those cut by the enzyme used, will be present on at least some fragments of DNA (Fig. 22.23). However, because restriction sites are not truly distributed at random, some fragments will be too large to be cloned and some genes will contain clustered multiple restriction sites and will be destroyed even in a partial digest. For total coverage, another library should be made with another restriction enzyme.

Screening a Library by Hybridization

Gene libraries can most easily be screened by hybridization with a DNA probe.

After cloning all the possible genes from an organism into a library, the next step is to identify the gene of interest. Libraries are often screened for the gene of interest by DNA/DNA hybridization using a DNA probe. The probes themselves are generally derived from two sources. Cloned DNA from a related organism is often used to screen a library. The stringency of the hybridization conditions must be adjusted to allow for a greater or lesser percentage of mismatches, depending on how closely related the two organisms are. Another possibility is to synthesize an artificial probe, using the base sequence deduced from the amino acid sequence of the corresponding protein. This assumes that the protein has been purified and that a partial amino acid sequence from the N-terminal region is available. The DNA probe is labeled for detection by autoradiography, fluorescence or chemical tagging as described in Ch. 21. Probes generally range from 100 to 1000 bases long, although shorter probes may sometimes be used. At least 80% matching over a 50 base stretch is needed for acceptable hybridization and identification.

The **target DNA** (i.e. the DNA from the library to be probed) is denatured and bound to a nitrocellulose or nylon membrane. The membrane is then incubated with the labeled probe. After washing away excess probe, the membrane is screened by the chosen detection system, e.g. autoradiography as illustrated in Fig. 22.24.

Screening a Library by Immunological Procedures

Gene libraries may also be screened for the proteins expressed from the genes, rather than the DNA sequences.

Instead of looking for DNA/DNA hybrids to identify the gene of interest from a library, the protein itself can be identified by **immunological screening**. This method relies on the production of the protein encoded by the gene of interest and therefore assumes that the cloned gene is efficiently expressed under the experimental conditions. That is, each of the library inserts must have transcriptional and translational start sequences as well as stop sequences. Usually, the library vector supplies these sequences, since the promoters from the genomic DNA will not usually be cloned still attached to the genes they control (see below). When the protein is expressed, it is detected by binding to an **antibody**. This means that antibody to the encoded protein (or a closely related protein from another organism) must be available.

In order to screen an expression library, the bacteria expressing the library inserts are grown on master plates and samples of each bacterial colony are transferred to a suitable membrane. The cells are lysed and the released proteins are attached to the membrane. The membrane is then treated with a solution of the appropriate antibody. After excess primary antibody is washed away, a second antibody that is specific for the primary antibody is added. This will bind any primary antibody it encounters (Fig. 22.25). This secondary antibody carries the detection system, such as alkaline phosphatase, which converts a colorless substrate, such as X-phos, to a colored product (see Ch. 21). If X-phos is used, the region on the membrane where the secondary antibody is bound turns blue. The blue spots must be aligned with the original bacterial colonies. The DNA from the bacteria containing the insert encoding the protein of interest can then be isolated.

The reason for using two different antibodies is to allow flexibility and amplify the signal. Antibodies to the protein of interest are made by injecting a rabbit with the

antibody Protein made by the immune system to recognize and bind to foreign proteins or other macromolecules
immunological screening Screening procedure that relies on the specific binding of antibodies to the target protein
target DNA DNA that is the target for binding by a probe during hybridization or the target for amplification by PCR

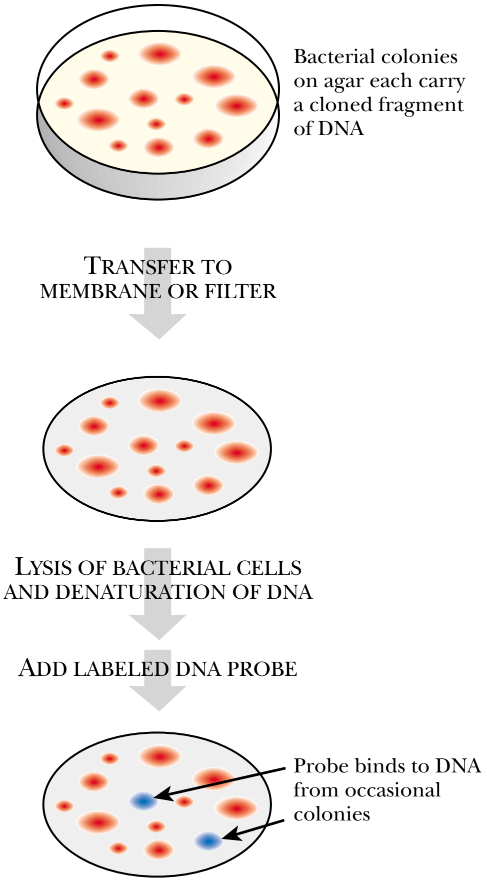

Bacterial colonies on agar each carry a cloned fragment of DNA
TRANSFER TO MEMBRANE OR FILTER
LYSIS OF BACTERIAL CELLS AND DENATURATION OF DNA
ADD LABELED DNA PROBE
Probe binds to DNA from occasional colonies

FIGURE 22.24 ***Screening a DNA Library by Probing***

The first step in screening a DNA library is to grow colonies of bacteria containing the library inserts on agar. A large number of different transformed bacteria are grown, so that all genes in the library have a reasonable chance of being present. Next, the bacterial colonies are transferred to a membrane or filter. The filter is applied to the top of the bacterial colonies and carefully lifted off. A portion of each bacterial colony will stick to the filter while the rest of the colony stays on the agar plate. Once on the filter, the bacteria are lysed open and the DNA is denatured. The single-stranded DNA stays bound to the filter, and the majority of the bacterial components are washed away. The library filters are covered with a solution of a radioactively labeled single-stranded DNA probe, allowed to hybridize, and then the excess probe is washed away. Placing a piece of photographic film over the filter identifies the hybrid molecules. When the probe hybridizes to a library insert, a black spot appears on the photographic film. By lining up the original bacterial colonies with the photographic film, the corresponding library insert can be isolated from the bacteria.

protein, and isolating all the antibodies from a sample of the rabbit's blood. Producing an antibody is costly and a long process, so instead of directly conjugating this antibody to the detection system, a second antibody is produced in another animal, such as a goat. The secondary antibody recognizes all rabbit antibodies; therefore, it can be used for any primary antibody made in a rabbit. The secondary antibodies are available from companies and are relatively inexpensive. Secondary antibodies also amplify the signal, since usually two secondary antibody molecules bind to each primary antibody. Double the color will be produced using the two antibody system.

Non-coding introns are a nuisance in many procedures. They can be avoided by making cDNA copies of the gene using reverse transcriptase.

Cloning Complementary DNA Avoids Introns

Most eukaryotic genes have intervening sequences of non-coding DNA (introns) between the segments of coding sequence (exons). In higher eukaryotes, the introns

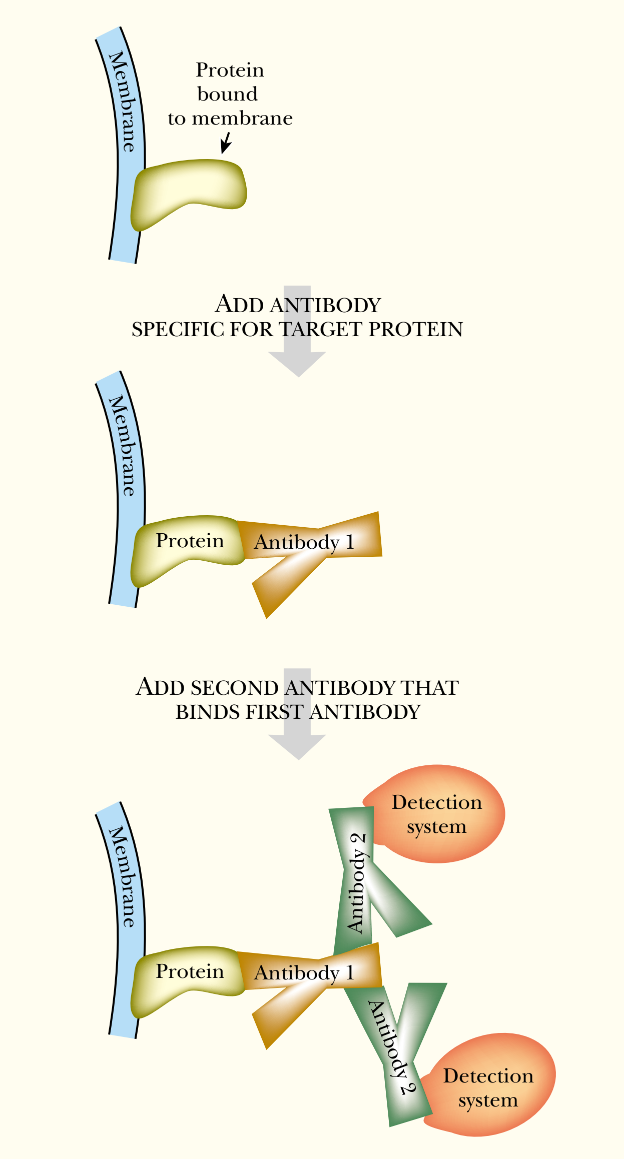

Membrane
Protein bound to membrane
ADD ANTIBODY SPECIFIC FOR TARGET PROTEIN
Membrane
Protein
Antibody 1
ADD SECOND ANTIBODY THAT BINDS FIRST ANTIBODY
Membrane
Detection system
Antibody 2
Protein
Antibody 1
Antibody 2
Detection system

FIGURE 22.25
Immunological Screening of a DNA Library

Bacteria carrying a library are grown on agar, transferred to a membrane and lysed. Released proteins are bound to the membrane. This figure shows only one attached protein, but in reality, a large number of different proteins will be present. These proteins include both those from the library as well as many bacterial proteins. The membrane is incubated with a primary antibody that only binds the protein of interest. Any non-specifically bound antibody is washed away. Finally a second antibody that binds the primary antibody and that also carries a detection system is added.

are often longer than the exons and the overall length of the gene is therefore much larger than the coding sequence. This creates two problems. First, cloning large segments of DNA is technically difficult; plasmids with large inserts are often unstable and transform poorly. Secondly, bacteria cannot process RNA to remove the introns and so eukaryotic genes containing introns cannot be expressed in bacterial cells. Using a DNA copy of mRNA, known as **complementary DNA or cDNA**, solves both problems since the mRNA has already been processed so that all the introns are removed.

To make a cDNA library, the mRNA must be isolated and used as a template. The library generated therefore reflects only those genes expressed in the particular tissue under the chosen conditions. First, total RNA is extracted from a particular cell culture, tissue or specific embryonic stage. The messenger RNA from eukaryotic cells is normally isolated from the total RNA by taking advantage of its poly(A) tail. Since adenine base pairs to thymine or uracil, columns containing **oligo(U)** or **oligo(dT)**

complementary DNA (cDNA) DNA copy of a gene that lacks introns and therefore consists solely of the coding sequence. Made by reverse transcription of mRNA
oligo(dT) Stretch of single-stranded DNA consisting solely of dT or deoxythymidine residues
oligo(U) Stretch of single-stranded RNA consisting solely of U or uridine residues

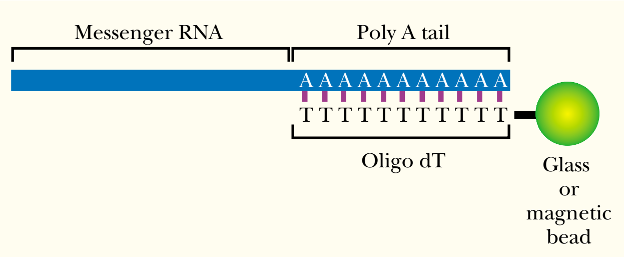


FIGURE 22.26 *Purifying mRNA by Oligo(dT)*

In order to isolate only messenger RNA from a sample of eukaryotic tissue, the unique features of the mRNA molecule are used. Only mRNA has a poly(A) tail, a long stretch of adenines following the coding sequence. The poly(A) tail of the mRNA will bind by base pairing to an oligonucleotide consisting of a long stretch of deoxythymidine residues—oligo(dT). The oligo(dT) is attached to glass or magnetic beads, which consequently bind mRNA specifically. Other RNAs will not bind to the beads, and can be washed from the column.

tracts bound to a solid matrix are used to bind the mRNA by its poly(A) tail. The mRNA is retained and the other RNA is washed through the column. The mRNA is then released by eluting with a buffer of high ionic strength that disrupts the H-bonding of the poly(A) tail to the oligo(dT) (Fig. 22.26). A variation of this approach is the use of magnetic beads with attached oligo(dT) tracts. After binding the mRNA the beads are separated magnetically.

To generate cDNA the enzyme reverse transcriptase, originally found in retroviruses (see Ch. 17), is added to the mRNA. This enzyme will make a complementary DNA (cDNA) strand using the mRNA as template. Several further steps are required to generated a double-stranded cDNA copy of the original mRNA (Fig. 22.27). Ribonuclease H, which only recognizes RNA, is used to remove the mRNA strand of the mRNA/cDNA hybrid molecule leaving a single-stranded cDNA. DNA polymerase I is then used to synthesize the second DNA strand. Any remaining single-stranded ends are trimmed off by S1 nuclease, which is an exonuclease specific for single-stranded regions of DNA. [Such single-stranded ends mostly result from oligo(dT) primers binding in the middle of the mRNA poly(A) tail.] The resulting double-stranded cDNA molecules can be isolated and cloned into an appropriate vector, resulting in a **cDNA library**. Since each mRNA has a different sequence, convenient restriction sites are generally added at each end. This is done by attaching linkers—short pieces of DNA that have restriction sites compatible with those in the multiple cloning site of the vector. Not only is cDNA easier to handle, because the cloned fragments are much shorter than the original eukaryotic genes, but the cDNA versions of eukaryotic genes can often be successfully expressed in bacteria.

Chromosome Walking

As a result of genetic investigation, we may know the approximate chromosomal location of a particular gene. Using this information to clone the gene is referred to as positional cloning. One of the simplest versions of this is **chromosome walking**, a method based on hybridization. This approach is used when one segment of DNA has already been cloned and the neighboring genes are of interest.

The chromosome that has the target gene can be identified and isolated by fluorescence *in situ* hybridization (FISH) followed by fluorescence activated sorting as described in Ch. 21. The chromosome of interest is then cut into manageable fragments with a suitable restriction enzyme. The original cloned fragment of DNA is used as the

Once a region on a chromosome has been cloned, neighboring regions can be accessed by moving along the chromosome using known sequences as probes.

cDNA library Collection of genes in their cDNA form, lacking introns
chromosome walking Method for cloning neighboring regions of a chromosome by successive cycles of hybridization using overlapping probes

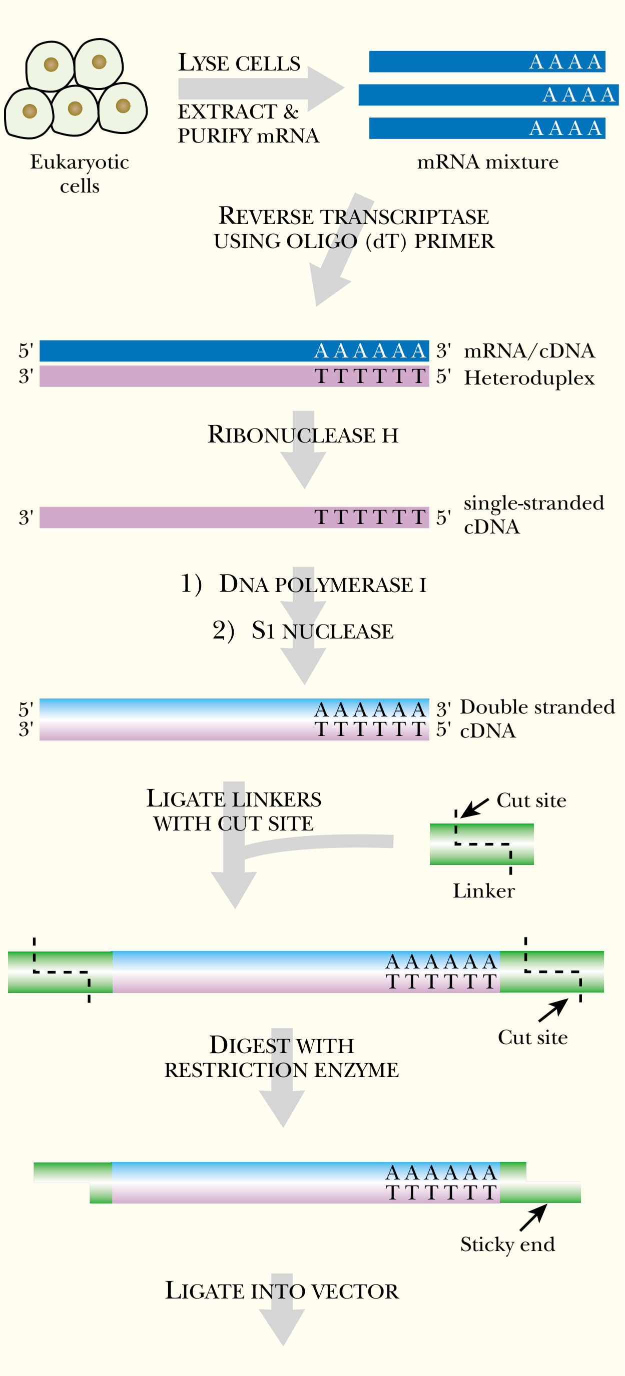


FIGURE 22.27 ***Making a cDNA Library from Messenger RNA***

First, eukaryotic cells are lysed and the mRNA is purified. Next, reverse transcriptase plus primers containing oligo(dT) stretches are added. The oligo(dT) hybridizes to the adenine in the mRNA poly(A) tail and acts as a primer for reverse transcriptase. This enzyme makes the complementary DNA strand so forming an mRNA/cDNA hybrid molecule. The mRNA strand is digested with ribonuclease H and DNA polymerase I is added to synthesize the opposite DNA strand, thus creating double-stranded cDNA. S1 nuclease trims off single-stranded ends. Since each mRNA has a different sequence, linkers must be ligated to the ends of the cDNA to allow convenient insertion into the cloning vector. After addition, the linkers are digested with the appropriate restriction enzyme and the cDNA is ligated into the vector. The resulting hybrid DNA molecules are then transformed into bacteria, so giving the final cDNA library.

first probe. Each fragment is tested for hybridization to an initial probe made from the segment that was already cloned. Obviously, this will overlap at least one or two neighboring fragments. These overlap fragments are then cloned and used as probes in a second cycle of hybridization, and so on. For simplicity, we have shown a walk in one direction only in Figure 22.28, though obviously in real life we would walk along the chromosome in both directions from the original starting point.

Each cycle of hybridization identifies segments of DNA that overlap the fragments isolated earlier and used as probes. By moving outwards from the starting position step by step, the whole chromosome can be mapped and cloned.

Cloning by Subtractive Hybridization

Subtractive hybridization is a technique used to isolate a DNA segment that is missing from one particular sample of DNA. Obviously, a second DNA sample that contains the fragment of interest is necessary. Suppose that a hereditary defect is due to the deletion of the DNA for a particular gene. A sample of DNA from the appropriate chromosome of the afflicted individual will lack this particular segment of DNA. To find the missing DNA, the corresponding chromosome from a healthy (wild type) individual is isolated. For example, the ***dmd* gene**, for **Duchenne muscular dystrophy**, is located in the Xp21 band, close to the middle of the short or p-arm of the X-chromosome. Using light microscopy to analyze chromosomal banding patterns, a patient was found who had a deletion large enough that the Xp21 band was missing. Subtractive hybridization of the mutant chromosome with a normal chromosome, allowed the *dmd* gene to be cloned.

Cloned genes can sometimes by found by a negative approach. Hybridization is used to remove genes shared by two organisms, leaving behind only those that are unique.

To do subtractive hybridization, both the mutant and wild type DNA samples are cut into fragments of convenient size using a restriction enzyme. Then the two sets of fragments are hybridized together. This will give hybrid molecules for all regions of the DNA except the region of the deletion, which is present only in the wild type chromosome. If a large surplus of mutant DNA is used, all fragments of the wild type chromosome will be hybridized to mutant fragments except the region corresponding to the deletion, which will be left over. The single strands of this lone fragment will have to pair with each other. Thus, we have subtracted out all the segments of DNA that are not wanted.

In practice, some means to obtain the left over "deletion fragment" is required. One approach is to cut the two batches of DNA with different restriction enzymes. If the normal DNA is cut with restriction enzyme 1 and the mutant DNA is cut with restriction enzyme 2, any hybrid molecule will have non-matching ends. If there mutant DNA hybridized with itself, the ends will be matching, and will cut with restriction enzyme 2. If the gene of interest from the normal DNA self-hybridizes, then this will have compatible ends to restriction enzyme 1. This fragment can then be cloned into a vector using restriction enzyme 1. To make the procedure even easier, restriction enzyme 2 could leave a blunt end after cutting. Since blunt ends are so much harder to ligate, only a self-hybrid molecule flanked by the sticky ends of restriction enzyme 2 would be cloned into the vector. Only the self-paired fragment of wild type DNA would have the sticky ends of restriction enzyme 2 (Fig. 22.29).

Subtractive hybridization can also be used to isolate a set of genes that are expressed under particular conditions. Two batches of cells are grown, one under standard conditions and the other under the conditions being investigated. For example, one batch of mouse cells can be grown with all the necessary nutrients, and another set of mouse cells can be grown with only limited nutrients. The total RNA is isolated from both samples, then the mRNA is purified by hybridization to oligo(dT) as

***Dmd* gene** Gene responsible for Duchenne muscular dystrophy
Duchenne muscular dystrophy One of several inherited diseases affecting mucle function
subtractive hybridization Technique used to remove unwanted DNA or RNA by hybridization so leaving behind the DNA or RNA molecule of interest

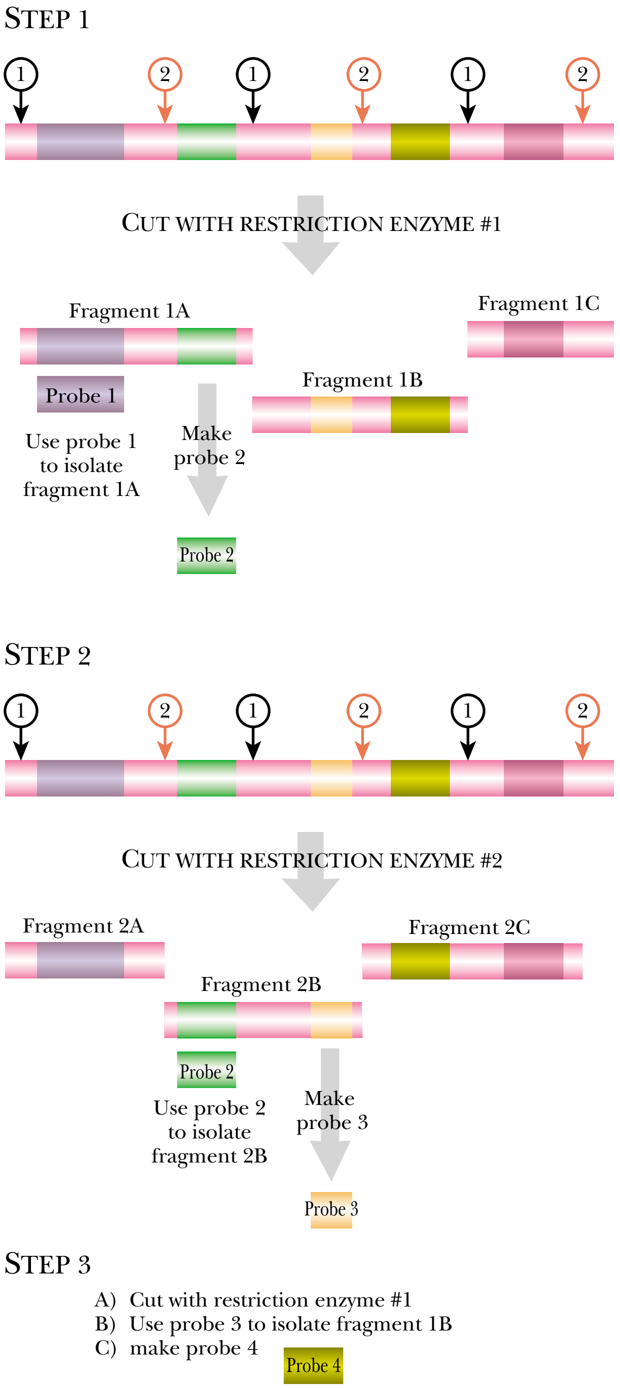

STEP 1
1
2
1
2
1
2
CUT WITH RESTRICTION ENZYME #1
Fragment 1A
Fragment 1C
Probe 1
Fragment 1B
Use probe 1 to isolate fragment 1A
Make probe 2
Probe 2
STEP 2
1
2
1
2
1
2
CUT WITH RESTRICTION ENZYME #2
Fragment 2A
Fragment 2C
Fragment 2B
Probe 2
Use probe 2 to isolate fragment 2B
Make probe 3
Probe 3
STEP 3
A) Cut with restriction enzyme #1
B) Use probe 3 to isolate fragment 1B
C) make probe 4
Probe 4

FIGURE 22.28 ***Chromosome Walking***

Chromosome walking utilizes overlapping fragments of a particular chromosome to isolate genes upstream and downstream from the original DNA fragment. The first step is to identify the region of the chromosome to which the probe hybridizes. In this example, probe #1 hybridizes to the purple region of the chromosome. When the chromosome is cut with restriction enzyme #1, fragment 1A will hybridize to probe #1 at one end. This allows fragment 1A to be isolated and sequenced and its downstream sequence is used to generate probe #2. To find the next segment of the chromosome, a different restriction enzyme is used. This time probe #2 will hybridize to fragment 2B. Once again the probe recognizes only the first half of this fragment. The downstream sequence of fragment 2B can then be determined, and this information can be used to make probe #3. Next, the chromosome is cut with restriction enzyme #1 again. Now probe #3 will hybridize with fragment 1B, whose downstream sequence can therefore be determined, and another probe, called probe 4 can be made. This procedure can be continued as far as desired, working in either direction.

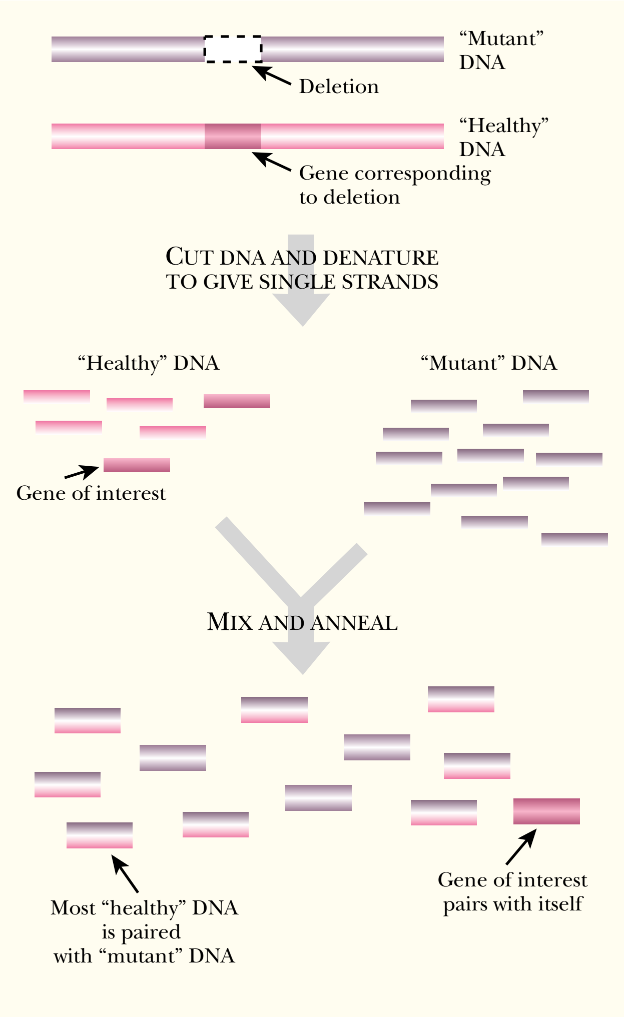

"Mutant" DNA
Deletion
"Healthy" DNA
Gene corresponding to deletion
CUT DNA AND DENATURE TO GIVE SINGLE STRANDS
"Healthy" DNA
"Mutant" DNA
Gene of interest
MIX AND ANNEAL
Most "healthy" DNA is paired with "mutant" DNA
Gene of interest pairs with itself

FIGURE 22.29 ***Cloning by Subtractive Hybridization***

The key to subtractive hybridization is to hybridize all the wild type or "healthy" DNA fragments (pink) with an excess of mutant DNA (purple). In this example, the mutant DNA is digested with restriction enzyme 1 and the wild type DNA is digested with restriction enzyme 2. Both samples are heated to separate the strands, forming a pool of single-stranded fragments. In order to ensure all the wild type DNA is hybridized to mutant DNA and not to itself, a large surplus of mutant DNA is mixed with a small amount of wild type DNA. The DNA is allowed to anneal yielding double-stranded DNA consisting of a mixture of mutant : mutant, mutant : wild type, and rare wild type : wild type molecules. Since the ratio of wild type DNA to mutant DNA was so low, theoretically the only molecules with two wild type strands should be those containing DNA that is missing from the mutant sample—i.e. the gene of interest. Because two different restriction enzymes were originally used to digest the different samples of DNA, these desired DNA molecules are the only ones that can be digested with restriction enzyme 2. This allows them to be cloned and captured.

described above. The standard sample will contain mRNA from genes expressed under normal nutrient conditions. The experimental sample will contain mRNA from genes only expressed when nutrients are limited. Limited nutrients may stimulate cells to manufacture their own nutrients, thus some mRNAs would be produced in a higher abundance than the other sample.

Subtractive hybridization can also be performed with two samples of mRNA from the same organism grown under different conditions.

The basic idea is that the standard mRNA is used to subtract out the corresponding mRNA molecules from the experimental sample. However, two mRNA molecules of the same sequence obviously cannot hybridize together directly. Therefore, the standard mRNA is first converted to the corresponding double-stranded cDNA by reverse transcriptase. The cDNA is bound to a filter and the experimental sample of

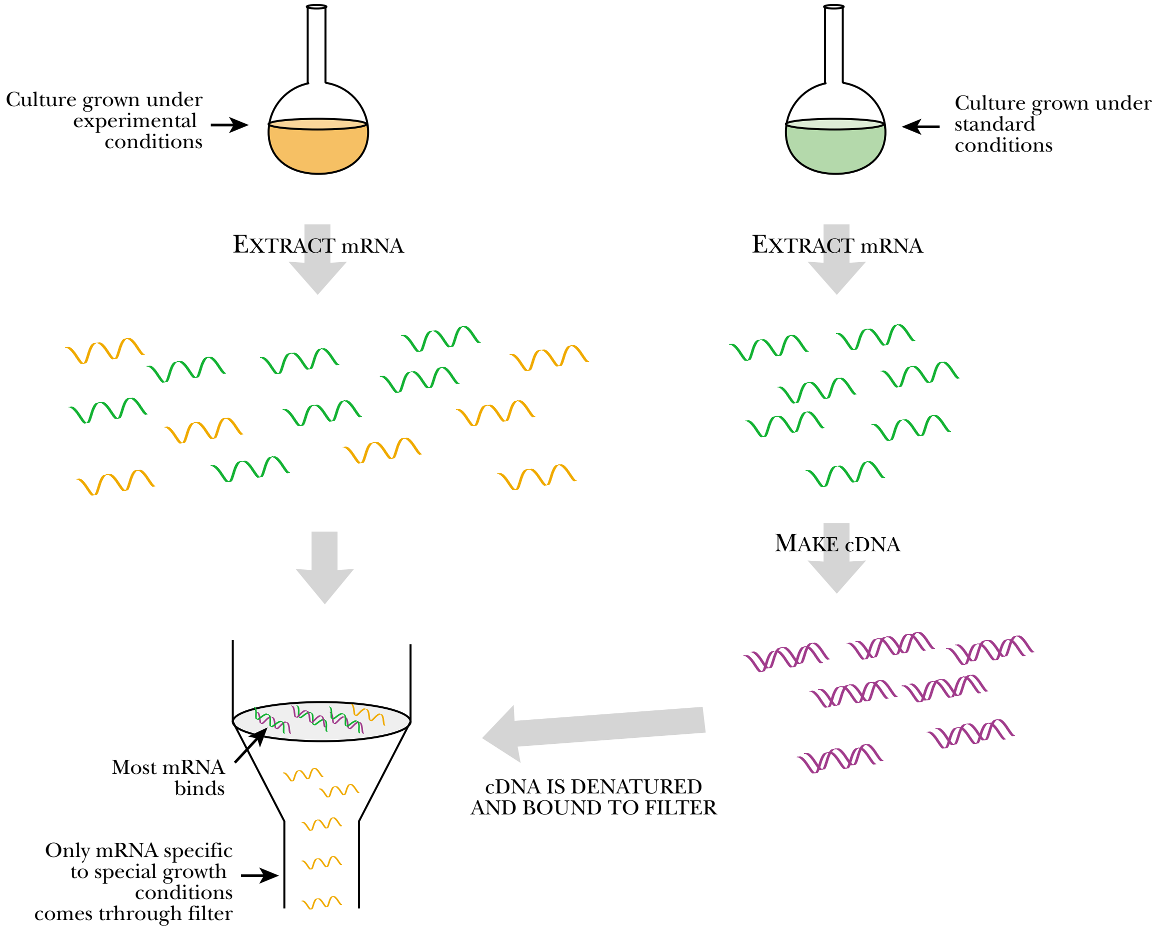

Culture grown under experimental conditions
Culture grown under standard conditions
EXTRACT mRNA
EXTRACT mRNA
MAKE cDNA
cDNA IS DENATURED AND BOUND TO FILTER
Most mRNA binds
Only mRNA specific to special growth conditions comes trhrough filter

FIGURE 22.30 ***Subtractive Hybridization Captures mRNA Expressed under Specific Conditions***

Two different cultures are grown, one under standard conditions (green) and one under experimental conditions (orange). The mRNA is isolated from each culture. To allow hybridization, one set of mRNA must be converted into double-stranded DNA by reverse transcriptase. The double-stranded cDNA is then denatured and bound to a filter. In this example, cDNA corresponding to the standard mRNA (green) is bound to the filter. The experimental mRNA (orange) is passed through the filter, where it binds to complementary single-stranded DNA from the standard conditions. If a gene is expressed highly under experimental conditions but not expressed (or present only in low amounts) during standard conditions, its mRNA will not be bound as no corresponding cDNA will be present on the filter. In practice, the mRNA that does not hybridize is pooled and then rehybridized to cDNA from the standard conditions. Repeating this step ensures that all the isolated mRNA is truly absent from the standard sample.

mRNA is passed through (Fig. 22.30). Messenger RNA corresponding to genes in the cDNA is retained by hybridization. Only the mRNA from genes that are expressed under the specific conditions of interest remains unbound and passes through the filter. This, in turn can now be converted to cDNA so giving a sample of those genes expressed under the particular conditions chosen.

Expression Vectors

Vectors may carry promoters and ribosome binding sites to mediate the expression of cloned genes.

Once a gene has been cloned into a vector it may or may not be expressed. If both structural gene and promoter were cloned on the same segment of DNA the gene may well be expressed. On the other hand, if only the structural gene was cloned then expression will depend on whether a promoter is provided by the plasmid. Vectors that

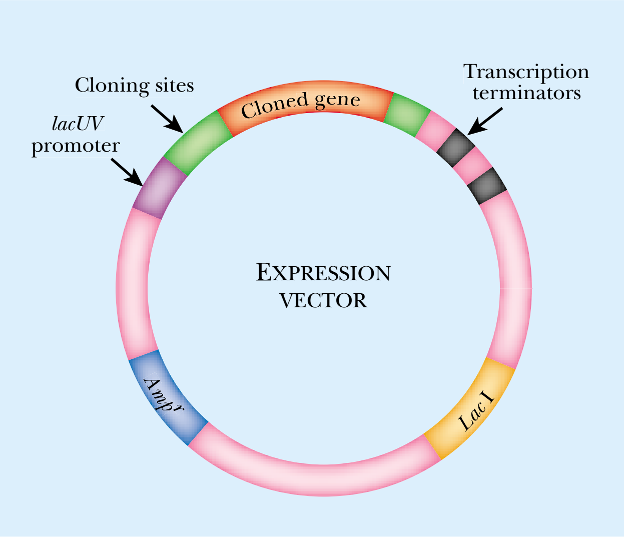


FIGURE 22.31 ***Expression Vectors Can Have Tightly Regulated Promoters***

An expression vector contains sequences upstream of the cloned gene that control transcription and translation of the cloned gene. The expression vector shown uses the *lac* promoter, which is very strong, but inducible. To stimulate transcription, an artificial inducer molecule called IPTG is added. IPTG binds to the LacI repressor protein which then detaches from the DNA. This allows RNA polymerase to bind. Before IPTG is added to the culture, the LacI repressor prevents the cloned gene from being expressed.

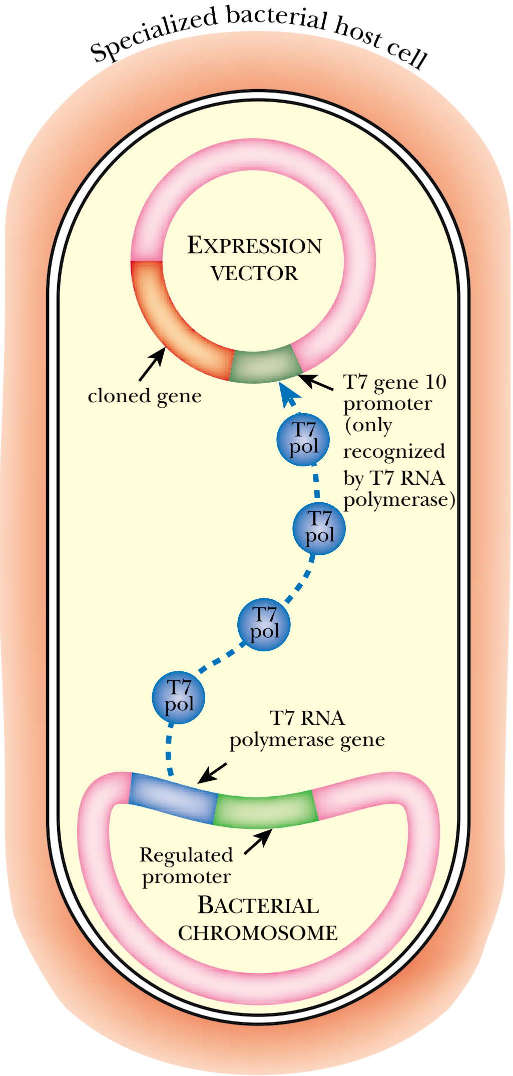


FIGURE 22.32 ***T7 RNA Polymerase System***

Specialized promoters can be used to control the expression of cloned genes. In the T7 RNA polymerase system, the cloned gene cannot be expressed unless the bacterial cell makes T7 RNA polymerase. The polymerase is produced by certain genetically engineered bacteria, which have the gene encoding it inserted into the chromosome. Expression of the T7 RNA polymerase gene is under control of the *lac* promoter, as described in the previous figure.

use blue and white screening (see above) place the cloned gene under control of the *lac* promoter, which lies upstream of the multiple cloning site.

Often, the objective of cloning a gene is to isolate high levels of the encoded protein. Purification of proteins has long been complicated because each protein folds up in an individualized manner and consequently behaves differently. To get around this problem the target protein is often tagged with another peptide that is easy to detect and/or purify. This allows purification and manipulation of many different proteins by the same procedures. Tagging is generally done at the genetic level—that is, an extra segment of DNA that codes for the tag is inserted beside the DNA coding for the target protein. This topic is discussed in detail in Ch. 26, Proteomics. For now it should be remembered that when we discuss expression of the "cloned gene", this will in practice often include extra sequences that specify tags or alter regulation to facilitate later analysis.

It is often helpful to deliberately control or enhance expression of a cloned gene, especially if high levels of the encoded protein are needed. **Expression vectors** are specifically designed to place the cloned gene under control of a plasmid-borne promoter. In practice the gene under investigation is normally first cloned in a general cloning vector and then transferred to the expression vector.

A variety of expression vectors exist with different promoters. The two basic alternatives are very strong promoters and tightly regulated promoters. Strong promoters are used when high levels of the gene product are required. Tightly regulated promoters are useful in physiological experiments where the effects of gene expression are to be tested under a variety of conditions.

Strong promoters are used to express high levels of proteins from cloned genes.

Some promoters are both strong and strictly regulated. These are useful when expressing large amounts of a foreign protein in a bacterial cell. Even if the foreign proteins are not actually toxic, the large amounts produced interfere with bacterial growth. Consequently, the bacteria are allowed to grow for a while before the foreign gene is turned on by addition of inducer. The bacteria then devote themselves to manufacture of the foreign protein.

The *lac* promoter of *E. coli* is inducible and certain mutant versions exist that are extremely strong promoters, such as the *lacUV* promoter. IPTG is an artificial inducer that turns on the *lac* promoter (see Ch. 9). However, repression by LacI, the *lac* repressor, is leaky—i.e. incomplete. Including the *lacI* gene on a multicopy cloning vector results in high levels of repressor, which turn off the cloned gene more effectively (Fig. 22.31).

Another strictly regulated promoter is the **lambda left promoter, P_L**. The **lambda repressor** or **cI protein** represses this promoter. If host cells contain a temperature sensitive version of the ***cI* gene**, such as *cI857*, then raising the temperature can alleviate repression. At 30°C the repressor is functional but at 42°C the repressor is inactivated.

Strong virus promoters, such as lambda or T7 promoters are useful for controlling the expression of cloned genes.

A third popular method is to place the gene under control by a strong promoter from **bacteriophage T7**. Such promoters are not recognized at all by bacterial RNA polymerase but only by T7 RNA polymerase. Transcription will only occur in specialized host cells that contain the gene for T7 RNA polymerase. Another regulated promoter, such as the lac promoter, in turn controls the expression of T7 RNA polymerase. Induction of the *lac* promoter induces synthesis of T7 RNA polymerase, which in turn, transcribes the cloned gene (Fig. 22.32). This provides both strict regulation and high-level expression.

bacteriophage T7 A bacteriophage that infects *E. coli* whose promoters are only recognized by its own RNA polymerase
***cI* gene** Gene encoding the lambda repressor or cI protein
cI protein Lambda repressor protein responsible for maintaining bacteriophage lambda in the lysogenic state
expression vector Vector specifically designed to place a cloned gene under control of a plasmid-borne promoter
lambda left promoter (P_L) One of the promoters repressed by binding of the lambda repressor or cI protein
lambda repressor (cI protein) Repressor protein responsible for maintaining bacteriophage lambda in the lysogenic state

CHAPTER TWENTY-THREE

The Polymerase Chain Reaction

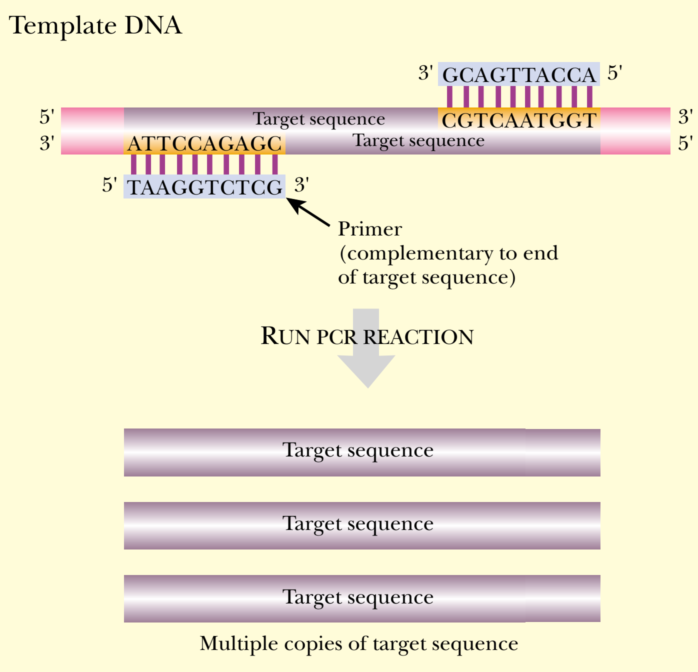


FIGURE 23.01 ***The Polymerase Chain Reaction (PCR)***

During PCR, two primers anneal to complementary sequences at either end of a target sequence on a piece of denatured template DNA. DNA polymerase synthesizes DNA, elongates the primers and makes two new strands of DNA, thus duplicating the original target sequence. In further cycles, the newly made DNA molecules are denatured in turn and duplicated by the same sequence of events, resulting in multiple copies of the original target sequence.

Fundamentals of the Polymerase Chain Reaction

> The PCR allows trace amounts of a DNA sequence to be amplified giving enough DNA for cloning, sequencing or other analyses.

Of all the technical advances in modern molecular biology, the **polymerase chain reaction (PCR)** is one of the most useful. The PCR provides a means of amplifying DNA sequences. Starting with incredibly tiny amounts of any particular DNA molecule, the PCR can be used to generate microgram quantities of DNA. PCR is sufficiently sensitive that it can amplify the DNA from a single cell into amounts sufficient for cloning or sequencing. Consequently, PCR is used in clinical diagnosis, genetic analysis, genetic engineering and forensic analysis. In particular, PCR has revolutionized and speeded up the whole area of recombinant DNA technology. Previously, cloned DNA was made by growing up bacterial cultures and extracting and purifying the DNA. PCR allows the rapid generation of large amounts of specific DNA sequences that are easier to purify and less damaged. In this chapter we will examine how the polymerase chain reaction works. As the name indicates, DNA polymerase is used to manufacture DNA using a pre-existing DNA molecule as template. Each new DNA molecule synthesized becomes a template for generating more, thus creating a chain reaction. The PCR actually amplifies only a chosen segment (the **target sequence**) within the original DNA template, not the whole template DNA molecule (Fig. 23.01).

The components involved in the polymerase chain reaction are as follows:

1. The original DNA molecule that is to be copied is called the template and the segment of it that will actually be amplified is known as the target sequence. A trace amount of the DNA template is sufficient.
2. Two **PCR primers** are needed to initiate DNA synthesis. These are short pieces of single-stranded DNA that match the sequences at either end of the target DNA segment. PCR primers are made by chemical synthesis of DNA as described in Ch. 21.

polymerase chain reaction (PCR) Amplification of a DNA sequence by repeated cycles of strand separation and replication
PCR primers Short pieces of single-stranded DNA that match the sequences at either end of the target DNA segment and which are needed to initiate DNA synthesis in PCR
target sequence Sequence within the original DNA template that is amplified in a PCR reaction

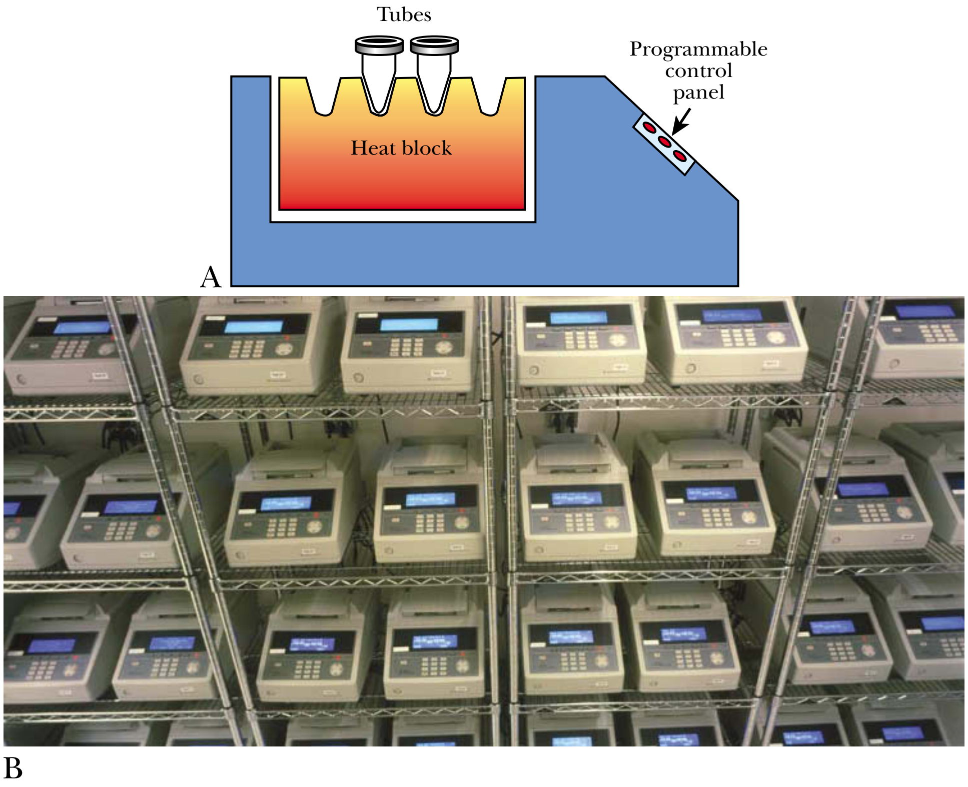


FIGURE 23.02 *PCR Machine or Thermocycler*

(A) The thermocycler or PCR machine can be programmed to change temperature rapidly. The heat block typically changes from a high temperature such as 90°C (for denaturation) to 50°C (for primer annealing), then back to 70°C (for DNA elongation) in a matter of minutes. This may be repeated for many cycles. (B) Rows of GeneAmp PCR machines copying human DNA at the Joint Genome Institute, in Walnut Creek, California, a collaboration between three of the US Department of Energy's National Laboratories. Credit: David Parker, Science Photo Library.

3. The enzyme DNA polymerase is needed to manufacture the DNA copies. The PCR procedure involves several high temperature steps so a heat resistant DNA polymerase is required. This came originally from heat resistant bacteria living in hot springs at temperatures up to 90°C. **Taq polymerase** from ***Thermus aquaticus*** is most widely used.
4. A supply of nucleotides is needed by the polymerase to make the new DNA. These are supplied as the nuceoside triphosphates.
5. Finally we need a **PCR machine** to keep changing the temperature (Fig. 23.02). The PCR process requires cycling through several different temperatures. Because of this, PCR machines are sometimes called **thermocyclers**.

The requirement for primers means that some knowledge of the sequence of the DNA template is needed. As described in Chapter 24, ever-increasing quantity of genomic DNA sequences is now available. Unknown sequences are dealt with in a variety of ways (for some specialized approaches see below). However, since binding of a primer need not be perfect, related sequences can often be used successfully, especially if longer primers are used.

PCR needs primers to start DNA synthesis which means that we must know some DNA sequence in or close to the region of interest.

When Kary Mullis invented PCR in 1987, he used normal DNA polymerase. Since the temperature needed to separate DNA into single strands destroys this enzyme, he had to add a fresh dose of polymerase to each tube every cycle! Luckily, heat resistant DNA polymerase was purified from *Thermus aquaticus* just a year or two later. Taq polymerase can be added to the reaction mixture at the beginning and survives all of the heating steps. It actually requires a high temperature to manufacture new DNA.

PCR machine See thermocycler
Taq polymerase Heat resistant DNA polymerase from *Thermus aquaticus* that is used for PCR
thermocycler Machine used to rapidly shift samples between several temperatures in a pre-set order (for PCR)
Thermus aquaticus Thermophilic bacterium found in hot springs and used as a source of thermostable DNA polymerase

Kary Mullis Invents PCR after a Vision

"Science, like nothing else among the institutions of mankind, grows like a weed every year. Art is subject to arbitrary fashion, religion is inwardly focused and driven only to sustain itself, law shuttles between freeing us and enslaving us."—Kary Mullis

Kary Mullis won the Nobel Prize in Chemistry in 1993 for developing the polymerase chain reaction (PCR). The PCR is one of modern biology's most useful techniques and has been used in virtually every area of molecular biology and biotechnology. Kary Mullis is one of sciences true eccentrics. In addition to molecular biology he has also contributed to other areas of science. While a doctoral candidate working on bacterial iron transport, he published an article entitled "The Cosmological Significance of Time Reversal" (*Nature 218*:663 (1968)), which deals with his notion that about half of the mass in the universe is going backward in time.

Kary Mullis invented PCR while working as a scientist for the Cetus Corporation. He conceived the idea while cruising in a Honda Civic on Highway 128 from San Francisco to Mendocino in April 1983. Mullis recalls seeing the polymerase chain reaction as clear as if it were up on a blackboard in his head. In lurid pink and blue. He pulled over and started scribbling. One basic ingredient of the PCR is that it amplifies DNA by constant repetition—rather like the computer programs Mullis was then involved in writing. Kary Mullis was given a $10,000 bonus by Cetus, who at first failed to realize the significance of the discovery. Later they sold the technology to Roche for $300,000,000.

In 1999, Kary Mullis mentioned the computer DNA connection again, "It is interesting that biochemistry developed alongside computers. If computers had not come along at about the same time as the structure of DNA was discovered, there would be no biochemistry. You always needed the computer to process the information. Without it we would have rooms and rooms full of monks writing out the sequences."

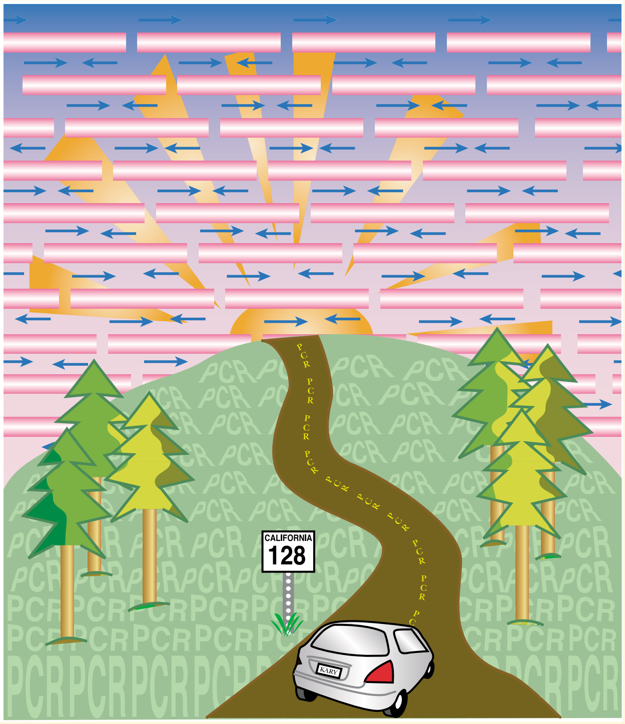


FIGURE 23.03 ***Kary Mullis Sees PCR in a Vision***

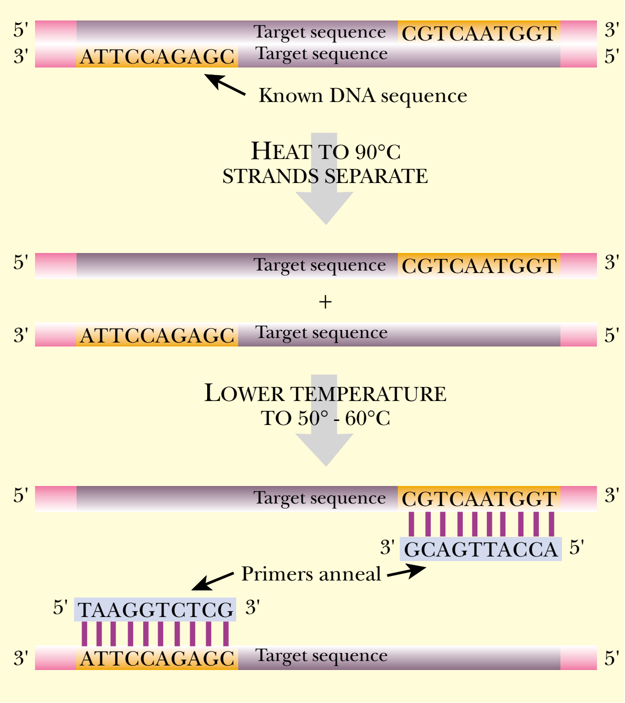


FIGURE 23.04 ***Denaturing the Template and Binding the Primers***

In the steps of PCR, a very small amount of template DNA is heated to 90°C, which separates the two strands of the double helix. When the temperature is lowered to 50–60°C, the primers can anneal to the ends of the target sequence. Since the primer is present in large excess over the template DNA, essentially all template strands will bind to primers rather than re-annealing to each other.

Cycling Through the PCR

PCR is a procedure involving multiple cycles of DNA strand separation, binding of primers, and synthesis of new DNA.

The first step of the PCR is to separate the strands of the template DNA by heating the template DNA to 90°C or so for a minute or two. Although the primers are present from the beginning, they cannot bind to the template DNA at 90°C. So, the temperature is dropped to around 50°C to 60°C, allowing the primers to anneal to the complementary sequences on the template strands (Fig. 23.04). Although the illustration shows 10 base primers, in real life they would be longer, say 15 to 20 bases. A longer primer is more specific for binding to the exact target sequence.

Next, the temperature is maintained at 70°C for a minute or two to allow the thermostable polymerase to elongate new DNA strands starting from the primers (Fig. 23.05). Remember that DNA polymerase cannot initiate the synthesis of a new strand, but can only elongate. In a living cell, RNA primers are used (Ch. 5) but artificial primers of single-stranded DNA are perfectly acceptable *in vitro*. Note that DNA synthesis goes from 5′ to 3′ for both new strands. This gives two partly double stranded pieces of DNA. Notice that the two new strands are not as long as the original templates. They are each missing a piece at the end where synthesis started. However, they are double-stranded over the region that matters, the target sequence. The cycle of events is then repeated many times.

The second cycle is shown in Figure 23.06. There are now four partly double-stranded pieces of DNA. Note again that although they vary in length, they all include double-stranded DNA from the target region.

As the cycles continue, the single strand overhangs are ignored and are rapidly outnumbered by segments of DNA containing only the target sequence. During the third cycle (Fig. 23.07), the first two pieces of double-stranded DNA that correspond exactly to the target sequence are made. These do not have any dangling single-stranded ends. Once past the first two or three cycles, the vast majority of the product is double-stranded target sequence with flush ends. Finally, the DNA generated is run on an agarose gel to assess the size of the PCR fragment.

FIGURE 23.05 ***Elongation of New Strand by Taq Polymerase***

Once the primers have annealed to the template, the temperature is increased to 70°C. This is the optimum temperature for the thermostable Taq polymerase to elongate DNA. The polymerase synthesizes new strands of DNA using the 3′ end of the primer as a starting point. A pool of nucleotide precursors is also necessary for this step of the reaction.

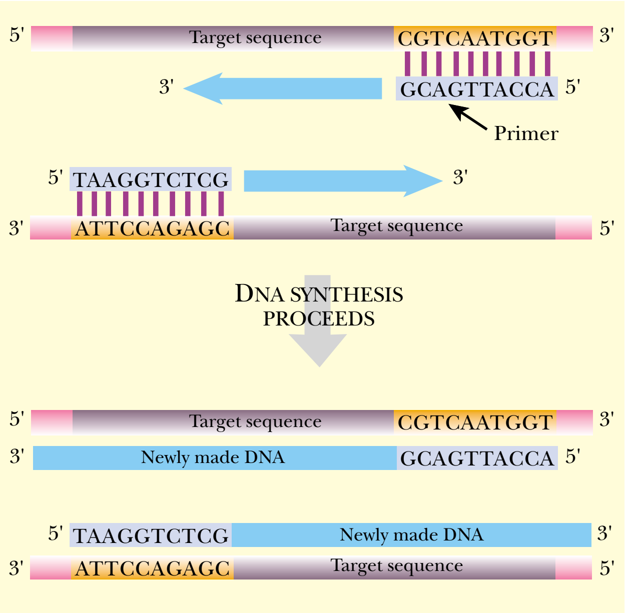
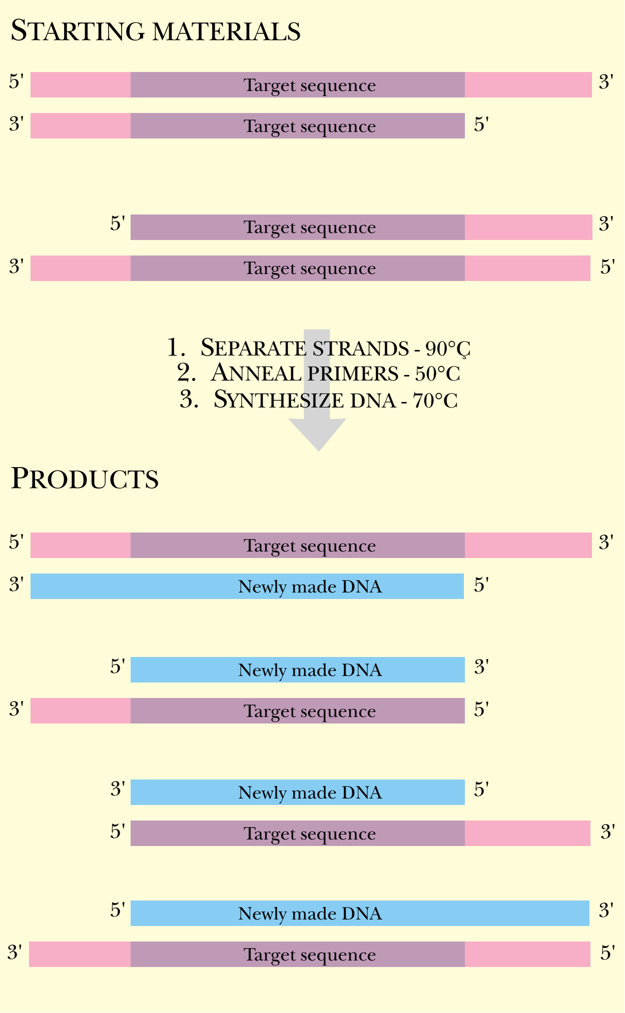


FIGURE 23.06 ***The Second Cycle of the PCR***

The entire cycle is repeated starting with the two DNA pieces produced in the first cycle. The two double-stranded pieces of DNA are denatured into four single-stranded pieces at 90°C. The temperature drops to 50°C in order for the primers to anneal. Finally, the polymerase extends the DNA at 70°C to convert the four single-stranded templates into double-stranded DNA.



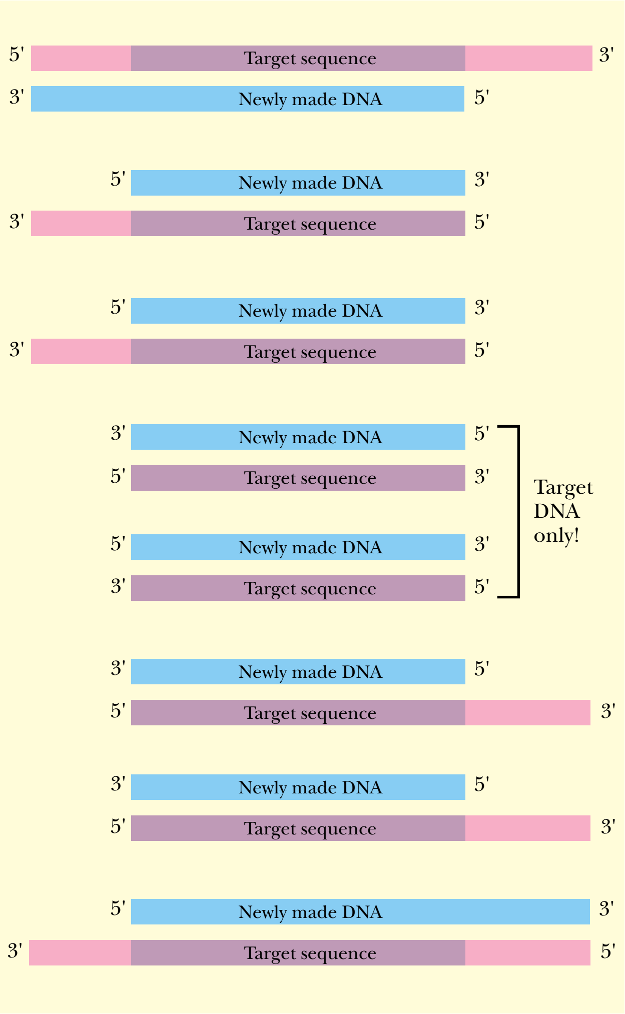


FIGURE 23.07 ***The Third Cycle of the PCR***

The products from the second cycle go through the same process as before. The four double-stranded pieces are denatured into eight single-stranded pieces. The primers anneal and DNA polymerase makes the complementary strands. After this cycle, the number of sequences containing only the target DNA grows exponentially, far exceeding any other product shown in this figure.

Degenerate Primers

The major problem with PCR is obvious. In order to make the PCR primers, some sequence information is required, at least at the ends of the target sequence. **Degenerate primers** are used when partial sequence information is available, but the complete sequence is unknown. For example, we may possess the sequence for a gene from one organism and be interested in obtaining the corresponding gene from another organism. If two organisms are related, their DNA sequences for a particular gene will be close, although rarely identical. Furthermore, the genetic code is degenerate and several codons can encode the same amino acid (see Ch. 8). In particular, many codon families share the first two bases and vary only in the third position. Since the sequence of the protein, rather than the DNA, is most important for function, most of the variation between closely related genes is in the third codon position.

Therefore, degenerate or redundant DNA primers are made that have a mixture of all possible bases in every third position. A degenerate primer is actually a mixture of closely related primers (Fig. 23.08). Presumably one of the primers in this mixture will recognize the DNA of the gene of interest. In addition, a perfect match is not really necessary. If, say, 18 or 19 of 20 bases pair up, a primer will work quite well. Many segments of DNA have been amplified successfully by PCR using sequence data from close relatives.

If only protein sequences are available, degenerate primers are used for PCR.

degenerate primer Primer with several alternative bases at certain positions

```
Partial sequence of polypeptide:

    Met--Tyr--Cys--Asn--Thr--Arg--Pro--Gly

Possible codons in DNA:

    ATG  TAC  TGT  AAT  ACT  AGA  GCT  GGT
         TAT  TGC  AAC  ACC  AGG  GCC  GGC
                        ACA       GCA  GGA
                        ACG       GCG  GGG

Corresponding redundant primer:

    ATG  TAC  TGT  AAT  ACT  AGA  GCT  GGT
           T    C    C    C    G    C    C
                          A         A    A
                          G         G    G


Bases in the third codon position are shown in red.
The redundant primer consists of a mixture of
primers with these bases varied as shown.
```

FIGURE 23.08 ***Degenerate DNA Primers***

Degenerate primers are used if only partial DNA sequence information is available. Often, as here, a short amino acid sequence from a protein is known. Because many amino acids are encoded by several alternative codons, the deduced DNA coding sequence is ambiguous. For example, the amino acid tyrosine is encoded by TAC or TAT. Hence the third base is ambiguous and when the primer is synthesized a 50 : 50 mixture of C and T will be inserted at this position. This ambiguity occurs for all the bases shown in red, resulting in a pool of primers with different, but related sequences. Hopefully, one of these primers will have enough complementary bases to anneal to the target sequence that is to be amplified.

Degenerate DNA primers must also be used if only a protein sequence is available. In this case, the protein sequence is translated backwards to give the corresponding DNA sequence. Due to the degeneracy of the genetic code, several possibilities will exist for the sequence of DNA that corresponds to any particular polypeptide sequence. Again, most of the ambiguity is in the third codon position. This ambiguous sequence may be used to make degenerate primers, as before (Fig. 23.08). Although proteins are rarely sequenced in their entirety nowadays, short stretches of N-terminal sequence are often obtained. After separating proteins, automated N-terminal sequencing may yield a sequence of the first dozen or more amino acids. This is often sufficient to allow design of a degenerate probe for hybridization screening of a gene library (see Ch. 22) or degenerate primers for PCR.

Inverse PCR

Another approach that uses incomplete sequence information to amplify a target gene is **inverse PCR**. In this case a sequence of part of a long DNA molecule, say a chromosome, is known. The objective is to extend the analysis along the DNA molecule into the unknown regions. To synthesize the primers for PCR, the unknown target sequence must be flanked by two regions of known sequence. The present situation is exactly the opposite of that. To circumvent this problem, the target molecule of DNA is converted into a circle. Going around a circle brings you back to the beginning. In effect, even though, only one small stretch of sequence is known, the circular form allows you to have that one region on both sides of the target sequence.

Performing PCR on a circularized DNA template allows access to neighboring regions of unknown sequence.

A restriction enzyme, usually one that recognizes a six-base sequence, is used to make the circle. This enzyme must not cut into the known sequence, therefore, eventually, this enzyme will cut either upstream or downstream from the known region.

inverse PCR Method for using PCR to amplify unknown sequences by circularizing the template molecule

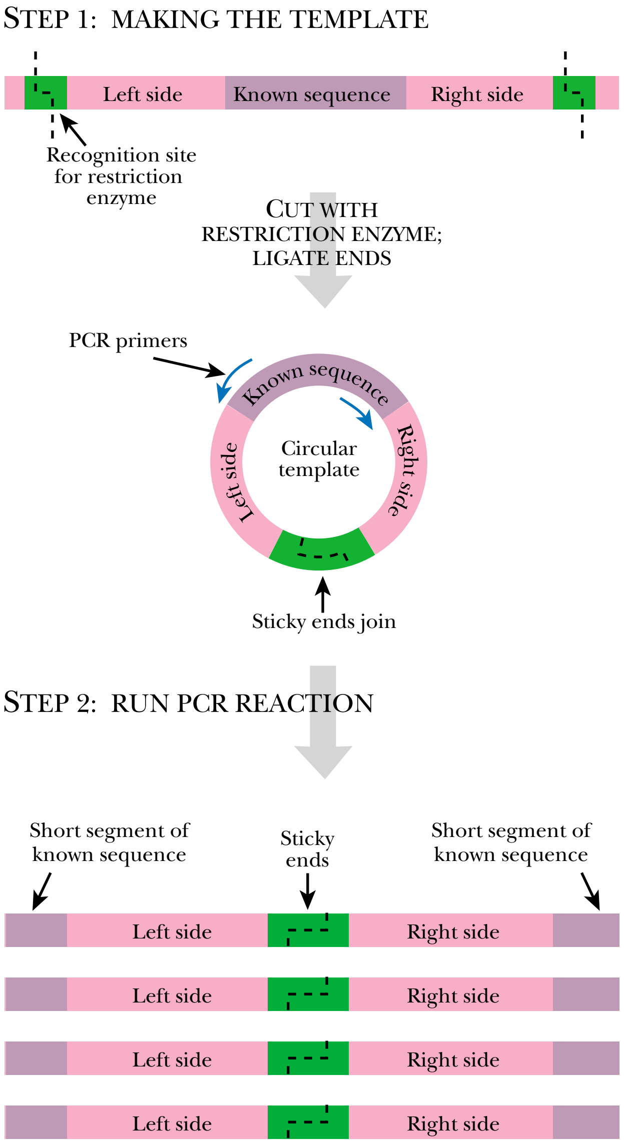


FIGURE 23.09 ***Inverse PCR***

Inverse PCR allows unknown sequences to be amplified by PCR provided that they are located next to DNA whose sequence is already known. The DNA is cut with a restriction enzyme that does not cut within the region of known sequence, as shown in Step I. This generates a fragment of DNA containing the known sequence flanked by two regions of unknown sequence. Since the fragment has two matching sticky ends, it may be easily circularized by DNA ligase. Finally, PCR is performed on the circular fragments of DNA (Step 2). Two primers are used that face outwards from the known DNA sequence. PCR amplification gives a single linear product that includes unknown DNA from both left and right sides. This PCR product can now be cloned and/or sequenced.

The resulting fragment will have unknown sequence first, the known sequence in the middle, followed by more unknown sequence. The two ends of the fragment will have compatible sticky ends that are easily ligated together to make a circle of DNA (Fig. 23.09). Two primers corresponding to the known region and facing outwards around the circle are used for PCR. Synthesis of new DNA will proceed around the circle clockwise from one primer and counter-clockwise from the other. Overall, inverse PCR gives multiple copies of a segment of DNA containing some DNA to the right and some DNA to the left of the original known region.

Adding Artificial Restriction Sites

Artificial restriction sites are often added to the ends of PCR products to aid in cloning.

Once a segment of DNA has been amplified by PCR it may be sequenced (see Ch. 24) or cloned. For cloning it is often convenient to use restriction enzymes to generate sticky ends on both insert and vector (see Ch. 22). However, it is unlikely that such sites will be located just at the ends of any particular target sequence. One way to create convenient restriction cut sites at the end of PCR fragments is to incorporate them into the primers. When designing the primers, artificial

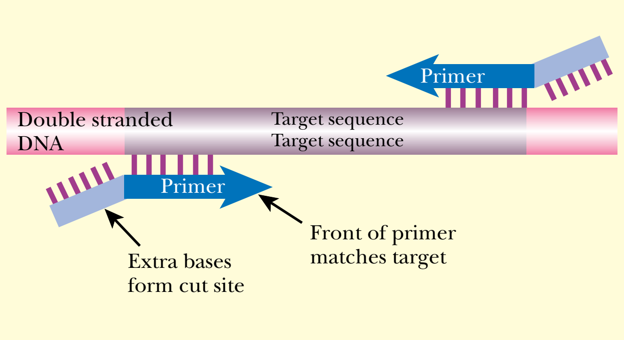


FIGURE 23.10 ***Incorporation of Artificial Restriction Sites***

Primers for PCR can be designed to have non-homologous regions at the 5′ end that contain the recognition sequence for a particular restriction enzyme. After PCR, the amplified product has the restriction enzyme site at both ends. If the PCR product is digested with the restriction enzyme, this generates sticky ends that are compatible with a chosen vector.

restriction enzyme recognition sites are added at the far ends of the primers (Fig. 23.10). As long as the primer has enough bases to match its target site, adding a few extra bases at the end will not affect the PCR reaction. The bases making up the restriction site get copied and appear on the ends of all newly manufactured segments of DNA. After the PCR reaction has been run, the PCR fragment is cut with the chosen restriction enzyme to generate sticky ends. The fragment is then cloned into a convenient plasmid.

TA Cloning by PCR

Single base overhangs may be used to clone PCR products.

Another approach to cloning PCR products takes advantage of the properties of the Taq polymerase itself. This enzyme adds a single adenosine to the 3′-ends of double-stranded DNA. This reaction does not depend on the sequence of the template or primers. The **TA cloning** procedure exploits this terminal transferase activity, which is shared by Taq polymerase and several other thermophilic DNA polymerases. Thus most of the DNA molecules amplified by Taq polymerase possess single 3′-A overhangs (Fig. 23.11). Consequently, they can be directly cloned into a vector that has matching single 3′-T overhangs on both ends (after being linearized). The same **TA cloning vector** can be used to clone any segment of amplified DNA. For that matter, DNA from other sources can have single 3′-A overhangs added to its ends by using Taq polymerase and can then be cloned by the same mechanism. This procedure is especially useful when convenient restriction sites are not available.

Randomly Amplified Polymorphic DNA (RAPD)

Randomly Amplified Polymorphic DNA, or RAPD, is usually found in the plural as RAPDs and is pronounced "rapids," partly because it is a quick way to get a lot of information about the genes of an organism under investigation. The purpose of RAPDs is to test how closely related two organisms are. In practice, DNA samples from unknown organisms are compared with DNA from a previously characterized organism. For example, traces of blood from a crime scene may be compared to

randomly amplified polymorphic DNA (RAPD) Method for testing genetic relatedness using PCR to amplify arbitrarily chosen sequences
TA cloning Procedure that uses Taq polymerase to generate single 3′-A overhangs on the ends of DNA segments that are used to clone DNA into a vector with matching 3′-T overhangs
TA cloning vector Vector with single 3′-T overhangs (in its linearized form) that is used to clone DNA segments with single 3′-A overhangs generated by Taq polymerase

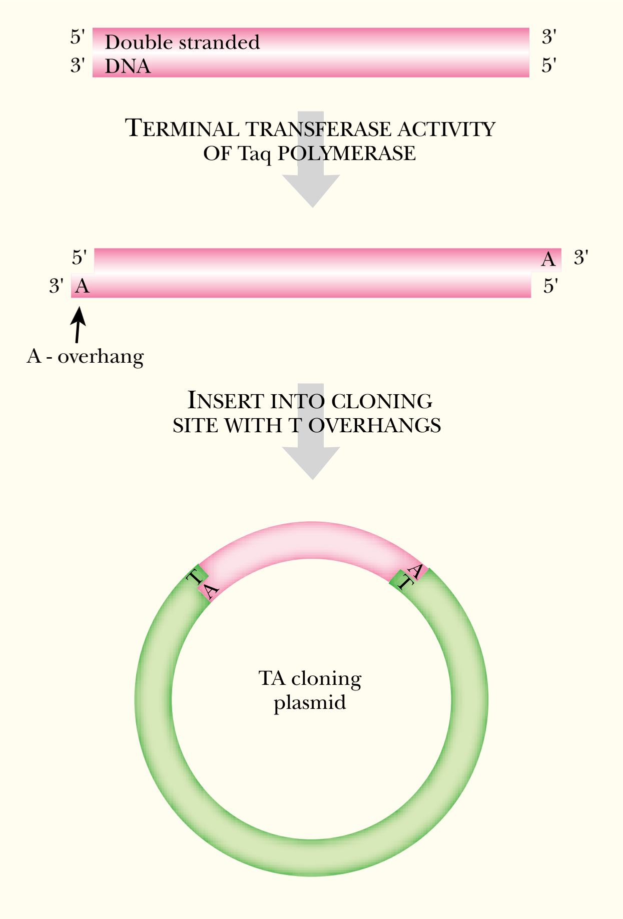


FIGURE 23.11 ***TA Cloning***

When Taq polymerase amplifies a piece of DNA during PCR, the terminal transferase activity of Taq adds an extra adenine at the 3′ end of the PCR product. The TA cloning vector was designed so that when linearized it has single 5′ thymidine overhangs. The PCR product can be ligated into this vector without the need for special restriction enzyme sites.

possible suspects, or disease-causing microorganisms may be related to known pathogens to help trace an epidemic.

PCR may be performed with arbitrary primers. Comparing results from two samples of DNA reveals their relatedness.

The principle of RAPDs is statistically based. Given any particular five-base sequence, such as ACCGA, how often will this exact sequence appear in any random length of DNA? Since there are four different bases to choose from, one in every 4^5 (or $4 \times 4 \times 4 \times 4 \times 4 = 1{,}024$) stretches of five bases will—on average—be the chosen sequence. Any arbitrarily chosen 11-base sequence will be found once in approximately every 4 million bases. This is approximately the amount of DNA in a bacterial cell. In other words, any chosen 11-base sequence is expected to occur by chance once only in the entire bacterial genome. For higher organisms, with much more DNA per cell, a longer sequence would be needed for uniqueness.

For RAPDs, the arbitrarily chosen sequence should be rare but not unique. PCR primers are made using the chosen sequence and a PCR reaction is run using the total DNA of the organism as a template. Every now and then a primer will find a correct match, purely by chance, on the template DNA (Fig. 23.12). For PCR amplification to occur there must be two such sites facing each other on opposite strands of the DNA. The sites must be no more than a few thousand bases apart for the reaction to work well. The likelihood of two correct matches in this arrangement is quite low.

In practice, the length of the primers is chosen to give five to 10 PCR products. For higher organisms, primers of around 10 bases are typical. The bands from PCR are separated by gel electrophoresis (see Ch. 21) to measure their sizes. The procedure is repeated several times with primers of different sequence. The result is a diagnostic

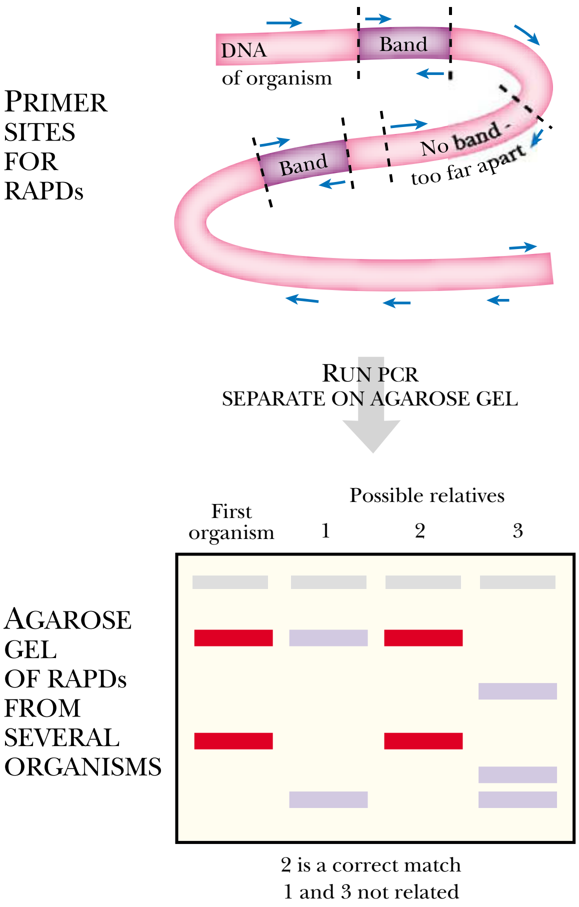


FIGURE 23.12 ***Randomly Amplified Polymorphic DNA***

The first step of RAPD analysis is to design primers that will bind to genomic DNA at random sites that are neither too rare or too common. In this example, the primers were sufficiently long to bind the genomic DNA at a dozen places. For PCR to be successful, two primers must anneal at sites facing each other but on opposite strands. In addition, these paired primer sites must be close enough to allow synthesis of a PCR fragment in a reasonable time. In our example, there are three pairs but only two of these pairs were close enough to actually make the PCR product. Consequently, this primer design will result in two PCR products as seen in the first lane of the gel (marked "First organism"). The same primers are then used to amplify genomic DNA from other organisms that are suspected of being related. In this example, suspect #2 shows the same banding pattern as the first organism and is presumably related. The other two suspects do not match the first organism and are therefore not related.

pattern of bands that will vary in different organisms, depending on how closely they are related. Although we do not know in which particular genes the PCR bands originate, this does not matter in measuring relatedness. Diagnosis therefore relies on having a primer (or set of primers) that reliably give a band of a particular size with the target organism and give different bands with other organisms, even those closely related. RAPD results using such a primer are shown in Figure 23.13. Grey mold, due to *Botrytis cinerea*, is one of the most destructive infections of strawberries and also attacks other plants. Classical diagnosis involves culturing the fungus on nutrient agar. It is slow and difficult due to the presence on the plants of other harmless fungi, which often grow faster in culture. As can be seen, RAPD analysis clearly identifies the pathogens from other related fungi, including other species from the genus *Botrytis*.

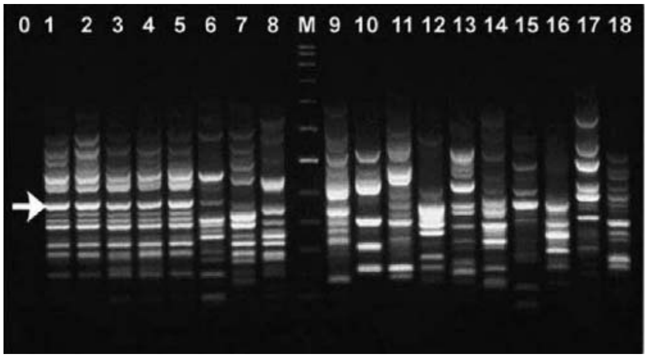


FIGURE 23.13 ***Identification of Fungal Pathogens by RAPD***

RAPD banding patterns generated using the 10 base primer, AACGCGCAAC, on genomic DNA of five closely related strains of the pathogen *Botrytis cinerea* (lanes 1–5), three other strains from the genus *Botrytis* (lanes 6, 7 & 8) and several less related harmless fungi *Alternaria* (9), *Aspergillus* (10), *Cladosporium* (11), *Epicoccum* (12), *Fusarium* (13), *Hainesia* (14), *Penicillium* (15), *Rhizoctonia* (16 & 17), and the host plant, strawberry (18). Lane 0: negative control (no DNA). Lane M: molecular mass marker. From: Rigotti et al., FEMS Microbiology Letters (2002) 209:169–174.

Reverse Transcriptase PCR

Reverse transcription followed by PCR allows cloning of genes starting from the messenger RNA.

The coding sequence of most eukaryotic genes is interrupted by intervening sequences, or introns (see Ch. 12 for introns and RNA processing). Consequently, the original version of a eukaryotic gene is very large, difficult to manipulate and virtually impossible to express in any other type of organism. Since mRNA has had the introns removed naturally, it may be used as the source of an uninterrupted coding sequence that is much more convenient for engineering and expression. This involves converting the RNA back into a DNA copy, known as **complementary DNA (cDNA)** by **reverse transcriptase** (see Ch. 22). Thus, when "cloning" eukaryotic genes the cDNA version is often used (rather than the true chromosomal gene sequence) as this lacks the introns. Once the cDNA has been made, PCR can be used to amplify the cDNA and generate multiple copies (Fig. 23.14). This combined procedure is referred to as **reverse transcriptase PCR (RT-PCR)** and allows genes to be amplified and cloned as intron-free DNA copies starting from mRNA.

RT-PCR has other uses. A specific mRNA molecule is only made when the gene for that protein is turned on and expressed. Therefore extraction and purification of the mRNA gives several mRNA copies of every gene that is being expressed under the particular growth conditions. RT-PCR can then performed on the mixture of mRNA using PCR primers that match some particular gene of interest. If this gene was expressed under the specific growth conditions, a PCR product will be produced, whereas, if the gene was switched off, none of this particular mRNA will be present and no band will be generated (Fig. 23.15). Carrying out RT-PCR on an organism under different growth conditions reveals when the gene under scrutiny was switched on. This allows analysis of which environmental factors bring about expression of any chosen gene.

complementary DNA (cDNA) Version of a gene that lacks the introns and is made from the corresponding mRNA by using reverse transcriptase
reverse transcriptase Enzyme that starts with RNA and makes a DNA copy of the genetic information
reverse transcriptase PCR (RT-PCR) Variant of PCR that allows genes to be amplified and cloned as intron-free DNA copies by starting with mRNA and using reverse transcriptase

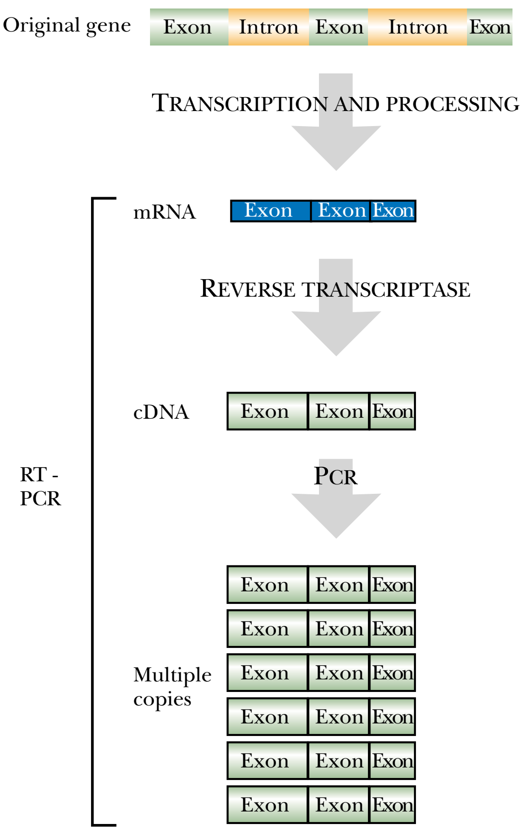


FIGURE 23.14 ***Reverse Transcriptase PCR***

RT-PCR is a two-step procedure that involves making a cDNA copy of the mRNA, then using PCR to amplify the cDNA. First, a sample of mRNA (which lacks introns) is isolated. Reverse transcriptase is used to make a cDNA copy of the mRNA. The cDNA sample then amplified by PCR. This yields multiple copies of cDNA without introns.

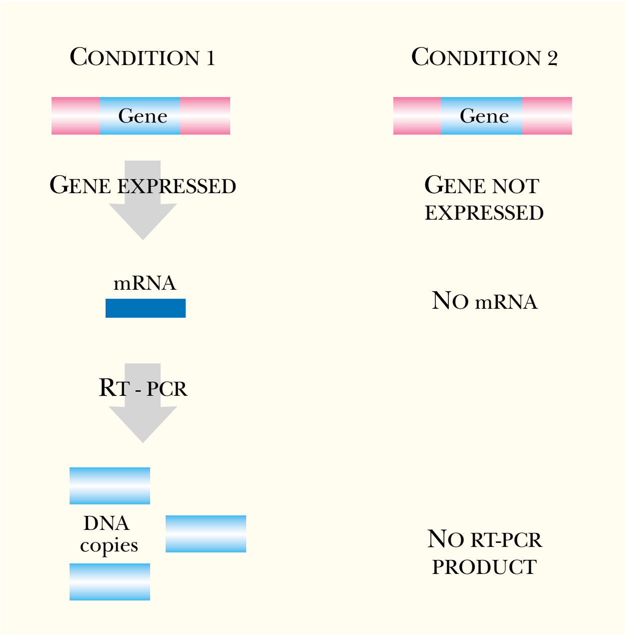


FIGURE 23.15 ***RT-PCR for Gene Expression***

RT-PCR can be used to determine whether or not mRNA corresponding to a particular gene is present. In other words, gene expression may be tested for an organism grown under two different conditions. In this example, the gene of interest is expressed in condition 1 but not in condition 2. Therefore in condition 1 mRNA from the gene of interest is present and reverse transcriptase generates the cDNA. The PCR primers specific for this gene can now bind to the cDNA and PCR will amplify a DNA band corresponding to the original mRNA. In condition 2 the mRNA is absent and so the RT-PCR procedure does not generate the corresponding DNA band.

The mixture of mRNA made by a cell can be surveyed by a PCR based approach.

Differential Display PCR

Differential display PCR is used to specifically amplify messenger RNA from eukaryotic cells. The technique is valuable because it allows the researcher to assess the expression of many different mRNA molecules simultaneously. This technique is a

differential display PCR Variant of RT-PCR that specifically amplifies messenger RNA from eukaryotic cells using oligo(dT) primers

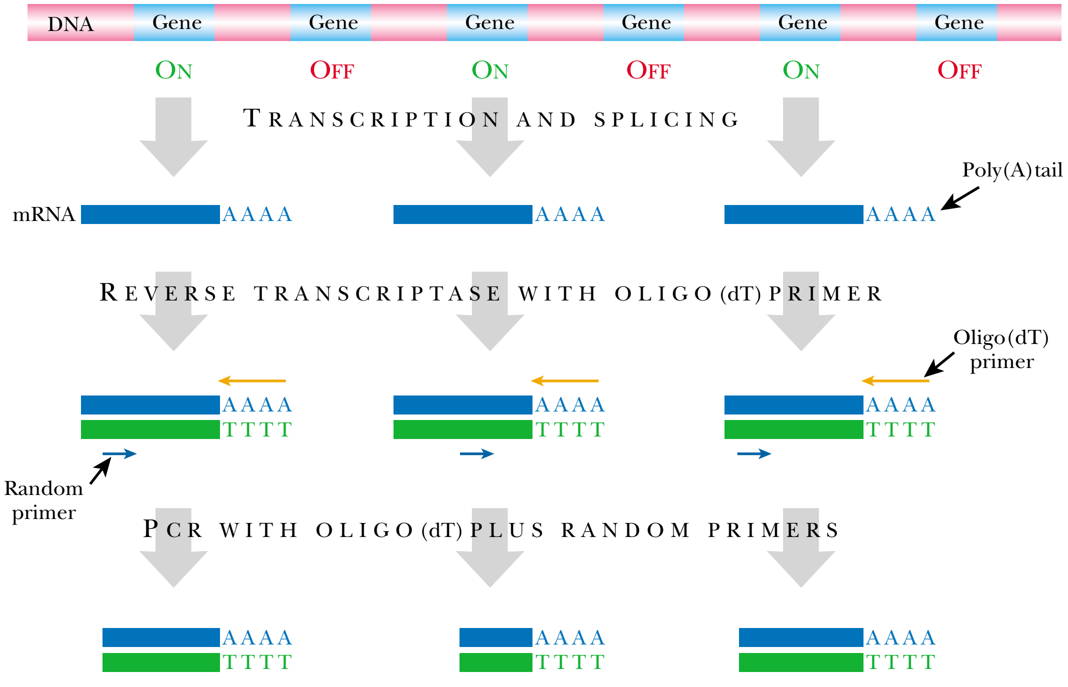


FIGURE 23.16 ***Differential Display PCR***

Differential Display PCR allows simultaneous measurement of the expression of many different mRNA molecules. The example shows that under the conditions used, three of the genes are turned on and three are turned off. Those mRNA molecules that are expressed are converted to cDNA by reverse transcriptase using an oligo(dT) primer, then amplified by PCR. The first PCR primer is oligo(dT) and so binds to the poly(A) sequences. The second PCR primer is a mixture of random sequences, calculated to anneal approximately once per cDNA. These primers ensure that many different cDNA molecules are amplified rather than just one. In this example, three PCR products are produced, corresponding to the original genes that are expressed.

combination of RAPD (see above) with RT-PCR and has one clever modification of its own, the use of oligo(dT) primers. Since almost all eukaryotic mRNA molecules have a 3′-tail of poly(A), an artificial primer made only of dT will base pair to this tail. This method of PCR allows the researcher to compare two different growth conditions on many different genes, rather than just one gene as in RT-PCR.

As in RT-PCR the RNA is extracted from the cells and the corresponding cDNA is made by the use of reverse transcriptase. Then a PCR reaction is run with two primers (Fig. 23.16):

1. An oligo(dT) primer that binds to the 3′ end of all cDNA copies of messenger RNA
2. Since the sequences at the other end of the mRNA molecules are unknown, the second primer is actually a mixture of random primers similar to those used in RAPDs.

These two primers ensure that there are not too many or too few amplified fragments.

The result is the amplification of many different DNA segments corresponding to each of the messenger RNA molecules in the original mixture. As usual, gel electrophoresis is used to separate the different components. This gives a series of DNA bands corresponding to each of the mRNAs being made in the cells that were analyzed. If the growth conditions are then altered, the pattern of DNA bands will change. In many cases multiple bands will appear, and multiple bands will disappear, thus allowing multiple random genes to be analyzed rather than just a single gene of interest as in RT-PCR.

Rapid Amplification of cDNA Ends (RACE)

Rescuing the "lost" ends of cloned genes may be done by a complex PCR based procedure.

Using only reverse transcriptase, full-length cDNA copies may be hard to get, especially from mRNA that is present only in very low amounts or unusually long. Reverse transcriptase often fails to reach the end of a long RNA template due to hindrance by RNA secondary structure. Thus the 5′-end is often incomplete. Consequently some means of recovering the complete cDNA is needed. The **RACE** technique generates the complete cDNA in two halves; hence the name **rapid amplification of cDNA ends**. It is necessary to know part of the internal sequence of the mRNA/cDNA in order to design the internal primers; therefore, the technique is generally used when an incomplete cDNA was isolated by other techniques such as library screening (see Ch. 22). The RACE procedure is essentially a modification of RT-PCR, but unique so-called **anchor sequences** are added to each end of the cDNA to facilitate the PCR portion of the reaction (Fig. 23.17).

The 3′-reaction of RACE-PCR primes reverse transcriptase to synthesize a DNA copy from the poly(A) tail of the mRNA by using an oligo(dT) primer that has a unique anchor sequence at the 5′ end. Since the internal sequence is known, an internal primer is designed so that PCR will amplify from the poly(A) tail to the middle of the gene. In the 5′-reaction, the internal primer is used to initiate DNA synthesis using reverse transcriptase. Next, an artificial poly(A) tail is added to the 3′ end of the DNA by terminal transferase and dATP. The same oligo(dT)/anchor primer as used to initiate the 3′-reaction is then used again during the PCR amplification cycles for the 5′-reaction. The anchor sequence primer and internal primers are generally designed to include convenient restriction sites to allow further cloning and sequencing.

PCR in Genetic Engineering

A whole plethora of modifications has been made to the basic PCR scheme. Some of these are used to alter genes rather than to analyze them. We can divide these modifications into two broad categories: a) rearranging large stretches of DNA and b) changing one or two bases of a DNA sequence. The latter is discussed below as "directed mutagenesis".

As already illustrated, we can amplify any segment of DNA provided we have primers that match its ends. Such segments of DNA may be joined or rearranged in a variety of ways. In order to make a hybrid gene, segments of two different genes must be amplified by PCR and then joined together. There are several protocols that vary in their details. But the crucial point is to use an **overlap primer** that matches part of both gene segments (Fig. 23.18).

Using overlap primers whose sequence matches the ends of two DNA molecules allows them to be fused end to end by PCR.

The PCR reaction is run using a primer for the front end of the first gene, a primer for the rear end of the second gene, and the overlap primer. The result is a hybrid gene. Some variants of this **"molecular sewing"** make the two halves separately and mix and join them later; other versions of this technique mix all three primers plus both templates in a single large reaction. By making hybrid genes using components from various sources it is sometimes possible to work out in detail which regions of a gene or protein are responsible for precisely which properties. The approach can also be used in biotechnology to construct artificial genes made up of modules from different sources.

anchor sequence Sequence added to primers or probes that may be used for binding to a support or may incorporate convenient restriction sites, primer binding sites for future manipulations, or primer bindings sites for subsequent PCR reactions
molecular sewing Creation of a hybrid gene by joining segments from multiple sources using PCR
overlap primer PCR primer that matches small regions of two different gene segments and is used in joining segments of DNA from different sources
RACE See rapid amplification of cDNA ends
rapid amplification of cDNA ends (RACE) RT-PCR-based technique that generates the complete 5′ or 3′ end of a cDNA sequence starting from a partial sequence

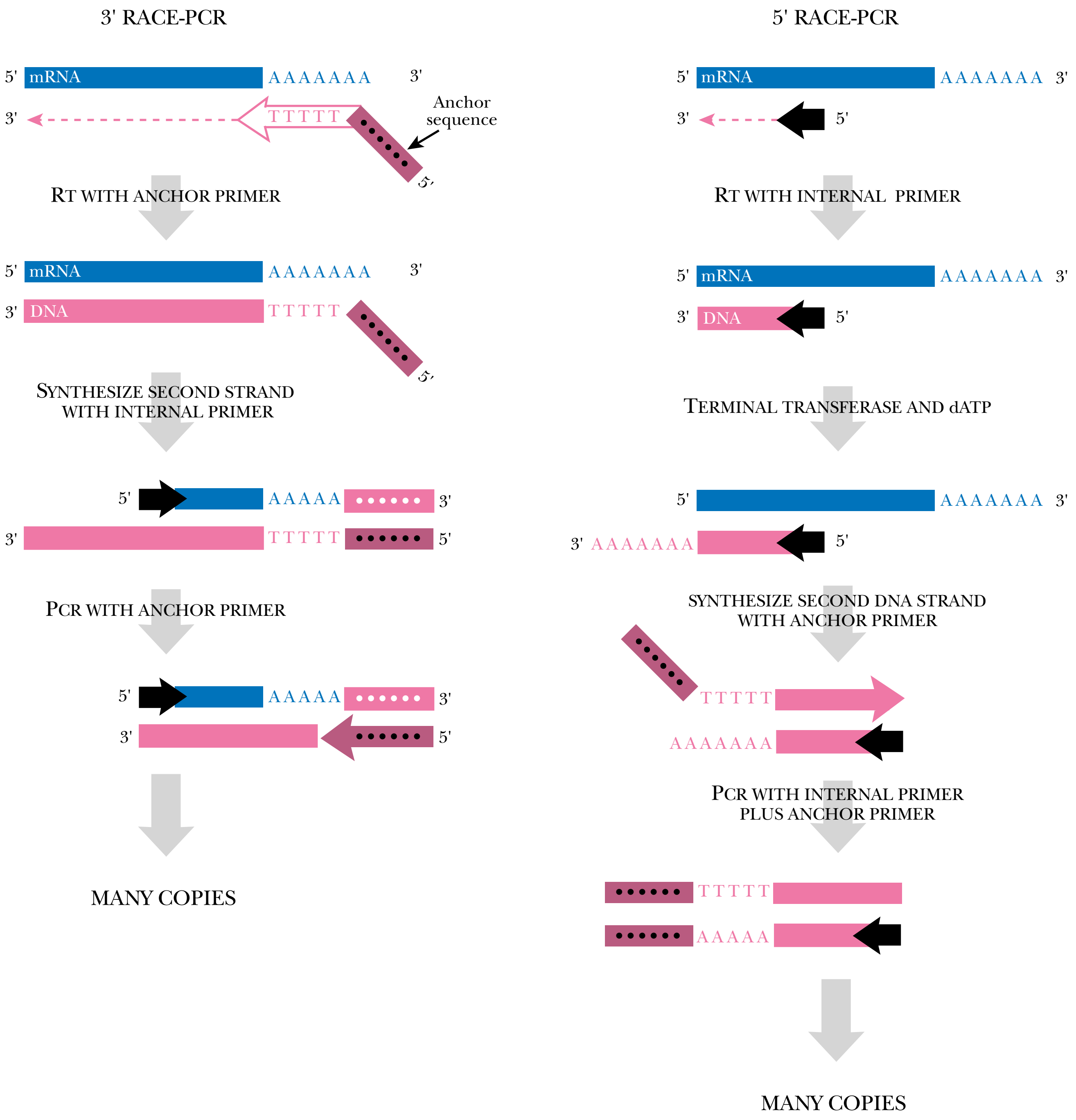


FIGURE 23.17 ***Rapid Amplification of cDNA Ends (RACE)***

RACE can be used to isolate the 5′ and/or 3′ ends of a cDNA that is incomplete. The method to amplify the 3′ end of the cDNA is shown on the left side of the figure. This requires an oligo(dT) primer that has an anchor sequence at the 5′ end. This primer is used with reverse transcriptase to make the mRNA : DNA hybrid molecule. The mRNA portion of this is removed and a second strand of DNA is synthesized. Instead of priming the second strand from the beginning of the DNA, an internal primer closer to the 3′ end of the gene is used. The same internal primer and a primer corresponding to the anchor sequence are then used in a standard PCR reaction to amplify just the 3′ end of the cDNA.

The right side of the figure shows how the 5′ end of a cDNA is isolated. An internal primer is designed to prime reverse transcriptase and make a hybrid molecule of mRNA : DNA. In order to add a primer binding site upstream of the end of the hybrid molecule, the enzyme, terminal transferase, is added together with dATP. This enzyme adds a run of adenines to the 3′ end of the DNA half of the hybrid. The mRNA half of the hybrid is then removed and replaced with DNA by using an oligo(dT) primer carrying the anchor sequence. The oligo(dT) binds the newly synthesized poly(A) stretch on the DNA, and primes the polymerase to make a cDNA. Subsequent PCR using the internal primer and the anchor primer amplify only the 5′ end of the cDNA.

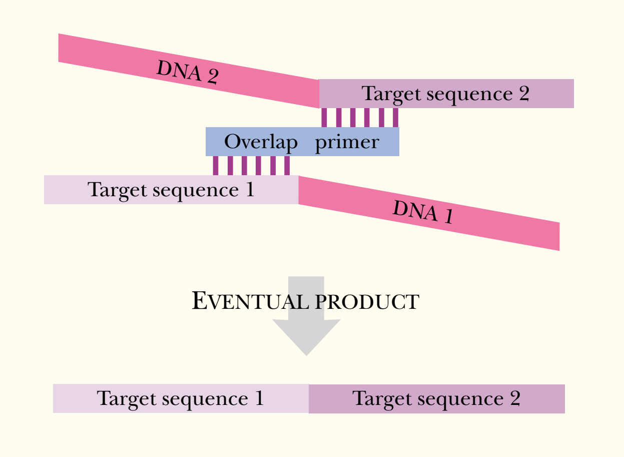


FIGURE 23.18 ***Synthesis of Hybrid Gene by Using Overlap Primers***

Overlapping primers can be used to link two different gene segments. In this scheme, the overlapping primer has one end with sequences complementary to target sequence 1, and the other half similar to target sequence 2. The PCR reaction will create a product with these two regions linked together.

Directed Mutagenesis

Mutations may be inserted artificially into DNA by altering a few bases in a PCR primer.

The term **directed mutagenesis** refers to a wide variety of *in vitro* techniques that are used to deliberately change the sequence of a gene. Several of these techniques use a PCR approach. The most obvious way to change one or two bases in a segment of DNA is to synthesize a PCR primer that carries the required alterations. Consider the sequence AAG CCG G**A**G GCG CCA. Suppose we wish to alter the A in the middle of this sequence to an T. Then we make a PCR primer with the required base alteration; that is, AAG CCG G**T**G GCG CCA. This mutant primer is used as one of a pair of PCR primers to amplify the appropriate segment of DNA using wild type DNA as template. The PCR product will contain the desired mutation, close to one end. As long as this primer is long enough to bind to the correct location on either side of the mutation, the DNA product will incorporate the change made in the primer. The mutant PCR product must then be reinserted into the original gene at the correct place.

Another less controlled way to introduce mutations into PCR products is to use manganese. Taq polymerase requires magnesium ions for proper function. Replacement of magnesium by manganese allows the enzyme to continue to synthesize DNA but the accuracy is greatly reduced. This approach introduces random base changes and yields a mixture of different mutations from a single PCR reaction. The error rate depends on the manganese concentration, so it is possible to get single or multiple mutations as desired.

Engineering Deletions and Insertions by PCR

Overlap primers may be used to precisely insert segments of foreign DNA into a chromosome or other DNA molecule.

PCR is widely used to generate DNA cassettes that can be introduced into chromosomes by homologous recombination. In this procedure, a convenient marker gene, usually an antibiotic resistance gene, is inserted into the chromosome of the host organism where it replaces any chosen gene. In order to target the incoming cassette to the correct location it must first be flanked with DNA sequences homologous to DNA both upstream and downstream of the chosen gene. This is done by using PCR primers that overlap the resistance cassette and also contain about 40–50 bp of DNA homologous in sequence to the target location (Fig. 23.19). The cassette is transformed into the host organism and is inserted into the chromosome by homologous recombination. Antibiotic resistance is then used to select those organisms that have gained the cassette. This approach may be used to generate deletions of any desired chromosomal gene and works especially well in yeast and bacteria.

directed mutagenesis Deliberate alteration of the DNA sequence of a gene by any of a variety of artificial techniques

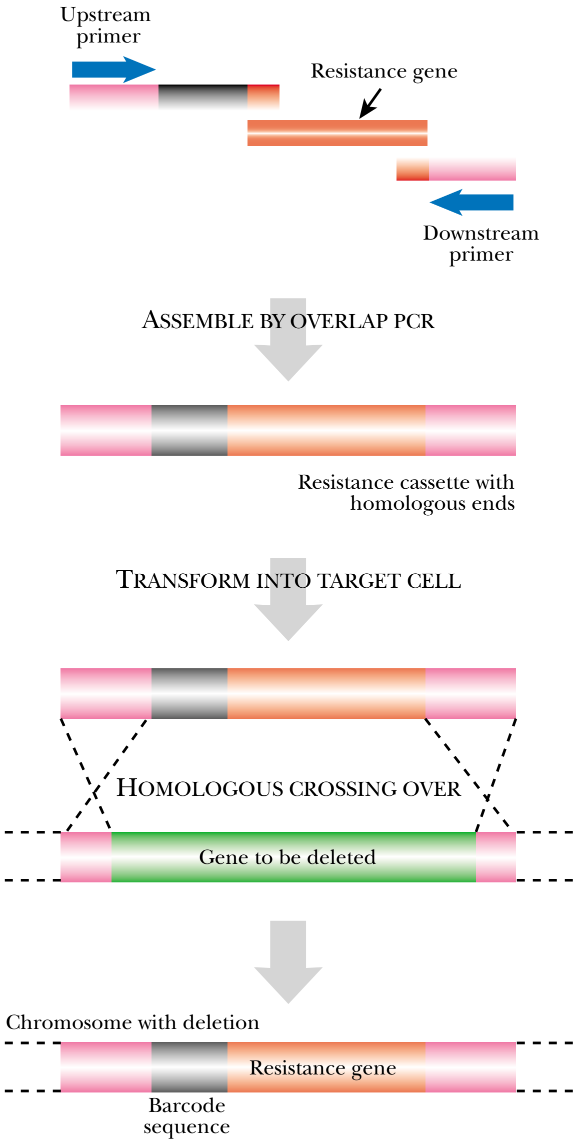


FIGURE 23.19 ***Generation of Insertion or Deletion by PCR***

In the first step, a specifically targeted cassette is constructed by PCR. This contains both a suitable marker gene and upstream and downstream sequences homologous to the chromosomal gene to be replaced. The engineered cassette is transformed into the host cell and homologous crossing over occurs. Recombinants are selected by the antibiotic resistance carried on the cassette.

A collection of yeast strains deleted for all approximately 6,000 known genes has been generated by this procedure. Each strain has had a single coding sequence replaced by a cassette comprising the *npt* gene plus a barcode sequence. The *npt* gene encodes neomycin phosphotransferase which confers resistance to neomycin and kanamycin on bacteria and resistance to the related antibiotic geneticin on eukaryotic cells, such as yeast. A barcode sequence is a unique sequence of around 20 bp that is included as a molecular identity tag. Each insertion has a unique barcode sequence allowing it to be tracked and identified. Such barcode or zipcode sequences are increasingly being used in high volume DNA screening projects where it is necessary to keep track of many similar constructs.

Clearly, the above procedure also generates an insertion of whatever gene is carried on the cassette. Thus any foreign gene may be inserted by this approach. There is no need to delete a resident gene if the objective is the insertion of an extra gene. All that is necessary is that the incoming gene must be flanked with appropriate lengths of DNA homologous to some location on the host chromosome.

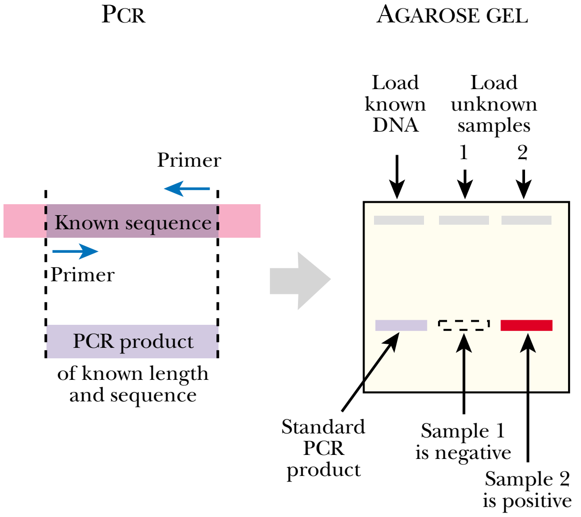
Traces of DNA can be amplified by PCR and used for diagnosis or forensic analysis.

Use of PCR in Medical Diagnosis

PCR can be used to identify an unknown sample of blood, tissue, or hair. First, a specific sequence of DNA must be determined from the organism. For example, if we want

FIGURE 23.20 ***PCR is Used to Diagnose Genetic Relatedness***

PCR can determine the identity of an unknown DNA sample. In this figure, two unknown DNA samples are isolated and amplified using PCR. The primers used to test these DNA samples are specific for a known sequence (pink). In sample 1, the primers did not anneal and no PCR product was made. Therefore, the sample must not contain any DNA with the known sequence (pink). Sample 2 showed a PCR product of the predicted size for the known sequence, therefore, the primers must have annealed and amplified this sequence.



to determine whether the unknown sample of blood is from a human, than a unique sequence from humans must be determined. Since many different genomes are currently being sequenced, this data is easily obtained. Next, two primers must be designed and synthesized using the sequence information. Small samples of DNA of unknown origin can then be tested by PCR using these primers. After the PCR reaction the DNA generated is run on an agarose gel to separate it according to size. If the DNA sample tested contains the target or known sequence, the PCR product will be of the predicted length (e.g., Fig. 23.20; unknown sample No. 2). If the test DNA is not from the same organism, no band will be generated (e.g., Fig. 23.20; unknown sample No. 1). The key to this experiment is the primers and how well they anneal to the target sequence. The primers may bind to closely related sequences, but the DNA made by PCR can be sequenced to determine whether or not this occurred. Increasing the temperature in which the primers anneal to the template can increase the stringency for this reaction also.

Clearly, PCR can be used in a variety of diagnostic tests. For example, visible symptoms of AIDS only appear a long time after infection, often several years. However, using PCR primers specific for sequences found only in the HIV genome, scientists can test for HIV DNA in blood samples, even when no symptoms are apparent. Another example is tuberculosis. Unlike many bacteria, *Mycobacterium*, which causes this disease, grows very slowly. Originally, to test for tuberculosis, the bacteria were cultured on nutrient plates, but this test took nearly a month. In contrast, PCR identification of mycobacterial DNA can be done in a day. Faster medical diagnoses are critical to help prevent the spread and progression of these diseases.

PCR is a powerful tool for amplifying small amounts of DNA. The DNA from 1/100th of a milliliter of human blood contains about 100,000 copies of each chromosome. If the target sequence for PCR is 500 base pairs, then there is about one-tenth of a picogram (10^{-12} gram) by weight of a target sequence. A good PCR run will amplify the target sequence and yield a microgram (10^{-6} gram) or more of DNA. A microgram may not seem much but is plenty for complete sequencing or cloning. Obviously, it is possible to identify an organism from an extremely small trace of DNA-containing material. In fact, the DNA from a single cell can be used to amplify a specific target sequence. This technique has revolutionized the criminal justice system by allowing highly accurate identification of individuals from very small samples.

Environmental Analysis by PCR

It is possible to extract DNA directly from environmental samples, such as soil or water, without bothering to isolate and culture the living organisms that contain it first.

Such environmental samples of DNA may be amplified by PCR. Because PCR is so sensitive, DNA can be amplified and sequenced from microorganisms that are present in such very low numbers that they cannot be detected by other means. Furthermore, it is not necessary for the microbial cells to be culturable or even viable. If specific PCR primers are used, it is possible to amplify genes from a single bacterium out of the billions that might be present in an environmental sample. As explained in Ch. 20, molecular based classification of organisms is based primarily on the sequence of the ribosomal RNA. Consequently, for identification of microorganisms from environmental samples by PCR, primers to the gene for 16 S ribosomal RNA are usually used. One fascinating result of environmental PCR analysis has been the discovery of many novel microorganisms that have never been cultured or identified by any other means. Such "microorganisms" are known only as novel ribosomal RNA sequences and it is presumed that the corresponding organisms do actually live and grow in the environment even though they do not grow in culture in the laboratory.

DNA sequences may be amplified by PCR directly from environmental samples.

By using primer sets specific for any chosen gene, PCR also allows us to check whether or not that particular gene is present in the environment being sampled. For example, suppose that we wish to know whether or not there are microorganisms capable of photosynthesis in a lake. A sample of lake water would be analyzed by PCR using primers specific for a gene that encodes an essential component of the light-harvesting mechanism. [Primer sets that are based on genes involved in a specific metabolic pathway are called "metabolic primers".] If a positive result is obtained we may conclude that there are organisms capable of photosynthesis in the lake. However, this approach does not tell us whether these organisms are actually growing by this mechanism or even still alive. A further step in analysis is to extract RNA from the environment and subject it to RT-PCR. This converts any messenger RNA present in the sample into the corresponding cDNA, which is then amplified. This reveals whether the corresponding genes are being actively transcribed, although, strictly speaking we still do not know if the corresponding enzyme or protein is present.

RT-PCR allows detection of RNA from environmental samples and so reveals whether the target gene is being transcribed.

Community profiling by PCR involves assessing the abundance and diversity of bacteria in an environment. In the simplest approach, total genomic DNA is first isolated from an environmental sample. All of the bacterial 16 S rRNA genes in that sample are then amplified by PCR, cloned and sequenced. The sequences are analyzed to identify the bacteria present in that environment. Metabolic profiling using primers specific for genes in particular metabolic pathways may also be performed. The relative abundance of the various organisms in the environment may then be estimated by hybridization.

It is also possible to isolate useful genes directly by environmental PCR—an approach sometimes referred to as eco-trawling. DNA isolated directly from the environment is amplified by PCR and the PCR fragments are cloned into a suitable plasmid that will allow expression of any successfully captured genes. The plasmids are transformed into a suitable bacterial host cell and the captured genes are expressed. The PCR primers are generally chosen to correspond to some known gene of interest. This results in variant versions of that particular gene, which were present in the environment sampled, being obtained. For example, primers corresponding to the ends of DNA polymerase could be used with DNA extracted from a sample of water from a hot spring. The desired result would be genes encoding novel DNA polymerases able to function at high temperature. This approach is obviously well suited to finding variants of known enzymes that function under novel or extreme conditions.

Genes encoding useful proteins may be cloned from environmental samples without knowing which organism they came from.

Rescuing DNA from Extinct Life Forms by PCR

Since any small trace of DNA can be amplified by PCR and then cloned or sequenced, some scientists have looked for DNA in fossils. Stretches of DNA long enough to yield valuable information have been extracted from museum specimens such as Egyptian mummies and fossils of various ages. In addition, DNA has been extracted from mammoth and plant remains frozen in the Siberian permafrost. This data has helped in studying molecular evolution and is discussed more fully in Ch. 20.

FIGURE 23.21 ***Mosquito Preserved in Amber***

Photo by Karen Fiorino.

DNA may be amplified from fossil material and used in identification.

In the sci-fi best seller, "Jurassic Park", the DNA was not obtained directly from fossilized dinosaur bones. Instead, it was extracted from prehistoric insects trapped in amber (Fig. 23.21). The stomachs of bloodsucking insects would contain blood cells complete with DNA from their last victim, and if preserved in amber, this could be extracted and used for PCR. DNA has indeed been extracted from insect fossils preserved in amber. However, the older the fossil, the more decomposed the DNA will be. Normal rates of decay should break the DNA double helix into fragments less than 1,000 bp long in 5,000 years or so. So, though we will no doubt obtain gene fragments from an increasing array of extinct creatures, it is unlikely that any extinct animal will be resurrected intact.

Realtime Fluorescent PCR

PCR has been modified for rapid diagnosis by using fluorescent dyes to follow DNA accumulation.

Recently methods have been developed that allow PCR reactions to be followed in real time by monitoring the increased emission from fluorescent probes (see Ch. 21). Instruments for combined PCR and fluorescence detection carry out the PCR reaction in glass capillary tubes. This allows the penetration of light to activate the fluorophore and the monitoring of the fluorescence emission as the PCR reaction is occurring. Modern instruments can monitor several different fluorescent dyes simultaneously, allowing several reactions to be run in the same tube.

DNA-binding fluorescent probes whose fluorescence increases upon binding to DNA are included in the PCR reaction mixture. As the amount of newly synthesized target DNA increases, the probe binds to the target DNA, and the fluorescence emission increases. The simplest DNA-binding fluorescent probes are not sequence specific. An example is the dye **SYBR® Green I** (Molecular Probes, Eugene, OR), with fluorescence emission at 520 nm. This binds only to double-stranded DNA and becomes fluorescent only when bound (Fig. 23.22).

SYBR® Green monitors the total amount of double-stranded DNA but cannot distinguish between different sequences. To be sure that the correct target sequence is being amplified a sequence specific fluorescent probe is needed. An example is the **TaqMan® probe** (Applied Biosystems, Foster City, CA). The TaqMan® probe consists of two fluorophores linked by a DNA sequence that will hybridize to the middle of the target DNA. **Fluorescence resonance energy transfer (FRET)** transfers the energy from the short-wavelength fluorophore on one end to the long wavelength fluorophore at the other end. This quenches the short wave emission (Fig. 23.23).

fluorescence resonance energy transfer (FRET) Transfer of energy from short-wavelength fluorophore to long-wavelength fluorophore so quenching the short wave emission
SYBR® Green I A DNA-binding fluorescent dye that binds only to double-stranded DNA and becomes fluorescent only when bound
TaqMan® probe Fluorescent probe consisting of two fluorophores linked by a DNA probe sequence. Fluorescence increases only after the fluorophores are separated by degradation of the linking DNA

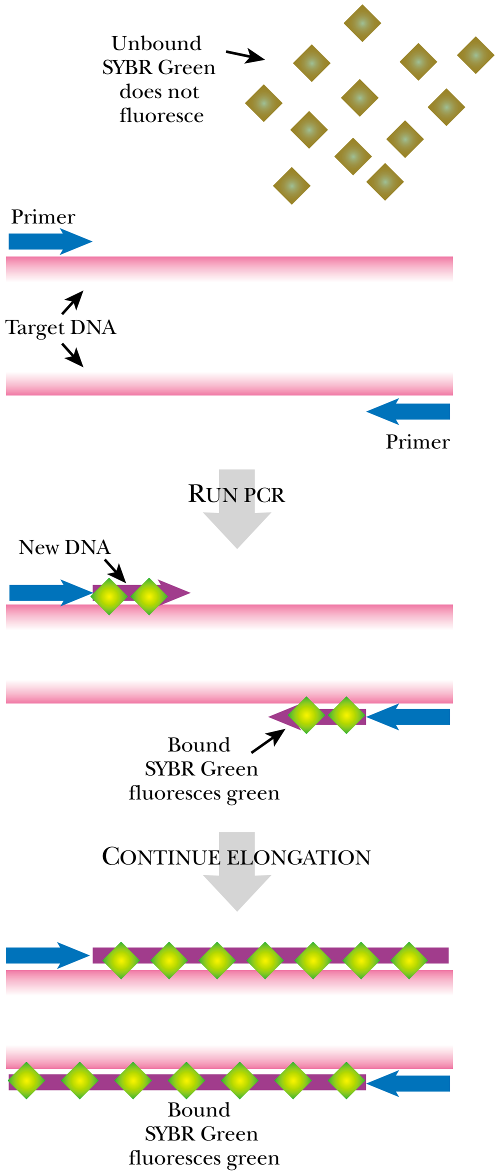


FIGURE 23.22 ***Realtime Fluorescent PCR with SYBR® Green***

When the fluorescent probe SYBR® green is present during a PCR reaction, it binds to the double-stranded PCR product and emits light at 520 nm. The SYBR® Green dye only fluoresces when bound to DNA. Hence, the amount of fluorescence correlates with the amount of PCR product produced. This allows the accumulation of PCR product to be followed through many cycles.

Specific probes can be included in fluorescent PCR procedures to ensure that only specific target DNA sequences give rise to fluorescence.

During PCR the TaqMan® probe binds to the target sequence after the denaturation step that separates the two DNA strands. As the Taq polymerase extends the primer during the next PCR cycle it will eventually bump into the TaqMan® probe. The Taq polymerase is not only capable of displacing strands ahead of it but also has a 5′-nuclease activity that degrades the DNA strand of the probe. This breaks the linkage between the two fluorophores and disrupts the FRET. The short-wavelength fluorophore is now free from quenching and its fluorescence increases. In this case the increase in fluorescence is directly related to the amount of the specific target sequence that has been amplified.

Inclusion of Molecular Beacons in PCR—Scorpion Primers

Molecular beacons may be used in conjunction with PCR primers to give a highly specific amplification plus detection system. **Scorpion primers** consist of a molecular

molecular beacon A fluorescent probe molecule that contains both a fluorophore and a quenching group and that fluoresces only when it binds to a specific DNA target sequence
Scorpion primer DNA primer joined to a molecular beacon by an inert linker. When the probe sequence binds target DNA, the quencher and fluorophore are separated allowing fluorescence

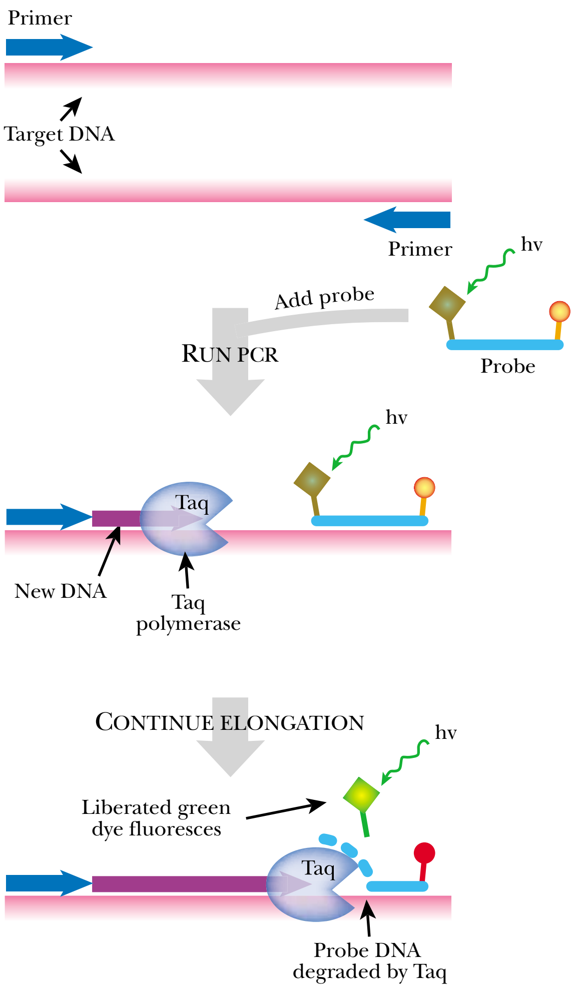


FIGURE 23.23 *Realtime Fluorescent PCR with TaqMan® Probe*

The TaqMan® probe has three elements: a short-wavelength fluorophore on one end (diamond), a sequence that is specific for the target DNA (blue), and a long-wavelength fluorophore at the other end (circle). The two fluorophores are so close that fluorescence is quenched and no green light is emitted. This probe is designed to anneal to the center of the target DNA. When Taq polymerase elongates the second strand during PCR, its nuclease activity cuts the probe into single nucleotides. This releases the two fluorophores from contact and abolishes quenching. The short-wavelength fluorophore can now fluoresce and a signal will be detected that is proportional to the number of new strands synthesized.

beacon (see Ch. 21) joined to a single-stranded DNA primer by an inert linker molecule (e.g. hexethylene glycol). When the beacon is in its hairpin structure, the **quencher** (e.g. methyl red) binds to the fluorophore (e.g. fluorescein) and prevents fluorescence. The loop portion of the stem and loop structure has sequences complementary to the target DNA, and constitutes the probe segment. When the probe sequence binds to target DNA, the hairpin is disrupted, the quencher and fluorophore are separated and fluorescence occurs.

During PCR, the Scorpion primer binds to the target DNA and is elongated by the Taq polymerase. The two strands are separated in the next denaturation cycle. The Scorpion probe sequence then hybridizes to the single-stranded DNA in the middle of the target sequence. This releases the fluorophore from the quencher and promotes fluorescence (Fig. 23.25).

Rolling Circle Amplification Technology (RCAT)

Several novel methods other than PCR have been proposed for amplifying DNA. These methods all use DNA polymerase to amplify DNA but they avoid the high temperature requirement for DNA denaturation and the consequent temperature cycling. Which of these methods will prove useful in the long run is still undecided. Perhaps

quencher Molecule that prevents fluorescence by binding to the fluorophore and absorbing its activation energy

On Site Diagnosis of Plant Disease by Realtime PCR

The development of realtime PCR has resulted in a great decrease in the time needed for detection of DNA from infectious microorganisms. Classical methods often required 3 or 4 days to isolate the microorganism and another week to confirm its identity—always assuming the pathogen can be cultured. Standard PCR methods not only cut the time needed to 2 to 3 days but also work directly on tissue samples without requiring that the microorganisms should be cultured. Realtime PCR with fluorescent detection has cut the time required for diagnosis even further, although the technique was originally lab-based and relatively expensive. However, portable realtime PCR machines have been developed recently that allow DNA identification in a couple of hours. An example is the Smart Cycler® TD made by Cepheid Corporation of Sunnyvale, California.

The Smart Cycler® has been used to diagnose plant diseases, on-site in the fields where the crops are growing. For example, watermelon fruit blotch is a bacterial disease that causes major economic losses of watermelon crops worldwide. The causative agent, *Acidovorax avenae* subsp. *citrulli* requires 10–14 days for diagnosis by classical procedures. Portable, realtime PCR allows on-site identification within an hour of taking samples from plants with suspected infections. This rapid diagnosis is of great value both in managing crop diseases and also in deciding whether quarantine is required when facing a possible outbreak of a transmissible plant disease that could threaten crops in other locations.

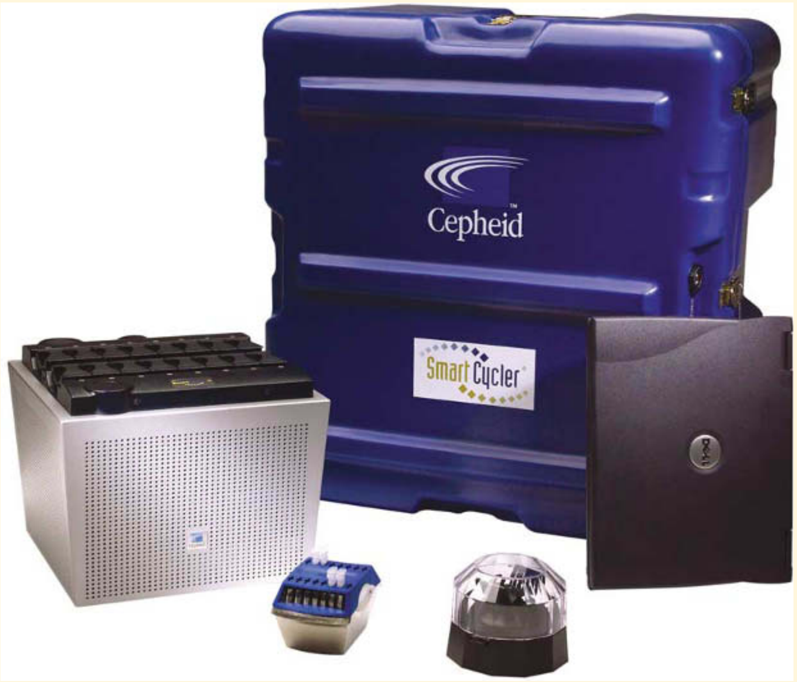


FIGURE 23.24 ***Portable Realtime PCR***

The Cepheid Smart Cycler® system has been used for rapid on-site detection of plant diseases. Courtesy Cepheid Corporation.

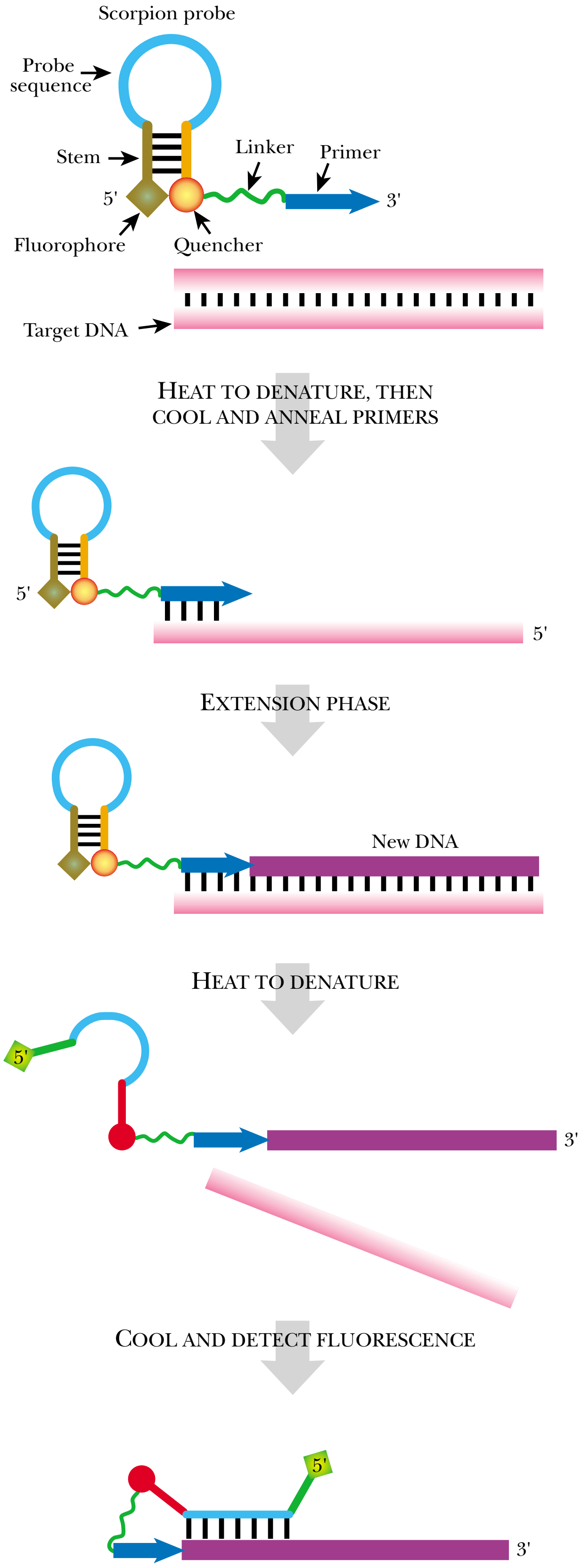


FIGURE 23.25 ***Scorpion Primer with Combined Fluorescent Probe***

Scorpion primers provide another method to detect the PCR product by fluorescence. The Scorpion probe has a stem loop structure that keeps the fluorophore molecule (diamond) in close proximity to the quencher (circle). The loop has a sequence complementary to the target DNA. The stem/loop is linked to a regular PCR primer designed to amplify the target DNA. During the extension step of PCR, the primer portion of the probe anneals to the template and Taq polymerase makes new DNA. During the next denaturation step, the whole probe plus new DNA strand become single-stranded. The loop (blue) can now anneal to the single-stranded target DNA, releasing the fluorophore from the quencher. The fluorescence emitted is a direct measure of the amount of PCR product produced.

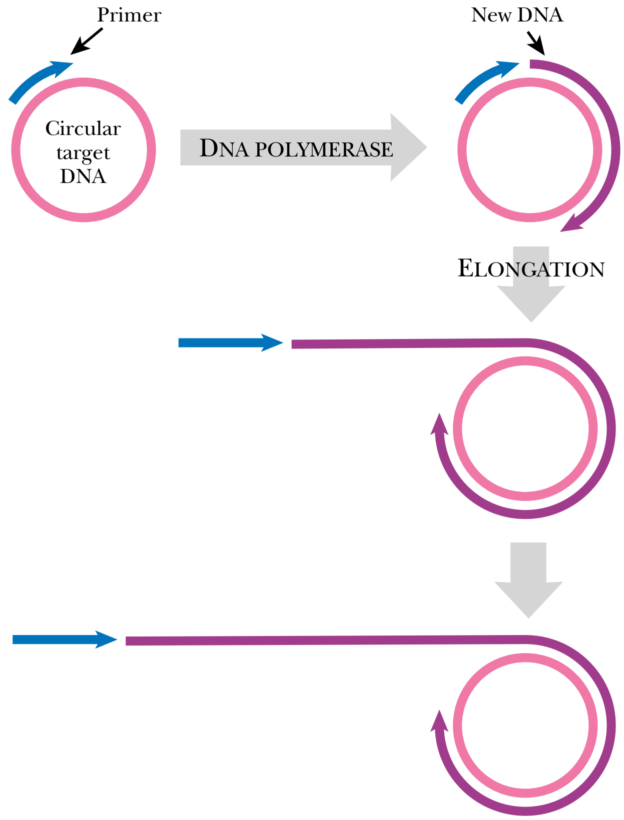


FIGURE 23.26 ***Rolling Circle Amplification of DNA—Linear Version***

Just like rolling circle replication in bacteria, RCAT produces many copies of a circular target DNA. This process only requires one primer for DNA polymerase, and the temperature does not need to be elevated. The result is a long linear piece of DNA.

Amplification of DNA by the rolling circle mechanism avoids the high temperature step of PCR.

the most promising is **Rolling Circle Amplification Technology (RCAT)** (marketed by Molecular Staging, New Haven, CT).

This technique is based on the rolling circle mechanism for DNA replication used by many plasmids and viruses (see Ch. 17); therefore, a circular DNA template is needed. This process occurs at a single temperature and does not require thermostable DNA polymerase. In linear RCAT, a DNA primer binds to the circular DNA, which is then copied many times to give a long single-stranded product that may consist of up to 100,000 tandem repeats of the target sequence (Fig. 23.26). Since the RCAT product remains attached to the original circular template, the method may be used in combination with DNA microarrays (see Ch.24 for DNA arrays).

In exponential RCAT (E-RCAT) a second primer is used that binds to the opposite strand—in other words it binds to the newly made linear strand. When this primer binds to neighboring tandem copies of the target sequence elongation results in strand displacement. This creates single-strand branches that in turn can bind primer number 1 and so be converted to double-stranded DNA. Alternate extension using two primers results in multiple branching of the amplified DNA (Fig. 23.27). The exponential version of RCAT can manufacture 10^{12} copies of each original circle in an hour and can detect a single target molecule. It is thus superior to PCR in both aspects.

rolling circle amplification technology (RCAT) Method based on rolling circle replication that uses DNA polymerase to amplify target DNA at normal temperatures

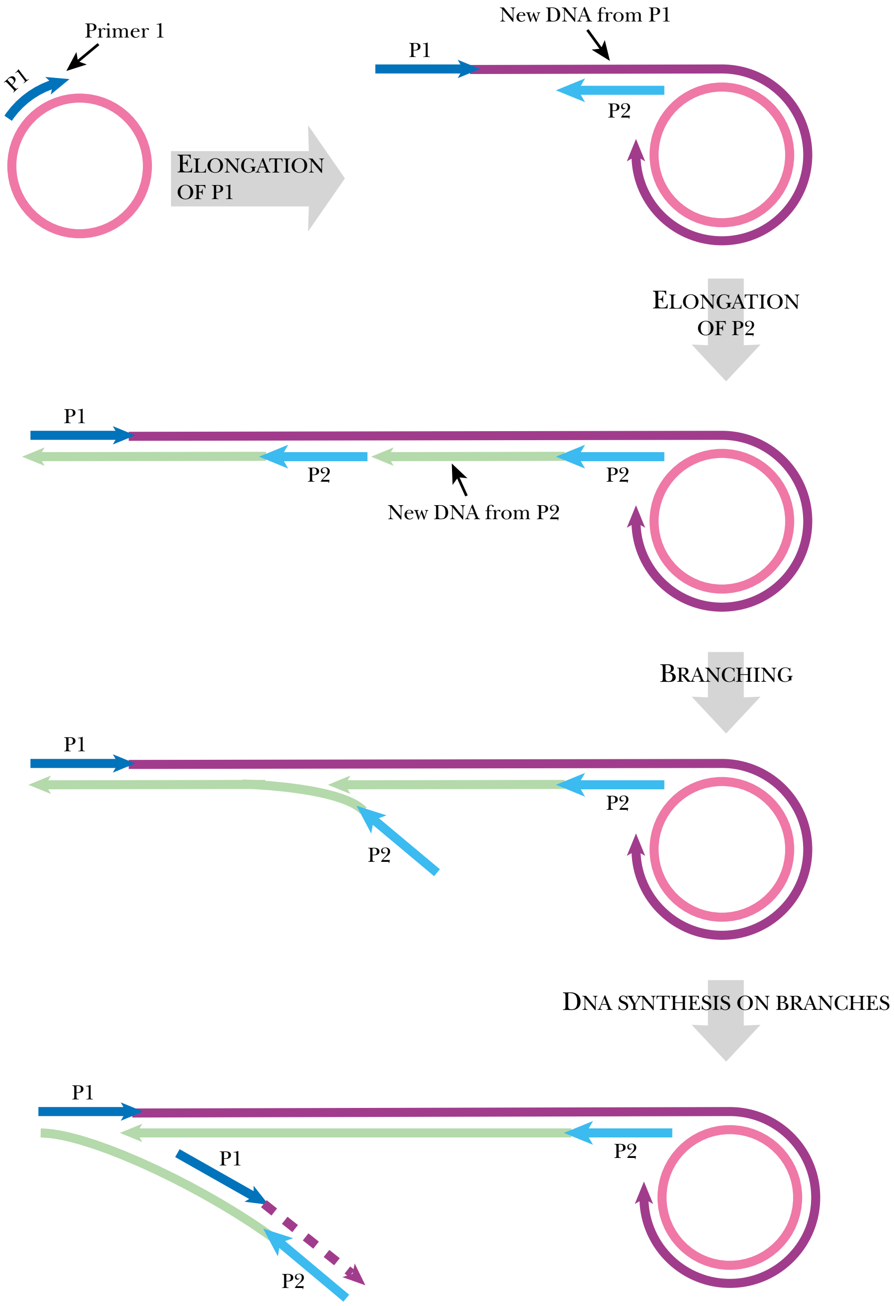


FIGURE 23.27 ***Rolling Circle Amplification of DNA—Exponential Version***

As in the linear version, E-RCAT reactions start with polymerase and the first primer (P1). Rolling circle replication produces tandem single-stranded copies of the circular target DNA. Next, a second primer (P2) anneals at the end of each tandem segment. Complementary strands of DNA are then synthesized starting from primer P2, which makes the entire DNA region double-stranded. As the DNA polymerase passes the end of a target segment, it displaces the DNA strand in front of it, making a single-stranded branch. The first primer, P1, can anneal to the inside end of the branch and polymerase makes the second strand.

The major limitation of RCAT is the need for a circular template. However, variants that involve preliminary circularization of a non-circular target molecule have been developed. As with PCR itself, the RCAT technique can also be combined with fluorescence detection.

CHAPTER TWENTY-FOUR

Genomics and DNA Sequencing

Introduction to Genomics

A second major revision and analysis of the human genome announced in 2004 has reduced the estimated number of human genes from 35,000 to 25,000.

The last few years have seen an information explosion in DNA sequences. Many whole bacterial genomes and a significant number of eukaryote genomes have been fully sequenced. In this chapter, we will consider first the approaches used to sequence DNA, including whole genomes. Secondly, we will consider the methods used to screen whole genomes for the presence of particular sequences or to measure global gene expression. Those investigatory approaches that deal with the whole genome, rather than single genes, define the emerging field of **genomics**.

The sequencing of the human genome revealed that we have approximately 35,000 genes in contrast to predictions of up to 100,000 from some quarters. Although the bacterium *Escherichia coli* has only around 4,000 genes, some bacteria have as many as 9,000 genes. Moreover, primitive animals such as roundworms have 18,000 genes and the plant, *Arabidopsis*, has 25,000. The discovery that humans possess scarcely three times as much genetic information as the most complex bacteria and less than twice as much as a worm was a humbling revelation. Despite the smaller than expected number of genes, the functions of most of our genes are still unknown. The analysis of gene expression and of protein function will be considered in the following two chapters. Here we will consider the sequence of the DNA itself.

DNA Sequencing—General Principle

DNA sequencing actually involves synthesizing DNA sub-fragments of all possible lengths and separating them on a gel.

Before searching for genes or comparing different DNA sequences, the DNA must first be sequenced. The overall approach first involves generating a reasonably sized template DNA by cloning or PCR. The actual sequencing involves generating sub-fragments of all possible lengths from this template. The sub-fragments are generated by DNA polymerase to differ in length by only one base pair so the fragments end at each of the base pairs in the original fragment. Therefore, if the template were 200 base pairs in length, there would be 200 different sub-fragments ranging from one base pair to 200 base pairs. These are then grouped according to which base they end in, and are separated by gel electrophoresis. Let's illustrate this using the eight-base sequence ACGATTAG as an example (Fig. 24.01). DNA polymerase generates eight sub-fragments of this example, ranging from the entire eight base pair piece to only a single nucleotide found at the beginning.

The four groups of fragments are separated by gel electrophoresis by running them on the same gel in four parallel lanes. Those fragments ending in A are run in the first lane, those ending in G in the second lane, and so on. The fragments will be separated according to their lengths and we will see a separate band for each fragment (Fig. 24.02). Starting at the bottom of the gel and reading upwards, we can read off the sequence directly.

The Chain Termination Method for Sequencing DNA

Dideoxy base analogs are used to terminate growing DNA chains.

We must now consider the detailed technical questions of how to actually make such fragments and, in particular, how to separate them into four groups depending on the last base. The method routinely used today is known as **chain termination sequencing** or **dideoxy sequencing**. Both names refer to the fact that dideoxy analogs of normal DNA precursors cause premature termination of a growing chain of nucleotides being made by DNA polymerase. This allows us to generate the fragments

chain termination sequencing Method of sequencing DNA by using dideoxynucleotides to terminate synthesis of DNA chains. Same as dideoxy sequencing

dideoxy sequencing Method of sequencing DNA by using dideoxynucleotides to terminate synthesis of DNA chains. Same as chain termination sequencing

genomics Study of genomes as a whole rather than one gene at a time

Original	These are grouped as follows:			
8 fragments	Ending in A	Ending in G	Ending in T	Ending in C
ACGATTAG		ACGATTAG		
ACGATTA	ACGATTA			
ACGATT			ACGATT	
ACGAT			ACGAT	
ACGA	ACGA			
ACG		ACG		
AC				AC
A	A			

FIGURE 24.01 ***Sequencing—Fragments of All Possible Lengths***

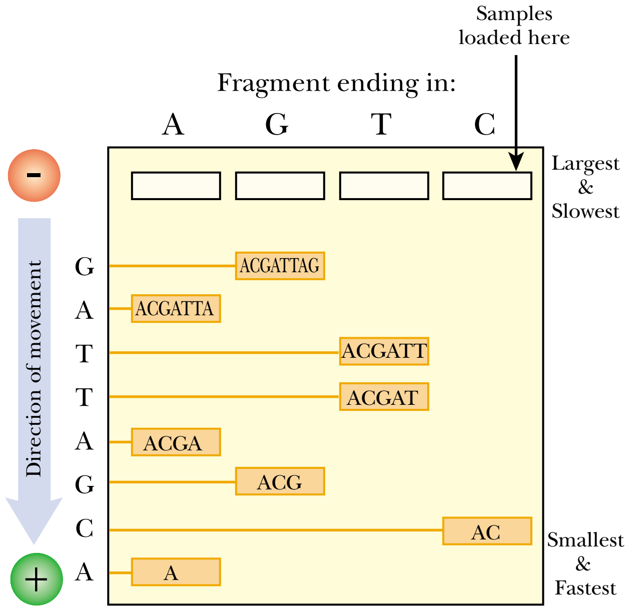


FIGURE 24.02 ***Principle of DNA Sequencing***

The sequence of a DNA fragment can be determined by generating sub-fragments from the original sequence as shown in Fig. 24.01. The sub-fragments are generated in four separate reactions, one for each of the four bases. Each reaction mixture is then separated by size using gel electrophoresis. The sequence can be read off, starting at the bottom of the gel and reading upwards.

of the segment of DNA to be sequenced. Using four different dideoxy analogs, one for each of the four bases, allows the generation of four sets of fragments in four separate reactions.

Like all DNA synthesis reactions, sequencing reactions require DNA polymerase, a region of single stranded DNA called the template, and a primer to which nucleotides are added (see Ch. 5 for details). The DNA polymerase will then elongate the primer and make a new DNA strand complementary to the template strand (Fig. 24.03). Whenever a base is added to the growing strand of nucleic acid, it is provided as the nucleoside triphosphate (NTP), a base linked to a sugar and three phosphate groups. The outermost two phosphate groups are lost when each nucleotide (sugar + phosphate + base) is added to the end of the growing DNA chain.

Since DNA contains the four bases, adenine, guanine, thymine and cytosine, DNA polymerase must be supplied with a mixture of the four deoxynucleoside triphosphates, dATP, dGTP, dTTP, and dCTP. When nucleotides are joined to make a nucleic acid, the phosphate group, which is attached to the 5′-carbon atom of the sugar of the incoming nucleotide, is linked to the 3′-hydroxyl group of the sugar belonging to the last nucleotide in the chain (Fig. 24.04). Or, in brief, they are polymerized in the 5′ to 3′ direction.

The 3′-hydroxyl group is critical for DNA polymerase to add more nucleotides to the growing chain of RNA or DNA. The sugar of RNA is ribose, which has hydroxyl groups on both the 2′ and 3′ carbons of the sugar ring. The sugar of DNA is called deoxyribose because it is missing an oxygen molecule relative to ribose. Deoxyribose has no hydroxyl group on the 2′-carbon atom of the ring. The elongation scheme shown

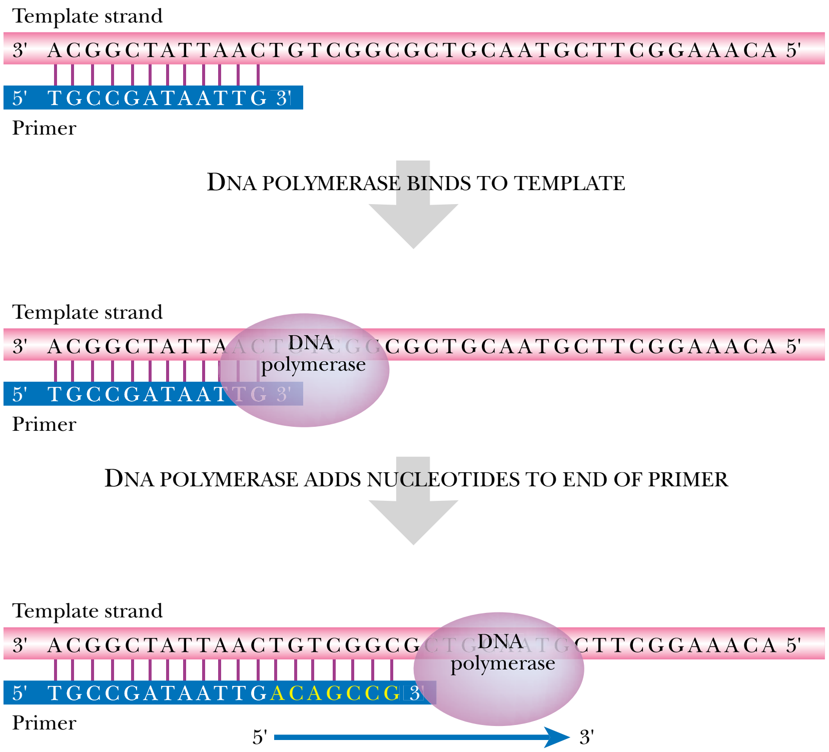

Template strand
3' ACGGCTATTAACTGTCGGCGCTGCAATGCTTCGGAAACA 5'
5' TGCCGATAATTG 3'
Primer
DNA POLYMERASE BINDS TO TEMPLATE
Template strand
3' ACGGCTATTAACTGTCGGCGCTGCAATGCTTCGGAAACA 5'
DNA polymerase
5' TGCCGATAATTG 3'
Primer
DNA POLYMERASE ADDS NUCLEOTIDES TO END OF PRIMER
Template strand
3' ACGGCTATTAACTGTCGGCGCTGCAATGCTTCGGAAACA 5'
DNA polymerase
5' TGCCGATAATTGACAGCCG 3'
Primer
5'
3'

FIGURE 24.03 ***Synthesis of DNA—Priming and Elongation***

During normal DNA synthesis, DNA polymerase reads the template strand and makes a new complementary strand of DNA. To get DNA synthesis started, a short oligonucleotide primer must anneal to the 3′ end of the template. DNA polymerase recognizes the 3′ end of the primer and adds incoming nucleotides to the 3′ end, hence synthesis occurs in a 5′ to 3′ direction.

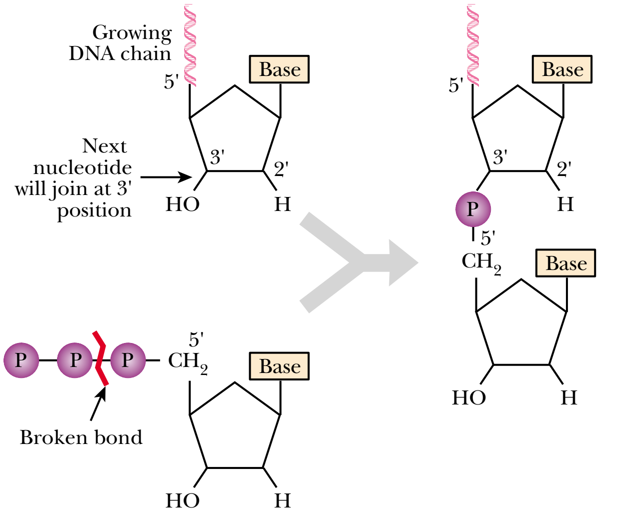


FIGURE 24.04 ***Synthesis of DNA—Phosphodiester Bonding***

DNA polymerase links nucleotides via phosphodiester bonds. When adding another nucleotide, DNA polymerase breaks the bond between the first and second phosphates of the deoxynucleoside triphosphate. The incoming nucleotide is then joined to the free 3′ hydroxyl of the growing DNA chain.

above works for both RNA and DNA since they both have hydroxyl groups at the 3′ position on their sugars. However, **dideoxyribose** is a sugar that lacks the oxygen of both the 2′ and the 3′ hydroxyl groups (Fig. 24.05). Nucleotides containing dideoxyribose can be incorporated into a growing nucleic acid chain, but that is literally the end of the chain. Incoming nucleotides must be added to the 3′-hydroxyl group of the previous nucleotide. But since dideoxyribose has no 3′-hydroxyl group, no further nucleotides can be added, and the chain is terminated.

Just as we use dG to refer to a normal deoxynucleotide with the base guanine, we use ddG to refer to the **dideoxynucleotide** with the base guanine. Consider what happens if ddGTP, dideoxyguanosine triphosphate, the dideoxyribose analog of dGTP, is added to a growing DNA chain. When the polymerase reaches the next G, it puts in ddG instead of dG. Then the chain is terminated (Fig. 24.06). During the sequencing

dideoxynucleotide Nucleotide whose sugar is dideoxyribose instead of ribose or deoxyribose
dideoxyribose Derivative of ribose that lacks the oxygen of both the 2′ and the 3′ hydroxyl groups

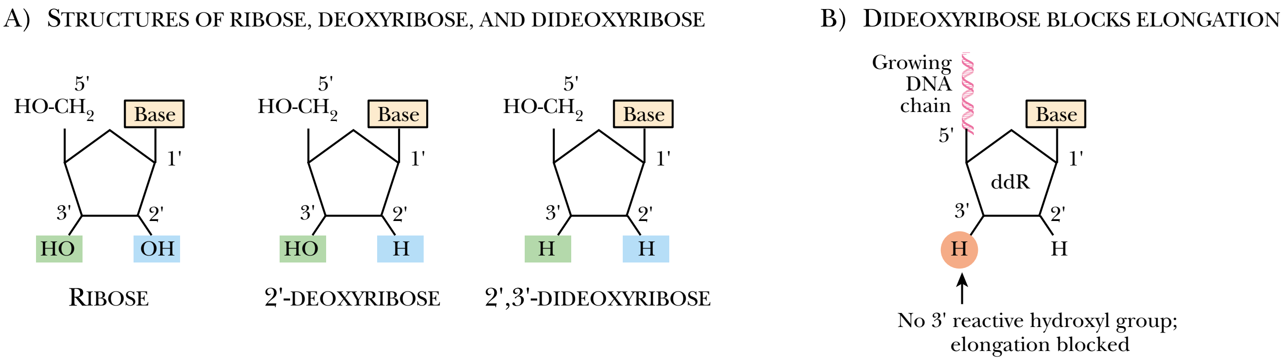

A) STRUCTURES OF RIBOSE, DEOXYRIBOSE, AND DIDEOXYRIBOSE
5'
HO-CH2
Base
1'
3'
2'
HO
OH
RIBOSE
HO
H
2'-DEOXYRIBOSE
H
H
2',3'-DIDEOXYRIBOSE
B) DIDEOXYRIBOSE BLOCKS ELONGATION
Growing DNA chain
5'
Base
ddR
1'
3'
2'
H
H
No 3' reactive hydroxyl group; elongation blocked

FIGURE 24.05 ***Dideoxyribose, Deoxyribose and Ribose***

(A) The structures of ribose, deoxyribose, and dideoxyribose differ in the number and location of hydroxyl groups on the 2′ and 3′ carbons. (B) DNA polymerase cannot add another nucleotide to a chain ending in dideoxyribose because its 3′ carbon does not have a hydroxyl group.

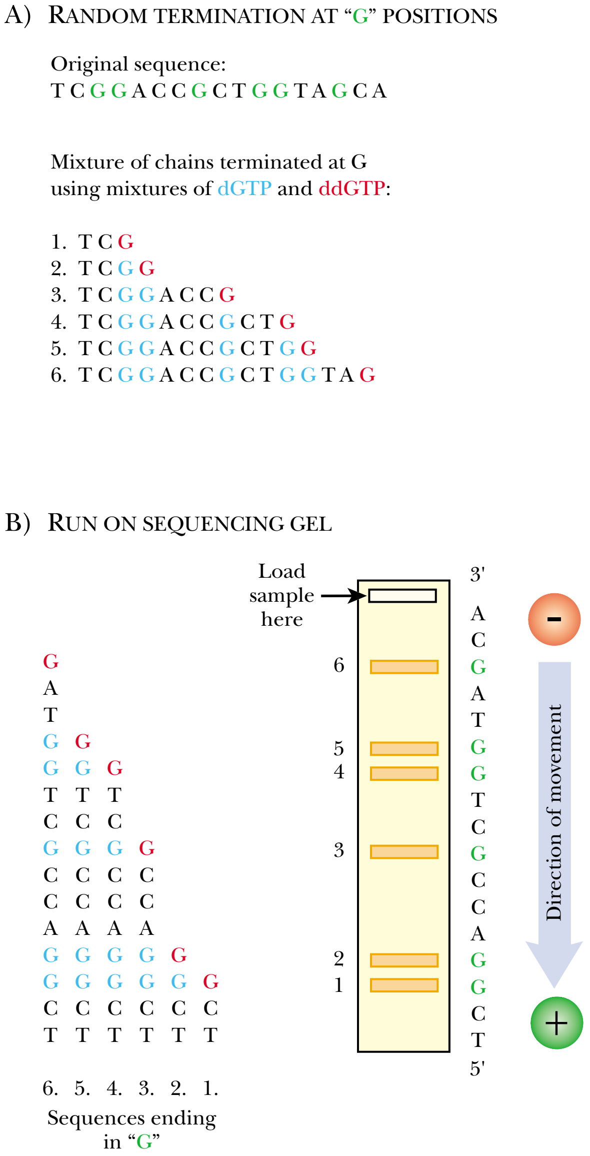

A) RANDOM TERMINATION AT "G" POSITIONS
Original sequence:
T C G G A C C G C T G G T A G C A
Mixture of chains terminated at G using mixtures of dGTP and ddGTP:
1. T C G
2. T C G G
3. T C G G A C C G
4. T C G G A C C G C T G
5. T C G G A C C G C T G G
6. T C G G A C C G C T G G T A G
B) RUN ON SEQUENCING GEL
Load sample here
6
5
4
3
2
1
3'
A C G A T G G T C G C C A G G C T
5'
-
+
Direction of movement
6. 5. 4. 3. 2. 1.
Sequences ending in "G"

FIGURE 24.06 ***Chain Termination by Dideoxynucleotides***

(A) During the sequencing reaction, DNA polymerase makes multiple copies of the original sequence. Sequencing reaction mixtures contain artificial dideoxynucleotides that terminate growing DNA chains. The example here shows the G reaction, which includes triphosphates of both deoxyguanosine (dG) and dideoxyguanosine (ddG). Whenever ddG is incorporated (shown in red), it causes termination of the growing chain. If dG (blue) is incorporated, the chain can continue growing. (B) When the sequencing reaction containing the ddG is run on a polyacrylamide gel, the fragments are separated by size. Each band directly represents a guanine in the original sequence.

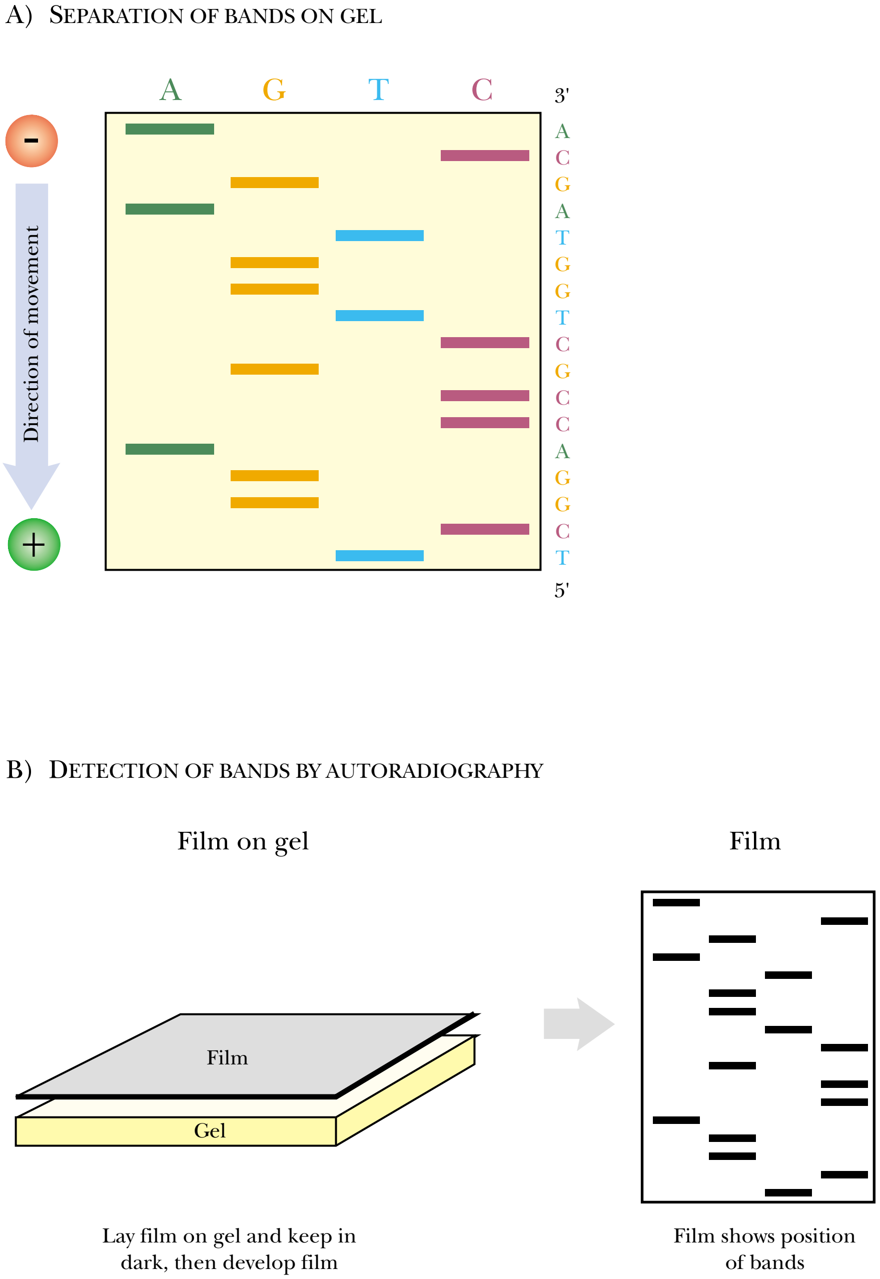

A) Separation of bands on gel
A
G
T
C
3'
A
C
G
A
T
G
G
T
C
G
C
C
A
G
G
C
T
5'
-
+
Direction of movement
B) Detection of bands by autoradiography
Film on gel
Film
Film
Gel
Lay film on gel and keep in dark, then develop film
Film shows position of bands

FIGURE 24.07 ***Separation and Detection of Fragments on Gel during DNA Sequencing***

(A) The products of the four separate sequencing reactions are run side by side on a polyacrylamide gel and the fragments of different sizes are separated by electrophoresis. (B) To detect the fragments, the gel is transferred to a piece of paper to give it strength, and dried so that the polyacrylamide does not stick to the film. After the gel is completely dry, a piece of photographic film is placed over the gel. The positions of the radioactive DNA fragments are revealed by the dark bands they produce on the film.

reaction, a mixture of dGTP and ddGTP are available for DNA polymerase. Sometimes DNA polymerase uses dG and sometimes ddG. The amount of ddG relative to dG is adjusted to yield a mixture of chains that are terminated by ddG at all positions where there is a G in the sequence. The reaction to generate the "G" lane on a sequencing gel contains normal deoxynucleotide triphosphates for the other three bases but both dG and ddG for guanosine. The other three sequencing reactions are set up similarly. In practice, the template, primers, radioactive dNTPs for all four bases, and DNA polymerase are all mixed together. This mixture is then distributed among four tubes, each with a different dideoxynucleotide.

DNA fragments are separated according to size on a polyacrylamide gel.

This mixture of fragments is separated according to their sizes by gel electrophoresis (Fig. 24.07A). Since the largest DNA sub-fragments generated by sequencing reactions are usually only 200–300 base pairs in length, the fragments are too small to be resolved by the large pores of agarose and must be separated using a polyacrylamide gel. The shortest pieces are closer to the bottom, as they move fastest during electrophoresis. These short pieces are generated when the ddNTP is incorporated shortly after DNA polymerase begins synthesis at the primer. A series of bands is produced, each corresponding to a piece of DNA of a particular length. The lengths reveal the positions of the G bases in the original DNA. To completely sequence DNA, the same thing is done simultaneously for all four bases, A, G, T, and C. So there are four reaction mixtures, each containing a specific ratio of normal deoxynucleotides to dideoxynucleotides. Each reaction contains a series of artificially terminated chains ending in one of the four bases. All four samples are loaded side by side onto a gel. Separation by electrophoresis gives four ladders representing each of the bases (Fig. 24.07A).

DNA fragments were originally detected by radioactive labeling, although nowadays fluorescent dyes are normally used as labels.

Some way is needed to detect the DNA bands. Originally, radioactive nucleotide precursors or radioactively labeled primers were incorporated into the sub-fragments generated by the polymerase. After separating the sub-fragments by electrophoresis, the polyacrylamide gel is dried and a sheet of photographic film is laid on top of the gel. The radioactive bands leave a black mark on the film, thus allowing the researcher to visualize the position of the original bands (Fig. 24.07B). As before, the position of each band corresponds to a chain of DNA of a particular length and reveals the position of one base. The sequence is read off directly from the bottom, combining results from all four bases. Several hundred bases of sequence can usually be obtained from one gel. Part of a real sequencing gel is shown in Figure 24.08. Instead of radioactivity, modern DNA sequencing techniques use fluorescently labeled nucleotide precursors or primers for detection (see below).

DNA Polymerases for Sequencing DNA

All DNA polymerases can elongate a primer that is annealed to a single-stranded DNA template. However, the characteristics needed for use in sequencing are more rigorous. First, the polymerase must have high processivity, that is, it must move a long way along the DNA before dissociating. Premature dissociation would give strands that ended at random before the dideoxynucleotide was incorporated. In addition, many DNA polymerases possess exonuclease activities (see Ch. 5). 5′ to 3′ exonuclease activity may be used to remove a strand of DNA ahead of the replication point. In contrast, 3′ to 5′ exonuclease activity is used to remove incorrect bases during proofreading. Such activities interfere with accurate sequencing as they might shorten the length of strands already synthesized.

Genetically engineered DNA polymerase with properties suitable for use in test tubes rather than cells is now used for DNA sequencing.

In practice, no natural DNA polymerase is entirely suitable for sequencing. The first DNA polymerase used was **Klenow polymerase**, which is DNA polymerase I from *E. coli* that lacks the 5′ to 3′ exonuclease domain. Klenow polymerase was originally obtained by protease digestion of purified DNA polymerase I but was later made by expression of a modified gene. Because Klenow polymerase has relatively low processivity, it can only be used to sequence around 250 bases per reaction. More recently, genetically modified DNA polymerase from bacteriophage T7 has been used. This is marketed as "**Sequenase**" and has high processivity, a rapid reaction rate, negligible exonuclease activity and the ability to use many modified nucleotides as substrates, thus making it perfect for sequencing reactions.

Producing Template DNA for Sequencing

High quality DNA sequencing requires purified single-stranded DNA to which the sequencing primer can bind. Originally, template DNA was from the bacterial virus **M13** that was engineered to contain the template sequences (Fig. 24.09). M13 virus is rod-shaped and contains a circle of single stranded DNA (ssDNA). Upon infecting an *E. coli* cell, the single-stranded viral DNA is converted to a double-stranded form, the **replicative form (RF)**. After replicating itself for a while, the RF then turns its efforts

Klenow polymerase DNA polymerase I from *E. coli* that lacks the 5′ to 3′ exonuclease domain
M13 Rod-shaped bacteriophage that infects *E. coli*, contains a circle of single stranded DNA, and is used to manufacture DNA for sequencing
replicative form (RF) Double-stranded form of the genome of a single-stranded DNA (or RNA) virus. The RF first replicates itself and is then used to generate the ssDNA (or ssRNA) to pack into the virus particles
Sequenase® Genetically modified DNA polymerase from bacteriophage T7 used for sequencing DNA

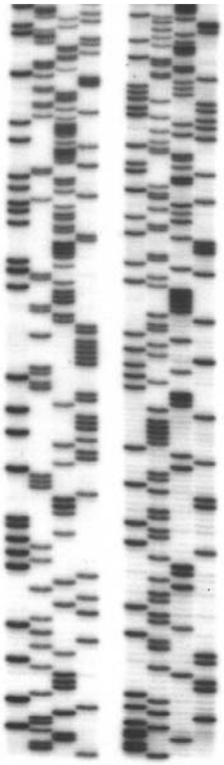


FIGURE 24.08
Autoradiograph of Real Sequencing Gel

The sequencing of two different DNA templates is shown. The two sequences each consist of four lanes that represent the four different bases. The sequence is read from the bottom of the gel toward the top. This gel was run by Kiswar Alam in the Author's laboratory.

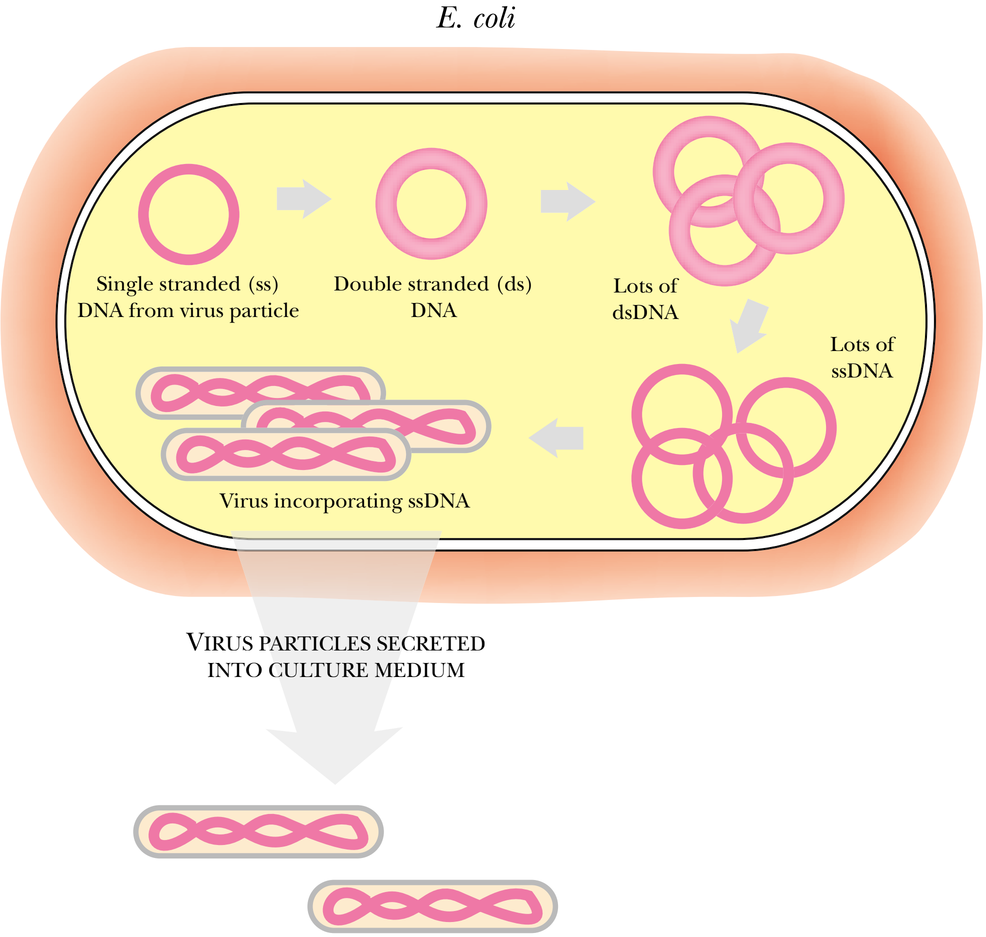

E. coli
Single stranded (ss) DNA from virus particle
Double stranded (ds) DNA
Lots of dsDNA
Lots of ssDNA
Virus incorporating ssDNA
VIRUS PARTICLES SECRETED INTO CULTURE MEDIUM

FIGURE 24.09 ***Single-Stranded DNA from Bacteriophage M13***

When M13 infects *E. coli*, the single-stranded viral DNA is converted to a double-stranded replicative form (RF). This RF then replicates so making many double-stranded copies. After the double-stranded form becomes abundant, large numbers of single-stranded copies are made. These are eventually packaged into viral particles that are secreted into the culture medium.

Single stranded DNA for use as a template can be generated by taking advantage of bacteriophage M13, which naturally contains ssDNA.

to manufacturing large numbers of single-stranded circles of DNA to pack into newly made virus particles.

Not only does M13 generate single-stranded DNA but it also purifies it. Unlike most viruses, M13 doesn't destroy the bacterial cells. Instead, the cells continuously secrete virus particles containing ssDNA into the surrounding medium. In addition, since the viral DNA does not integrate into the bacterial chromosome, only viral DNA gets packaged into the particles. Since the viral particles are secreted, they are easily isolated from the bacterial cells, and the DNA they contain can be extracted.

The template DNA that has unknown sequence is first cloned into the double-stranded replicative form of M13. Normally, an M13 vector that has already been engineered to contain a convenient multiple cloning site is used (Fig. 24.10). This multiple cloning site is contained within the N-terminal fragment of the *lacZ* gene of *E. coli*, which allows the use of blue/white screening to monitor insertion of the template DNA into the M13 vector (see Ch. 22 for details). Furthermore, the sequence to the side of the inserted DNA is already known and provides a starting point. This is essential, as the primer for sequencing must be complementary to a known sequence on the template strand in order to hybridize in the correct position. This engineered virus is used to infect *E. coli*, and virus particles containing single strands are manufactured in large quantities. Nowadays, bacterial plasmids containing the M13 origin of replication are used to manufacture single-stranded DNA. The use of intact virus is avoided and improved yields of DNA can be obtained more conveniently.

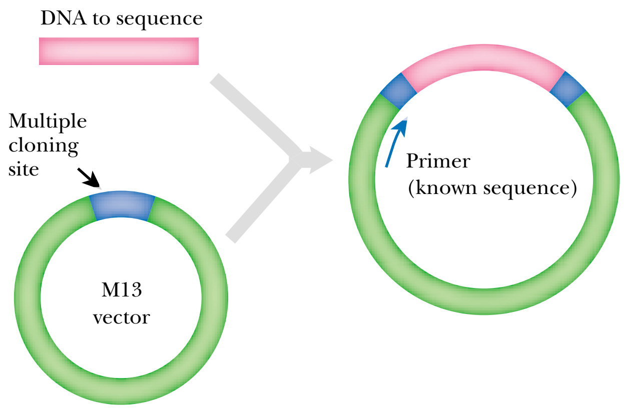


FIGURE 24.10 ***Sequencing Using M13-Based Vectors***

The use of M13 vectors allows the easy production of single-stranded template DNA. The DNA to be sequenced is inserted into the multiple cloning site (MCS) within the M13 vector. The MCS is located within the alpha fragment of the *lacZ* gene. When no insert is present functional β-galactosidase is made, which turns the *E. coli* host blue in the presence of X-gal. When an insert disrupts *lacZ*, no functional β-galactosidase is made and the cells stay white. This allows simple identification of cells carrying M13 vectors that have received DNA inserts. Sequencing is carried out using primers corresponding to M13 sequences just outside the cloned DNA.

A variety of technical improvements have made DNA sequencing a little less tedious. Using double-stranded DNA (dsDNA) directly for sequencing is more convenient than generating single strands. In reality, "double stranded" DNA sequencing involves a preliminary step, either heat or alkali treatment, to denature the dsDNA into single strands. Therefore, the actual sequencing reactions use single stranded DNA just as described above.

PCR products can be used for sequencing after separating into single strands.

In fact, it is now possible to completely avoid cloning DNA into either M13 or a plasmid vector by using PCR to generate segments of DNA (see Ch. 23). PCR products are linear double-stranded lengths of DNA and they can be directly sequenced after separation into single strands. As noted in Chapter 23, one drawback of PCR is that we need to know enough sequence on each side of the target DNA to construct primers for PCR. Hence, we cannot always avoid cloning.

Primer Walking along a Strand of DNA

Sequencing moderately long pieces of DNA was originally done by cutting the DNA into smaller segments with restriction enzymes and then sub-cloning each fragment separately into M13 or another vector. Nowadays, this has largely been replaced by **primer walking** (Fig. 24.11). This involves first sequencing the cloned DNA as far as possible using the primer belonging to the M13 or plasmid vector. Next, the newly obtained sequence information is used to design another primer. Sequencing is continued as far as this allows. Then another primer is made, and another, and so on until the end of the cloned DNA is reached.

Automated Sequencing

Today, the majority of sequencing is done using automated techniques. The main modification here is to use fluorescent dyes to label the DNA instead of radioactivity. Each of the four sequencing reactions is done just as before, but the DNA is labeled by

primer walking Approach to sequencing a long cloned DNA molecule by using successive primers located at stages along the molecule

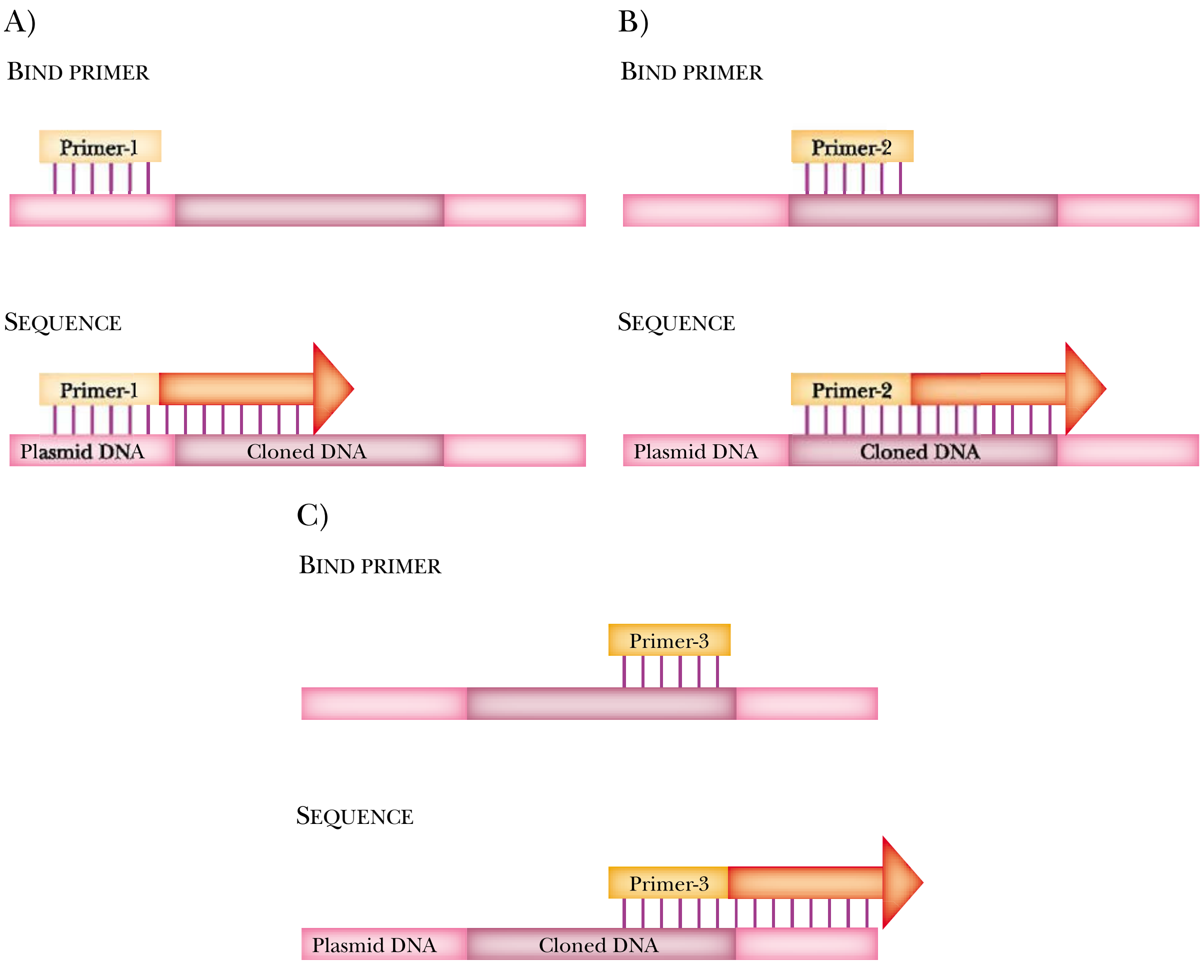


FIGURE 24.11 ***Primer Walking along a DNA Molecule***

When the DNA to be sequenced is too long to be sequenced by a single reaction, primer walking is used. First (A), the cloned DNA is sequenced as far as possible starting from a primer binding site within the vector. The sequence information obtained allows a second primer to be made that lies close to the far end of the known sequence. A second sequencing reaction with this primer provides a further stretch of sequence (B). This process is continued, using as many primers as necessary to cross the inserted DNA. Eventually, the sequence obtained corresponds to the vector (C). This tells the experimenter that the unknown DNA has been completely crossed and is now fully sequenced.

attaching a fluorescent dye to the primer before running the reactions. Although the same DNA primer is used for each of the four reactions, four fluorescent dyes of different colors are needed. The first color is used when carrying out the A-reaction, another one for G, another for T, and the fourth for C. When the sequencing gel is run, bands of four different colors are seen, a separate color for each base. In fact, since the bases are color coded, all four completed reactions can be run in the same track on the sequencing gel, as shown in Figure 24.12.

Automated sequencing relies on using four fluorescent dyes of different colors, one for each base. This allows all fragments to be run in a single gel lane, where they are scanned by a laser.

Rather than running the gel for a fixed period and then examining the bands afterwards, the gel is monitored while running. As the bands move down the gel they pass a laser and detector assembly. The laser beam scans the bands and the four different dyes fluoresce in different colors. A computer records the color of each band, and compiles the data into actual sequence. The first bands to be recorded will have run right through the gel and off the end while later bands are still passing the laser. Consequently, more bases can be read from a single sequencing reaction by the continuous flow approach. Automated sequencers have been improved by using capillary separation. This improves speed but more importantly has allowed the assembly of machines that may have as many as 96 sequencing reactions running simultaneously.

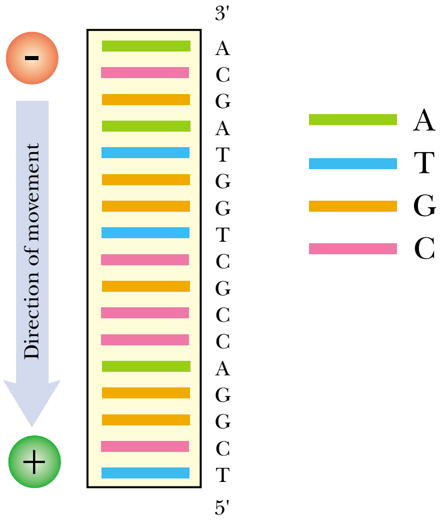


FIGURE 24.12 ***Automated Fluorescent DNA Sequencing***

Automated sequencing uses four different fluorescent dyes, one for each of the four bases. All four reactions are run in a single lane of the gel since the four bases are easily distinguished by their colors.

The Emergence of DNA Chip Technology

DNA arrays are used for a variety of purposes, including sequencing. Large numbers of probes are bound to the chip and hybridization with target DNA occurs on the chip surface.

Earlier DNA technology was largely based on gel electrophoresis, an approach that is both difficult to automate and labor intensive. DNA chips were developed to allow automated side-by-side analysis of multiple DNA sequences. In practice the simultaneous analysis of thousands of DNA sequences is possible. The first chip was introduced by a company called Affymetrix in California in the early 1990's. Since then DNA chips have been used for a variety of purposes including sequencing, detection of mutations and gene expression. DNA chips all rely on hybridization between single-stranded DNA permanently attached to the chip and DNA (or RNA) in solution. Many different DNA molecules are attached to a single chip forming an array of spots on a solid support (the chip). The DNA or RNA to be analyzed must be labeled, usually with fluorescent dyes. Hybridization at each spot is scanned and the signals are analyzed by appropriate software to generate colorful data arrays. Two major variants of the DNA chip exist. Earlier chips mostly used short oligonucleotides. However, it is also possible to attach full length cDNA molecules. Prefabricated cDNA or oligonucleotides may be attached to the chip. Alternatively, oligonucleotides may be synthesized directly onto the surface of the chip by a modification of the phosphoramidite method described in Chapter 21. Modern arrays may have 100,000 or more oligonucleotides mounted on a single chip.

The Oligonucleotide Array Detector

DNA arrays can detect the presence of multiple small fragments of DNA sequence. A computer then compiles the overall sequence.

The **oligonucleotide array detector** simultaneously detects and identifies lots of short DNA fragments (i.e., oligonucleotides). It can be used both for diagnostic purposes and for large scale DNA sequencing. The key principle involved is DNA-DNA hybridization (see Ch. 21).

Consider a piece of DNA of unknown sequence. This is denatured to give single strands and one of these is tested for hybridization to a known probe sequence of say, eight bases (an octonucleotide; e.g., CGCGCCCG). If the unknown DNA binds to the probe, then the probe sequence occurs somewhere in the complementary strand of the unknown DNA. The unknown DNA is then tested for hybridization to all other possible stretches of eight bases, one at a time, to see which are found.

oligonucleotide array detector Chip used to simultaneously detect and identify many short DNA fragments by DNA-DNA hybridization. Also known as DNA array or DNA chip

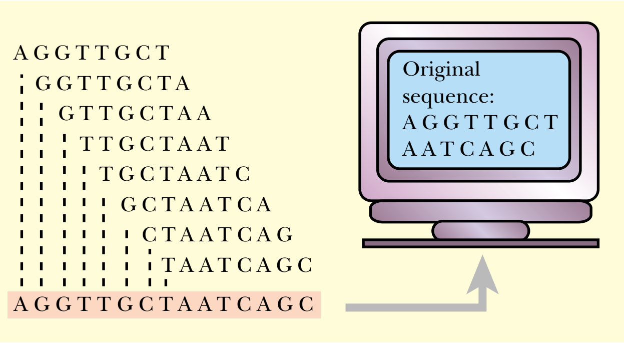


FIGURE 24.13 ***Deducing Sequence From Oligonucleotide Overlaps***

By aligning all of the eight-base pair probe sequences that hybridize to the unknown DNA, the computer can determine the sequence of the DNA.

In practice, the hybridizations are all carried out at once. There are actually 65,536 possible eight-base sequences. Samples of each of the eight-base sequences to be used as probes are arranged in a square array and anchored to the surface of a glass chip. The glass chip can then be dipped in a solution of the target DNA, which will hybridize simultaneously to all those eight-base sequences with which it has complementary sequences. Instead of nucleotide array detector, the array is simply called a **DNA chip**. The technology is so precise that an array about 1 cm square can carry up to a million nucleotide probe sequences.

For example, if the unknown sequence of DNA is TCCAACGATTAGTCG, then its complementary strand will be AGGTTGCTAATCAGC. Consequently, of all 65,536 possible eight-base sequences, only the following can hybridize with the original sequence:

```
AGGTTGCT  TAATCAGC  TGCTAATC  GCTAATCA
GTTGCTAA  GGTTGCTA  TTGCTAAT  CTAATCAG
```

Given this information, a computer program can test all possible overlaps for these eight-base sequences and generate the solution as shown in Figure 24.13.

To sequence a large piece of DNA, it is first broken into relatively small pieces and tagged with a fluorescent dye. The unknown DNA will bind to just a few of the many eight-base probes on the oligonucleotide array. The chip is then scanned by a laser, which locates the fluorescently tagged DNA. The positions to which it has bound are recorded. The computer then calculates the complete sequence of the unknown DNA. For simplicity, the oligonucleotide array illustrated in Figure 24.14 is designed to detect all possible four-base sequences. The "unknown" sequence, ACTGGC, contains three overlapping four-base sequences: ACTG (No. 1), CTGG (No. 2), and TGGC (No. 3). Their positions are shown on the array.

Arrays cannot deal well with repetitive sequences but are extremely good at checking for mutations.

The oligonucleotide array runs into difficulties if the target DNA contains repeated sequences. Therefore, conventional sequencing is performed on totally new DNA sequences. However, for diagnostic tests to check for hereditary defects (i.e., mutations in known genes) and for forensic analysis, the oligonucleotide array is faster and simpler. The first **GeneChip® array**, made by Affymetrix Corporation, was designed to detect mutations in the reverse transcriptase gene of the AIDS virus. A variety of others are now available both for genome analysis and for diagnostic purposes, such as checking for mutations in the *p53* or *BRCA1* cancer genes. DNA arrays are also used for the whole genome analysis of gene expression as described in Ch. 25.

DNA chip Chip used to simultaneously detect and identify many short DNA fragments by DNA-DNA hybridization. Also known as DNA array or oligonucleotide array detector
GeneChip® array The first brand of DNA chip, made by Affymetrix Corporation

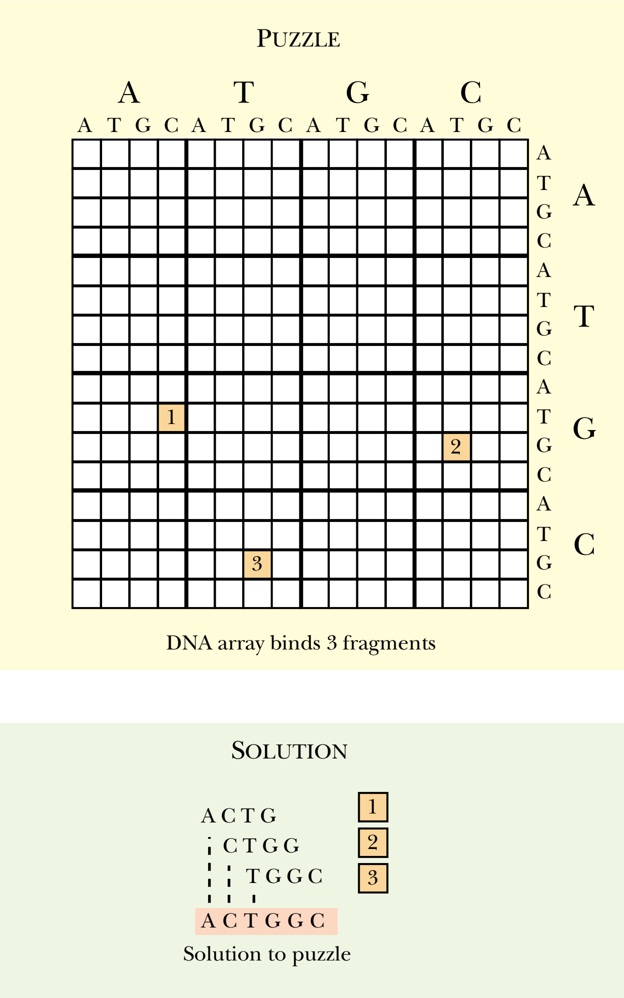

PUZZLE
A T G C
A T G C A T G C A T G C A T G C
A T G C A T G C A T G C A T G C
A T G C
1
2
3
DNA array binds 3 fragments
SOLUTION
A C T G
C T G G
T G G C
1
2
3
A C T G G C
Solution to puzzle

FIGURE 24.14 ***Sequencing by Oligonucleotide Array***

This example shows an oligonucleotide array that has every four base pair combination possible. The unknown DNA fragment is fluorescently tagged and allowed to hybridize to all the possible oligonucleotides on the chip. The first spot that hybridized to the unknown DNA has the sequence, ACTG. The second has the sequence, CTGG, and the third, TGGC. A computer assembles these into the correct overlapping order. The sequence is then determined based on this information.

Pyrosequencing

Pyrosequencing is used to analyze short regions of DNA by monitoring light release.

Pyrosequencing is a "mini-sequencing" method that can be automated. In practice it is used for short regions in multiple individuals rather than for sequencing long regions of unexplored DNA. The scheme relies on the generation of a light pulse, by a coupled reaction, each time a base is incorporated (Fig. 24.15). This means that each of the four bases must be added in turn, one at a time, to the reaction mix. If a light pulse is seen, the added base was incorporated (and therefore the sequence being analyzed contains the complementary base at this point). The bases are added as their nucleoside triphosphates (dNTPs). Each time a nucleotide is incorporated, pyrophosphate is released. ATP sulfurylase converts adenosine phosphosulfate (APS) plus pyrophosphate to ATP. This in turn allows firefly luciferase to oxidize luciferin and release a pulse of light. Unused dNTP and ATP are degraded by the enzyme apyrase before the next dNTP is added. Since both ATP and dATP are used by luciferase, dATP cannot be used for nucleotide incorporation. Instead, an analog that is used by DNA polymerase but not by luciferase is added. This is typically α-thio-dATP (in which the first phosphate is replaced by sulfate).

Pyrosequencing is especially useful if the DNA sequence is already known and variants that differ in one or a few bases are being compared. In this case, bases that correspond to the known sequence are added until one of them gives no light signal—indicating that a sequence alteration is present at that point. Then the other three bases are tried until one gives a response and reveals the base present in the particular sample or individual being analyzed.

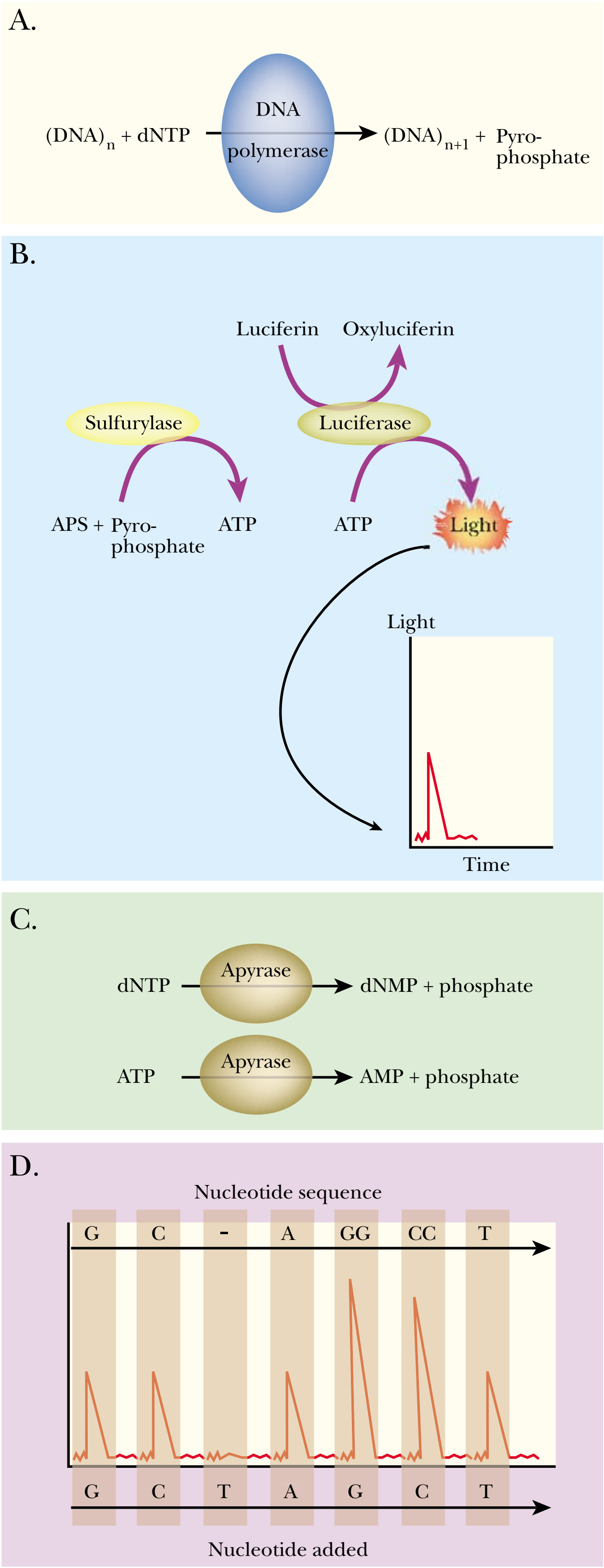


FIGURE 24.15 ***Principle of Pyrosequencing***

A. During each sequencing reaction, DNA is elongated by one nucleotide and pyrophosphate is released. B. The pyrophosphate is used together with adenosine phosphosulfate (APS) by ATP sulfurylase to generate ATP. Luciferase uses ATP plus luciferin and emits light. C. Apyrase removes unused triphosphates. D. An example of a short sequence generated by pyrosequencing. Modified after material kindly provided by Pyrosequencing AB, Uppsala, Sweden.

Nanopore Detectors for DNA

Nanopore detectors admit a single DNA strand through a tiny pore and sequence it as it passes through.

Nanotechnology is based on microscopic machinery that operates at the level of single molecules. **Nanopore detectors** for DNA contain extremely narrow pores that permit a single strand of DNA to pass through one at a time. As the DNA molecule transits the pore, a detector records its presence and its characteristics. Ultimately this should result in a novel method for sequencing individual DNA molecules one at a time. The advantages of nanopore technology are its high speed and its ability to handle long DNA molecules.

A practical nanopore detector consists of a channel in a membrane that separates two aqueous compartments. When a voltage is applied across the membrane, ions flow through the open channel. Since DNA is negatively charged, the DNA is pulled through the nanopore to the positive side. The DNA molecules enter the pore, and are pulled through in extended conformation, one at a time (Fig. 24.16). During the time the channel is occupied by the DNA, the normal ionic current is reduced. The amount of reduction depends on the base sequence (G > C > T > A), therefore, a computer can measure the current, and decipher the sequence based on the differences.

Initial demonstrations have used alpha-hemolysin from *Staphylococcus* as the channel and a lipid bilayer as the membrane. The mouth of the alpha-hemolysin channel is about 2.5 nm wide—roughly 10 atomic diameters. Double-stranded DNA can enter the pore mouth, but toward the middle, the channel narrows to less than 2 nm, which prevents dsDNA from going any further. The dsDNA remains stuck until the strands separate, allowing single-stranded DNA to pass through the length of the pore.

At present, nanopore detectors can tell apart two 20 nucleotide DNA strands that differ in a single base. It takes approximately a microsecond per base for DNA to transit the pore. Although single-stranded DNA molecules 1000 bases long have been successfully pulled through nanopores, they move so fast through the pores that it is difficult to detect individual bases. Estimates based on future technical improvements suggest the possibility of chips with 500 pores each reading 1,000 bases/second. This could in theory read a bacterial genome in around a minute and read the entire human genome (3×10^9 bases) in less than two hours.

Large Scale Mapping with Sequence Tags

Sequence tags are merely short regions of known sequence in known locations. They are needed when genomes contain vast amounts of non-coding DNA.

Ultimately, all the DNA sequence fragments of a genome must be correlated with a genetic map showing the location of the genes. However, in higher organisms, genes only comprise a small proportion of the DNA. Thus it is necessary to have a series of genetic markers other than genes themselves. We have already discussed sequence motifs such as RFLPs (Ch. 22) and VNTRs and microsatellites (see Ch. 4). However, even the combination of these motifs is not sufficiently specific to map a large genome. Therefore, the human genome was mapped largely by the use of **sequence tagged sites (STSs)** including **expressed sequence tags (ESTs)**.

A sequence tagged site is simply a short sequence (usually 100–500 bp) that is unique and can be detected by PCR. It is especially important to avoid the repetitive sequences that are found in large numbers on eukaryotic chromosomes. In fact, identification by PCR requires two short sequences to which the PCR primers hybridize, separated by a known length of DNA. The STS is amplified by PCR to give a DNA fragment of specific length (Fig. 24.17).

Expressed sequence tags come from transcribed regions of the DNA.

An expressed sequence tag (EST) is a special type of STS derived from a region of DNA that is expressed, i.e. transcribed into mRNA. Purified mRNA is used to generate the corresponding cDNA by reverse transcriptase. The cDNA is then amplified

expressed sequence tag (EST) A special type of STS derived from a region of DNA that is expressed by transcription into mRNA
nanopore detector Detector that allows a single strand of DNA through a molecular pore and records its characteristics as it passes through
sequence tagged site (STS) A short sequence (usually 100–500 bp) that is unique within the genome and can be easily detected, usually by PCR

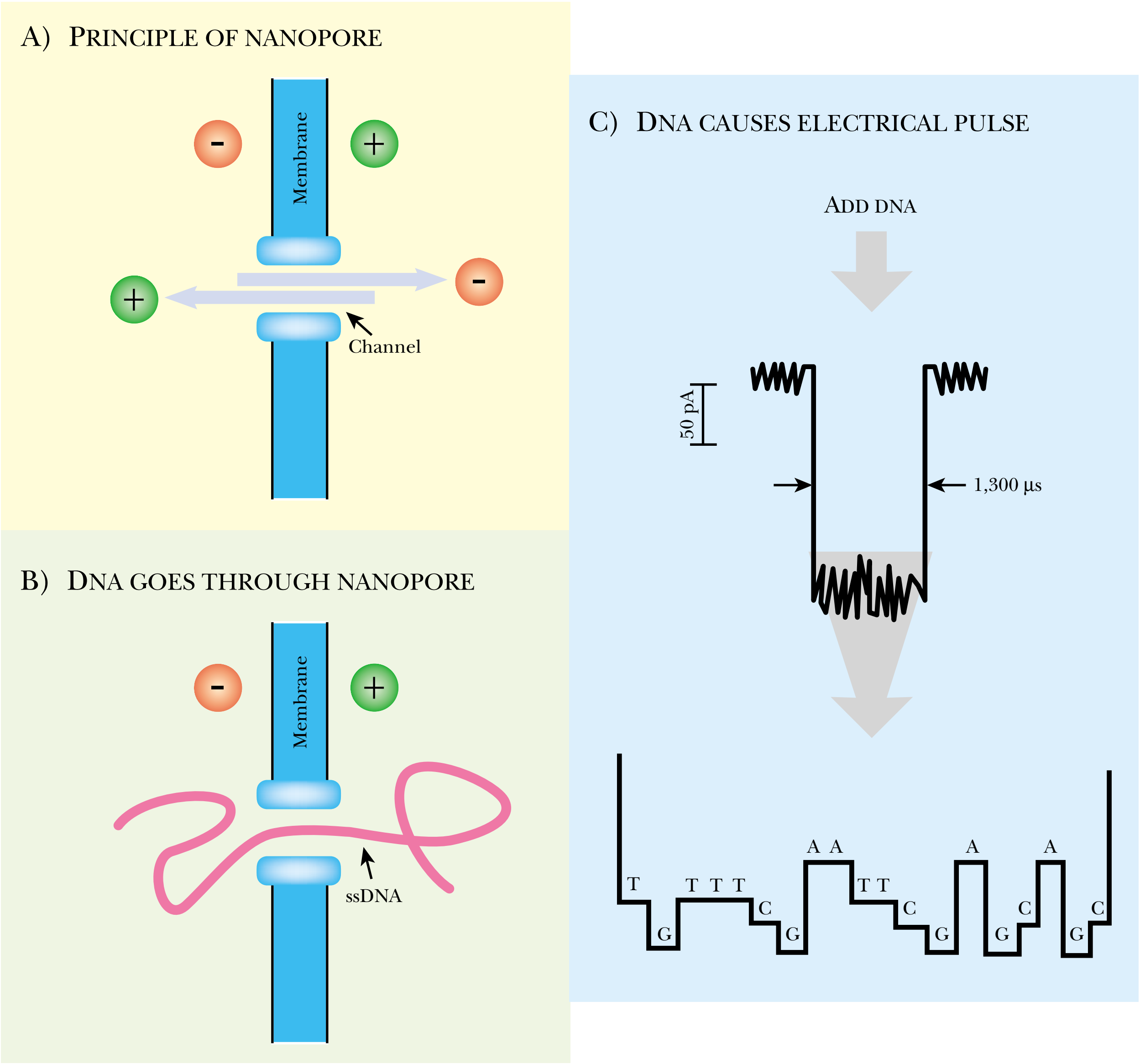


FIGURE 24.16 ***Principle of the Nanopore Detector***

(A) Nanopores are small openings in a membrane that only allow one molecule through at a time. The nanopore membrane separates two compartments of different charge. (B) Since there is a charge separation between compartments, negatively charged molecules like DNA can pass through the pore in an extended conformation. (C) While the DNA is passing through the pore, a detector measures how much the current, due to normal ion flow, is reduced. Since each base alters the current by different amounts, the detector can determine the sequence as the DNA passes through the pore.

by PCR. An EST represents only a small section of a gene, thus it is possible to have several ESTs from the same gene. To avoid such duplication, ESTs are usually derived from the 3′-untranslated regions of the mRNA.

Mapping of Sequence Tagged Sites

To map STS or EST sites a collection of chromosome fragments is examined for the presence of each STS or EST (Fig. 24.18). The fragments may be derived from a single chromosome or the whole genome. The chance that any two STSs will be on the same

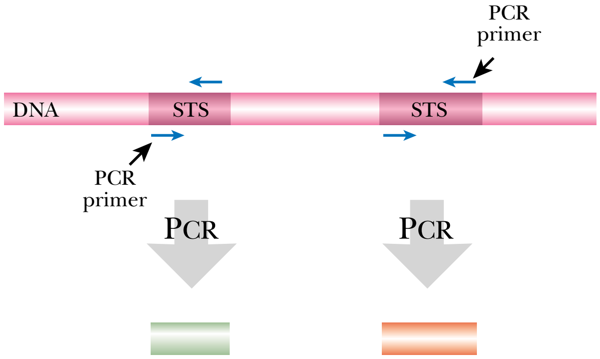


FIGURE 24.17 ***PCR Detection of Sequence Tagged Site***

Mapping large genomes requires many genetic markers. Those based solely on their unique sequences are known as sequence tagged sites (STS). These unique sequences may be amplified by PCR and mapped relative to each other. In practice each STS is defined by a pair of specific PCR primers at its ends that generate a PCR fragment of defined length.

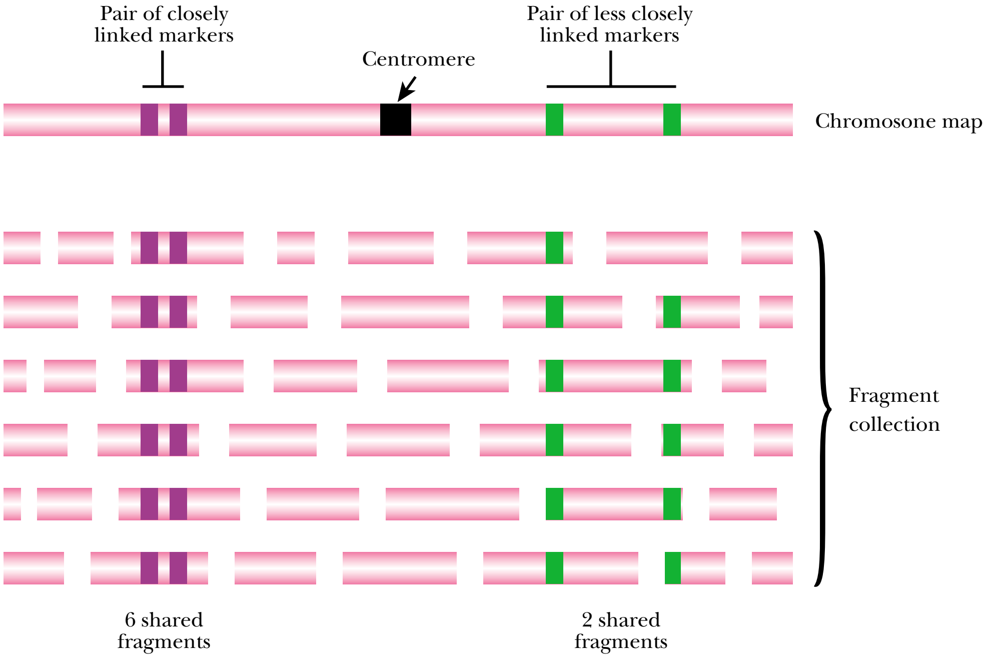


FIGURE 24.18 ***Mapping of Sequence Tagged Sites***

STS mapping is shown for four STS sites on a single chromosome. A variety of restriction enzyme digests are performed to cut the chromosome into many different sized fragments. The number of times two STS sequences are found on the same fragment reveals how close the two markers are to each other. In this example, the two purple STSs are found on the same fragment six times, and must be close to each other on the chromosome. The two green STSs are only found on the same fragment two times and are therefore further apart. The purple STSs are never found on the same fragment as the green STSs, therefore they must be far apart on the chromosome. Continuing to add different STSs will refine the map even more.

Sequence tags are mapped relative to each other by analysing how frequently tags are found together on the same chromosome fragments.

fragment depends how close they are on the original chromosome. Neighboring STSs will tend to be found together on many fragments whereas those further apart will only rarely be found on the same fragment. This type of data can be used to construct a linkage map for the STS sites examined.

The chromosome fragments to be examined were originally derived by cloning large segments of DNA into high capacity vectors such as yeast artificial chromosomes

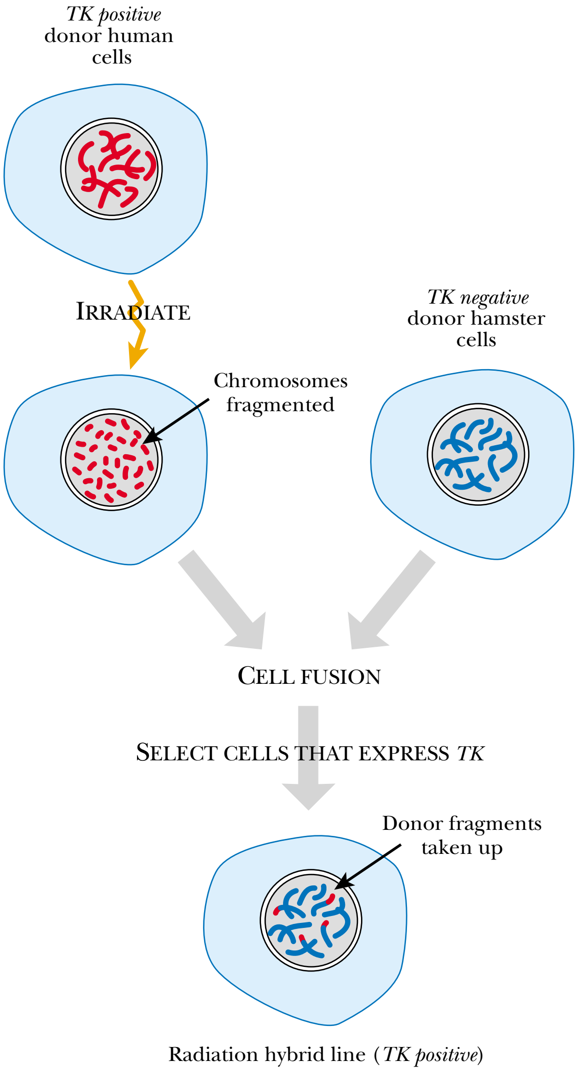

TK positive donor human cells
IRRADIATE
TK negative donor hamster cells
Chromosomes fragmented
CELL FUSION
SELECT CELLS THAT EXPRESS *TK*
Donor fragments taken up
Radiation hybrid line (*TK positive*)

FIGURE 24.19 ***Radiation Hybrid Mapping***

To determine how close STSs and ESTs are to each other, many large chromosome fragments must be analyzed. Radiation hybrid mapping allows large human chromosome fragments to be inserted into hamster cells. First, the human chromosomes, which have the gene for thymidine kinase (TK^+), are fragmented by irradiation. The human cells are then fused with hamster cells, which are TK^-. If a human cell and hamster cell fuse successfully, the hybrid should express thymidine kinase and can be selected by plating on selective medium. Random loss of human chromosome fragments occurs during this process. Consequently, each radiation hybrid cell line will contains a different set of human chromosome fragments, which can be screened for the presence of STSs and ESTs.

(YACs). Unfortunately, YAC clones often contain two or more segments of DNA from different original locations. In practice, locating STS and EST sites relative to each other has mostly been done by radiation hybrid mapping (Fig. 24.19). A **radiation hybrid** is a cell (usually from a rodent) that contains fragments of chromosomes from another species.

Radiation hybrids are cell lines that contain fragments of chromosomes from other eukaryotic cells.

To make a radiation hybrid, cultured human cells are irradiated with a lethal dose of X-rays or γ-rays. This treatment breaks the chromosomes into fragments. The dying human cells are then fused with hamster cells. Cell fusion is promoted with polyethylene glycol or by using Sendai virus. The resultant hybrid cells contain random selections of the human chromosome fragments. Typical fragments are 5–10 Mbp in size and each hamster cell contains about 15–35% of the human genome. The hybrid cells are screened to see which STSs or ESTs are found together—i.e. are on the same human chromosome fragments. The more often two STSs are found in the same hybrid cell, the closer they are linked on the original human chromosome before fragmentation. By the late 1990s, STS-based maps of over 30,000 sites had been constructed for the human genome. This gives a density of approximately one marker per 100 kbp of DNA.

radiation hybrid A cell (usually from a rodent) that contains fragments of chromosomes (generated by irradiation) from another species

Assembling Small Genomes by Shotgun Sequencing

Individual sequencing reactions give lengths of sequence that are several hundred base pairs long. A whole genome must be assembled from vast numbers of such short sequences. There are three approaches to whole genome assembly: shotgun sequencing, cloned contig sequencing, and the directed shotgun approach which is really a mixture of the first two.

Sequencing very large numbers of small fragments provides enough information to assemble a complete genome sequence—if your computer is powerful enough.

In **shotgun sequencing** the genome is broken randomly into short fragments (1 to 2 kbp long) suitable for sequencing. The fragments are ligated into a suitable vector and then partially sequenced. Around 400–500 bp of sequence can be generated from each fragment in a single sequencing run. In some cases, both ends of a fragment are sequenced. Computerized searching for overlaps between individual sequences then assembles the complete sequence. Overlapping sequences are assembled to generate **contigs** (Fig. 24.20). The term contig refers to a known DNA sequence that is contiguous and lacks gaps.

Since fragments are cloned at random, duplicates will quite often be sequenced. To get full coverage the total amount of sequence obtained must therefore be several times that of the genome to allow for duplications. For example, 99.8% coverage requires a total amount of sequence that is 6 to 8-fold the genome size. In principle all that is required to assemble a genome, however large, from small sequences is a sufficiently powerful computer. No genetic map or prior information is needed about the organism whose genome is to be sequenced. The original limitation to shotgun sequencing was the massive data handling that is required. The development of faster computers has overcome this problem. Nowadays, a more important issue is that repetitive sequences create ambiguities.

The bacterium *Haemophilus* had the honor of being the first organism to be totally sequenced.

The first bacterial genome to be sequenced, that of *Haemophilus influenzae*, was deduced from just under 25,000 sequences averaging 480 bp each. This gave a total of almost 12 million bp of sequence—six times the genome size. Computerized assembly using overlaps resulted in 140 regions of contiguous sequence—i.e. 140 contigs.

The gaps between the contigs may be closed by more individualistic procedures. The easiest method is to re-screen the original set of clones with pairs of probes corresponding to sequences on the two sides of each gap. Clones that hybridize to both members of such a pair of probes presumably carry DNA that bridges the gap between two contigs. Such clones are then sequenced in full to close the gaps between contigs. However, many of the gaps between contigs are due to regions of DNA that are unstable when cloned, especially in a multicopy vector. Therefore a second library in a different vector, often a single-copy vector such as a lambda phage, is often used during the later stages of shotgun cloning. Pairs of end-of-contig probes are used to screen the new library for clones that hybridize to both probes and carry DNA that bridges the gap between the two contigs (Fig. 24.21A). A third approach, that avoids cloning altogether, is to run PCR reactions on whole genomic DNA, using random pairs of PCR primers corresponding to contig ends. A PCR product will result only if the two contig ends are within a few kb of each other (Fig. 24.21B).

Race for the Human Genome

As described above, the first completely sequenced genomes of living cells were from bacteria with genomes consisting of a few million base pairs. The genomes of most higher animals and plants contains a thousand fold more DNA and their sequencing therefore presents greater problems. The original aim of the **Human Genome Project**

contig A stretch of known DNA sequence that is contiguous and lacks gaps
Human Genome Project Program to sequence all the DNA making up the human genome
shotgun sequencing Approach in which the genome is broken into many random short fragments for sequencing. The complete genome sequence is then assembled by computerized searching for overlaps between individual sequences

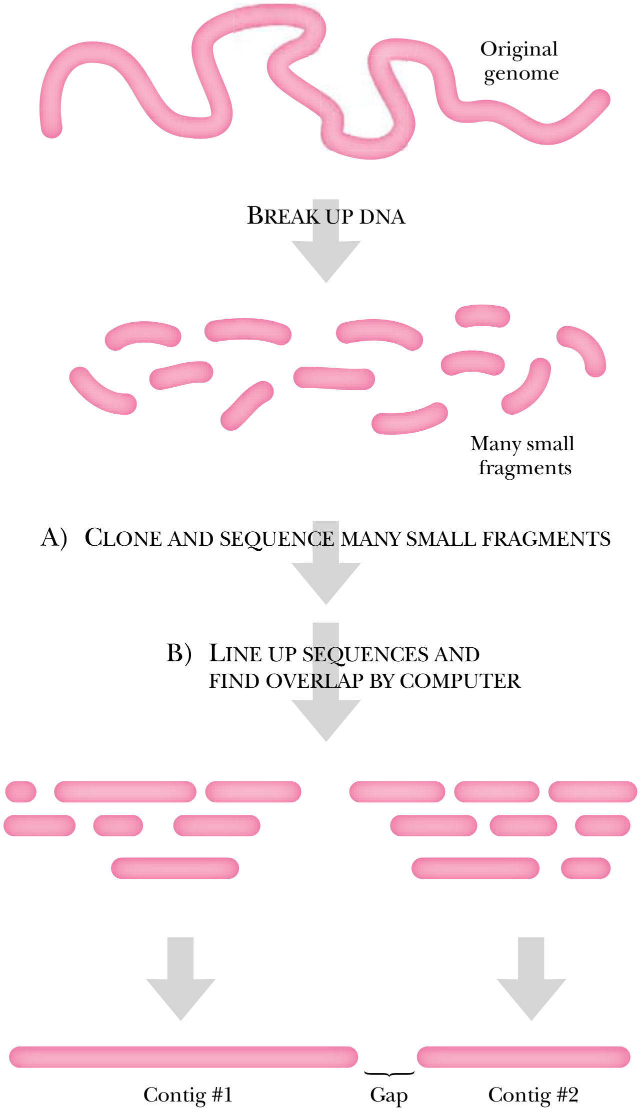

Original genome
BREAK UP DNA
Many small fragments
A) CLONE AND SEQUENCE MANY SMALL FRAGMENTS
B) LINE UP SEQUENCES AND FIND OVERLAP BY COMPUTER
Contig #1
Gap
Contig #2

FIGURE 24.20 ***Shotgun Sequencing***

The first step in shotgun sequencing an entire genome is to digest the genome into a large number of small fragments suitable for sequencing. All the small fragments are then cloned and sequenced. Computers analyze the sequence data for overlapping regions and assembled the sequences into several large contigs. Since some regions of the genome are unstable when cloned, some gaps may remain even after this procedure is repeated several times.

was to sequence the DNA making up the entire human genome. When it first began in 1990, the official Human Genome Project was to be completed by 2005. In 1990, it cost $10 to determine each base of sequence, but technological advances reduced the cost of sequencing to 50 cents per base by the late 1990s. Therefore, the Human Genome Project was to cost roughly 1.5 billion dollars.

The first draft of the human genome appeared in early 2001.

However, in 1998 an upstart private venture, Celera Genomics, led by Craig Venter, claimed that it would finish the job by the end of 2001 at a cost of a mere $200,000. To prove it was serious, Celera Genomics sequenced the entire genome of the fruit fly, *Drosophila*, between May and December of 1999. Three million random sequence reads were assembled to give the 180 Mb genome, demonstrating the feasibility of the shotgun approach. The private and public Human Genome Projects jointly announced that the first draft of the human genome was complete in June 2000. In fact, the Celera sequence was 99% complete whereas the public project was only 85% done. A working draft was published in February of 2001.

The official Human Genome Project proceeded by cloning large fragments of human DNA, mostly in yeast artificial chromosomes (YACs) and bacterial artificial chromosomes (BACs), and mapping them to their chromosomal locations. Only then were the large mapped fragments broken up for shotgun sequencing. Although this makes assembly easier, cloning consumes time and money. Instead, Celera used the

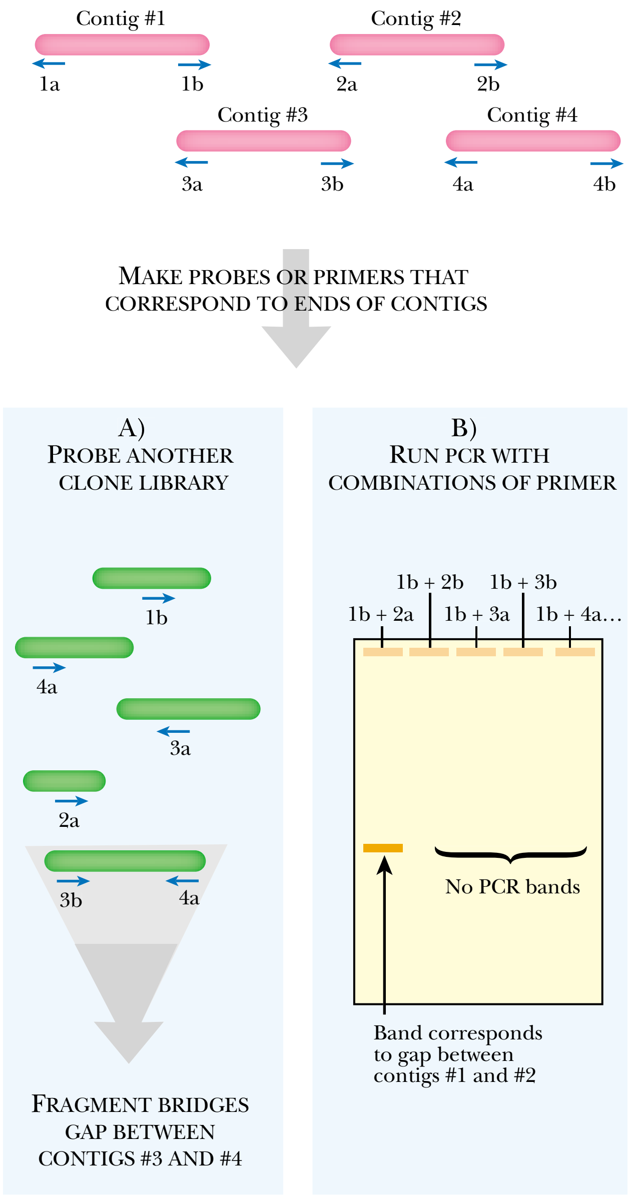

Contig #1
1a
1b
Contig #2
2a
2b
Contig #3
3a
3b
Contig #4
4a
4b
MAKE PROBES OR PRIMERS THAT CORRESPOND TO ENDS OF CONTIGS
A)
PROBE ANOTHER CLONE LIBRARY
1b
4a
3a
2a
3b
4a
FRAGMENT BRIDGES GAP BETWEEN CONTIGS #3 AND #4
B)
RUN PCR WITH COMBINATIONS OF PRIMER
1b + 2b
1b + 3b
1b + 2a
1b + 3a
1b + 4a…
No PCR bands
Band corresponds to gap between contigs #1 and #2

FIGURE 24.21 ***Closing Gaps between Contigs***

To identify gaps between contigs, probes or primers are made that correspond to the ends of the contigs (pink). In (A) a new library of clones (green) is screened with end-of-contig probes. Clones that hybridize to probes from two sides of a gap are isolated. In this example a probe for the end of contig #3 (3b) and the beginning of contig #4 (4a) hybridize to the fragment shown. Therefore, the sequence of this clone should close the gap between contig #3 and #4. The second approach uses PCR (B). PCR primers that correspond to the ends of contigs are combined in random pairs and used to amplify genomic DNA. If the primer pair is within a few kilobases of each other, a PCR product is made and can be sequenced.

directed shotgun sequencing approach described below. Automated sequencers sequenced vast numbers of small fragments. Finally, a computer examined all the random segments of sequence for overlaps and assembled them into contigs and, ultimately, into complete chromosomal sequences. Even with the help of an STS map, this method relies on colossal amounts of computing power to cope with the millions of separate short sequences. Such computers were not available until recently and the success of the Celera method is largely due to the rapid increase in computer power over the last few years.

Now that the human genome has been sequenced, the next major task is to identify all the genes and elucidate their functions. The claim that once we have sequenced ourselves we will understand all human disease is doubtful. We have known for years the complete gene sequences of several viruses, including HIV, yet no complete cure has emerged. Deducing the function of a protein given only the DNA sequence that encodes it is hazardous at best. Although DNA sequences are very useful, a great deal of experimental work must also be done to understand inherited defects.

Assembling a Genome from Large Cloned Contigs

Very large genomes may be broken up into large fragments which are cloned using YACs or BACs and then sequenced by the shotgun approach.

Cloning fragments that are as large as possible, but still are able to be manipulated *in vitro* is the first step to assembling a large genome from cloned contigs. The genome is first broken up into large fragments that are cloned to give a library of overlapping pieces. These fragments are inserted into high capacity vectors such as yeast artificial chromosomes (YACs) or bacterial artificial chromosomes (BACs), which may carry several hundred kb of DNA (see Ch. 22). Each fragment is then analyzed separately by shotgun sequencing. Hopefully this results in a complete contiguous sequence. Overall this approach yields a set of what are effectively large "cloned contigs". This approach may be used for eukaryotic genomes that contain vastly more DNA than bacteria. This is the approach that was taken by the official government sponsored human genome project.

After sequencing the individual cloned fragments, the next problem is to identify the overlapping regions among the clones. As described above for closing the gaps left after shotgun sequencing, hybridization and PCR methods may be used to identify the overlapping fragments. Other methods include screening the cloned contigs for similarities in restriction profiles and repetitive elements.

However, comparing large numbers of clones by such methods is slow and tedious when tackling very large genomes. The human genome of 3×10^9 bp would give 10,000 cloned fragments of 300,000 bp (the maximum size for BAC/YAC inserts)—even without the 6 to 8-fold redundancy necessary to ensure complete coverage. In order to sequence each of these large clones, approximately 600 reactions would have to be performed, assuming about 500 base pairs per reaction. If 80,000 clones were constructed, about 48 million different sequences would have to be assembled into the complete genome.

Assembling a Genome by Directed Shotgun Sequencing

For random shotgun sequencing of the human genome it would be necessary to sequence about 70 million small stretches of about 500 bp. This would give the necessary redundancy for 99.8% coverage of 3×10^9 bp. With 100 automatic sequencers generating 1000 sequences per day, this could be done in 700 days—i.e. roughly two years.

Using a map of sequence tagged sites greatly helps in the computations needed to assemble a genome from shotgun sequencing.

The critical issue is the assembly of these sequences into contigs and ultimately into complete chromosomes. The vast amount of computer time plus the uncertainties due to repetitive sequences make this approach prohibitive as it stands. However, if an STS map is used as a framework, then assembly becomes possible. In fact, this is the successful approach taken by Craig Venter of Celera genomics to complete the human genome ahead of schedule.

Sixty million sequences were generated from a library of fragments averaging 2 kb inserted into a multicopy plasmid vector. Another ten million sequences were from another library of larger pieces (10 kb) in a different vector. The 10 kb library is especially important in dealing with repeated sequences, since most of these are around 5 kb in size (or smaller) and can be entirely contained within a 10 kb fragment. The 2 kb library would not contain the entire repetitive region, and correctly aligning the numerous clones for this region would be impossible. Using the end-sequences of the 10 kb fragments allows the assembly process to avoid making incorrect overlaps between two identical repetitive sequences that are actually in different locations (Fig. 24.22).

Survey of the Human Genome

The sequence of the human genome still has a few gaps. These are mostly in the highly repetitive and highly condensed heterochromatin, which contains few coding

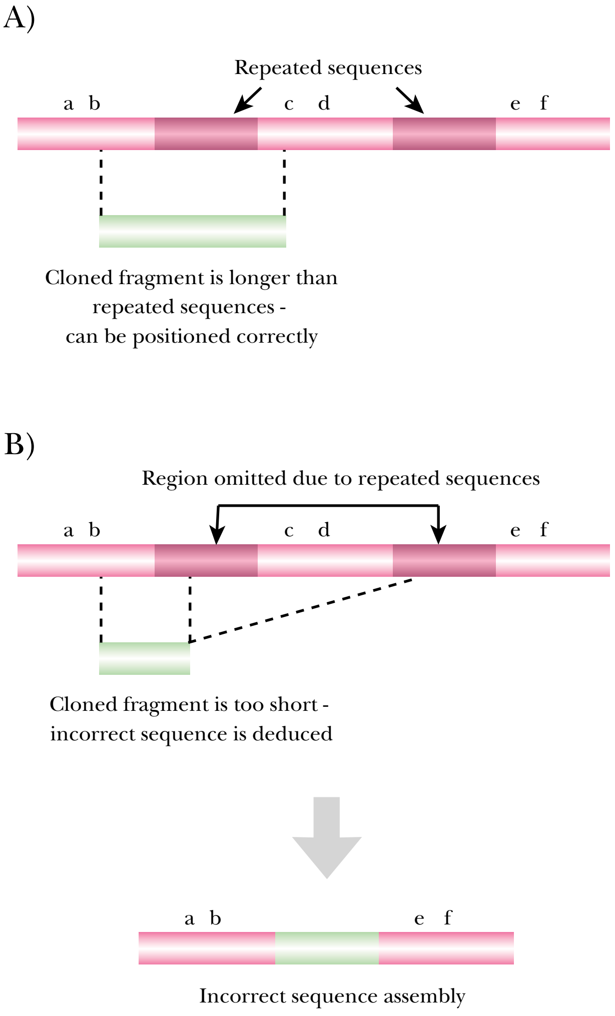

A)
Repeated sequences
a b
c d
e f
Cloned fragment is longer than
repeated sequences -
can be positioned correctly
B)
Region omitted due to repeated sequences
a b
c d
e f
Cloned fragment is too short -
incorrect sequence is deduced
a b
e f
Incorrect sequence assembly

FIGURE 24.22 ***Avoiding Incorrect Overlaps between Repetitive Sequences***

(A) If a clone is large enough to completely contain repetitive sequences, then the unique flanking sequences are able to position that clone within the genome. (B) If a cloned fragment is short relative to repetitive sequences, errors may result. Here the end of the green cloned fragment could align with either the first or second repetitive sequence; therefore, the sequence labeled c-d may be omitted. Determining the spacing and number of repetitive elements is virtually impossible with short clones.

Mammalian genomes have about 3 thousand million base pairs, but only 1% codes for proteins.

Repetitive sequences account for half the human genome.

sequences (see Ch. 4). The total estimated size of the human genome is 3,200 million (3.2×10^9) base pairs of DNA or 3.2 **Gigabase pairs** (Gbp; 1 Gbp = 10^9 base pairs) of which 2.95 Gb is euchromatin. A typical page of text contains about 3,000 letters. So the human genome would fill about a million pages. Most DNA from higher organisms is non-coding DNA, including intergenic regions, introns, repetitive sequences and so forth. About 28% of human DNA is transcribed into RNA but, since primary transcripts include introns, only a mere 1.25% is sequence that actually codes for proteins. On average, the introns are longer in human DNA than in other organisms sequenced so far. There are both AT-rich regions and GC-rich regions in the human genome. Curiously, the zones of GC-rich sequence have a higher density of genes and the introns are shorter. The significance of this is unknown.

Over half the human genome consists of repeated sequences (these are discussed in Ch 4). Some 45% is parasitic DNA (SINEs—13%; LINEs—20%; defunct retroviruses—8% and DNA-based transposons—3%). Repeats of just a few bases (microsatellites, VNTRs, etc.) account for 3% and duplications of large genome segments for 5%. Much of the genome resembles a retro-element graveyard, with only scattered outcrops of human information. The junk DNA tends to accumulate near the ends and close to the centromeres of the chromosomes.

How many genes the human genome contains may seem to be a simple question; however, computer algorithms to find genes are far from perfect, especially when surveying DNA that is mostly non-coding. Although the best estimates are probably

gigabase pair (Gbp) 10^9 base pairs

TABLE 24.01 Homology of Predicted Human Proteins

	%
No homology	1
Prokaryotes only	<1
Eukaryotes plus prokaryotes	21
Eukaryotes (including animals)	32
Animals (including vertebrates)	24
Vertebrates only	22

around 30,000 to 40,000 genes, analysis of the same human genome sequence has resulted in estimates of from 25,000 to 70,000 genes. Many predicted genes could be inactive pseudogenes and conversely, many genes may be overlooked especially if they consist of small exons interrupted by many long introns. Furthermore, different computer analyses may assign a particular exon sequence to different genes. Although all protein encoding genes must be transcribed, unfortunately the converse is not true. Non-gene sequences are transcribed relatively frequently and thus the presence of a transcript does not confirm the existence of a genuine gene. Determining the exact number of human genes will require very detailed analysis using a combination of laboratory and computer methods.

Identifying genes unambiguously is especially difficult in genomes with large amounts of non-coding DNA.

Are humans really more complex than other organisms? The revelation that humans only have around 25,000 genes rather than the previously estimated 100,000 upset many people. [Actually, the estimate of 100,000 human genes was little more than a guess based on inflated self-importance.] The lowly nematode worm *Caenorhabditis,* with approximately 18,000 genes, therefore has half as much genetic information as humans. The mouse, and presumably most of our fellow mammals, have essentially the same number of genes as humans. Those who apparently feel that human pride depends on having more genetic information than other organisms have retreated behind the claim that humans have more gene products than other organisms. This claim is based on the observation that alternative splicing generates multiple proteins from single genes and is more frequent in higher animals. Even so, most genes do not undergo alternative splicing and there is no particular reason to believe that humans indulge in more alternative splicing than other mammals—especially the chimpanzee with whom we share some 98.5% of our DNA sequence. Quibbling about which animal has most genes is in any case now moot. Sequencing has revealed that the genome of the rice plant contains 40,000 to 50,000 genes—some 10,000 more than humans. So it is the flowering plants who represent the peak of evolution, not us!

The highest numbers of genes are found in flowering plants, not animals.

Comparing the sequences of the 30,000 to 40,000 predicted human genes with other organisms yields some surprising results (Table 24.01). Less than 10% of our protein families are specific to vertebrates. More than 90% of the identifiable domains that make up human proteins are related to those of worms and flies. Most novel genes include previously evolved domains and thus appear to result from the re-shuffling of ancient modules.

Comparing the sequence of the human genome with other genomes originally suggested that just over two hundred human genes were apparently borrowed from bacteria during relatively recent evolution. Homologs of the genes were absent from the genomes of flies, worms and yeast, but were often found in bacteria as well as other vertebrates. Several independent horizontal transfer events were proposed as responsible. However, further phylogenetic analysis suggests that most of these are actually present in more ancient eukaryotes. In particular, homologs of many of these genes have been found in EST databases from lower eukaryotes whose whole genomes had not been fully sequenced at the time, such as the slime mold *Dictyostelium*.

It has been long known that humans have a poor sense of smell. The genome sequence has revealed why—many of our smell sensors are defective.

Comparison of gene families is beginning to reveal insights into behavior. Apes and monkeys have a poor sense of smell compared to most other mammals, and humans are the most defective. Most mammals have over a thousand closely related genes for olfactory receptors. These are the detector proteins that bind and detect molecules responsible for odors in the nasal lining. In the mouse essentially 100% of olfactory receptor genes are intact and functional, in chimps and gorillas 50% and in humans only 30%.

Not all genes code for proteins. Several thousand human genes produce non-coding RNA (rRNA, tRNA, snRNA, snoRNA, etc.,—see Ch. 12). Such non-coding RNA genes lack open reading frames and are often short. Such genes are therefore difficult to identify by computer searches (unless of course the sequence of the non-coding RNA or a homolog from another organism is already known). In addition, non-coding RNAs lack the poly(A) tails characteristic of mRNA so are absent from cDNA or EST libraries. There are about 500 human tRNA genes—fewer than in the worm *Caenorhabditis*! The human rRNA genes for 18S, 28S and 5.8S rRNA are found as cotranscribed units. Tandem repeats of these are found on the short arms of chromosomes 13, 14, 15, 21 and 22 giving a total of about 200 copies of these rRNA genes. The 5S rRNA gene is found separately, but also in tandem repeats, the longest cluster being on chromosome 1, near the telomere of the long, q-arm. There are 200–300 genuine 5S rRNA genes and at least 500 pseudogenes.

Genes for non-translated RNA are hard to find just by computer searching.

Both locating their genes and telling apart related sequences and pseudogenes from true functional copies are even more difficult for the other non-coding RNAs. At present several genes for expected non-coding RNAs are still missing, yet at the same time, many related sequences of uncertain function have been located.

Sequence Polymorphisms: SSLPs and SNPs

Although the human genome is sequenced, analysis of different genomes can lead to a variety of discoveries. A **polymorphism** is simply a difference in DNA sequence between two related organisms, e.g. two individual humans. Polymorphisms may be divided into those consisting of base changes and those where there is a difference in the length of the corresponding region of DNA. Polymorphisms could explain why people have different appearances, different susceptibility to diseases, and even perhaps different personality traits.

An **SNP** ("snip") or **single nucleotide polymorphism** is genomics terminology for a single base change between two individuals. In the human genome, there are several hundred thousand SNPs within genes and vastly more in non-coding DNA. SNPs are generally detected by use of hybridization using DNA chips (see above). If a SNP lies within a restriction recognition site it will cause an RFLP (restriction fragment length polymorphism; see Ch. 22). However, most SNPs do not give rise to RFLPs because they do not lie within a restriction cut site.

Many single base differences are found between the genomes of individuals of the same species.

SSLP stands for **simple sequence length polymorphism**. This is a general term that applies to any DNA region consisting of tandem repeats that vary in number from individual to individual. It includes VNTRs, microsatellites, and other tandem repeats that have already been discussed (Ch. 4).

Different human individuals differ by approximately one base change every 1,000–2,000 bases. This amounts to around 2.5 million SNPs over the whole genome. About 60,000 known SNPs fall within the exons of genes. Therefore, the genetic diversity in the human population is much smaller than would be expected. Despite their smaller population, chimpanzees show much more genetic diversity than humans. The most likely explanation is that after splitting off from chimps about 5 million years ago, humans went through a genetic bottleneck. Modern humans probably emerged about 100,000 years ago from a small initial population (See African Eve, Ch. 20); therefore, the genetic diversity in the beginning was very low.

polymorphism A difference in DNA sequence between two related individual organisms
simple sequence length polymorphism (SSLP) Any DNA region consisting of tandem repeats that vary in number from individual to individual, including VNTRs, microsatellites, and other tandem repeats
single nucleotide polymorphism (SNP) A difference in DNA sequence of a single base change between two individuals

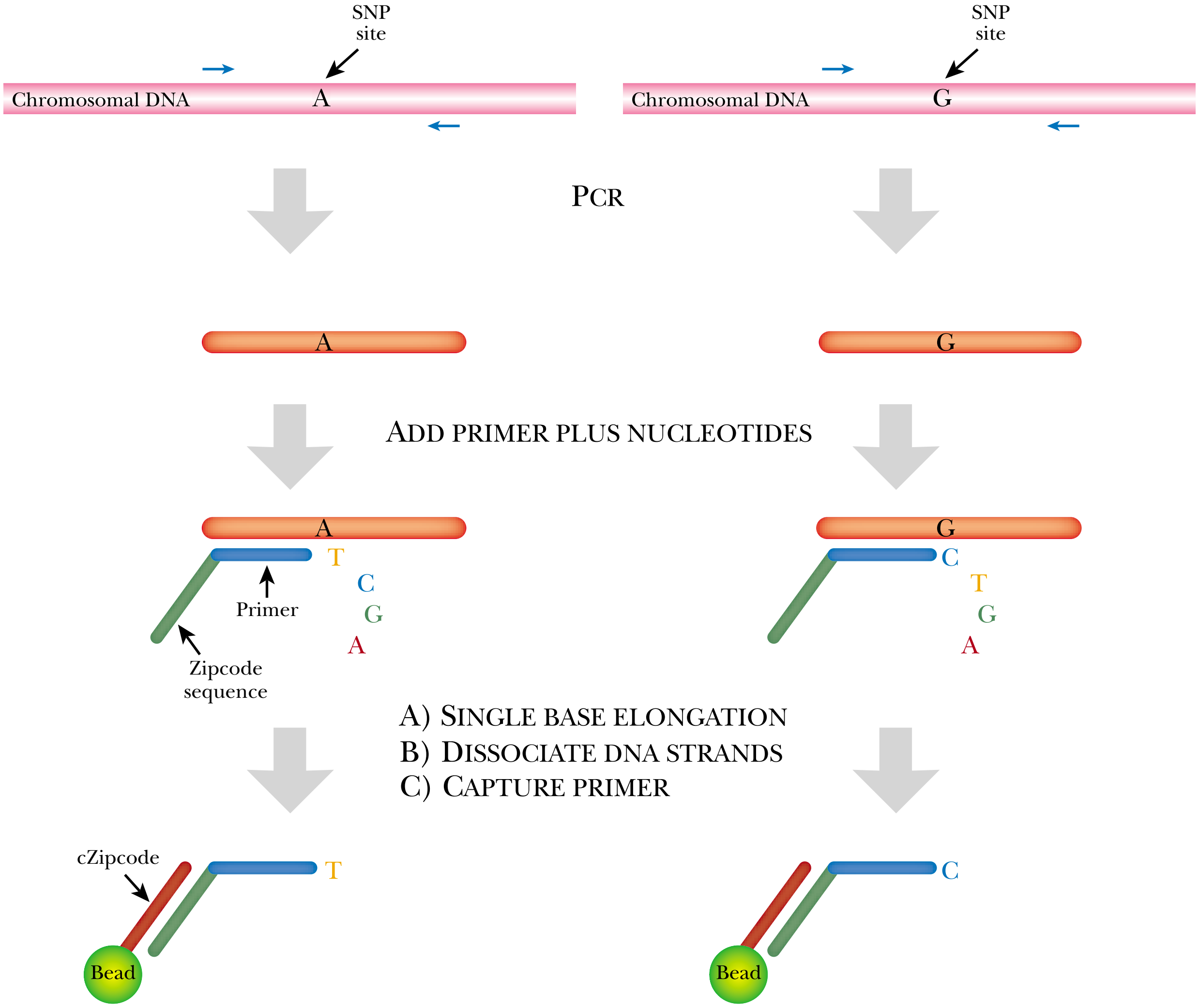


FIGURE 24.23 ***Zipcoded SNP Analysis by Single Base Extension***

A segment of DNA that includes an SNP site is generated by PCR (only a single strand of the DNA is shown here, for simplicity). Single base extension is performed with a primer that binds one base in front of the SNP. Person I has an A at the SNP site and therefore T is incorporated; in person II, a G results in incorporation of C. The incorporated bases are labeled with different fluorescent dyes. The elongated primer is then trapped by binding of its Zipcode sequence to the complementary cZipcode, which is attached to a bead or other solid support.

SNP analysis looks for single base changes in critical regions of the genome.

SNP analysis is used increasingly to screen for possible hereditary defects and also to test for individual variation in genes that affect the response to pharmaceuticals. A variety of methods are in use, but most start by using PCR to generate the region of DNA that contains the polymorphism. The sequence must then be determined—but only for a single base at one precise location. Thus it is not necessary to sequence the whole PCR fragment. Instead a single base extension reaction is performed, using a primer that binds just in front of the polymorphism site plus specifically labeled dideoxynucleotides. Thus each of A, T, G and C could be labeled with fluorescent dyes of different colors. Using dideoxynucleotides ensures that the primer is only elongated by a single base—the one that is complementary to the base at the polymorphism site. The elongated primer will now be fluorescently labeled, and its color will reveal which base was present in the SNP (Fig. 24.23).

Pharmacogenomics—Genetically Individualized Drug Treatment

The genetic differences between individuals may cause significant differences in their reactions to certain drugs or clinical procedures. Cytochrome P450 plays a major role in the oxidative degradation of many foreign molecules, including a wide range of pharmaceuticals. Cytochrome P450 is actually a family of several closely related enzymes, whose substrate range varies so providing protection against a wide range of different foreign molecules. For example, the CYP2D6 isoenzyme oxidizes drugs of the tricyclic antidepressant class. Any given cytochrome P450 may have multiple allelic variants, some of which may show altered activity. Thus the *CYP2D6* gene has several alleles with lowered activity and also one (a duplication) with increased activity (Table 24.02). Such alleles may be present at different frequencies in different human populations. Patients who possess low activity alleles metabolize the corresponding drugs much more slowly and are consequently not only more sensitive to the desired effects of the drug but also more likely to show toxic side effects. In such cases, individual SNP analysis of patients can reveal which allele is present before administering the drug. The drug dosage can then be adjusted to the individual patients genetic constitution. The new and rapidly expanding field that relates individual genotypes to pharmaceutical treatment is known as pharmacogenomics.

TABLE 24.02 Allele Frequencies for Cytochrome P450 CYP2D6

		European	Asian	African
*CYP2D6*2xN*	Duplication	1–5	0–2	2
*CYP2D6*5*	Deletion	2–7	1	2
*CYP2D6*10*	Unstable enzyme	1–2	6	4
*CYP2D6*17*	Lower affinity for drugs	0	51	6

Artificial zipcode sequences are used to keep track of the vast number of different SNPs.

In practice large numbers of SNP analyses are often run in parallel. One way to sort these out is to use so-called Zipcode sequences attached to the primers. Each SNP is allocated a different Zipcode sequence that can be specifically bound by using the complementary sequence or cZipcode. The cZipcode sequence is attached to a solid support or a polystyrene bead. Different cZipcode sequences may be attached to color coded beads that are later separated by a FACS (fluorescence activated cell sorter—see Ch. 21) or attached to a solid surface forming an array.

Gene Identification by Exon Trapping

Exons can be experimentally isolated and identified by using their flanking splice sites.

In eukaryotes, the actual coding sequences only account for a minority of the DNA. Given a large stretch of DNA sequence, how are the genes identified? Although computer algorithms exist to analyze sequences, the method known as **exon trapping** allows the experimental isolation of coding sequences. This method relies on the fact that exons are flanked by splice recognition sites that are used during RNA processing to

exon trapping Experimental procedure for isolating exons by using their flanking splice recognition sites

Personal Genomics

To take full advantage of pharmacogenomics requires knowledge of individual DNA sequence differences, at least for those genes directly relevant to the type of clinical treatment proposed. At present specific genes can be sequenced on a need-to-know basis. However, it has been suggested that everyone should get their complete genome sequenced individually.

The three factors involved are time, technology and cost. The human genome project took about 10 years, cost $3,000 million and has given a consensus sequence based on 10 different people. Perlegen Sciences in California has used microarray based sequencing to provide individual genome sequences to approximately 25 people (as of August 2002) at a cost of $1.5 million each. It is believed that miniaturization combined with high throughput technology could perhaps reduce the cost to $50,000 per person in a year or two. Other, more futuristic projections, based on trends rather than present technology, suggest that within 10 years customers will be able to buy their own genome sequences for the same price as a flat screen TV.

So what will your personal DNA sequence reveal? We know a reasonable amount about hereditary defects due to single genes (such as cystic fibrosis or sickle cell anemia). However, the genetic factors involved in conditions such as heart disease, obesity, cancer, life expectancy and mental disorders are more complex and due to multiple interacting genes, many of which have yet to be identified. Even though interpretation will be a problem, it will doubtless be more economical to get your whole genome sequenced than pay for lots of individual tests for each gene whose effects are understood.

Once everyone takes their own DNA sequence home to analyze on their personal computer, we will presumably see an orgy of self diagnosis. Will we also see an outbreak of numerology with people claiming to find messages from aliens encoded in their genomes?

splice out the introns (see Ch. 12 for details of splicing). Introns can be spliced using an *in vitro* system; therefore, a length of DNA containing the splice recognition site can be identified. Consequently, exon trapping can be used even if the sequence of the DNA is unknown, although in this case we will not know the relative order in the original DNA of the exons that are isolated.

During exon trapping, the DNA to be analyzed must first be cloned into a special vector that can replicate both in *E. coli* and in suitable animal cells. The vector carries an artificial mini-gene consisting of just two exons and an intervening intron, together with a promoter and poly(A) tail recognition site (Fig. 24.24). The intron contains a multiple cloning site for cloning lengths of unknown DNA. The pSPL vectors, as they are called, use a simian virus 40 (SV40) origin of replication as well as an SV40 promoter and tail site for the mini-gene. These vectors can replicate in modified monkey cells (COS cells) that contain a defective SV40 genome integrated into a host chromosome.

DNA containing the exons to be trapped is cut into segments using an appropriate restriction enzyme. These segments are inserted into the multiple cloning site within the intron on the pSPL vector (Fig. 24.25). The plasmid is then transformed into the COS monkey cells. The mini-gene will be expressed and the RNA primary transcript will be spliced. If an extra exon was cloned into the middle of the mini-gene, it will be present in the mRNA, which will therefore be longer. To isolate the trapped exon, the mRNA is converted to cDNA and then PCR is used to amplify the region containing the trapped exon. This technique will have to be used in conjunction with sequence analysis to identify all the different exons within the human genome.

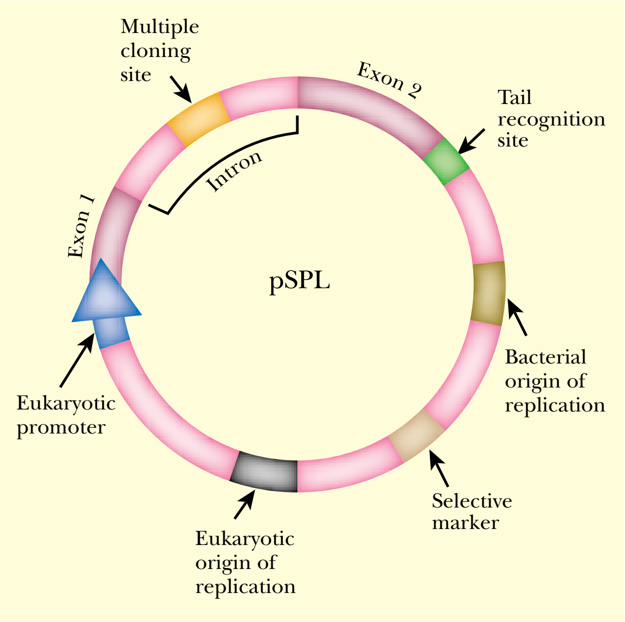


FIGURE 24.24 ***Exon Trapping Vector***

The pSPL vector is used to identify exons within regions of suspected coding DNA. The vector has both bacterial and eukaryotic origins of replication so that it can be grown in both *E. coli* and animal cells. The multiple cloning site is within an intron sequence that is flanked by two exons. This region of the vector can be transcribed into RNA because it contains eukaryotic promoter and eukaryotic poly(A) tail sequences.

A major update to the human genome in 2004 resulted in the number of estimated human genes dropping from around 35,000 to around 25,000. Build 34 of the human genome contains 22,287 protein-encoding genes with an average of 1.5 alternative transcripts and 10 exons per gene. The total coding sequence is 34 Mbp or 1.2% of the euchromatin (i.e., the genome excluding the highly condensed and unsequenced regions, especially in the centromeres and telomeres).

Bioinformatics and Computer Analysis

The field of **bioinformatics** deals with the computerized analysis of large amounts of sequence data. A variety of websites are now available for online searching and manipulation of sequences (Table 24.03).

Vast amounts of genetic data are now becoming available. Computer analysis of this data has essentially created a new field of enquiry.

Both in molecular biology and other areas, vast amounts of information are accumulating in computer data banks. **Data mining** is the use of computer programs to find useful information by filtering or sifting through the data. Hence, intelligent software designed for data mining is sometimes known as “siftware”. **Genome mining** is the application of this approach to genomic data banks. There are several stages to genome mining:

1. Selection of the data of interest.
2. Preprocessing or “data cleansing”. Unnecessary information is removed to avoid slowing or clogging the analysis.
3. Transformation of the data into a format convenient for analysis.
4. Extraction of patterns and relationships from the data.
5. Interpretation and evaluation.

bioinformatics The computerized analysis of large amounts of biological sequence data
data mining The use of computer analysis to find useful information by filtering or sifting through large amounts of data
genome mining The use of computer analysis to find useful information by filtering or sifting through large amounts of biological sequence data

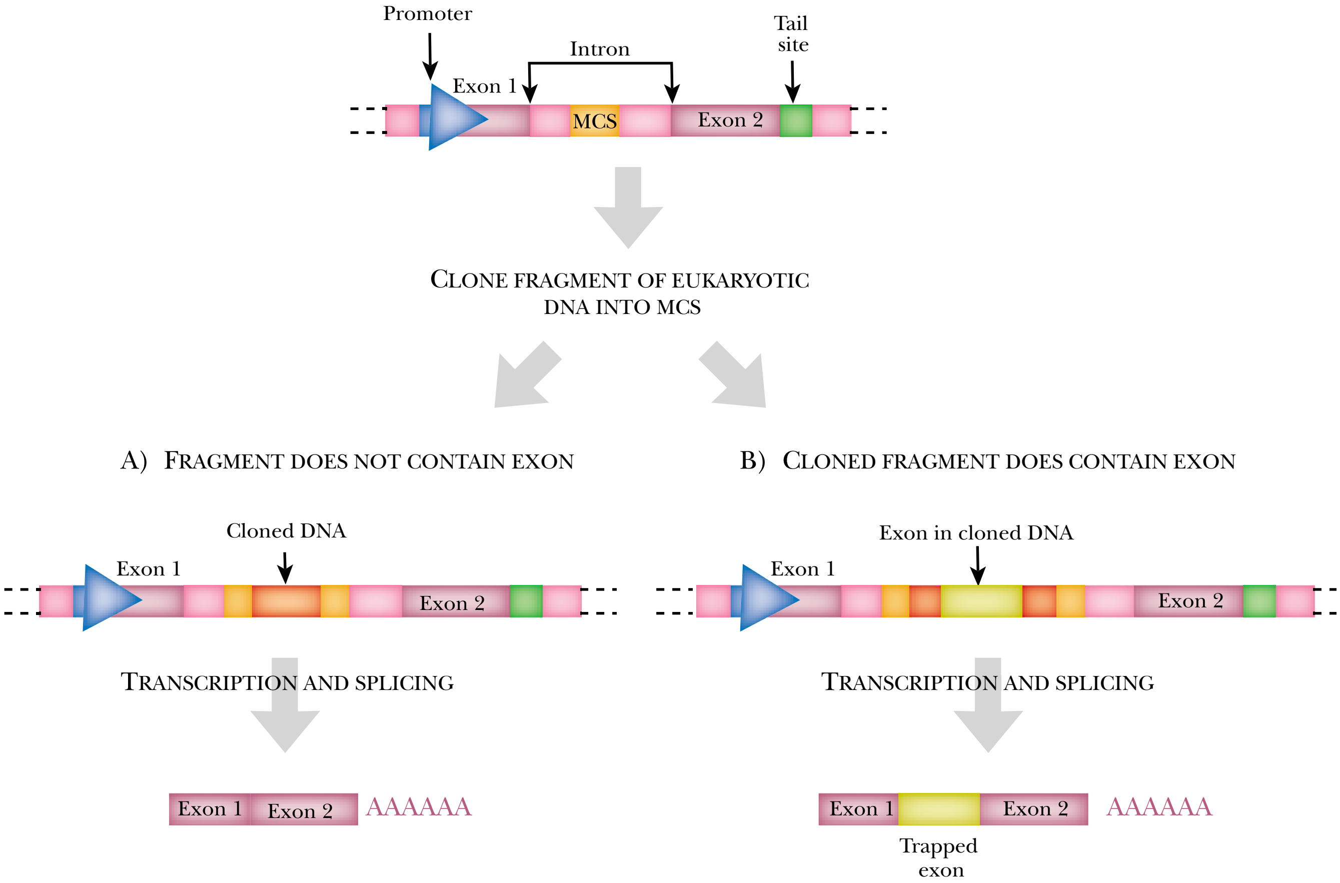

Promoter
Intron
Tail site
Exon 1
MCS
Exon 2
CLONE FRAGMENT OF EUKARYOTIC DNA INTO MCS
A) FRAGMENT DOES NOT CONTAIN EXON
B) CLONED FRAGMENT DOES CONTAIN EXON
Cloned DNA
Exon 1
Exon 2
Exon in cloned DNA
Exon 1
Exon 2
TRANSCRIPTION AND SPLICING
TRANSCRIPTION AND SPLICING
Exon 1
Exon 2
AAAAAA
Exon 1
Exon 2
AAAAAA
Trapped exon

FIGURE 24.25 ***Exon Trapping Procedure***

In order to determine if a length of DNA includes an exon, the unknown DNA is cloned into the pSPL vector. The multiple cloning site is within a mini-gene that is transcribed and spliced when the vector is in animal cells. If the unknown DNA does not contain an exon, the mRNA transcript will be the same size as in the original vector. If the unknown DNA does contain an exon, the mRNA transcript will be longer. Any exons discovered by this method can be amplified using PCR for cloning and/or sequencing.

TABLE 24.03 Some Bioinformatics Websites

GenBank and linked databases
http://www.ncbi.nlm.nih.gov/Entrez/
http://www.ncbi.nlm.nih.gov/genome/guide/human/

Institute for Genomics Research (TIGR)
http://www.tigr.org/tdb

Genome Database (GDB) (human genome)
http://gdbwww.gdb.org

European Bioinformatics Institute (Including EMBL & Swissprot)
http://www.ebi.ac.uk/

Flybase (*Drosophila* genome)
http://flybase.bio.indiana.edu:82

RCSB Protein Data Bank
http://www.rcsb.org/pdb/

PIR Protein Information Resource (PIR)
http://www-nbrf.georgetown.edu/pir/searchdb.html

A variety of analyses may be performed on DNA sequences. Some simple examples are as follows:

A. Searching for related sequences. Any DNA sequence may be compared with other sequences available in the data banks. Searches can also be run on protein sequences after translation of coding DNA. If another protein is found with a related sequence this may give some idea of the function of the protein under investigation. Of course, this assumes that the function of the other protein has already been deciphered! Another major use of sequence comparisons is to trace the evolution both of individual genes and of the organisms that carry them (see Ch. 20).

B. Codon bias analysis can locate coding regions. Due to third base redundancy and the preferential use of some codons over others (in coding regions but not in random, intergenic DNA), there are differences in codon frequency between coding and non-coding DNA. A codon bias index can be computed that gives a reasonable first estimate of whether a stretch of DNA is likely to be coding or non-coding.

C. Searching for known consensus sequences. A variety of short consensus sequences or sequence motifs are known. Analysis of DNA sequences may reveal promoters, ribosome binding sites (in prokaryotes only), terminators and other regulatory regions. Inverted repeats in DNA imply possible stem and loop structures, which are often sites for the binding of regulatory proteins. Analysis of protein sequences may indicate binding sites for metal ions cofactors, nucleotides, DNA etc.

Despite the vast amount of information available from the analysis of DNA sequences, we still need to investigate how genes are regulated at the genome level and how the encoded proteins function. Just as the totality of genetic information is known as the genome, so the sum of the transcribed sequences is the transcriptome and the total protein complement of an organism is the proteome. These are discussed in Chapters 25 and 26 respectively.

CHAPTER TWENTY-FIVE

Analysis of Gene Expression

Introduction

Gene expression may be examined in a variety of ways, both at the level of individual genes and, increasingly in recent years, at the level of the whole genome. By analogy with genomics, the sum total of an organisms RNA transcripts are sometimes referred to as the **transcriptome**. Here we will first consider individual genes and then cover approaches to screening expression of large numbers of genes simultaneously. This is known as transcriptome analysis and, with proteomics and metabolomics (see Ch. 26), makes up the area of **functional genomics**. Of the plethora of newly coined terms ending in -ome, perhaps the nicest is the "unknome" proposed by Mark Gerstein of Yale University. This consists of the large proportion of genes with no currently known function!

Monitoring Gene Expression

Expression of most genes results in their transcription to give RNA followed by translation of the RNA to give the final gene product, protein. In addition there are a few genes that produce non-coding RNA (such as tRNA or rRNA) and so have RNA as the final gene product. Although housekeeping genes are needed all the time, most genes are expressed only under certain environmental conditions or in particular tissues or at certain stages of the developmental cycle, as discussed in Chapters 9 and 10. Measurement of gene expression means estimating the level of gene product synthesized. Since most genes vary in expression under different conditions it is necessary to measure the level of gene expression under a variety of conditions.

It is possible to monitor gene expression directly by measuring the levels of protein or RNA. Proteins may be detected by running cell extracts on polyacrylamide gels or by antibody-based assays. If the protein is an enzyme the enzyme activity may be assayed. Direct detection and assay of proteins as gene products is deferred until Ch. 26, which covers proteomics.

Gene expression is measured by monitoring the RNA or protein products that are made.

Here we will consider the monitoring of gene expression at the transcriptional level. The transcriptional expression of a gene may be estimated by measuring the level of mRNA directly. This may be done by hybridization (Northern blotting) using a DNA probe specific for the sequence of the gene under investigation. Hybridization has already been discussed in Ch. 21. The use of fluorescent probes has greatly increased the sensitivity of Northern hybridization; nonetheless, for accurate measurement of the expression of individual genes under many different conditions, using gene fusions with reporter genes is preferable.

Reporter Genes for Monitoring Gene Expression

Genes whose products are convenient to assay are used as "reporters".

Genes that are used in genetic analysis because their products are easy to detect are known as **reporter genes**. They are often used to report on gene expression, although they may also be used for other purposes, such as detecting the location of a protein or the presence of a particular segment of DNA.

Suppose that a DNA molecule, such as a cloning plasmid, has been inserted into a new bacterial host cell or a transgene has been inserted into the chromosome of a new animal host. Antibiotic resistance genes are often used to monitor whether the DNA is indeed in the intended location. Thus antibiotic resistance genes may be regarded as reporter genes (Fig. 25.01). As already discussed in Ch. 22, after transformation of the plasmid into the target cells, they are treated with the antibiotic. Those

functional genomics The study of the whole genome and its expression
reporter gene Gene that is used in genetic analysis because its product is convenient to assay or easy to detect
transcriptome The total sum of the RNA transcripts found in a cell, under any particular set of conditions

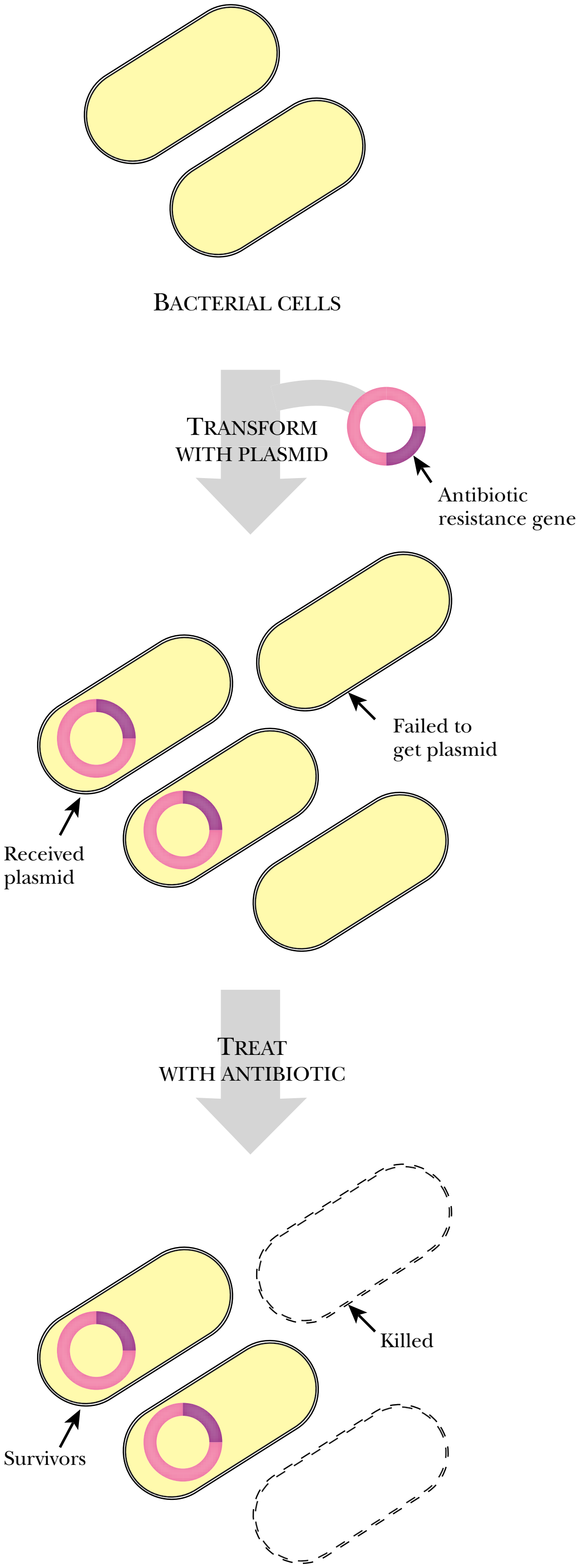


FIGURE 25.01 ***Antibiotic Resistance as a Reporter Gene***

Antibiotic resistance genes are included on plasmids in order to determine whether the plasmids are present in a cell. When bacteria are transformed with plasmid DNA those that get a plasmid that carries an antibiotic resistance gene will survive when treated with the antibiotic whereas those cells that fail to get a plasmid will be killed.

that receive the plasmid become antibiotic resistant; those not getting the antibiotic resistance gene are killed. An antibiotic resistance gene carried on the same fragment of DNA can also report whether a transgene has integrated into another DNA molecule such as a chromosome or a virus.

Easily Assayable Enzymes as Reporters

Beta-galactosidase and alkaline phosphatase can be assayed using substrates that generate colored or fluorescent products.

The most widely used reporter gene for monitoring gene expression is the ***lacZ* gene** encoding **β-galactosidase** (Fig. 25.02). This enzyme normally splits lactose, a compound sugar found in milk, into the simpler sugars glucose and galactose. However, β-galactosidase will also split a wide range of compounds of galactose (i.e., **galactosides**) both natural and artificial (Fig. 25.02). The two most commonly used artificial galactosides are ONPG and X-gal. **ONPG (*o*-nitrophenyl galactoside)** is split into *o*-nitrophenol and galactose. The *o*-nitrophenol is yellow and soluble, so it is easy to measure quantitatively. **X-gal (5-bromo-4-chloro-3-indolyl β-D-galactoside)** is split into galactose plus the precursor to an indigo type dye. Oxygen in the air converts the precursor to an insoluble blue dye that precipitates out at the location of the *lacZ* gene.

Another widely used reporter gene is the ***phoA* gene** that encodes **alkaline phosphatase**. This enzyme cleaves phosphate groups from a broad range of substrates (Fig. 25.03). Like β-galactosidase, alkaline phosphatase will use a variety of artificial substrates:

1. ***o*-Nitrophenyl phosphate** is split, releasing yellow *o*-nitrophenol.
2. **X-phos** (5-bromo-4-chloro-3-indolyl phosphate) consists of an indigo dye precursor joined to phosphate. After the enzyme splits this, exposure to air converts the dye precursor to a blue dye, as in the case of X-gal.
3. **4-Methylumbelliferyl phosphate** releases a fluorescent compound when the phosphate is removed.

Light Emission by Luciferase as a Reporter System

The enzyme luciferase emits light when it reacts with its substrate, luciferin.

A more sophisticated reporter gene encodes **luciferase** (Fig. 25.04). This enzyme emits light when provided with a substrate known as **luciferin**. Luciferase is found naturally in assorted luminous creatures from bacteria to deep-sea squid. The ***lux* genes** from bacteria and the ***luc* genes** from fireflies produce different brands of luciferase, but both work well as reporter genes. The luciferins used by the different types of luciferase are chemically different. Bacterial luciferase uses the reduced form of the cofactor FMN (flavin mononucleotide) as its luciferin. Oxygen and a long chain aldehyde (R-CHO) are also needed. Both the reduced FMN and the aldehyde are oxidized.

alkaline phosphatase Enzyme that cleaves phosphate groups from a broad range of substrates
β-galactosidase Enzyme that splits lactose and other compounds of galactose
galactoside Compound of galactose, such as lactose, ONPG or X-gal
***lacZ* gene** Gene encoding β-galactosidase; widely used as a reporter gene
***luc* gene** Gene encoding luciferase from eukaryotes
luciferase Enzyme that emits light when provided with a substrate known as luciferin
luciferin Chemical substrate used by luciferase to emit light
***lux* gene** Gene encoding luciferase from bacteria
4-methylumbelliferyl phosphate An artificial substrate that is cleaved by alkaline phosphatase, releasing a fluorescent molecule
ONPG (*o*-nitrophenyl galactoside) Artificial substrate that is split by β-galactosidase, releasing yellow *o*-nitrophenol
***o*-nitrophenyl phosphate** Artificial substrate that is split by alkaline phosphatase, releasing yellow *o*-nitrophenol
***phoA* gene** Gene encoding alkaline phosphatase; widely used as a reporter gene
X-gal (5-bromo-4-chloro-3-indolyl β-D-galactoside) Artificial substrate that is split by β-galactosidase, releasing a blue dye
X-phos 5-bromo-4-chloro-3-indolyl phosphate, an artificial substrate that is split by alkaline phosphatase, releasing a blue dye

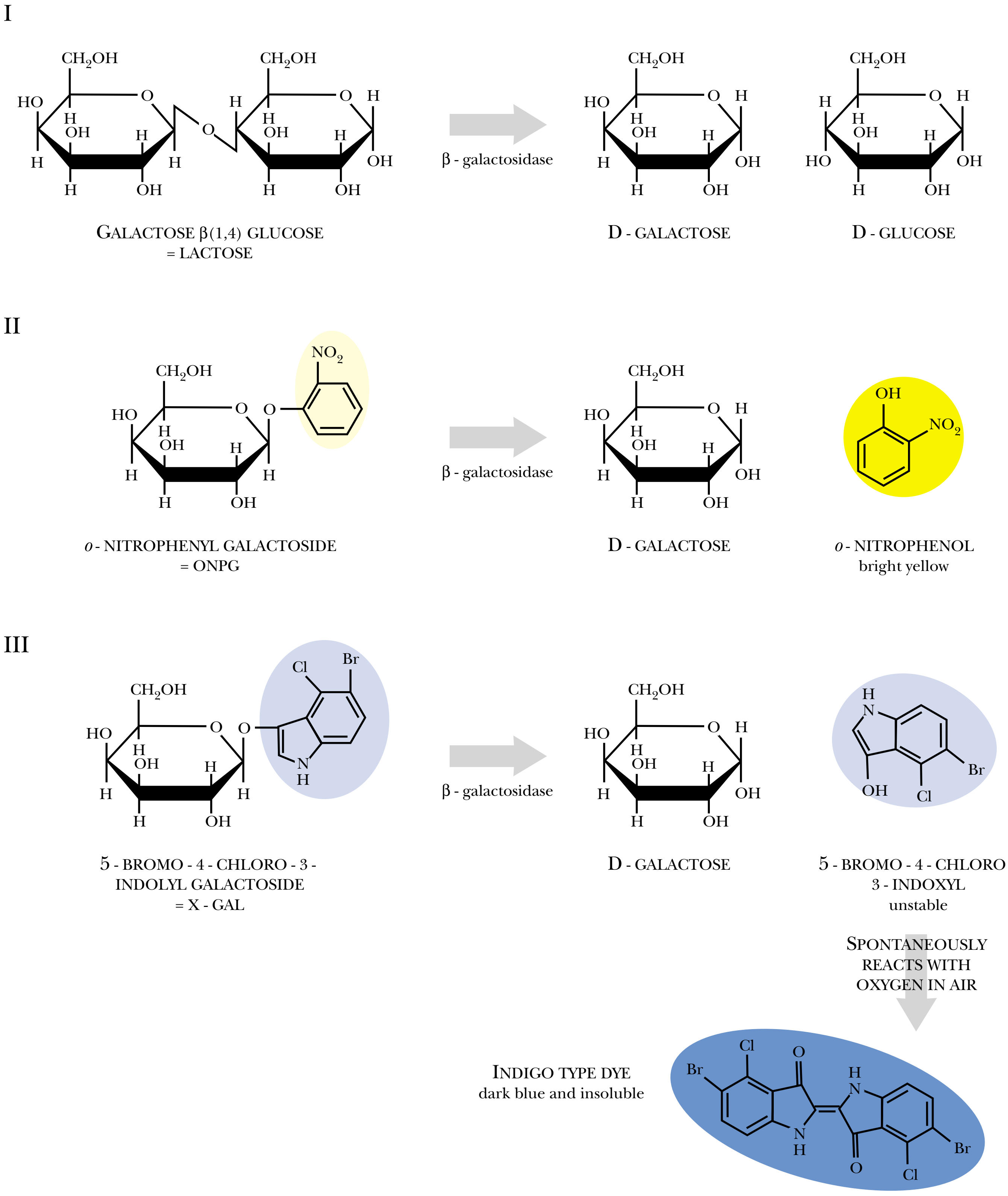


FIGURE 25.02 ***Substrates Used by β-Galactosidase***

The enzyme, β-galactosidase, normally cleaves lactose into two monosaccharides, glucose and galactose. β-galactosidase also cleaves two artificial substrates, ONPG and X-Gal, releasing a group that forms a visible dye. ONPG releases a bright yellow substance called *o*-nitrophenol, whereas X-gal releases an unstable group that reacts with oxygen to form a blue indigo dye.

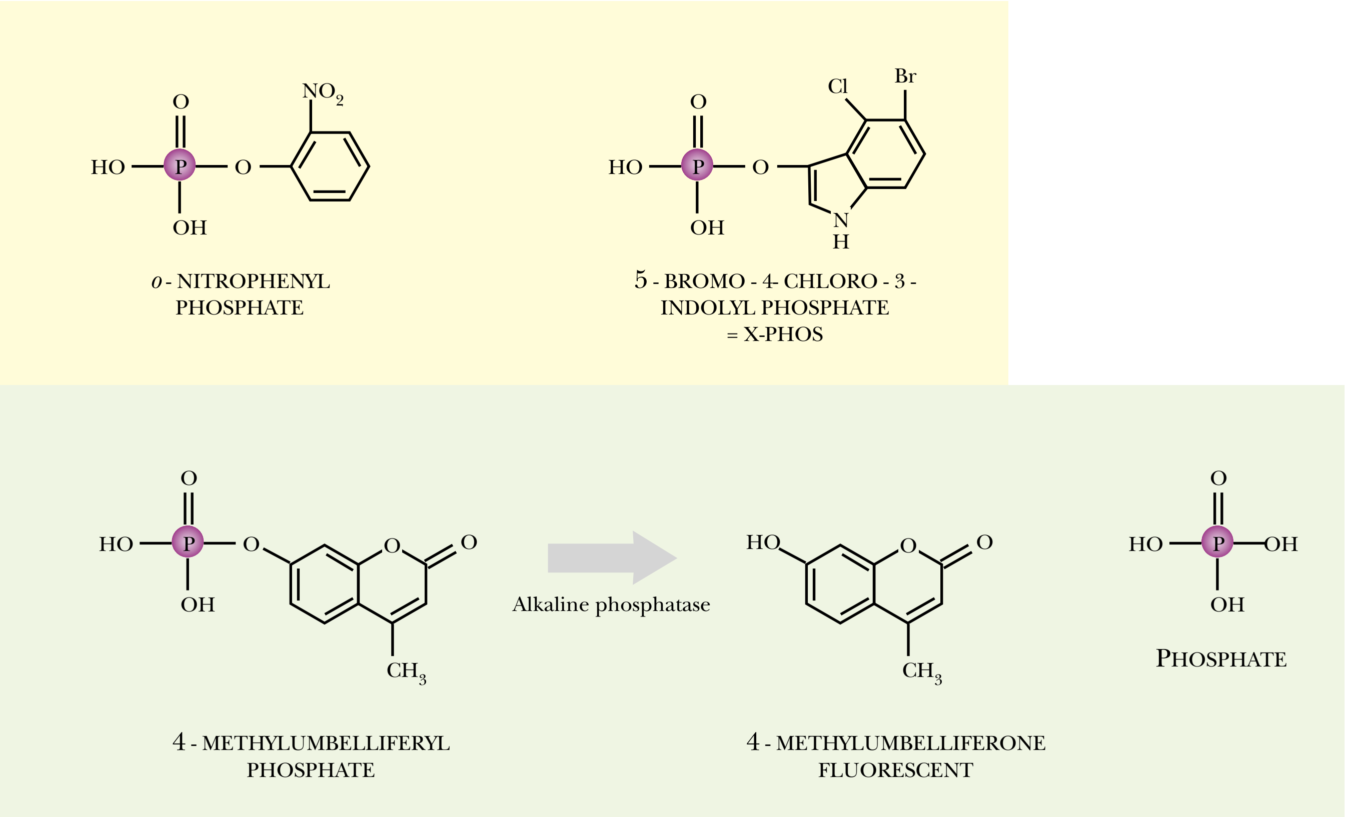

HO
P
O
OH
NO2
o - NITROPHENYL PHOSPHATE
Cl
Br
N
H
5 - BROMO - 4- CHLORO - 3 - INDOLYL PHOSPHATE = X-PHOS
CH3
4 - METHYLUMBELLIFERYL PHOSPHATE
Alkaline phosphatase
4 - METHYLUMBELLIFERONE FLUORESCENT
PHOSPHATE

FIGURE 25.03 ***Substrates Used by Alkaline Phosphatase***

Alkaline phosphatase removes phosphate groups from various substrates. When the phosphate group is removed from *o*-nitrophenyl phosphate, a yellow dye is released. When the phosphate is removed from X-phos, further reaction with oxygen produces an insoluble blue dye as for X-gal. Additionally, alkaline phosphatase releases a fluorescent molecule when the phosphate is removed from 4-methylumbelliferyl phosphate.

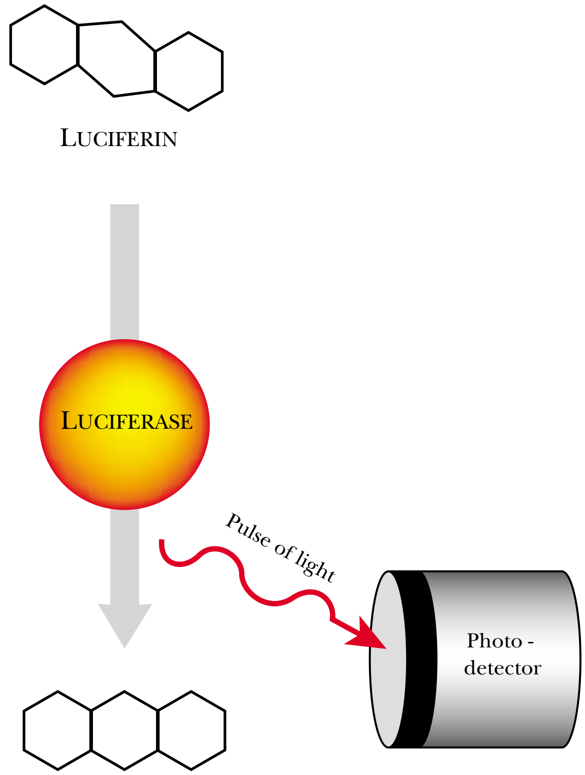


FIGURE 25.04 ***Luciferase Degrades Luciferin and Emits Light***

Luciferase is an enzyme that alters the structure of luciferin. When the structure is altered, a pulse of light is emitted, which is detected by a photodetector. The luciferin shown in this figure is FMN (flavin mononucleotide), which is used by bacterial luciferases.

$$\text{R-CHO} + \text{FMNH}_2 + \text{O}_2 \rightarrow \text{R-COOH} + \text{FMN} + \text{H}_2\text{O} + h\upsilon$$

Different groups of eukaryotes make several chemically distinct luciferins that are used solely for light emission. Firefly luciferase requires ATP as well as oxygen and firefly luciferin.

$$\text{luciferin} + \text{O}_2 + \text{ATP} \rightarrow \text{oxidized luciferin} + \text{CO}_2 + \text{H}_2\text{O} + \text{AMP} + \text{diphosphate} + h\upsilon$$

If DNA carrying a gene for luciferase is incorporated into a target cell, it will emit light only when the appropriate luciferin is added. Although high-level expression of luciferase can be seen with the naked eye, usually the amount of light is small and must be detected with a sensitive electronic apparatus such as a luminometer or a scintillation counter.

Green Fluorescent Protein as Reporter

The products of most reporter genes are enzymes that must be assayed in some manner. Unlike the products of most reporter genes, **green fluorescent protein (GFP)** is not an enzyme, and it does not need a non-protein cofactor for it to fluoresce. GFP is a stable and non-toxic protein from jellyfish that can be visualized by its inherent green fluorescence. Consequently, GFP can be directly observed in living tissue without the need for adding any reagents. Nearly 2,000 years ago, the Roman author Pliny noted that the slime from certain jellyfish would generate enough light when rubbed on his walking stick to help guide his steps in the dark.

Green fluorescent protein does not need a substrate or a cofactor. It emits green light after illumination with long-wave UV.

The original GFP came from the jellyfish *Aequorea victoria*. Wild type GFP is excited by long wavelength UV light (excitation maximum 395 nm) and emits at 510 nm in the green. A variety of genetically engineered variants of GFP are also in use. Many of these are available as the Living Colors$^{\text{TM}}$ series from Clontech Corporation. Some of these were chosen for showing higher fluorescence and/or emitting at a different wavelength. Thus enhanced GFP (EGFP), enhanced yellow fluorescent protein (EYFP) and enhanced cyan fluorescent protein (ECFP) can be detected simultaneously using an argon-ion laser plus a detector with appropriate filters. A recent addition is a "Red GFP" (DsRed2—really a shade of orange). Other modifications include adapting GFP for high-level expression in different organisms by altering the codon usage (i.e., by changing bases in the redundant third codon position). Humanized versions of GFP exist that are adapted for use in cultured human cell lines.

GFP can be used to follow gene expression or to localize proteins inside the cell.

Fusions of regulatory regions and promoters to the *gfp* gene have been used to monitor the expression of many genes, especially in living animals. The nematode, *Caenorhabditis*, and the zebrafish are both transparent and so GFP can be used to follow differential gene expression in different internal tissues of living animals. Transgenic mice, rabbits, monkeys and several plants have been engineered that have the *gfp* gene inserted into the host genome (Fig. 25.05).

In addition to monitoring gene expression, GFP is widely used to localize proteins within the cell (Fig. 25.06). Gene fusions are constructed that yield a hybrid protein. These are normally designed so that the GFP protein is attached at the C-terminal end of the protein under investigation. The fluorescence due to GFP will reveal the subcellular location of the target protein. For example, a fusion of Red GFP (DsRed) to the targeting sequence from subunit VIII of cytochrome c oxidase is located in the mitochondrial inner membrane. Fusions between actin or tubulin and GFP are used to study cell architecture.

Gene Fusions

Reporter genes can be used to track the physical location of a segment of DNA or to monitor gene expression. In particular reporter genes are often incorporated into gene fusions where they are used to follow the level of expression of the target gene. Many

green fluorescent protein (GFP) A protein, originally from a jellyfish, whose green fluorescence makes it useful as a reporter molecule

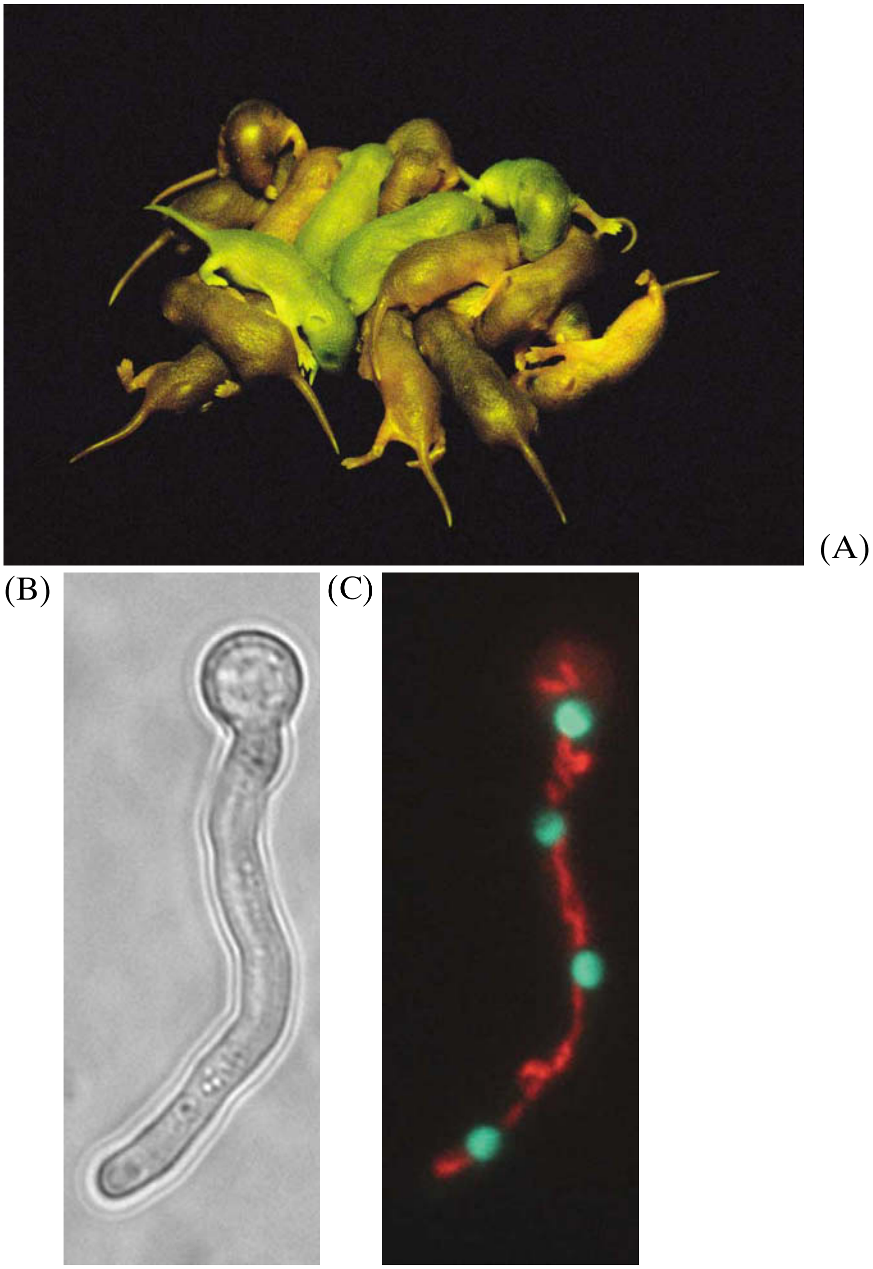


FIGURE 25.05 ***Transgenic Organisms with Green Fluorescent Protein***

The gene for GFP has been integrated into the genome of animals, plants and fungi. After exposure to long wavelength UV the organisms emit green light.
A) Transgenic mice with GFP among normal mice from the same litter. The *gfp* gene was injected into fertilized egg cells to create these mice. GFP is produced in all cells and tissues except the hair. Credit: Eye of Science, Photo Researchers, Inc.
B) Phase contrast and
C) Fluorescence emission of germlings of the fungus *Aspergillus nidulans*. Original GFP was used to label the mitochondria and a red GFP variant (DsRed) for the nucleus. From: Toews et al., Current Genetics 45 (2004):383–389.

Gene fusions are used to monitor genes whose products are difficult to assay. Reporter genes are fused to the regulatory region of the target gene.

genes have products that are complicated or tedious to assay by direct measurement or may even be unknown. To avoid this, the original gene product is replaced by fusing its regulatory region to the structural region of a reporter gene. The target gene is cut between its regulatory region and coding region. The same is done with the reporter gene. Then the regulatory region of the gene under investigation is joined to the coding region of the reporter gene (Fig. 25.07). This hybrid structure is a **gene fusion**.

The gene fusion will be controlled the same way the original target gene was controlled, but instead of making the original gene product, it makes the enzyme belonging to the reporter gene. The cells carrying the fusion can be grown under an immense number of different conditions and assayed for the reporter enzyme. This approach, especially using *lacZ* and β-galactosidase, is widely used in surveying gene regulation (Fig. 25.08). In addition, gene fusions may be used to test for possible effects of regulatory genes. A mutation that inactivates a regulatory gene is introduced into the cell

gene fusion Structure in which parts of two genes are joined together, in particular when the regulatory region of one gene is joined to the coding region of a reporter gene

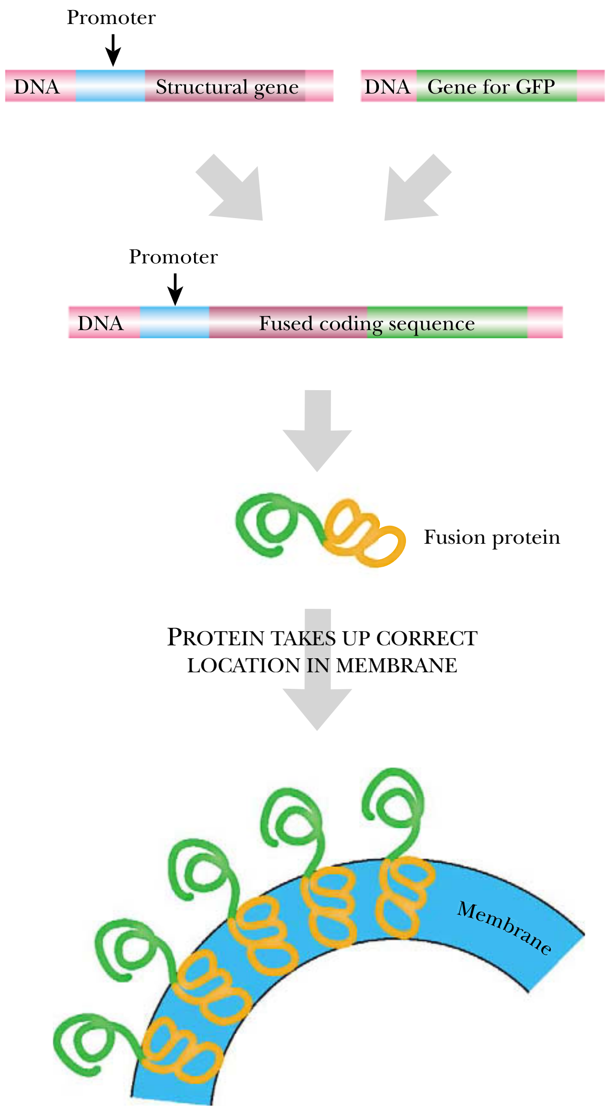


FIGURE 25.06 ***GFP for Protein Localization***

GFP can be used to reveal where a protein is localized within the cell. The first step is to fuse the GFP gene in frame with all or part of the structural gene that encodes the protein of interest. The fused construct is then expressed in a host cell. The cells are excited with long wavelength UV light and visualized under the microscope. If the protein is normally located in the membrane, as in this example, the cell membrane will fluoresce green under the microscope.

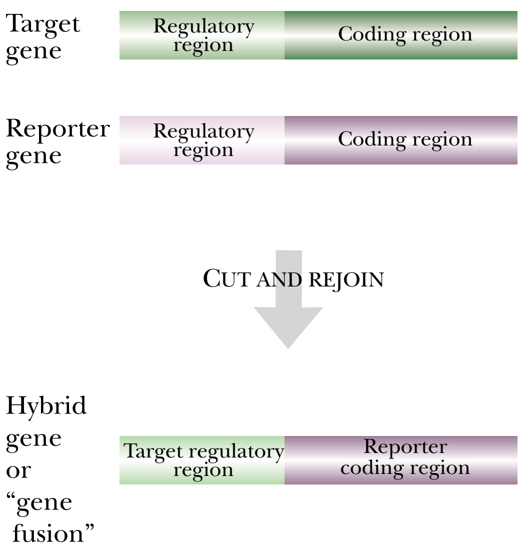


FIGURE 25.07 ***Construction of a Gene Fusion***

Creating gene fusions help in investigating how genes are regulated. The regulatory region of the target gene is joined to the reporter gene coding region. The reporter enzyme will now be made under conditions where the target gene would normally be expressed.

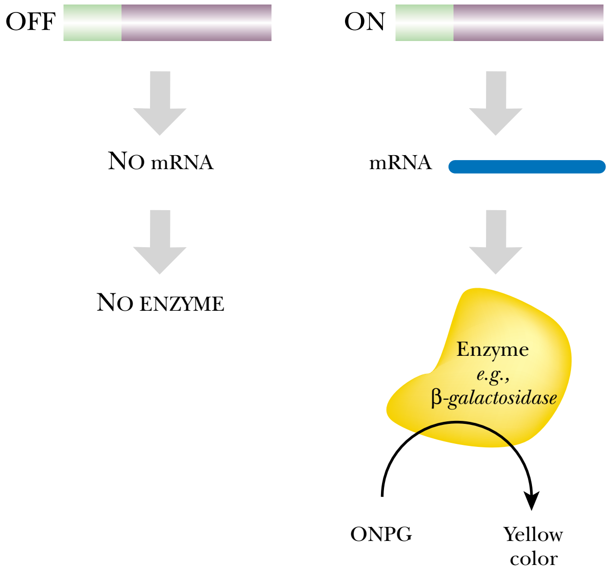


FIGURE 25.08 ***Using Gene Fusions to Survey Regulation***

The regulatory region of the target gene (green) controls the expression of the reporter structural gene (purple). Therefore, assaying the level of reporter enzyme reveals how the target gene would be expressed under the chosen conditions. In this example, the reporter gene is *lacZ* and the level of expression is monitored by the breakdown of ONPG to release yellow nitrophenol.

carrying the gene fusion. Expression is then measured under appropriate conditions. A series of regulatory genes suspected of controlling the target gene may be rapidly surveyed by this approach.

Many bacteria, such as *E. coli*, already possess a wild type copy of reporter genes such as the *lacZ* or *phoA* genes. In these cases the wild type version of the gene must be deleted from the chromosome before the gene fusions are used. Strains of *E. coli* deleted for the *lac* operon or for *phoA* are readily available.

Gene fusions can be made between the coding region of *lacZ* or other reporter genes and the regulatory region of genes whose functions are completely unknown. They are then used to survey possible environmental conditions to see to what stimulus the target gene responds. Alternatively they may be tested for possible effects of mutations in regulatory genes. This may give some clue as to the role of the unknown gene.

Deletion Analysis of the Upstream Region

Regulatory regions may be analysed by deleting segments of the DNA and checking for effects on gene expression.

The upstream regulatory region of a gene often contains several sites for regulatory proteins as well as the promoter region where RNA polymerase binds. These regulatory sites often enhance or suppress the expression of the gene under a variety of conditions. To determine the function of the regulatory elements, it is often helpful to construct a series of altered upstream regions in which presumed binding sites have been eliminated. The simplest way to do this is to remove successive segments from the 5′-end of the upstream region. Originally, restriction enzymes were used to create the deletions. However, finding convenient restriction sites was always a problem. Nowadays PCR is usually used to generate upstream regions of different lengths. A variety of PCR primers are constructed to different points within the upstream region. Each of the primers is paired with a primer within the beginning of the structural gene to create successively smaller segments of the upstream region.

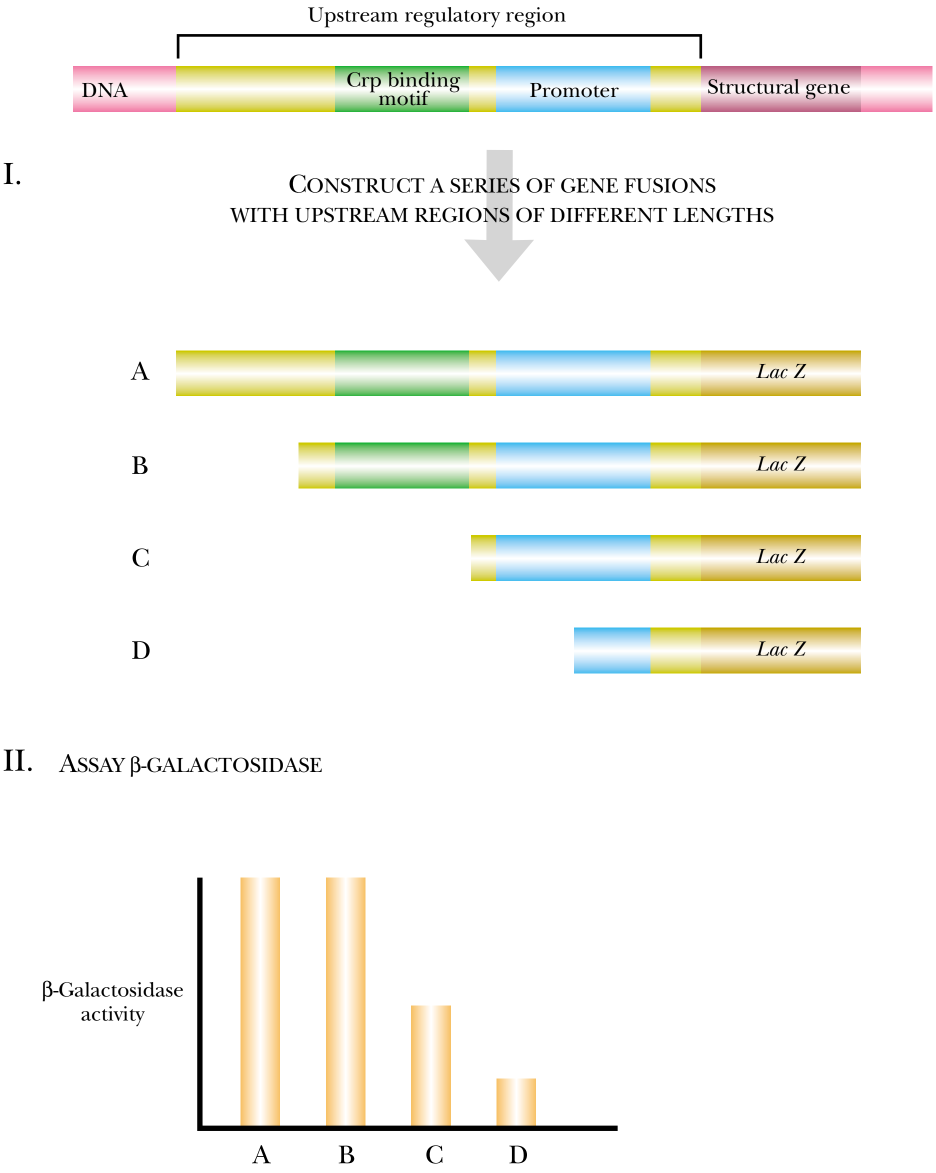
These engineered upstream regions are then tested for possible alterations in gene expression and regulation. They may be examined directly for binding regulatory proteins or by gene fusion analysis. For example, suppose we have an upstream region whose sequence reveals a binding motif for Crp, the *E. coli* cAMP receptor protein, in the 5′ region of the promoter (Fig. 25.09). This region would be removed and both the original and the truncated upstream region would be tested for expression both by using a *lacZ* fusion and by binding of Crp protein by gel retardation.

Locating Protein Binding Sites in the Upstream Region

The upstream region of a gene usually contains sequences in the DNA that control gene expression. Most of these regulatory sequences are sites where regulatory pro-

FIGURE 25.09 ***Deletion Analysis of Upstream Region***

In order to determine which parts of the upstream regulatory region affect the expression of a gene, a series of deletions of the upstream region are made. The upstream region is usually also surveyed for possible binding sites or regulatory motifs. This example shows two potential regulatory regions, a Crp binding site and a putative promoter region. A set of deletions are constructed such that smaller and smaller segments of the upstream region are present. These are fused with the reporter gene, *lacZ*. Next, the constructs are expressed in cells, and the activity of β-galactosidase is assayed. In this example, the whole upstream region (construct A) has high activity. Removal of the far end (B) has negligible effect, suggesting that no important sequences lie in this region. When the Crp binding site is removed, the activity decreases by half, suggesting that Crp regulates gene expression (C). When half of the putative promoter is deleted (D), the β-alactosidase activity is almost zero. These results confirm that the two presumed sites do control the activity of the reporter gene, and therefore also control the original structural gene.



If a regulatory protein binds to a segment of DNA it will slow the migration of the DNA through a gel.

teins bind, although others may be involved in bending DNA or forming stem and loop structures in either the DNA itself or the RNA. After sequencing a gene and its regulatory region, the sequence may be analyzed for known protein binding motifs. However, many cases are known of computer predicted binding sites that do not actually function in real life. Deletion analysis of these upstream sites determines how they affect gene expression, but do not show if a protein actually binds to the site. Consequently, even after a presumed binding site has been found, the binding of the regulatory protein must be confirmed experimentally.

One approach to deciding if a particular regulatory protein affects gene expression is to test directly if the protein binds to DNA from the upstream region. To do this we need both purified DNA from the regulatory region and also purified regulatory protein. The mobility of a fragment of DNA in a gel is altered upon binding protein. This procedure is therefore known either as an electrophoretic **mobility shift** assay, **bandshift** assay or **gel retardation** assay (Fig. 25.10).

bandshift assay Method for testing binding of a protein to DNA by measuring the change in mobility of DNA during gel electrophoresis. Same as gel retardation or mobility shift assay

gel retardation Method for testing binding of a protein to DNA by measuring the change in mobility of DNA during gel electrophoresis. Same as bandshift assay or mobility shift assay

mobility shift assay Method for testing binding of a protein to DNA by measuring the change in mobility of DNA during gel electrophoresis. Same as bandshift assay or gel retardation

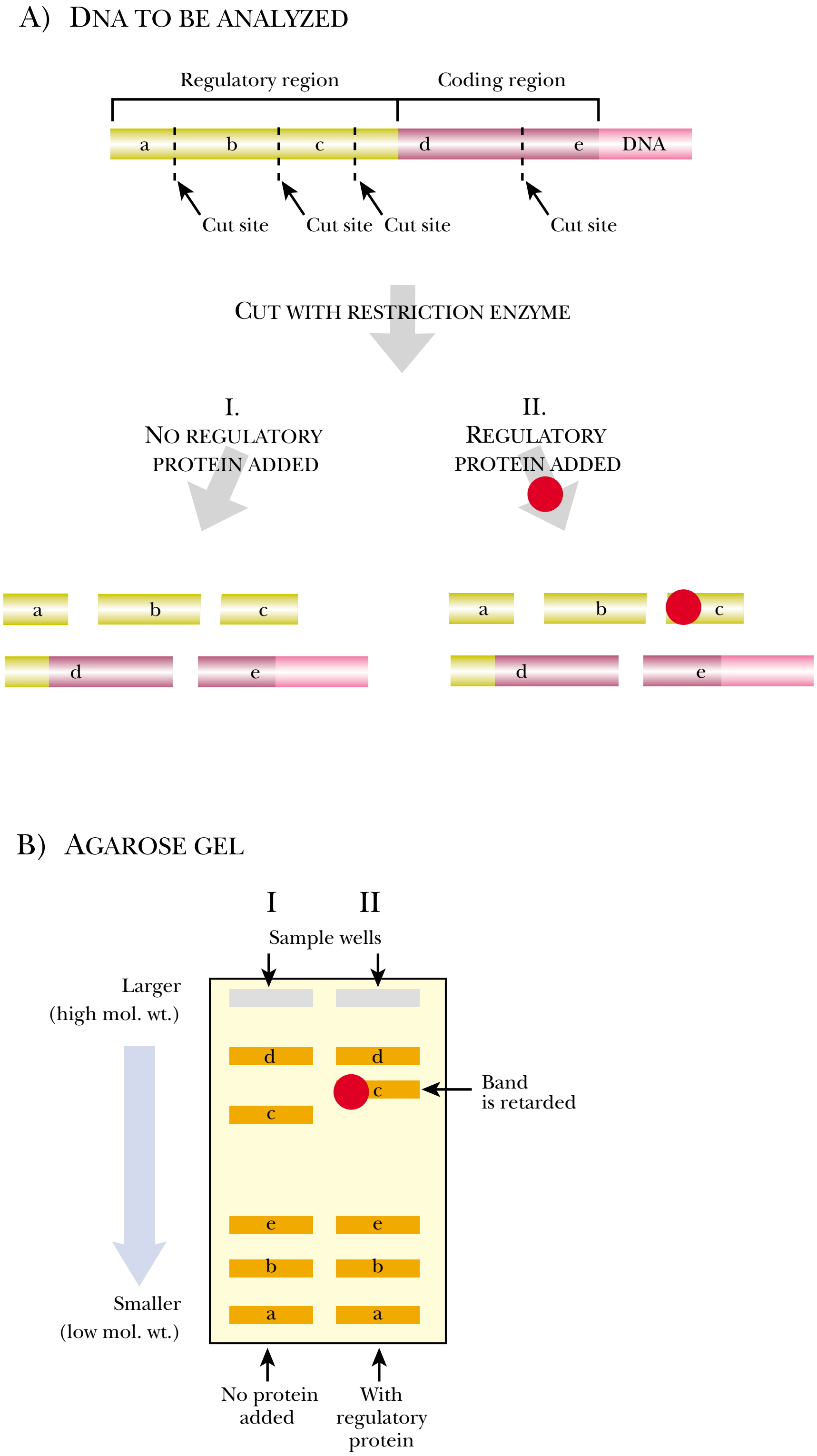


FIGURE 25.10 ***Gel Retardation to Assess Protein Binding to DNA***

Gel retardation assays determine whether or not a specific regulatory protein actually binds to the DNA in the regulatory region of a gene. The regulatory and coding regions of a gene are cut into various fragments by a restriction enzyme. The fragments are divided into two samples. In one (I), no regulatory protein is added, whereas in the other sample (II), purified regulatory protein is mixed with the DNA. The DNA fragments are separated by size using agarose gel electrophoresis. If the regulatory protein binds to the DNA, that fragment is heavier and travels slower. It is therefore retarded relative to its position in the absence of protein. In this example, fragment c has a binding site for the regulatory protein and its band is retarded.

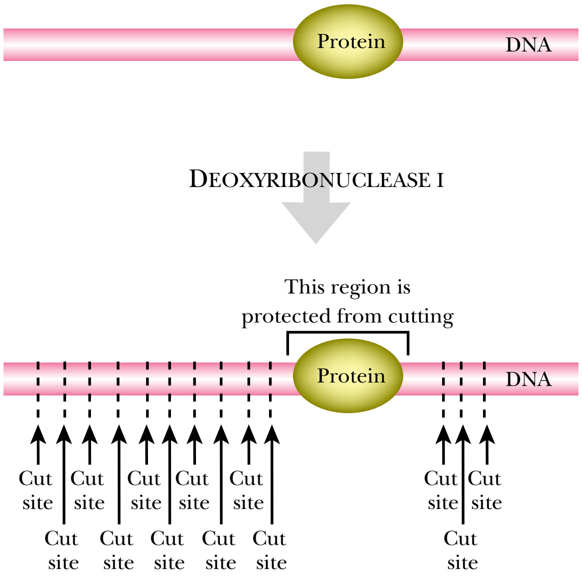


FIGURE 25.11 ***Footprint Analysis—Procedure***

To identify the exact location of a protein binding site, the fragment that contains the binding site is first isolated. The fragment is mixed with the protein, and then treated with DNase. This enzyme cuts randomly along the length of the DNA fragment, but will not be able to cut in the region where the protein binds.

First, the DNA carrying the gene and its upstream region is cut with a convenient restriction enzyme to get a series of fragments. After cutting the DNA, it is split into two samples. To one of these is added the protein being tested. Both samples are then run side by side on a non-denaturing agarose gel. If the protein binds to one of the DNA fragments, the complex formed will be larger and run slower than the original DNA; i.e., that fragment will be retarded. To visualize the DNA fragments after running the gel, the DNA should be "labeled" (usually by making it radioactive or fluorescent) before starting the experiment.

An average protein has a molecular weight of about 40,000. A segment of DNA of 1,000 base pairs has a molecular weight of about 700,000. If a typical protein is bound to a length of DNA much bigger than this, the relative change in size, and therefore in mobility, would be 5 percent or less. Such a small change would impossible to observe. Consequently, for gel retardation analysis, restriction enzymes are chosen to give segments of DNA in the range of 250 to 1,000 base pairs. An alternative to using restriction enzymes is to use PCR to generate fragments from the upstream region of a gene. Primers can be chosen to generate any segment suspected of harboring a binding site. The PCR fragments can then be examined one at a time for binding to the test protein.

When proteins are bound to DNA they protect the region on the DNA that they cover from chemical attack.

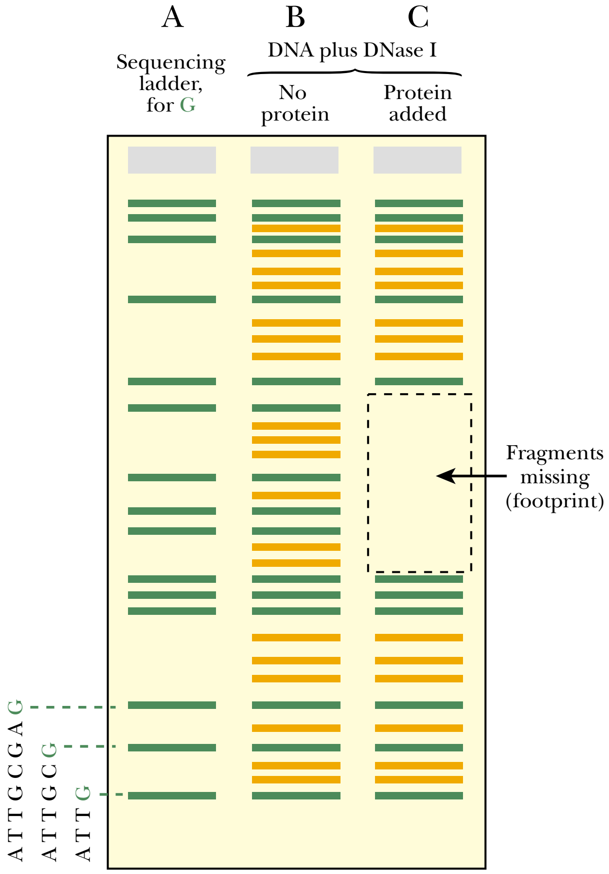
Gel retardation reveals which segment of DNA binds a protein. To locate the binding site more precisely, a **footprint** analysis is performed. In footprinting, the fragment of DNA that binds the protein is labeled at one end with radioactivity or fluorescence. As before, the sample of DNA is split into two portions and the protein is mixed with one batch. Both portions of the DNA are then treated with a small amount of a reagent that breaks DNA strands. **Deoxyribonuclease I (DNase I)** is often used because it is relatively non-specific and cuts DNA between any two nucleotides. Other chemical reagents that attack DNA may also be used. In either case, the DNA is attacked and degraded except in the region covered, and thus protected, by the protein (Fig. 25.11).

Only a small amount of DNase is used, just enough to cut each molecule of DNA once on average, in a random position. Consequently the sample of protected DNA will have certain fragments missing. In contrast, cutting a sample of unprotected DNA will give rise to a series of fragments of all possible lengths, varying by a single base pair. When the two samples are run on a gel side by side, we see a "footprint" (Fig. 25.12). In practice, the footprint is run side by side with a sequencing reaction (see Ch. 24), which allows matching the footprint with the DNA sequence.

deoxyribonuclease I (DNase I) Non-specific nuclease that cuts DNA between any two nucleotides. Often used in footprint analysis
footprint Method for testing binding of a protein to DNA by its protection of DNA from chemical degradation

FIGURE 25.12 ***Footprint Analysis—Results***

During the footprinting reaction, a sample of the DNA fragment containing the binding site is mixed with purified protein and DNase I. The DNase I cuts the DNA between any two nucleotides. The amount of DNase I is controlled so that each DNA fragment is only digested once. Since DNase does not cut the DNA where the protein is bound, sub-fragments of those particular lengths are absent in the sample containing the protein. The samples are run on a sequencing gel to separate the fragments, which differ by as little as one base pair. When Lane B (no protein) and Lane C (plus protein) are compared, it can be seen that Lane C shows no bands in the boxed region. Therefore the regulatory protein bound to DNA in this region and protected it from cutting by the DNase. Alignment with a sequencing ladder (lane A) allows the precise region of binding to be deduced.



Location of the Start of Transcription by Primer Extension

The start site for transcription may be located by isolating mRNA and using reverse transcriptase to make complementary DNA.

To fully understand the transcriptional regulation of a gene we need to know where transcription actually starts. **Primer extension** allows precise location of the start of transcription to the exact nucleotide. This approach involves binding of an artificial primer to messenger RNA. The primer, which is a DNA oligonucleotide, is then extended by reverse transcriptase, thus synthesizing DNA that is complementary to the mRNA (Fig. 25.13).

Since mRNA is needed, cells must first be grown under conditions where the gene of interest is highly expressed. The total mRNA is then extracted. An artificial DNA primer is synthesized so that it is complementary to a sequence close to the suspected start of transcription. The primer should be specific for the gene of interest and should therefore hybridize only with mRNA from this gene. The primer is hybridized to the mRNA and is extended by using reverse transcriptase.

The DNA/RNA hybrid is denatured and run on the same type of denaturing gel used in DNA sequencing (see Ch. 24). The extended DNA product must be labeled in some manner to allow visualization, either by radioactivity or fluorescence. The primer itself may be labeled or labeled nucleotides may be used during the extension step. The same primer as used for primer extension is also used to carry out a sequencing reaction using DNA corresponding to the transcribed region as template. The primer extension product will migrate to the same place size as the sequencing fragment that represents the precise transcriptional start site.

primer extension Method to locate the 5′ start site of transcription by using reverse transcriptase to extend a primer bound to mRNA so locating the 5′-end of the transcript

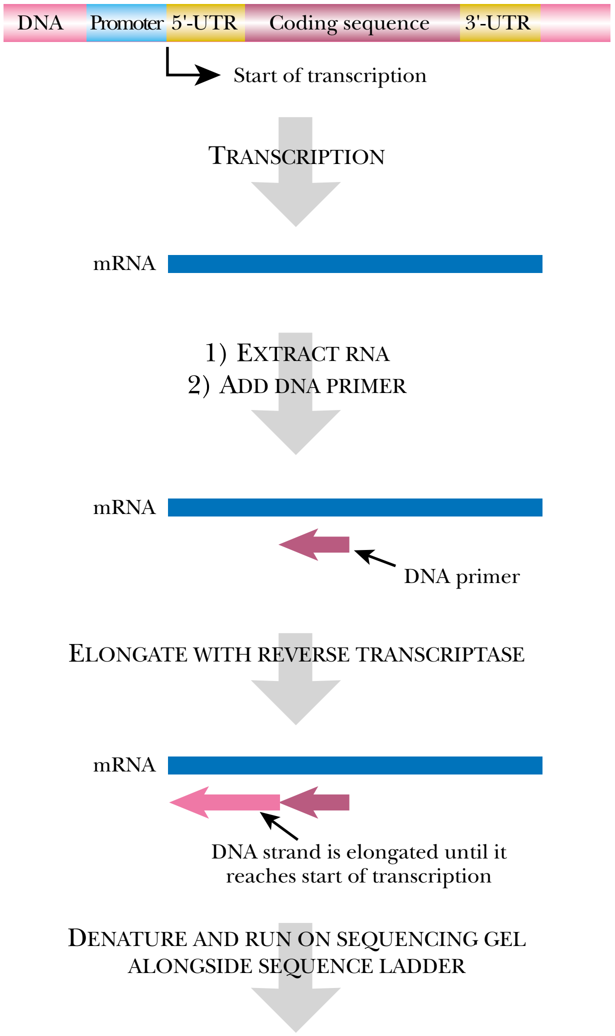

DNA
Promoter
5'-UTR
Coding sequence
3'-UTR
Start of transcription
TRANSCRIPTION
mRNA
1) EXTRACT RNA
2) ADD DNA PRIMER
mRNA
DNA primer
ELONGATE WITH REVERSE TRANSCRIPTASE
mRNA
DNA strand is elongated until it reaches start of transcription
DENATURE AND RUN ON SEQUENCING GEL ALONGSIDE SEQUENCE LADDER

FIGURE 25.13 ***Primer Extension Reveals Start of Transcription***

First, messenger RNA is isolated from cells that are expressing the gene of interest. A primer specific to the gene of interest is added and anneals to the mRNA. Reverse transcriptase makes a complementary DNA strand from the primer to the 5′ end of the mRNA (i.e. the start of transcription). The length of the primer extension DNA strand is found by running it on a gel next to a sequencing ladder of the same region of DNA. This allows determination of the exact start site.

Location of the Start of Transcription by S1 Nuclease

Another way to find the start of transcription is by hybridizing the mRNA to the corresponding DNA and cutting away the single stranded overhangs with S1 nuclease.

Another method to locate the start of transcription uses **S1 nuclease**. This is an endonuclease from *Aspergillus oryzae* that cleaves single-stranded RNA or DNA but does not cut double-stranded nucleic acids. The DNA carrying the suspected start of transcription must first be cloned onto a suitable plasmid vector. The DNA is digested with a restriction enzyme that yields a fragment containing the presumed start site. A more recent approach is to generate the fragment using PCR (see Ch. 23), thus avoiding the need for cloning. In either case the DNA must be denatured to give single strands before S1 analysis. An alternative is to clone the DNA fragment into an M13 vector, which gives single-stranded DNA directly (see Ch. 24 for the use of M13 to make ssDNA for sequencing).

The single-stranded DNA fragment is labeled on its 5′-end and then denatured and hybridized to the corresponding mRNA (Fig. 25.14). The 5′-end of the mRNA corresponds to the site where transcription started. DNA beyond this point remains single-stranded. The sample is next split into two portions. S1 nuclease is mixed with one half and degrades the single-stranded overhangs (DNA at one end, RNA at the other). The two samples of DNA, with and without nuclease treatment, are compared by running side-by-side on a denaturing gel. This allows the difference in length, and hence the location of the start site, to be estimated.

S1 nuclease Endonuclease from *Aspergillus oryzae* that cleaves single-stranded RNA or DNA but does not cut double-stranded nucleic acids

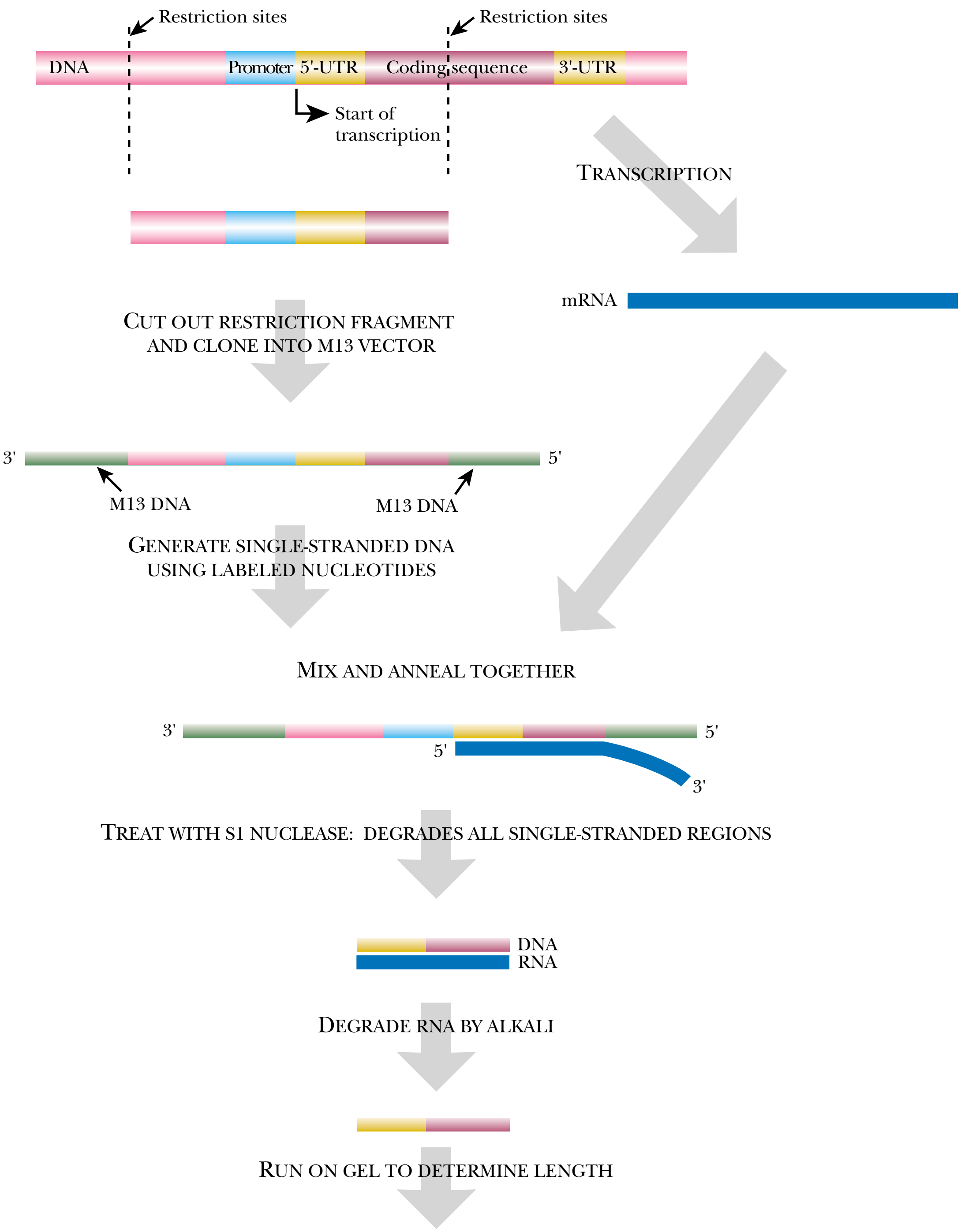

Restriction sites
Restriction sites
DNA
Promoter
5'-UTR
Coding sequence
3'-UTR
Start of transcription
TRANSCRIPTION
mRNA
CUT OUT RESTRICTION FRAGMENT AND CLONE INTO M13 VECTOR
3'
5'
M13 DNA
M13 DNA
GENERATE SINGLE-STRANDED DNA USING LABELED NUCLEOTIDES
MIX AND ANNEAL TOGETHER
3'
5'
5'
3'
TREAT WITH S1 NUCLEASE: DEGRADES ALL SINGLE-STRANDED REGIONS
DNA
RNA
DEGRADE RNA BY ALKALI
RUN ON GEL TO DETERMINE LENGTH

FIGURE 25.14 ***Locating Start of Transcription by S1 Nuclease***

The first step in mapping the transcriptional start site by S1 nuclease treatment is to clone the upstream region of the gene into an M13 vector. Next, single-stranded M13 DNA is prepared using labeled nucleotide precursors for use as a probe. The labeled single-stranded DNA is mixed with the total cellular mRNA. The mRNA with sequence complementary to the DNA will hybridize with the DNA. S1 nuclease is added to the mixture to digest all the single-stranded RNA and DNA. All that is left is the DNA:RNA hybrid, which is isolated from the degraded nucleotides by precipitation. The DNA portion of the hybrid is isolated by alkali treatment and the length determined by agarose gel electrophoresis.

Modification of this technique allows **S1 nuclease mapping** to be used for locating the 3′-end of a transcript. In this case, the DNA probe includes the fragment with the suspected transcriptional stop site. S1 nuclease may sometimes degrade the ends of the RNA/DNA hybrid slightly or may not fully digest the single-stranded regions. Consequently, the S1 nuclease method is not as accurate as primer extension. However, primer extension cannot locate the 3′-end of a transcript and S1 nuclease mapping is the best way to achieve this.

Transcriptome Analysis

The transcriptome is the mixture of RNA that results from transcribing the genome.

The transcriptome refers to all of the various RNA molecules that result from transcription in a particular cell. Usually, interest is focused on the messenger RNA but, strictly speaking, the transcriptome also includes non-coding RNA. Unlike the genome, the transcriptome varies as different genes are expressed under different conditions. Transcriptome analysis attempts to measure the levels of all transcribed RNAs simultaneously. Several techniques are available for monitoring multiple RNA molecules.

Differential display PCR has already been described (Ch. 23). This method identifies mRNA molecules (after conversion to cDNA) that are expressed differentially in two samples. If sufficient different primer combinations are used, differential display PCR can in theory identify all the transcripts that are present. Although widely used, the results are often not quantitative and since PCR fragments of the same size may be generated from more than one gene this approach may provide ambiguous results. More recently, real time PCR has been developed (see Ch. 23 for details). This allows rapid monitoring of mRNA levels and hence of gene expression. The use of DNA microarrays and SAGE (see below) are more accurate for simultaneous monitoring of large numbers of mRNAs.

DNA Microarrays for Gene Expression

DNA microarrays can be used to detect gene expression by hybridizing the array to messenger RNA.

We have outlined the principles of the **DNA microarray** in Chapter 24 while discussing its use in sequencing DNA and in the diagnostic detection of particular DNA sequences. In these cases, DNA immobilized on the chip hybridizes to the target DNA fragments in the sample to be analyzed. Since DNA microarrays work by hybridization, they can also be used to monitor RNA. Microarrays are very expensive and analysis of the data is highly labor intensive, despite computerized analysis. If only one or a few genes are the objects of interest, other methods such as performing Northern hybridization to detect mRNA or using a reporter gene to measure the level of transcription are more appropriate.

For total transcriptome analysis, the solid support (i.e. the "chip") must carry DNA sequences complementary to all of the possible mRNA molecules that a cell might express. The DNA is robotically printed onto a nylon membrane or a glass slide. Current technology can print about 100,000 spots of DNA per cm^2, with glass slides capable of carrying higher densities than nylon membranes. The mRNA to be surveyed must be extracted and labeled, either with a radioactive isotope or more often with a fluorescent dye. After binding to the chip, the intensity of each spot gives an estimate of the amount of each mRNA present. The pattern of mRNA expression in different samples of mRNA, from cells grown under different conditions can be compared. If the two samples of RNA are labeled with different fluorescent dyes (e.g., Cy3 which is green and Cy5 which is red), both RNA samples may be hybridized to the same DNA chip. Red spots will show genes expressed under condition one, green spots will show genes expressed under condition two and yellow spots result when genes are expressed under both conditions (Fig. 25.15; Fig. 25.16).

cDNA arrays use the cDNA versions of whole genes.

In practice, two types of DNA microarray are used for binding mRNA, arrays of cDNA or arrays of oligonucleotides. Arrays of cDNA use PCR products generated from a cDNA library as the immobilized DNA. One problem is the existence of gene families (e.g., the globin family), whose individual members are highly homologous and

DNA microarray or DNA array DNA chip used to simultaneously detect and identify many short RNA or DNA fragments by hybridization. Also known as DNA chip or oligonucleotide array detector

S1 nuclease mapping Method using S1 nuclease to locate the 5′-end or 3′-end of a transcript

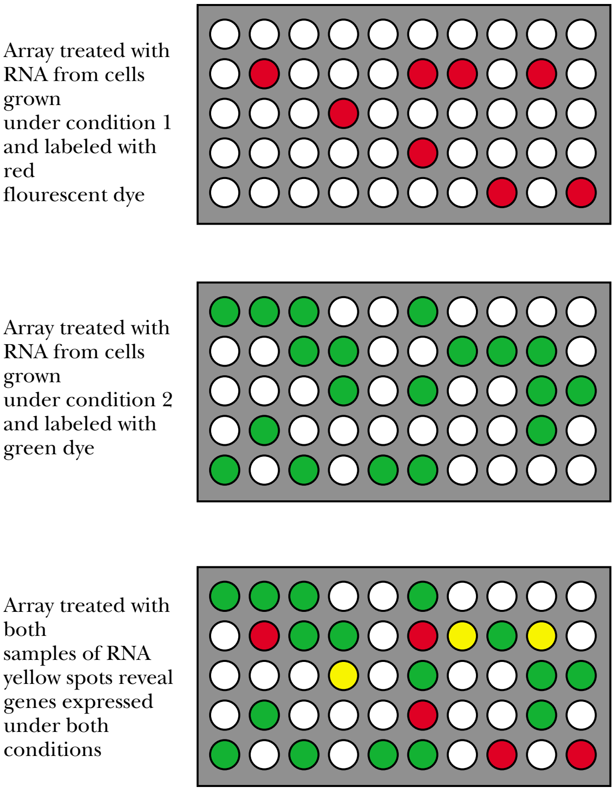

Array treated with RNA from cells grown under condition 1 and labeled with red flourescent dye
Array treated with RNA from cells grown under condition 2 and labeled with green dye
Array treated with both samples of RNA yellow spots reveal genes expressed under both conditions

FIGURE 25.15 ***DNA Chip Showing Detection of mRNA by Fluorescent Dyes***

DNA chips can monitor many different mRNAs at one time. Each spot on the grid has a different DNA sequence attached. To determine which genes are expressed under which conditions, mRNA is isolated. In this case mRNA isolated from cells grown under two different conditions is labeled with two different fluorescent dyes. Under condition one (red dye), eight different mRNAs hybridized to DNA spots on the chip. Under condition two, nineteen different mRNAs were seen (green dye). Since two different color dyes were used, both samples can be analyzed on the same chip. In this case, the mRNAs that are expressed under both conditions give yellow spots.

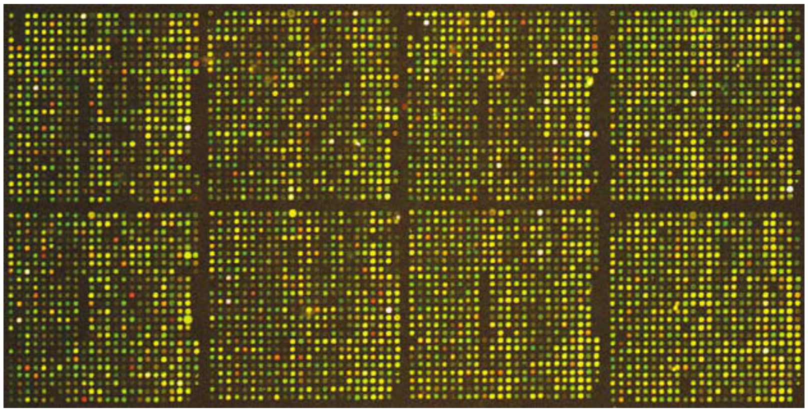


FIGURE 25.16 ***Hybridization of mRNA to a 19,200 Element Array***

RNA from related human colon carcinoma and reference cell-lines was reverse transcribed and the cDNA was labeled with Cy-5 (red) and Cy-3 (green), respectively. The cDNA was then hybridized to a microarray containing 19,200 distinct human cDNA clones. Genes expressed by the cancer cells are shown in red and those from the normal cells are green. Yellow spots indicate expression in both cell lines. Credit: Hegde et al., The Institute for Genomic Research, Rockville, MD.

may cross-hybridize. To avoid this, it is normal to use sequences from the 3′-end of the cDNA, which often include part of the 3′-untranslated region of the mRNA transcript. Non-coding sequences diverge much more than coding sequences, and so are much less likely to cross-hybridize.

Oligonucleotide arrays use short synthetic segments of single-stranded DNA.

Oligonucleotide arrays use synthetic segments of single-stranded DNA, usually 20–25 nucleotides long, as the immobilized DNA. To make such oligonucleotides it is obviously necessary to know the sequences of the genes to be monitored. A particular sequence of n bases will occur, by chance, in DNA every 4^n bases. For a mammalian genome with 3×10^9 bases, n must be at least 16 for a sequence to be unique. It is safer to make oligonucleotides longer than this minimum and, for example, the GeneChip® arrays made by Affymetrix Corporation use 25-mers. Furthermore, multiple different oligonucleotides are included at different locations on the chip for each mRNA.

The oligonucleotides of the GeneChip® array are synthesized directly on the chip (Fig. 25.17). This is done using the techniques of photolithography developed for use

oligonucleotide array DNA array used to simultaneously detect and identify many short RNA or DNA fragments by hybridization. Also known as DNA array or DNA chip

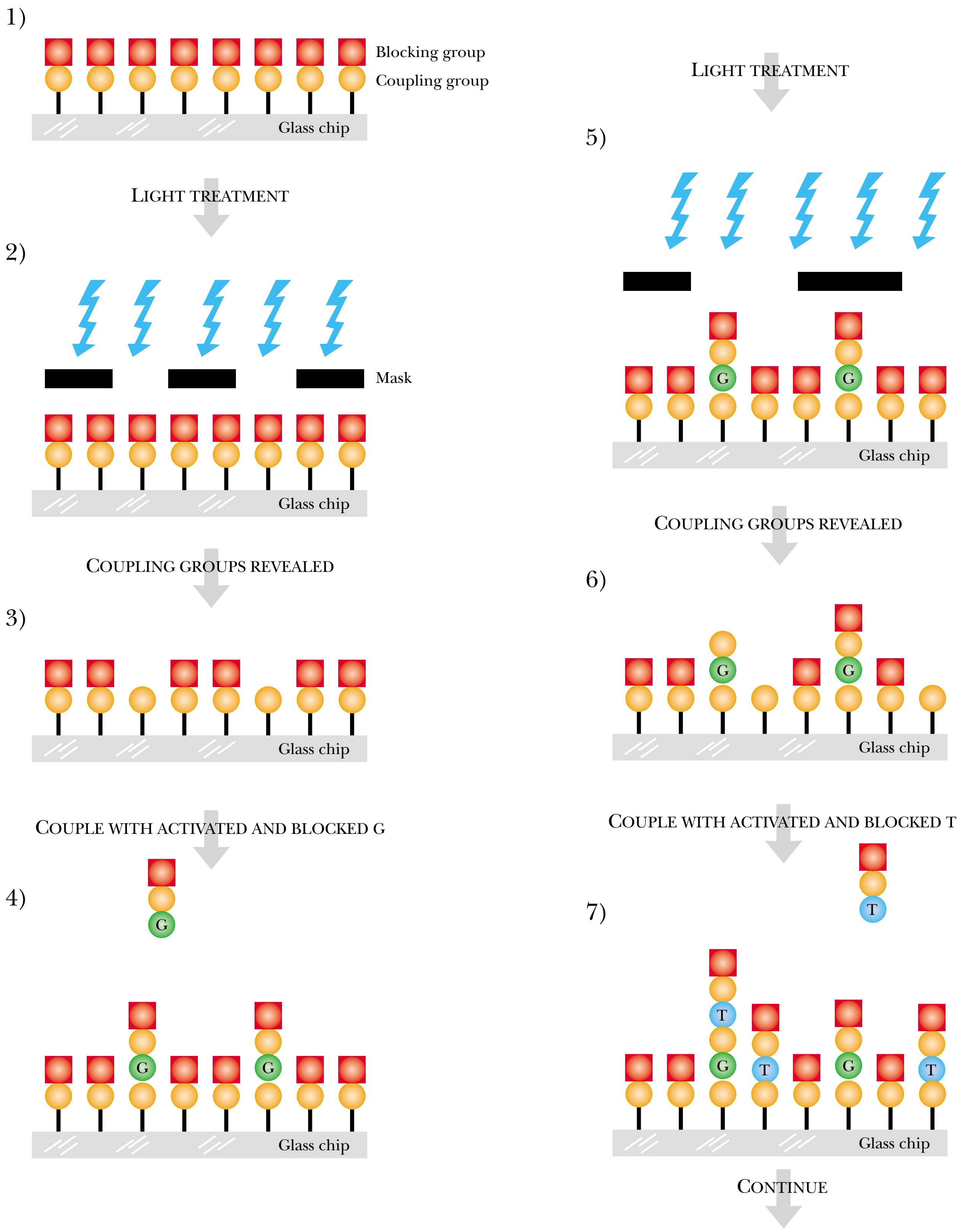


FIGURE 25.17 ***On-Chip Synthesis of Oligonucleotides***

Arrays may be created by chemically synthesizing oligonucleotides directly on the chip. First reactive groups are linked to the glass chip and blocked. Then each of the four nucleotides is added in turn (in this example, G is added first, then T). A mask covers the areas that should not be activated during any particular reaction. Light activates all the groups not covered with the mask, and a nucleotide is added. The cycle is repeated with the next nucleotide.

Improved On-Chip Synthesis of Oligonucleotides with Virtual Masks

The original on-chip procedure for making microarrays uses physical chrome and glass masks. A chip that uses oligonucleotides of length N needs 4N such masks. This results in both a high cost and lengthy construction time for the array. Avoiding physical masks greatly reduces fabrication cost and allows greatly increased flexibility in designing custom arrays. The NimbleGen corporation has recently introduced a proprietary maskless technology into microarray synthesis. The physical mask is replaced by a computer generated "virtual mask" which controls a digital micromirror array. This is an array of tiny, individually addressable polished mirrors that can be positioned either to direct the UV light source onto a known position in the array or direct light away from the array. By coordinating the addition of protected phosphoramidite precursors and the sequence of illumination it is possible to make custom arrays with more than 200,000 separate oligonucleotide probes.

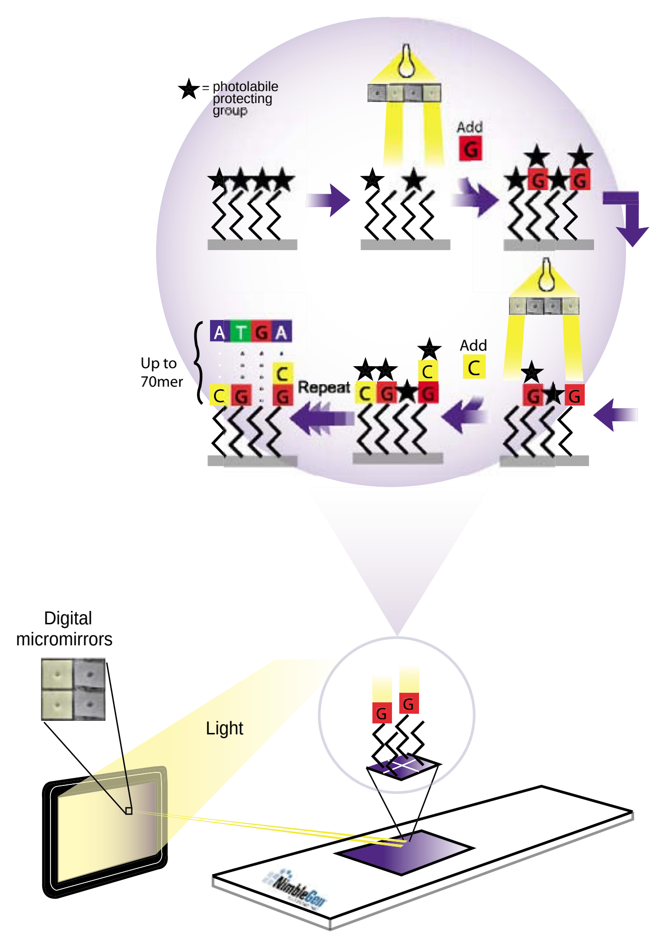


FIGURE 25.18 ***Virtual Mask Process for On-Chip Synthesis of Oligonucleotides***

NimbleGen builds its arrays using photo deprotection chemistry with its Maskless Array Synthesizer (MAS) system. At the heart of the system is a Digital Micromirror Device (DMD; Texas Instruments, Inc.), employing a solid state array of miniature aluminum mirrors to pattern up to 786,000 individual pixels of light. The DMD creates "virtual masks" that replace the inflexible physical chromium masks used in traditional arrays.

in building computer chips. A glass slide is first covered with a reactive group. This is then covered with a photosensitive blocking group that can be removed by light. In each synthetic cycle, those sites where a nucleotide will not be attached are covered with a mask. Those sites where a particular nucleotide (say, A) is to be attached are illuminated to remove the blocking group. The nucleotide is then added and is chemically coupled to the exposed sites. Only one kind of nucleotide can be added at a time, as it will couple to all exposed sites. Also, the other end of the added nucleotide must be blocked before addition and coupling. The cycle is repeated with another nucleotide (say, T). This cycling process is repeated with different masking patterns and different nucleotides until the required oligonucleotides are finished.

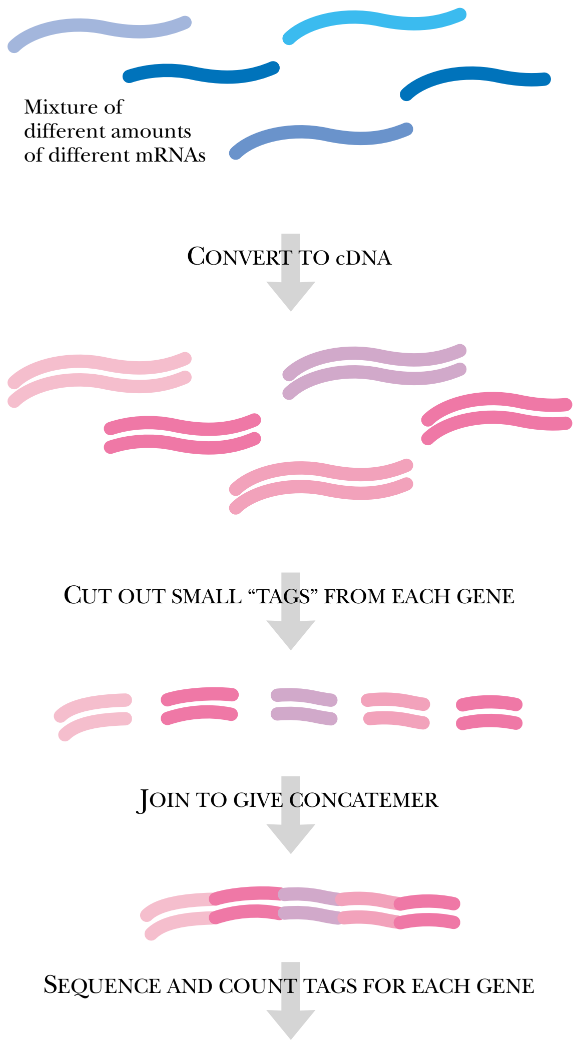


FIGURE 25.19 ***SAGE—The Principle***

To analyze the total mRNA expressed in a cell, small sequences from each mRNA are converted to complementary DNA and linked together into one long concatemer, which is sequenced. Each of the segments represents a single mRNA; therefore, the number of repeats of each segments correlates with the level of expression of the corresponding gene in the cell.

Serial Analysis of Gene Expression (SAGE)

If all the mRNA molecules in a cell were joined end to end and then sequenced, this would reveal how many copies of each mRNA were present—hence the level of gene expression.

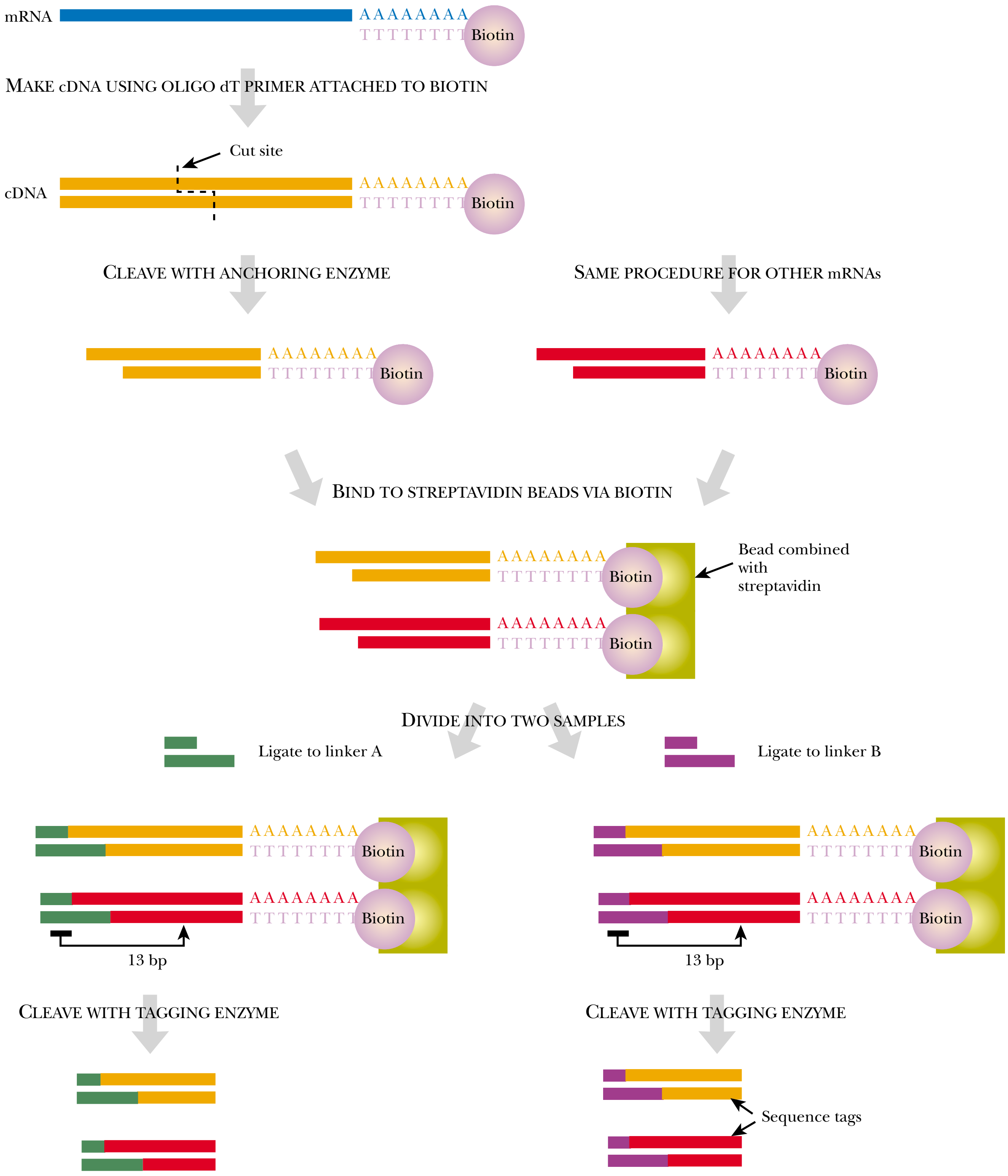
A DNA sequencing approach can also measure the expression level of multiple mRNA molecules simultaneously. The basic idea is to sequence all of the mRNA in a cell and then examine the accumulated sequence to see how many copies of each mRNA are represented. To actually do this the mRNA molecules must be joined end to end to give a single giant concatenated molecule, which is converted to DNA for sequencing. The term **serial analysis of gene expression (SAGE)** refers to this large concatemer, which contains every expressed gene. The number of copies of each repeat in the concatemer indicates the level of gene expression. To make the approach feasible, only a short sequence from each mRNA is sequenced. The DNA concatemer thus contains many linked sequence tags of approximately 10 bases each (Fig. 25.19).

The first step in SAGE is to extract all the mRNA from a eukaryotic cell and convert it into cDNA using an oligo(dT) primer that hybridizes to the poly(A) tail of the mRNA (Fig. 25.20). The oligo(dT) primer also carries a **biotin** tag that can be bound by the protein **streptavidin**. The cDNA is cleaved by a restriction enzyme (the "anchor-

biotin Vitamin that is widely used to label or tag nucleic acids in molecular biology because it may be bound very tightly by avidin or streptavidin

serial analysis of gene expression (SAGE) Method to monitor level of multiple mRNA molecules by sequencing a DNA concatemer that contains many serially linked sequence tags derived from the mRNAs

streptavidin Protein from *Streptomyces* that binds biotin extremely tightly and specifically. Used in detection procedures for molecules labeled with biotin


mRNA
AAAAAAAA
TTTTTTTT
Biotin
MAKE cDNA USING OLIGO dT PRIMER ATTACHED TO BIOTIN
Cut site
cDNA
AAAAAAAA
TTTTTTTT
Biotin
CLEAVE WITH ANCHORING ENZYME
SAME PROCEDURE FOR OTHER mRNAs
AAAAAAAA
TTTTTTTT
Biotin
AAAAAAAA
TTTTTTTT
Biotin
BIND TO STREPTAVIDIN BEADS VIA BIOTIN
AAAAAAAA
TTTTTTTT
Biotin
AAAAAAAA
TTTTTTTT
Biotin
Bead combined with streptavidin
DIVIDE INTO TWO SAMPLES
Ligate to linker A
Ligate to linker B
AAAAAAAA
TTTTTTTT
Biotin
AAAAAAAA
TTTTTTTT
Biotin
13 bp
AAAAAAAA
TTTTTTTT
Biotin
AAAAAAAA
TTTTTTTT
Biotin
13 bp
CLEAVE WITH TAGGING ENZYME
CLEAVE WITH TAGGING ENZYME
Sequence tags

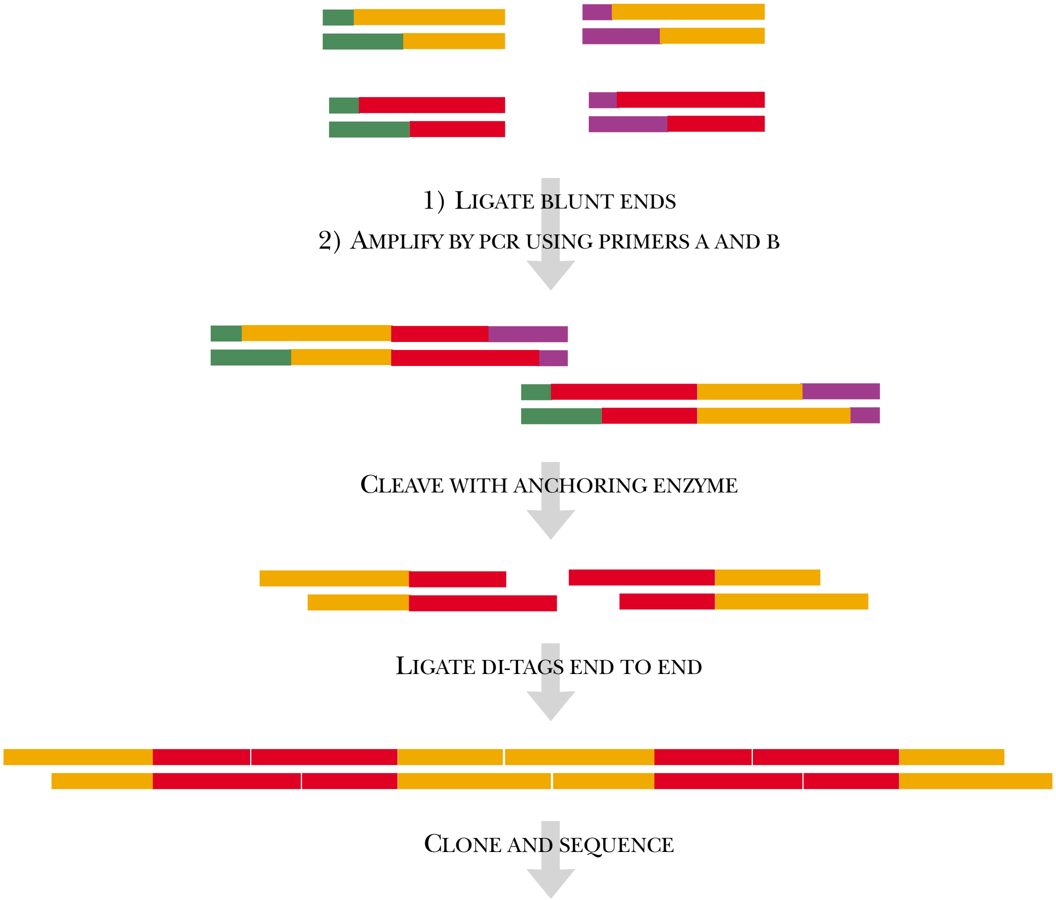

1) LIGATE BLUNT ENDS
2) AMPLIFY BY PCR USING PRIMERS A AND B
CLEAVE WITH ANCHORING ENZYME
LIGATE DI-TAGS END TO END
CLONE AND SEQUENCE

FIGURE 25.20 ***SAGE—The Procedure***

The first step in making long concatemers of expressed sequences involves isolating the total cellular mRNA and making the corresponding cDNA. The total mRNA is bound via its poly(A) tail to an oligo(dT) primer linked to biotin. It is then converted to cDNA using reverse transcriptase.

The cDNAs are then truncated to short, tagged sequences. First, the cDNAs are cleaved with a restriction enzyme known as the anchoring enzyme. This generates a pool of shortened cDNA averaging 256 bp long, with some longer and others shorter. These are isolated using streptavidin, which binds to the biotin tag on the poly(A) tail end of the cDNA. This mixture is divided into two samples and each is ligated to a different linker. This linker has two features: (a) its overhang matches the overhang generated previously by the anchoring enzyme, and (b) it has a recognition site for a type II restriction enzyme (known as the tagging enzyme). Each sample is cut with the tagging enzyme. This enzyme recognizes the sequence in the linker, but actually makes a blunt end cut downstream in the cDNA sequence. This generates two pools of small cDNA sequence tags with different linkers.

Finally, the sequence tags are joined into one long sequence. First, fragments are linked by blunt-end ligation. Then PCR primers complementary to the linkers are used to amplify only those ligated molecules that have linker A and linker B flanking two different sequence tags. The PCR products are digested with the anchoring enzyme to remove the linkers and generate sticky ends. These are ligated and the resulting fragment is cloned and sequenced.

ing enzyme") with a 4 bp recognition site and which generates sticky ends. This gives fragments of average length 256 bp, which are then bound to streptavidin-coated magnetic beads. This method generates a library of 3′ ends, many just containing the 3′ UTR region. The 3′ UTR sequence is more divergent, therefore, mRNAs from highly homologous families can be more readily distinguished using this procedure.

When this approach is put into practice, the RNA is converted to DNA for sequencing and only a short segment from each RNA is sequenced.

The collected fragments are divided into two samples. The two sets of fragments are ligated to two different artificial linkers that contain recognition sites for a **type II restriction enzyme** (the "tagging enzyme"). Type II enzymes cut a fixed number of bases away from their recognition sites. A favorite choice is FokI which cuts 13 bp downstream and so leaves 9 bp of the original mRNA sequence (the "tag") plus the 4 bp anchoring enzyme site attached to the linker. The fragments are then blunt-end ligated head to tail, to give structures containing two mRNA-derived tags flanked by linker A and linker B. This structure is used as the target for PCR using two primers, a forward primer that binds to linker A and a backward primer that binds to linker B. (Other structures are formed during the ligation, but only those with the desired structure will be amplified by using this pair of PCR primers.) The PCR products are cleaved with the anchoring enzyme to give tag-dimers with sticky ends and these are ligated to give the DNA concatemer. Finally, the DNA concatemer is cloned and sequenced. The tags are identified and counted to indicate the relative levels of the original mRNA molecules.

type II restriction enzyme Type or restriction enzyme that cuts a fixed number of bases away from its recognition site

CHAPTER TWENTY-SIX

Proteomics: The Global Analysis of Proteins

Introduction to Proteomics

Gel Electrophoresis of Proteins

Two Dimensional PAGE of Proteins

Western Blotting of Proteins

Mass Spectrometry for Protein Identification

Protein Tagging Systems

Full-Length Proteins Used as Fusion Tags

Self Cleavable Intein Tags

Selection by Phage Display

Protein Interactions: The Yeast Two-Hybrid System

Protein Interaction by Co-Immunoprecipitation

Protein Arrays

Metabolomics

Introduction to Proteomics

Today we know the complete genome sequences for many microorganisms and a handful of higher organisms including ourselves. However, we still have no idea what the function of most genes might be. The availability of this novel ocean of ignorance to explore has led some scientists to designate the 21st century as the post-genomic or proteomic era. For most genes, proteins are the final gene products. Proteins are made of 20 different amino acids and vary greatly in their 3D structure. They also vary in both function and stability, as described in Ch. 7. Consequently, elucidating the role of proteins, especially on a large scale, is in many ways more difficult than for nucleic acids. The term **proteome** was originally defined as the total set of proteins encoded by a genome. Alternatively, it may be viewed as the total protein complement of an organism. The term **translatome** is sometimes used to refer to all the proteins that are present in a cell under any particular set of conditions, that is, those that have actually been translated. This is distinct from the proteome, which consists of all those proteins that are potentially available. The relationship between genome and proteome/translatome is not simply linear. Many proteins are processed and modified by other proteins; therefore, the final protein complement depends on complex interactions between proteins that may vary depending on the growth conditions.

Proteomics involves surveying the global protein composition of a cell or organism.

Although gene expression is often assessed by monitoring transcription, mRNA levels do not always correspond to the levels or activity of the final gene products, the encoded proteins. The disparity between transcriptome and proteome may be summarized as follows:

a. Some RNA molecules are non-coding and do not give rise to any protein products.

b. Some primary RNA transcripts undergo alternative splicing; therefore, the gene may give rise to multiple protein products.

c. Levels of mRNA may not correlate with protein levels due to differential rates of mRNA translation or degradation.

d. The activity of many proteins is regulated after translation by addition or removal of acetyl, phosphate, AMP, ADP-ribose, or other groups.

e. The activity of many proteins is altered after translation by chemical modification of amino acid residues.

f. Many proteins are processed after translation, e.g., by proteolytic cleavage or addition of sugar or lipid residues to give glycoproteins or lipoproteins.

g. Proteins themselves may be degraded and vary greatly in stability.

The various modifications listed above may all vary depending on the growth conditions and the activities of other genes and/or proteins. Thus in practice it may be necessary to monitor not only the level but also the activity of proteins.

Different proteins differ from one another much more in both structure and function than nucleic acid molecules.

The fundamental problem in proteomics is the individuality of different proteins. Since proteins differ in structure, stability, solubility, charge, activity, etc., traditionally each individual protein had to be purified and assayed by a different procedure. Proteomics requires the parallel analysis of multiple proteins and relies on methods that are applicable to many different proteins and are not affected by variations in 3-D structure etc. In practice, separation of proteins for proteomics usually relies on denaturation followed by 2-D gel electrophoresis. Identification is often based on recently developed improvements in mass spectrometry, especially the MALDI/TOF

proteome The total set of proteins encoded by a genome or the total protein complement of an organism
translatome The total set of proteins that have actually been translated and are present in a cell under any particular set of conditions

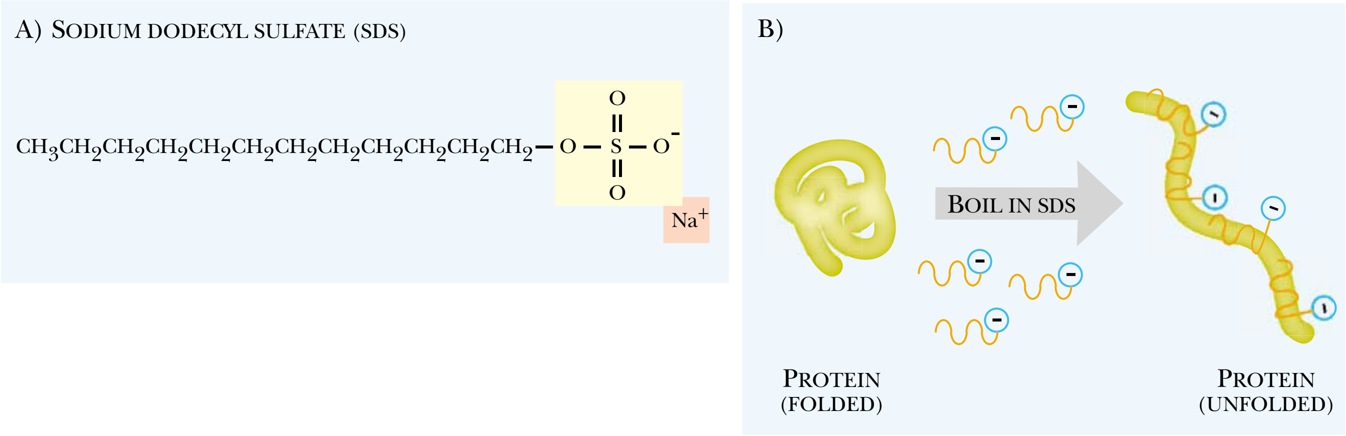

A) SODIUM DODECYL SULFATE (SDS)
$CH_3CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2-O-S(=O)_2-O^-$
Na^+
B)
BOIL IN SDS
PROTEIN (FOLDED)
PROTEIN (UNFOLDED)

FIGURE 26.01 ***Denaturation of Protein by SDS***

(A) The structure of SDS is amphipathic, that is, the long hydrocarbon tail is hydrophobic and the sulfate is hydrophilic. (B) The first step to separate a mixture of proteins by size is to boil the sample. Boiling proteins in a solution of SDS destroys the tertiary structure of the protein. The hydrophobic portion of SDS coats the polypeptide backbone and prevents the protein from refolding. The hydrophilic group of SDS keeps the protein soluble in water, and the negative charges repel each other, which also helps to keep the protein from refolding. The result is an unfolded protein with a net negative charge that is proportional to its molecular weight.

approach. Generic methods for purification of proteins rely on attaching the same tag molecule (e.g., His-tag, FLAG or GST) to many different proteins and then binding the tag.

Gel Electrophoresis of Proteins

Because nucleic acids are all negatively charged they all move towards the positive electrode during electrophoresis (Ch. 21). However, proteins are not so convenient. Some of the amino acids from which proteins are built have a positive charge, some have a negative charge, and most are neutral. So, depending on its overall amino acid composition, a protein may be positive, negative or neutral.

If a mixture of native (i.e. non-denatured) proteins is run on a gel, some move towards the positive electrode, others towards the negative electrode whereas neutral proteins scarcely move at all. Consequently, the samples are usually started in the middle of the gel, rather than at the end. This is referred to a native protein electrophoresis and is sometimes used to purify proteins without inactivating them. After running, the gel is stained with reagents specific to the protein of interest. For example, an enzyme that gives a colored product may be located in this manner.

When proteins are separated by size using gel electrophoresis, they are first denatured and coated with negatively charged detergent molecules.

To separate proteins on the basis of molecular weight, they are first boiled in a solution of the detergent **sodium dodecyl sulfate (SDS)**. Boiling in detergent destroys the folded 3-D structure of the protein; that is, the protein is denatured. The SDS molecule has a hydrophobic tail with a negative charge at the end. The tail wraps around the backbone of the protein and the negative charge dangles in the water. The protein is unrolled and covered from head to toe with SDS molecules, which give it an overall negative charge (Fig. 26.01). In addition, the disulfide bonds that help maintain the tertiary structure of some proteins and/or hold protein subunits together must be disrupted for proper denaturation. This is done by adding small sulfhydryl reagents, usually β-mercaptoethanol ($HS—CH_2CH_2OH$):

sodium dodecyl sulfate (SDS) Detergent used to unfold proteins and cover them with negative charges for electrophoresis

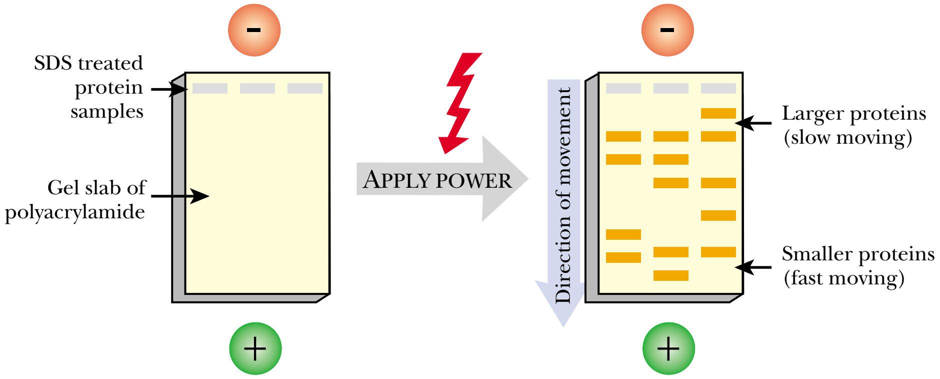

SDS treated protein samples
Gel slab of polyacrylamide
-
+
APPLY POWER
Direction of movement
Larger proteins (slow moving)
Smaller proteins (fast moving)

FIGURE 26.02 ***SDS Polyacrylamide Gel Electrophoresis***

Proteins treated with SDS can be separated by size using gel electrophoresis. Since all the proteins have a net negative charge, the proteins are repelled by the negative cathode and attracted to the positive anode. As the proteins move toward the anode, the polyacrylamide meshwork obstructs and slows the larger proteins but allows the smaller proteins to move faster. In consequence, the distance traveled in a given time is proportional to the log of the molecular weight. After separation, the proteins are visualized with a dye such as Coomassie blue or a silver compound.

$$\text{Protein-S—S-Protein} + 2\ \text{SH—CH}_2\text{CH}_2\text{OH} \rightarrow$$
$$2\ \text{Protein-SH} + \text{HO—CH}_2\text{CH}_2\text{—S—S—CH}_2\text{CH}_2\text{OH}$$

Furthermore, the number of negative charges bound is proportional to the length of the protein. Thus proteins can be separated according to size by running them through a gel (Fig. 26.02). Because proteins are a lot smaller on average than DNA or RNA, the gel is made of the artificial polymer polyacrylamide, which gives smaller gaps in its meshwork than agarose. The technique is thus known as **PAGE** or **polyacrylamide gel electrophoresis**. After running, the gel is stained to visualize the protein bands. The two favorite choices are **Coomassie Blue**, a blue dye that binds strongly to proteins, or silver compounds. Silver atoms bind very tightly to proteins and yield black or purple complexes. Silver staining is more sensitive and, of course, more expensive.

Two Dimensional PAGE of Proteins

Much larger numbers of proteins can be separated by gel electrophoresis in two directions.

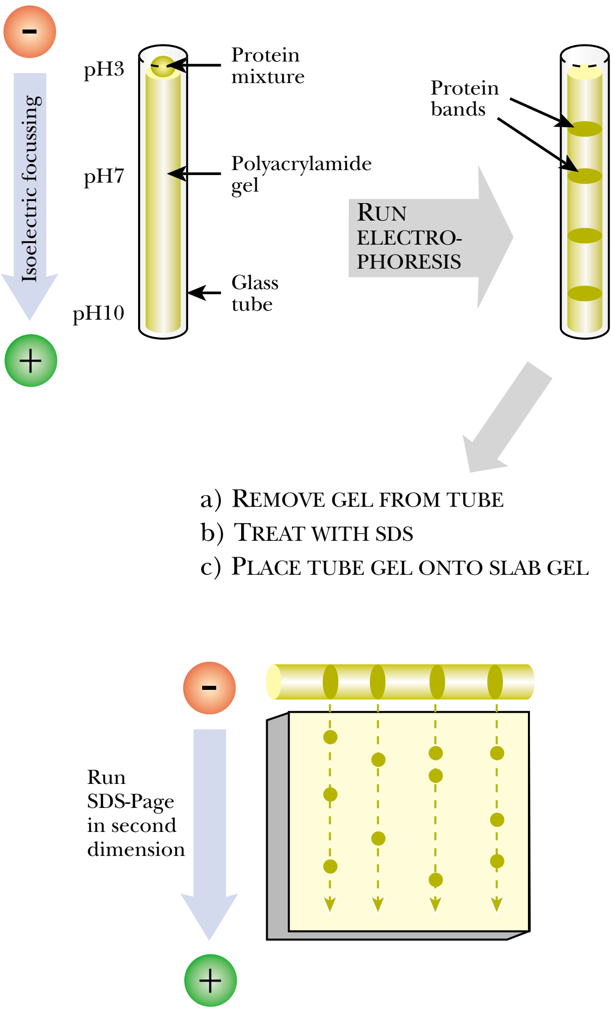
Separation of large numbers of proteins is normally done by two-dimensional polyacrylamide gel electrophoresis (2D PAGE). The proteins are separated by charge in the first dimension and then by size in the second dimension. **Isoelectric focusing** is used in the first dimension and separates native proteins based on their original charge. A pH gradient is set up along a cylindrical gel and proteins migrate until they find a position where their native charges are neutralized—the isoelectric point. Standard SDS-PAGE as described above is used in the second dimension and separates denatured proteins based on their molecular weight (Fig. 26.03).

Early 2D gels were able to resolve a 1,000 or so protein spots and were used to characterize the protein complement of bacteria such as *E. coli* where about 1,000 of the 4,000 genes are expressed at any given time (see Ch. 9 Fig. 9.01). More recently, large 2D gels with higher resolution have been developed that allow separation of over 10,000 spots and can be used to analyze the proteome of higher organisms (Fig. 26.04). After separation, the protein spots are cut out from the gel, digested by protease

Coomassie Blue A blue dye used to stain proteins
isoelectric focusing Technique for separating proteins according to their charge by means of electrophoresis through a pH gradient
PAGE Polyacrylamide gel electrophoresis. Technique for separating proteins by electrophoresis on a gel made from polyacrylamide
polyacrylamide gel electrophoresis (PAGE) Technique for separating proteins by electrophoresis on a gel made from polyacrylamide

FIGURE 26.03 ***Two-Dimensional Polyacrylamide Gel Electrophoresis***

The first step in separating large numbers of proteins in two dimensions is to separate them according to their inherent charge. The mixture of proteins is loaded onto a gel that has a gradient of increasing pH. An electric field is applied and the proteins move along the pH gradient until they reach the point at which their charges are neutralized. At this point, each band in the gel contains several different proteins with the same (or very similar) isoelectric point. The tube gel is removed from its tube and exposed to SDS to denature the proteins. It is then placed on a slab of polyacrylamide gel and traditional SDS-PAGE is run in the second dimension to separate the proteins by size. After staining, the result of 2D-PAGE is a square with small scattered dots representing individual proteins.



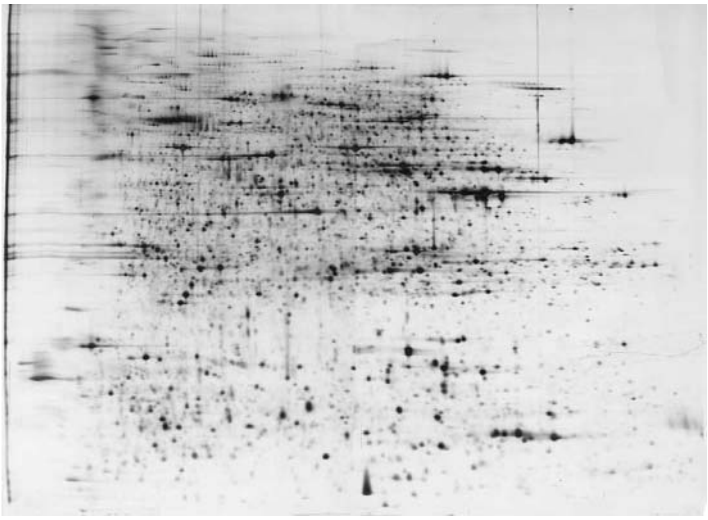


FIGURE 26.04 ***2D Protein Gel of Mouse Brain Tissue***

Soluble proteins were extracted from mouse brain and separated by 2D PAGE. The first dimension used isoelectric focusing and the second dimension used standard SDS-PAGE. The proteins were visualized by silver staining. Each spot on the gel is due to a separate protein. However, because proteins are frequently modified after synthesis, multiple spots sometimes arise from variants of the same original protein. Courtesy of Prof. Dr. Joachim Klose, Institut für Humangenetik, Humboldt-Universität, Berlin.

treatment and the resulting peptides are analyzed by mass spectrometry (see below). This allows unambiguous identification of each protein spot.

Proteins are visualized with a variety of dyes.

To detect the presence or absence of a particular protein, a sensitive silver stain is used. For quantitation of protein spots either Coomassie blue or fluorescent dyes are used to stain the gel. The gel is then scanned with a laser. A variety of fluorescent dyes are used, especially in large-scale proteomics work. SYPRO Orange and Red dyes bind proteins non-covalently. These are non-fluorescent in aqueous solution but fluoresce when in a non-polar environment, including when bound to protein-SDS complexes. It is also possible to covalently label two protein samples with differently colored dyes, say methyl-Cy5 and propyl-Cy3, and run them on the same gel. Differences between the samples may then be directly visualized (pure red and green spots indicate that a protein was found in one or other sample and a yellow spot indicates its presence in both).

Western Blotting of Proteins

Proteins may be detected by binding to a specific antibody.

Much like a Southern or Northern blot identifies one specific DNA or RNA (see Ch. 21), a **Western blot** allows detection of a single protein within a sample of many proteins. First, the sample of proteins is separated by size using SDS-PAGE or 2D-SDS-PAGE. The proteins are then electrophoretically transferred to a solid membrane such as nitrocellulose. Electrophoresis moves the proteins from the gel and onto the nitrocellulose where the proteins adhere (Fig. 26.05).

To detect a specific protein, an antibody to that protein must be available. An antibody can either be produced for the protein of interest or sometimes purchased commercially. The nitrocellulose membrane itself has many non-specific sites that can bind proteins, including antibodies. These sites must be blocked with a non-specific protein solution such as re-hydrated powdered milk. The primary antibody is added in the milk solution and binds to the protein of interest. The antibody protein complex is detected using a secondary antibody that has a label attached to it (Fig. 26.06). Often a reporter enzyme such as alkaline phosphatase is linked to the secondary antibody, and the addition of lumiphos or X-phos to the blot allows detection of the protein band (see Ch. 21).

Mass Spectrometry for Protein Identification

Automated mass spectrometry may be used to identify proteins in large numbers of samples.

Analysis of proteins and peptides by **mass spectrometry** relies on several recently developed techniques that are both extremely accurate and may be automated. The two most important are **matrix-assisted laser desorption-ionization (MALDI)** and **electrospray ionization (ESI).** Mass spectrometry (MS) measures the mass to charge ratio (*m/z*) of ions and allows derivation of the molecular weight. Before MALDI and ESI large heat-labile molecules such as proteins could not be analyzed by mass spectrometry.

In MALDI, gas-phase ions are generated from a solid sample by a pulsed laser. First, the sample protein or peptide is crystallized along with a matrix that absorbs at the wavelength of the laser. Matrix materials are usually aromatic acids such as 4-methoxy cinnamic acid. The laser excites the matrix material, which transfers the energy to the crystallized protein. The energy then releases ions, the size and charge of which are unique to each protein. The ions are accelerated by a high voltage electric field and travel in a vacuum through a tube to a detector. The **time-of-flight (TOF)**

electrospray ionization (ESI) Type of mass spectrometry in which gas-phase ions are generated from ions in solution
MALDI Matrix-assisted laser desorption-ionization. Type of mass spectrometry in which gas-phase ions are generated from a solid sample by a pulsed laser
mass spectrometry Technique for measuring the mass of molecular ions derived from volatilized molecules
matrix-assisted laser desorption-ionization (MALDI) Type of mass spectrometry in which gas-phase ions are generated from a solid sample by a pulsed laser
time-of-flight (TOF) Type of mass spectrometry detector that measures the time for an ion to fly from the ion source to the detector
Western blot Detection method in which an antibody is used to identify a specific protein

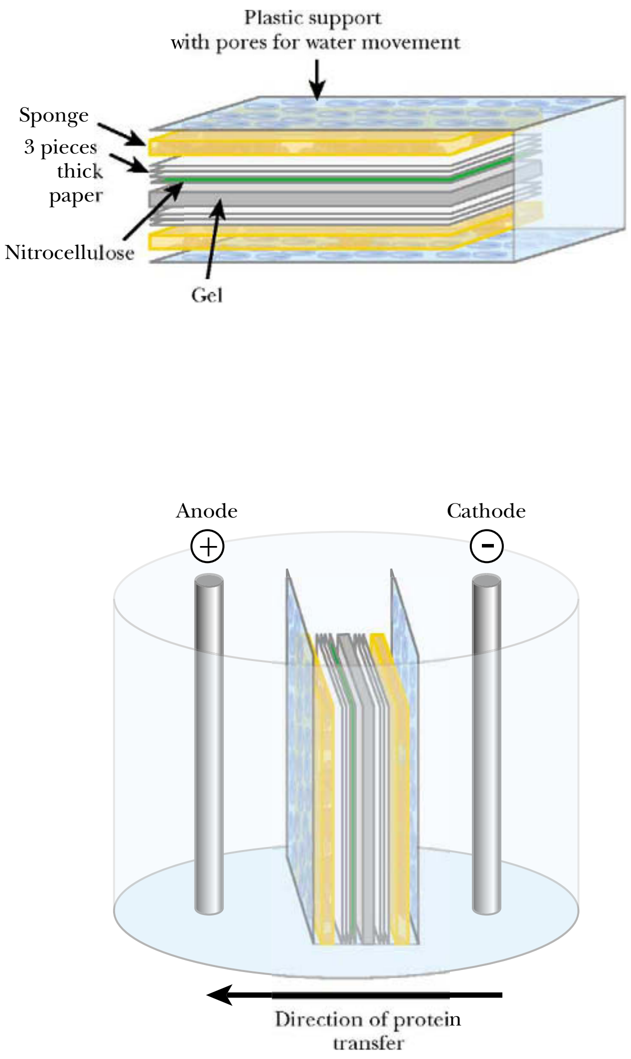


FIGURE 26.05
Electrophoretic Transfer of Proteins from Gel onto Nitrocellulose

A "sandwich" is assembled to keep the gel in close contact with a nitrocellulose membrane while in a large tank of buffer. The sandwich consists of the gel (gray) and nitrocellulose (green) between layers of thick paper and a sponge (yellow). The entire stack is squeezed between two solid supports so that none of the layers can move. The "sandwich" is transferred to a large tank filled with buffer to conduct the current. As in SDS-PAGE, the proteins are repelled by the negatively charged cathode and attracted to the positively charged anode. As the protein move out of the gel, they travel into the nitrocellulose where they adhere.

detector measures the time for an ion to fly from the ion source to the detector (Fig. 26.07). The time-of-flight is proportional to the square root of m/z. Typically molecular ions up to 100,000 daltons may be measured by MALDI/TOF. However, advances in instrumentation will probably increase this limit substantially in the near future.

The amount of sample needed for MALDI/TOF is now less than a picomole (10^{-12} mole). Mass resolution is 1 in 10,000 or better. Thus proteins separated by 2D-PAGE may be routinely identified by MALDI/TOF often after preliminary digestion to give a characteristic set of peptides. Post-translational modification of proteins may also be detected by shifts in molecular weight. Phosphate or sugar residues yield characteristic ion fragments and analysis after protease digestion may reveal the location of such groups within the protein.

Electrospray ionization (ESI) refers to the generation of gas-phase ions from ions in solution. A narrow capillary tube allows droplets of liquid to emerge into a strong electrostatic field (Fig. 26.08). The solvent evaporates and the droplet breaks up. Repeated evaporation and splitting of droplets eventually releases separate ions (either with single or multiple charges) that are accelerated towards a mass analyzer by an electric field. Mass analyzers such as quadrupole or ion-trap detectors are normally used with ESI mass spectrometers. The typical range for a singly charged ion is up to 5,000 daltons, but multiple charges allow heavier ions to be analyzed.

An advantage of ESI is that it can be directly coupled to liquid separation techniques such as capillary electrophoresis or HPLC (high performance liquid chromatography). Also, a parent ion can be isolated and fragmented into daughter

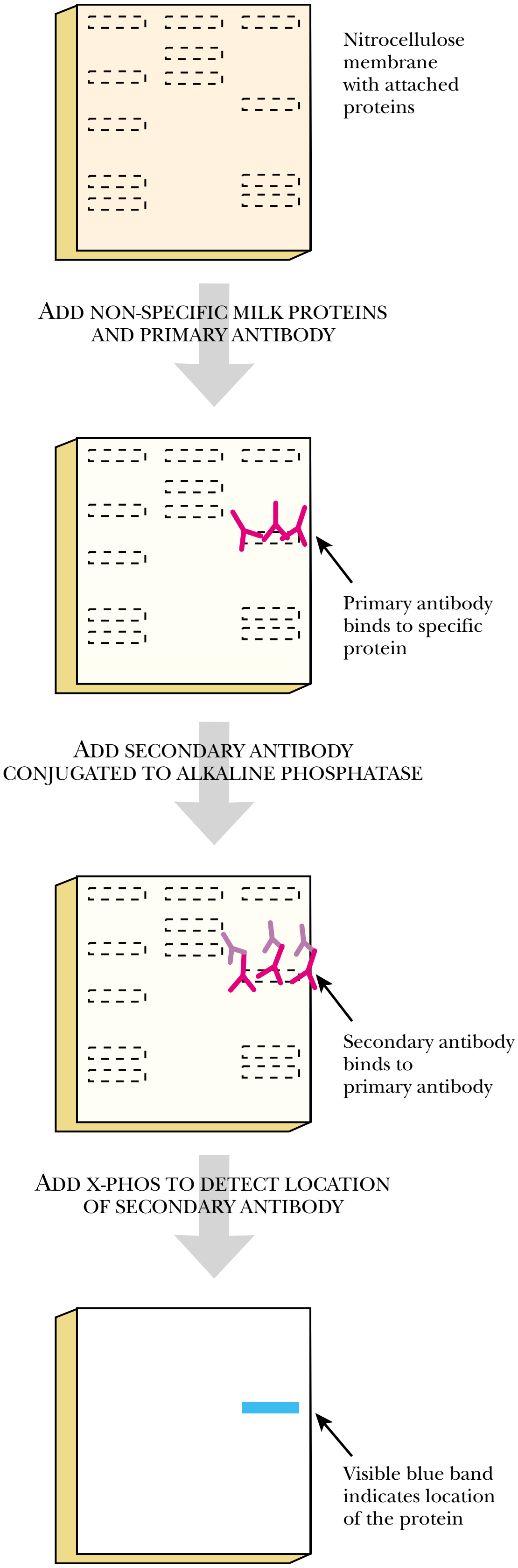

Nitrocellulose membrane with attached proteins
ADD NON-SPECIFIC MILK PROTEINS AND PRIMARY ANTIBODY
Primary antibody binds to specific protein
ADD SECONDARY ANTIBODY CONJUGATED TO ALKALINE PHOSPHATASE
Secondary antibody binds to primary antibody
ADD X-PHOS TO DETECT LOCATION OF SECONDARY ANTIBODY
Visible blue band indicates location of the protein

FIGURE 26.06 ***Western Blot***

After a mixture of proteins have adhered to the nitrocellulose membrane, one specific protein can be detected using an antibody. The antibody is added with a solution of milk proteins and incubated with the nitrocellulose membrane. The milk proteins bind to and block those regions of the nitrocellulose that do not have any protein attached. The primary antibody attaches only to the protein of interest. A secondary antibody with alkaline phosphatase attached binds specifically to the primary antibody, and allows the single protein band to be visualized.

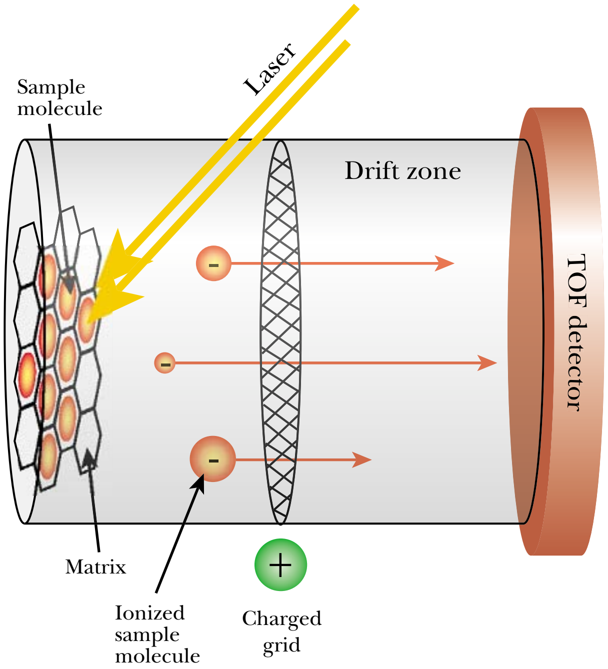


FIGURE 26.07 ***MALDI/TOF Mass Spectrometer***

Mass spectrometry can be used to determine the molecular weight of proteins. The proteins are crystallized in a solid matrix and exposed to a laser, which releases ions from the proteins. These travel along a vacuum tube, passing through a charged grid, which helps separate the ions by size and charge. The time it takes for ions to reach the detector is proportional to the square root of their mass to charge ratio (*m*/*z*). The molecular weight of the protein can be determined from this data.

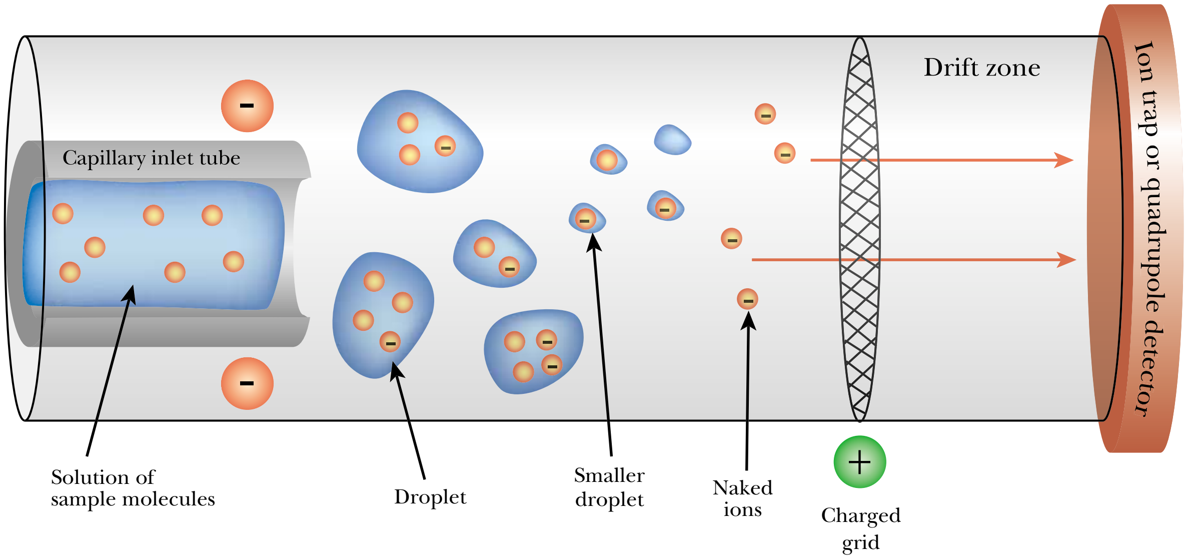


FIGURE 26.08 ***Electrospray Ionization Mass Spectrometer***

ESI mass spectrometry uses a liquid sample of the protein held in a capillary tube. After exposure to a strong electrostatic field, small droplets are released from the end of the capillary tube. A flow of heated gas within the drift zone helps evaporate the solvent and release small charged ions. The charged ions vary in size and charge and the pattern of ions produced is unique to each protein. The ions are further separated by size using a charged grid to either impede or help the flow toward the detector.

ions, so allowing more detailed analysis of molecules. This is known as **tandem mass spectrometry (MS/MS)**. It allows two parent ions with the same mass (e.g. two peptides with the same amino acid composition but different sequence) to be distinguished.

Protein Tagging Systems

A convenient way to identify and purify proteins is to tag them using a genetic approach.

Tagging of proteins with another easily recognized molecule allows many different proteins to be processed and purified by the same procedures. Protein tagging is usually done genetically, that is, the DNA encoding the protein is engineered to add an extra segment that codes for the tag. The gene must therefore first be cloned and carried on a suitable vector. The hybrid gene is then expressed in a suitable host organism, such as *Escherichia coli* or cultured mammalian cells, and the protein is purified by a method that binds to and isolates the tag sequence. Since the protein of interest is synthesized attached to the tag, it is purified along with the tag molecule. In many instances, the tag portion may then be removed. Alternatively, a protein array can by constructed by attaching the tagged protein to a chip via the tag (see below).

The His tag allows proteins to be purified by binding to a resin containing nickel ions.

The first protein tag to be widely used was the **polyhistidine tag (His tag)**, consisting of six tandem histidine residues. This may be added to the target protein at either the amino or carboxy terminus. The **His tag** binds very tightly to nickel ions; therefore, His tagged proteins are purified on a column to which Ni^{2+} ions are attached by a metal chelator (metal-chelate column chromatography) (Fig. 26.09).

Other short tags in wide use are the **FLAG®** and "Strep" tags. The FLAG tag is a short peptide (AspTyrLysAspAspAspAspLys) that is bound by a specific antibody. Anti-FLAG® antibody is available commercially (Immunex Corporation) and may be attached to a suitable resin for use in column purification. The "Strep" tag is a 10 amino acid peptide that mimics the 3-D structure of biotin. It is bound by the biotin-binding proteins **avidin** or **streptavidin** and was originally selected from a random oligopeptide library for its ability to bind to streptavidin.

Full-Length Proteins Used as Fusion Tags

Longer tags, consisting of entire proteins, are sometimes used to tag a specific protein. The tag protein coding sequence is normally placed 5′ to the gene of interest in order to ensure good translation initiation. Three of the most popular are **protein A** from *Staphylococcus*, **glutathione-S-transferase (GST)** from *Schistosoma japonicum*, and **maltose-binding protein (MBP)** from *E. coli*. These three proteins generally give fusion protein products that are stable, soluble and behave well during purification.

Proteins may be fused to a second protein chosen for its convenient properties.

The fusion proteins are purified on columns containing material that specifically bind the tag protein. A specific commercially available antibody binds protein A. After purification on an antibody column, the fusion protein can be eluted at pH 3. Glutathione-S-transferase binds to glutathione, a tripeptide. GST fusions are bound by glutathione-agarose columns and eluted with free glutathione. Maltose-binding protein

avidin Protein from egg-white that binds biotin very tightly
FLAG® tag A short peptide tag (AspTyrLysAspAspAspAspLys) that is bound by a specific anti-FLAG® antibody that may be attached to a resin for use in column purification of proteins
glutathione-S-transferase (GST) Enzyme that binds to the tripeptide, glutathione. GST is often used in making fusion proteins
His tag Six tandem histidine residues that are fused to proteins so allowing purification by binding to nickel ions that are attached to a solid support. Also known as polyhistidine tag
maltose-binding protein (MBP) Protein of *E. coli* that binds maltose during transport. MBP is often used in making fusion proteins
polyhistidine tag (His tag) Six tandem histidine residues that are fused to proteins so allowing purification by binding to nickel ions that are attached to a solid support
protein A Antibody binding protein from *Staphylococcus* that is often used in making fusion proteins
streptavidin Protein from *Streptococcus* that binds biotin very tightly
tandem mass spectrometry (MS/MS) Two successive rounds of mass spectrometry in which a parent ion is first isolated and then fragmented into daughter ions for more detailed analysis

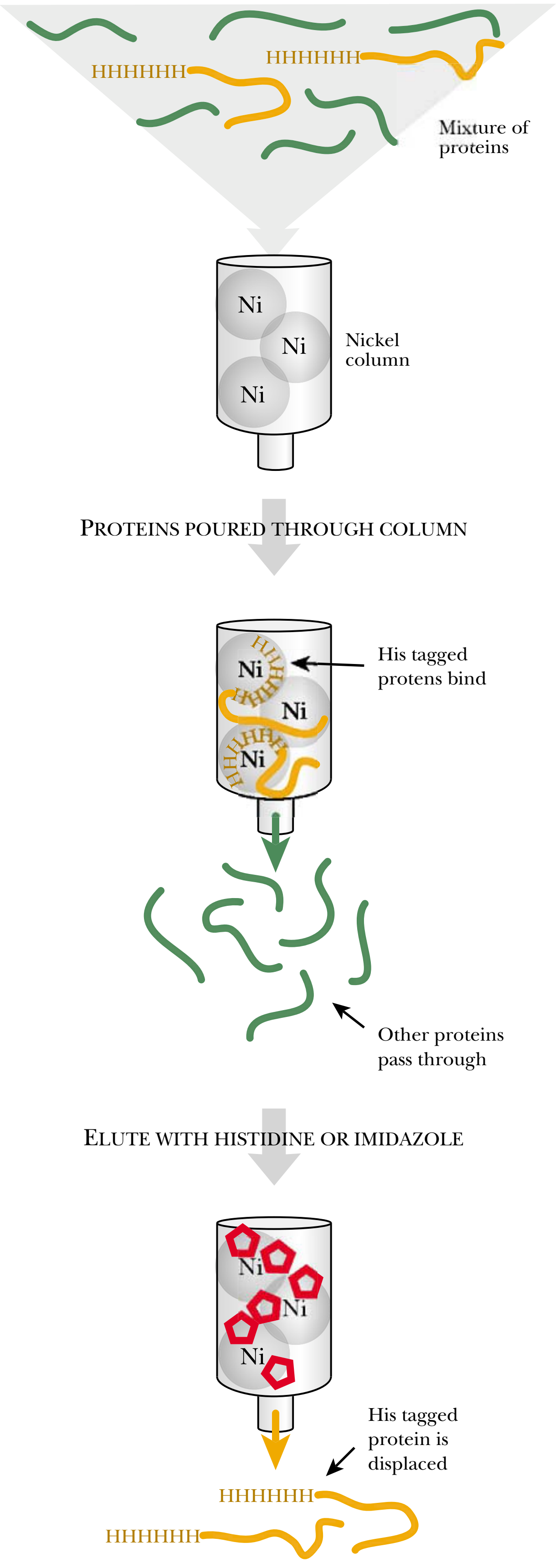

HHHHHH
HHHHHH
Mixture of proteins
Ni
Ni
Ni
Nickel column
PROTEINS POURED THROUGH COLUMN
His tagged protens bind
Other proteins pass through
ELUTE WITH HISTIDINE OR IMIDAZOLE
His tagged protein is displaced
HHHHHH
HHHHHH

FIGURE 26.09 ***Nickel Purification of His6 Tagged Protein***

To isolate a pure sample of one specific protein, the gene for the protein is linked to a tag sequence. In this example, the tag sequence encodes six histidines in a row. This engineered gene is then expressed in either bacteria or mammalian cells. The cells are harvested and lysed to release all the proteins. To isolate the tagged protein, the mixture of proteins is added to a nickel column. The nickel-coated beads bind the histidine tag, allowing all the other proteins to pass through the column. Next a solution containing histidine or imidazole is added to the column. The histidine or imidazole binds to the nickel-coated beads thus releasing the histidine tagged protein.

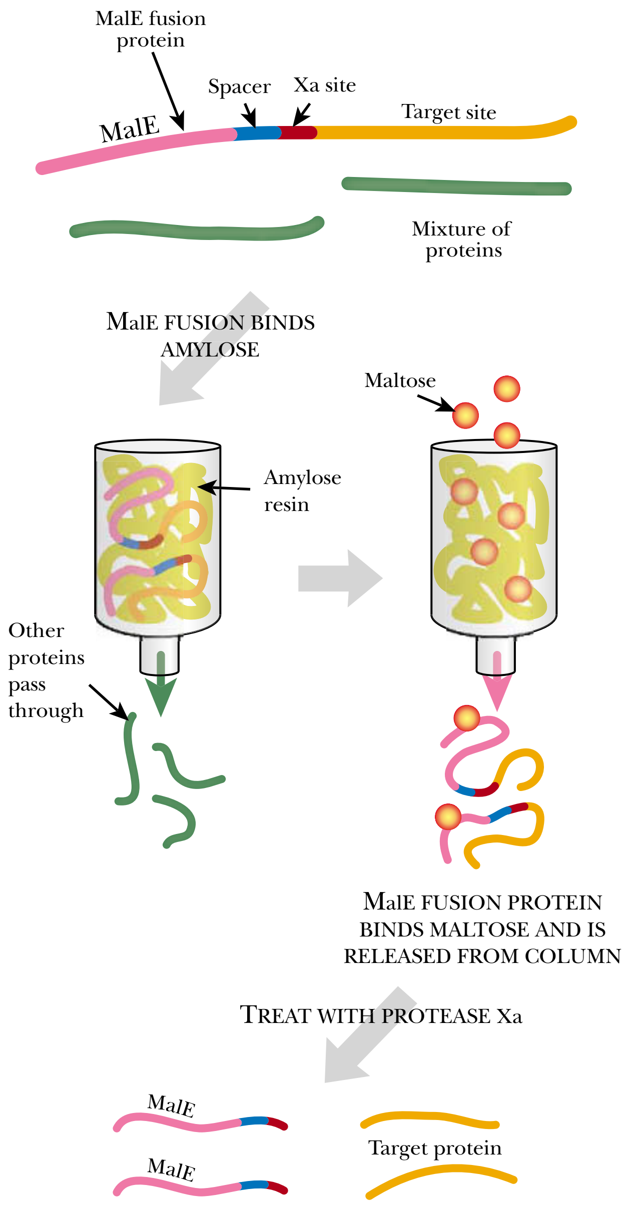


FIGURE 26.10 ***Protein Isolation using Fusions to Maltose-Binding Protein***

In order to isolate one specific protein from a mixture, the gene for maltose-binding protein can be genetically fused to the protein of interest. The fusion gene can then be expressed in an organism such as *E. coli*. The bacteria are lysed, and the mixture of proteins (top of figure) is isolated. The protein of interest has three new regions attached to it: the entire maltose binding protein, a cleavage site for Factor Xa, and a small spacer region. The protein mixture is added to an amylose column, where only the fusion protein binds via the maltose binding domain. Adding free maltose elutes the fusion protein. The eluted protein is treated with factor Xa, which cleaves the maltose binding protein section from the protein of interest.

binds maltose and the maltose polymer, amylose. An amylose column purifies proteins with an MBP tag. The bound fusion protein is then eluted with maltose, which preferentially binds to the amylose and releases the tagged protein (Fig. 26.10).

The pMAL vectors (New England Biolabs Inc., MA) are one example of a protein tagging system (Fig. 26.11). These vectors carry the *malE* gene of *E. coli*, encoding MBP. A spacer sequence encoding 10 Asn residues lies between *malE* gene and the polylinker. This insulates the maltose binding protein from the protein of interest. Adjoining this is the recognition site [Ile Glu (or Asp) Gly Arg] for a highly specific protease, factor Xa of the blood clotting system. After purification, the fusion protein is treated with a factor Xa and the two proteins are separated. Factor Xa itself must then be removed by binding it to benzamidine-agarose. The protein fusion is transcribed from the strong *tac* promoter and is translated using the initiation signals of *malE*. Variants exist with or without the *malE* signal sequence thus allowing for either secretion or cytoplasmic expression.

Sometimes, fusion proteins (or for that matter unmodified proteins) are degraded by host-cell proteases during purification. Protease inhibitors may be included, or in the case of a bacterial host such as *Escherichia coli*, protease-deficient mutants may be

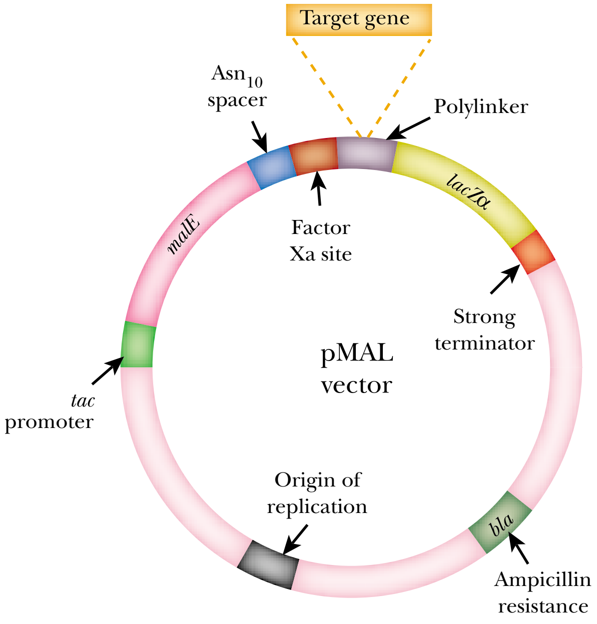


FIGURE 26.11 ***Maltose-Binding Protein Fusion Vector***

In order to manufacture an MBP-tagged protein, the target gene is cloned into the pMAL vector. The vector has a polylinker downstream from the *malE* gene for inserting the gene of interest. For correct translation, the target gene must be cloned in frame with the *malE* gene. The entire region between the strong *tac* promoter and the strong terminator produces a fusion protein consisting of MBP (MalE protein) linked to the target protein via a protease cleavage site. The *lacZα* sequence allows for blue/white color screening to detect DNA insertion (see Ch. 22).

used. Bacteria defective in several protease genes are useful, even though not all *E. coli* protease genes can be inactivated since some are essential for survival.

Self Cleavable Intein Tags

Tags based on inteins are convenient because they can be designed to cleave themselves off when no longer required.

A recent improvement in tagging systems eliminates the protease cleavage and purification step. Instead of a protease recognition site, the fusion protein cleaves itself after purification of the target protein. This approach is based on the properties of **inteins**, self-splicing intervening sequences that are found in proteins (see Ch. 12). The advantages are a reduction in the number of steps required for purification and the avoidance of expensive proteases that may sometimes cleave the protein of interest at other sites.

The Intein Mediated Purification with Affinity Chitin-binding Tag (IMPACT^{TM}) system from New England Biolabs depends on the self-splicing of an intein originally from the *VMA1* gene of *Saccharomyces cerevisiae*. This intein has been modified to undergo self-cleavage only at its N-terminus. This is triggered at low temperatures by thiol reagents such as dithiothreitol (DTT). At the C-terminal end of the intein is a small chitin binding domain (CBD) from the carboxy terminus of the chitinase A1 gene of *Bacillus circulans*. [Chitin is a structural polymer of N-acetyl glucosamine. It is found in most fungi and is the major structural component of the exoskeletons of insects.]

The gene encoding the target protein is inserted into a multiple cloning site (MCS) upstream of the gene encoding the intein plus chitin-binding domain (CBD). The fusion protein is purified by binding to a chitin column. While the fusion protein is still attached to the column, intein self-cleavage is induced by incubation at 4°C with DTT. The target protein is released while the intein plus the chitin binding domain remain bound to the column (Fig. 26.12).

Selection by Phage Display

Phage display allows us to find a gene by identifying the protein it encodes.

Most procedures in molecular biology deal with either genetic information (i.e., DNA or RNA) or gene products (i.e., proteins). Display protocols are designed to provide both the gene and its encoded protein together. The most common of these is

intein Self-splicing intervening sequence that is found in a protein

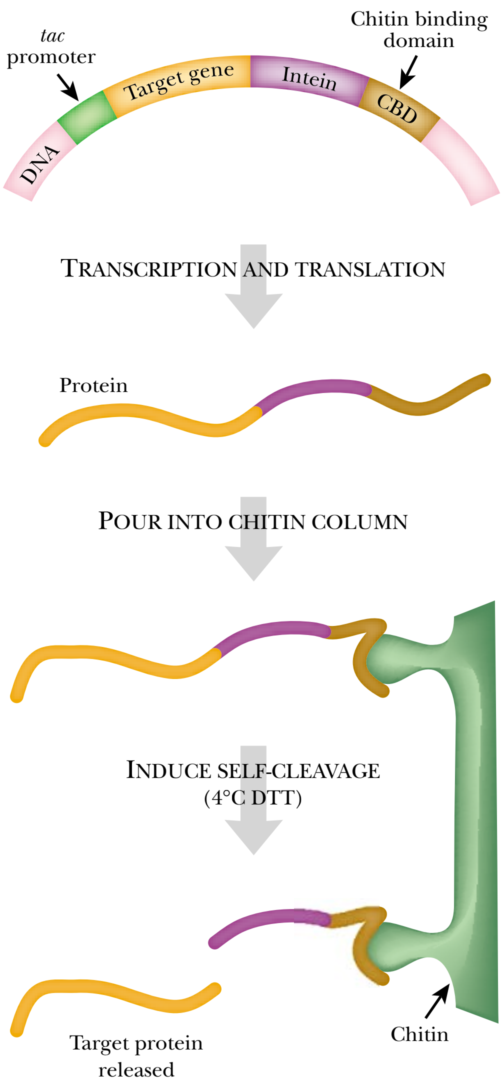


FIGURE 26.12 ***Intein Mediated Purification System***

Inteins that can self-cleave at the amino-terminus allow specific proteins to be purified and cleaved from a fusion protein in one step. First, the target gene is cloned upstream of the intein sequence and a chitin binding domain. The fusion gene is transcribed and translated in bacteria. The bacteria are lysed and release a mixture of proteins that are passed through a chitin column. The proteins with the chitin binding domain bind to the chitin and the remaining proteins pass through. Adding DTT and incubating at 4°C activates the intein to cleave itself from the target protein, which is therefore released from the column.

the **phage display** technique. Here a full-length protein or a shorter peptide is fused to a coat protein of a bacteriophage so as to be displayed on the outer surface of the virus particle. Meanwhile, the DNA encoding the fusion protein is carried inside the bacteriophage (Fig. 26.13).

Filamentous phage M13 is most popular choice for phage display but Lambda, T7 and T4 are also used. M13 is preferred since it is non-lytic and does not destroy the host bacteria during phage production. Instead, phage particles are secreted through the bacterial cell envelope. The absence of cell debris simplifies purification of the phage. Three M13 structural proteins have been used as expression platforms for the proteins of interest with gene III protein most popular. There are about 2,500 copies of the major coat protein (gene VIII protein) on the phage surface but only five copies of the minor coat protein (gene III protein). Having fewer copies of the displayed peptide on the surface of the phage avoids artifacts due to simultaneous binding of multiple polypeptides. Since the N-terminus of the M13 coat proteins is external and the C-terminal region interacts with the DNA inside the phage particle, the displayed peptide must be attached to the N-terminus of gene III.

The DNA encoding the protein or peptide is fused to the gene for the phage coat protein by a PCR-based technique. The linear PCR product is amplified and circularized to give a viral genome that is transformed into *E. coli* cells. The phages that are

The proteins to be screened are fused to virus proteins so that they appear on the outside surface of the virus particle.

phage display Fusion of a protein or peptide to the coat protein of a bacteriophage whose genome also carries the cloned gene encoding the protein. The protein is displayed on the outside of the virus particle and the corresponding gene is carried on the inside

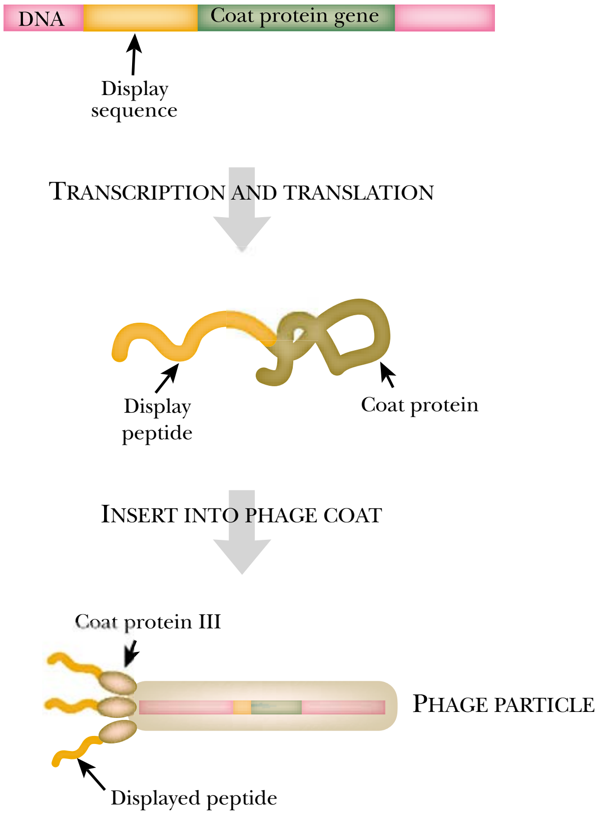


FIGURE 26.13 ***Principle of Phage Display***

In order to display a peptide on the surface of a bacteriophage, the DNA sequence encoding the peptide must be fused to the gene for a bacteriophage coat protein. In this example, the chosen coat protein is encoded by gene III of phage M13. Here the N-terminal portion will be on the outside of the phage particle whereas the C-terminus will be on the inside. Therefore the peptide must be fused in frame at the N-terminus to be displayed on the outside of the phage.

produced are slightly longer than the original wild type M13; nonetheless, they are packaged by extending the filamentous protein coat. The fusion proteins are expressed and the intruding peptides displayed on the surface of the M13 particles.

Many proteins have regions that bind to other proteins. These regions can be large or small, but usually a few amino acids are critical for the two proteins to interact properly. In order to identify the peptide sequence these binding sites recognize, **phage display libraries** are constructed. These consist of a large number of modified phages displaying a library of different peptide sequences. The first such libraries consisted of large collections of short random peptides. They are screened to find peptides that bind to specific molecules (the "target"), such as a particular antibody, enzyme or cell-surface receptor. The peptide of interest is found by a selection procedure referred to as **biopanning**. The phage display library is incubated with target molecules that are attached to a solid support (beads or membranes, etc.). Unbound phage is washed away. Bound phage is eluted and amplified by re-infecting cells of *E. coli*. Several cycles of binding and amplification will enrich for the phage that carries the peptide that binds most tightly to the target. Finally, individual clones are characterized by DNA sequencing (Fig. 26.14).

A mixture of viruses with different proteins displayed is screened for binding to an antibody or other specific binding protein.

Several commercially available peptide libraries are marketed by New England Biolabs under the trade name Ph.D. (for *Ph*age *D*isplay!). The Ph.D.-7 library consists of 2.0×10^9 random heptapeptide clones. It probably contains most of the theoretically possible amino acid heptamers (of which there are 20^7 = approximately 1.3×10^9). In contrast, the Ph.D.-12 library, also with 2.0×10^9 independent clones, only represents a small fraction of the $20^{12} = 4.1 \times 10^{15}$ 12-mers.

Full-length proteins can also be fused to phage coat proteins to produce a full-length phage display library. In principle, a gene library from any organism could be converted into a phage display library by insertion into a suitable phage coat protein gene. M13 is not practical for this type of library since the insert must be positioned between the signal sequence, which is needed for secretion and phage assembly, and

biopanning Method of screening a phage display library for a desired displayed protein by binding to a bait molecule attached to a solid support
phage display library Collection of a large number of modified phages displaying different peptide or protein sequences

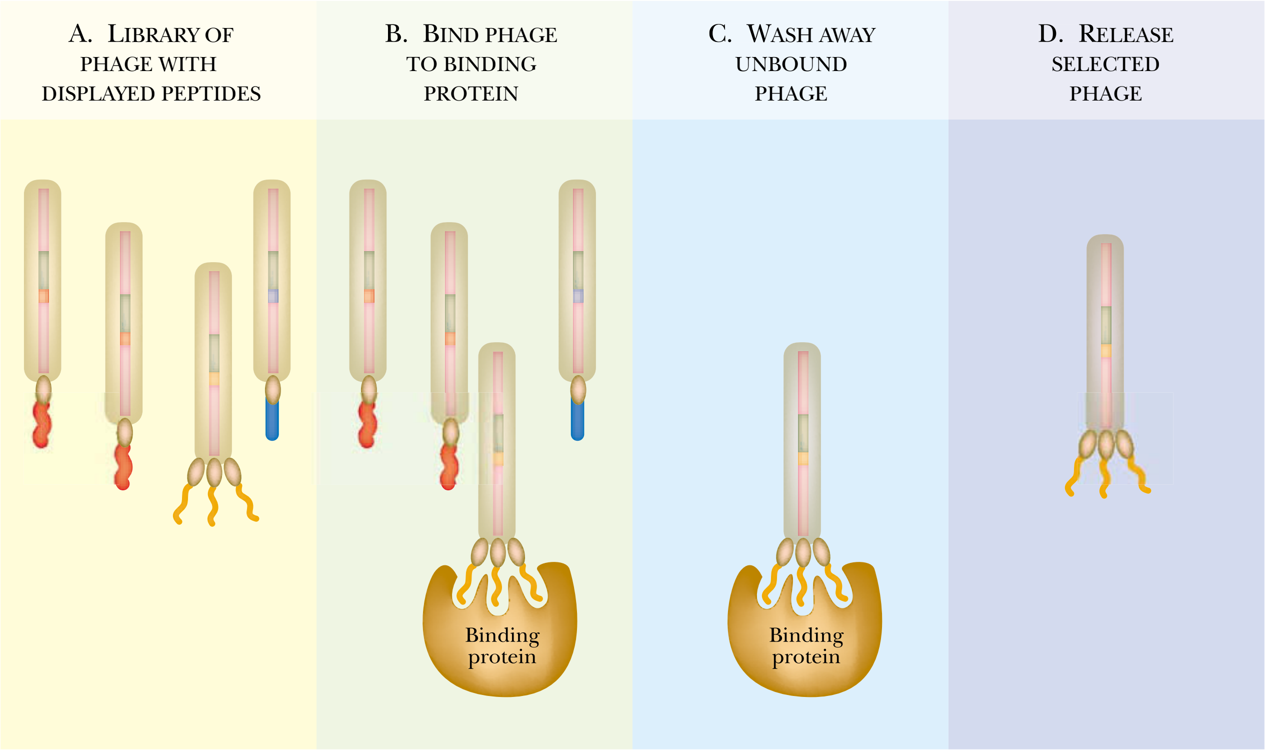

A. LIBRARY OF PHAGE WITH DISPLAYED PEPTIDES
B. BIND PHAGE TO BINDING PROTEIN
C. WASH AWAY UNBOUND PHAGE
D. RELEASE SELECTED PHAGE
Binding protein
Binding protein

FIGURE 26.14 ***Biopanning to Screen a Phage Display Library***

Biopanning is used to isolate peptides that bind to a specific target protein, which is usually attached to a solid support such as a membrane or column. The phage display library (A) is added to the binding protein (B). Those phage which display peptides that bind to the target protein will be retained (C) but the others are washed away. The phage that does recognize the binding protein can then be released, isolated and purified.

the N-terminus of the coat protein. Therefore, both ends of the insert must thus be in frame, and in addition, no stop codons must be present in the insert. In contrast, in T7 the C-terminal region of the coat protein is exposed on the outside. Using C-terminal coat protein fusions avoids the above problems and allows complete coding sequences to be inserted. Using the T7Select^{TM} system from Novagen, several cDNA libraries have now been biopanned to find proteins that bind to some chosen target molecule. For example, phage that display RNA-binding proteins have been isolated using RNA anchored to a solid support as bait (Fig. 26.15).

Other display systems use whole bacterial cells to carry the protein of interest. DNA sequences encoding polypeptides to be screened can be fused to the flagellin or pilin genes of *E coli*. The polypeptide library is then exposed on the cell surface attached to either the flagella or the pili. The phage display libraries are generally more convenient but one advantage of using bacteria is that the fluorescence-activated cell sorter (FACS) can sort the cells provided that the peptide target is labeled with a fluorescent dye.

Protein Interactions: The Yeast Two-Hybrid System

Many proteins recognize and bind to other proteins. The total of all protein-protein interactions is sometimes referred to as the **protein interactome** by those enthusiastic about “omics” terminology. Mass screening of such interactions has proven possible by means of the **“two-hybrid”** system. Interactome analysis is based on the idea of guilt

protein interactome The total of all the protein-protein interactions in a particular cell or organism
two-hybrid system Method of screening for protein-protein interactions that uses fusions of the proteins being investigated to the two separate domains of a transcriptional activator protein

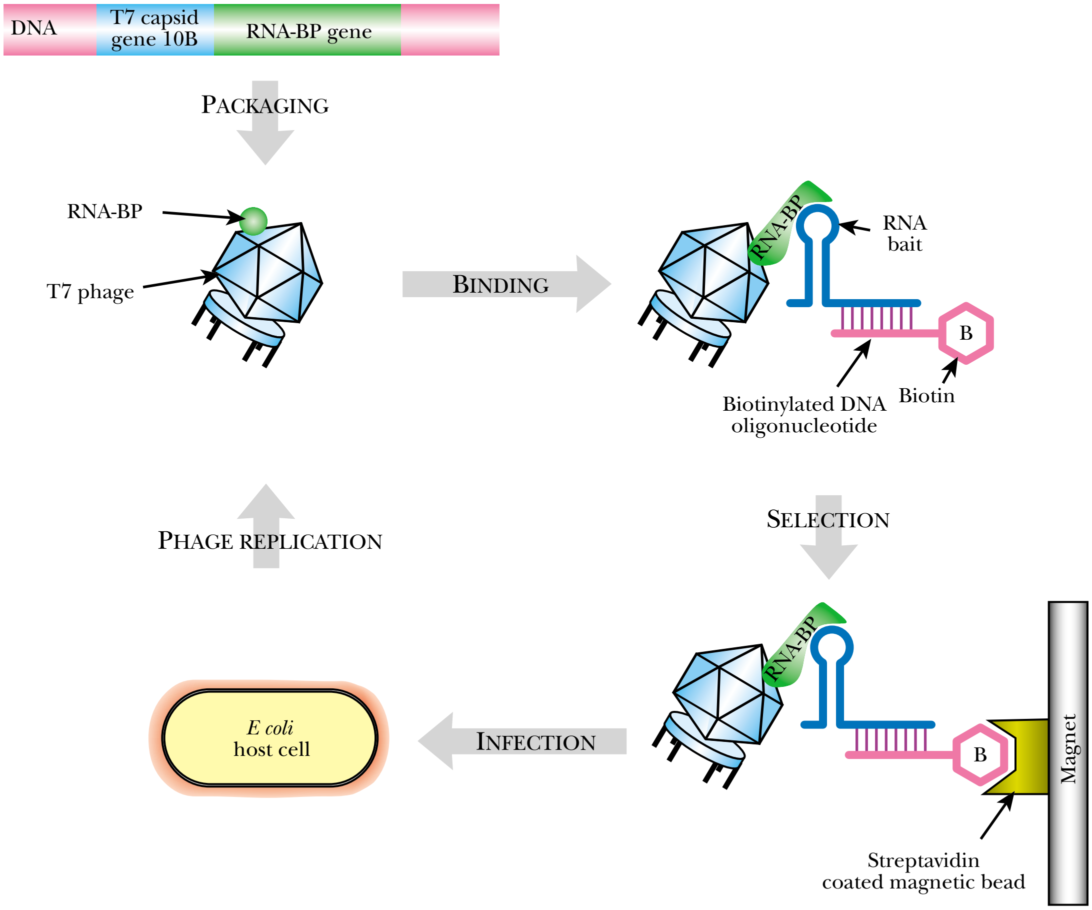


FIGURE 26.15 ***Biopanning for RNA-Binding Proteins***

To identify RNA-binding proteins (RNA-BP), a bait RNA (blue) is used that is linked to a biotinylated oligonucleotide. The bait RNA is incubated with a full-length phage display library. In this example, T7 gene 10B DNA is fused to a gene library that includes RNA binding protein genes (shown here in green). Those phage that express the full length RNA binding protein on the outside will bind to the RNA bait. This in turn is bound via the biotinylated oligonucleotide to magnetic beads coated with streptavidin. The captured phage is eluted from the magnetic beads by free biotin and is used to infect *E. coli*. Isolating the T7 DNA and sequencing the insert identifies the gene for the RNA binding protein.

by association. It is assumed that the binding of a novel protein to one that is well characterized may provide some hint as to function.

Two-hybrid analysis depends on the modular structure of transcriptional activator proteins. Many of these proteins consist of two domains, a DNA binding domain and an activation domain. The DNA binding domain (DBD) recognizes a specific sequence in the DNA upstream of a promoter and the activation domain (AD) stimulates transcription by binding to RNA polymerase (Fig. 26.17). Provided that the two domains interact, they will activate transcription. It is not usually necessary for the two domains to be covalently joined to form a single protein.

In the two-hybrid system, both the DBD domain and the AD domain are fused to two other proteins (X and Y in Fig. 26.17). These two hybrid proteins are referred to as the "**bait**" (DBD-X) and the "**prey**" (AD-Y). If the bait captures the prey, i.e. if proteins X and Y interact, a complex will form and the gene will be activated. A convenient reporter gene is used to monitor for a successful interaction.

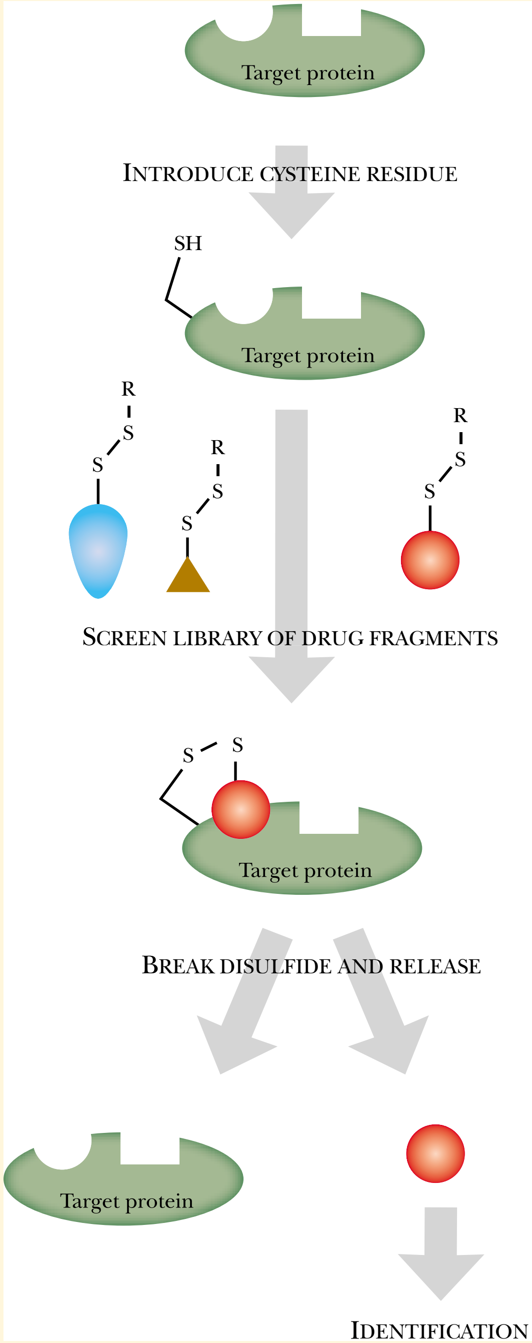
Proteins can be screened to see which other proteins they bind using two hybrid analysis.

The test proteins (bait and prey) are fused separately to the two halves of a transcription factor. If the bait and prey bind each other they will reassemble the transcription factor and activate the genes it controls.

bait The fusion between the DNA binding domain of a transcriptional activator protein and another protein as used in two-hybrid screening
prey The fusion between the activator domain of a transcriptional activator protein and another protein as used in two-hybrid screening

Tethering Technology Allows Small Molecule Screening

The converse problem is to isolate small molecules that bind to a given protein. In particular, a protein may have been chosen as a potential drug or antibiotic target. Consequently small molecules that bind to and inhibit the protein of interest are wanted. A library of small chemical molecules will be synthesized and screened for those that bind to the protein. This may be achieved by a variety of tedious procedures. However, a new approach known as tethering technology, developed by Sunesis corporation, allows greatly improved screening. This approach involves modifying the small molecule library by adding a chemical group that will act as a tether. The target protein is immobilized and also modified with a tethering group. The two tethering groups are chosen so that they will form a cross link if they come into close contact. In the diagram shown (Fig. 26.16), a disulfide linkage is formed between the two tethering groups. The small molecules carry a short side chain that ends in a masked sulfhydryl group and the protein has a cysteine residue engineered into it whose sulfhydryl group is exposed on the surface close to the binding pocket. When a small molecule fits the binding site on the protein, a cross link can form between the tethering groups and the small molecule is trapped. The small molecule is later released for identification.



Target protein
INTRODUCE CYSTEINE RESIDUE
SH
Target protein
R
S
S
R
S
S
R
S
S
SCREEN LIBRARY OF DRUG FRAGMENTS
S
S
Target protein
BREAK DISULFIDE AND RELEASE
Target protein
IDENTIFICATION

FIGURE 26.16 ***Tethering Technology***

A library of small molecules is modified by addition of a short side chain carrying a blocked sulfhydryl group. The target protein has a cysteine situated so as to provide another sulfhydryl group close to the binding site. The protein is immobilized and treated with the small molecule library. After binding of a small molecule to the protein, the tethering groups react to form a disulfide linkage. The other small molecules are rinsed away. The trapped small molecules are then released for identification.

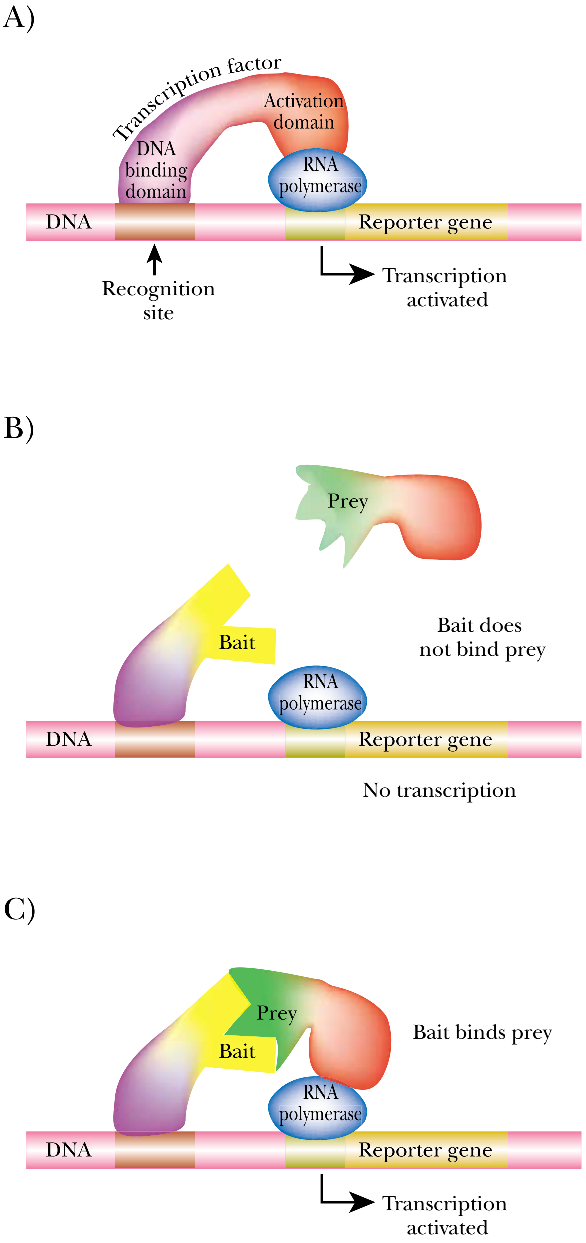


FIGURE 26.17 ***Principle of Two-Hybrid Analysis***

(A) Transcription of a yeast gene involves activation of RNA polymerase by a transcription factor with two different domains. The DBD (purple) recognizes upstream regulatory sites, and the AD (red) activates RNA polymerase to start transcription of the reporter gene. For two-hybrid analysis, two proteins (Bait and Prey) are fused separately to the DBD and AD domains of the transcription factor. The Bait protein is joined to the DBD, and the Prey protein to the AD. (B) Here, the Bait protein and Prey protein do not interact and the reporter gene is not turned on. (C) Here, the Bait binds the Prey, thus bringing the transcription factor halves together. The complex activates the RNA polymerase and the reporter gene is expressed.

Two-hybrid analysis was developed in yeast and is being used to generate a complete list of interactions between all 6000 or so yeast proteins. It is thus necessary to examine 6,000 × 6,000 combinations. To examine these potential interactions, each open reading frame in the yeast genome was amplified by PCR and cloned into two separate vectors, one carrying the DBD domain and one with the AD domain. Thus each yeast protein is tested as both bait and prey. The vectors are designed to give in-frame gene fusions of each ORF with the DBD domain and AD domains of a suitable transcriptional activator, such as GAL4 (Fig. 26.18). One vector has a multiple cloning site downstream of the GAL4-DBD and thus gives a 3′-fusion of GAL4-DBD to protein X (GAL4-DBD-X). The other vector has its MCS upstream of the GAL4-AD and gives a 5′-fusion of GAL4-AD and protein Y (Y-GAL4-AD).

The bait and prey fusion plasmids are transformed into yeast cells of different mating types. This results in two sets of approximately 6000 transformants. All possible matings are carried out between the two sets using a laboratory robot to manipulate the colonies. When the two yeast mate, the diploid cell will have a bait plasmid

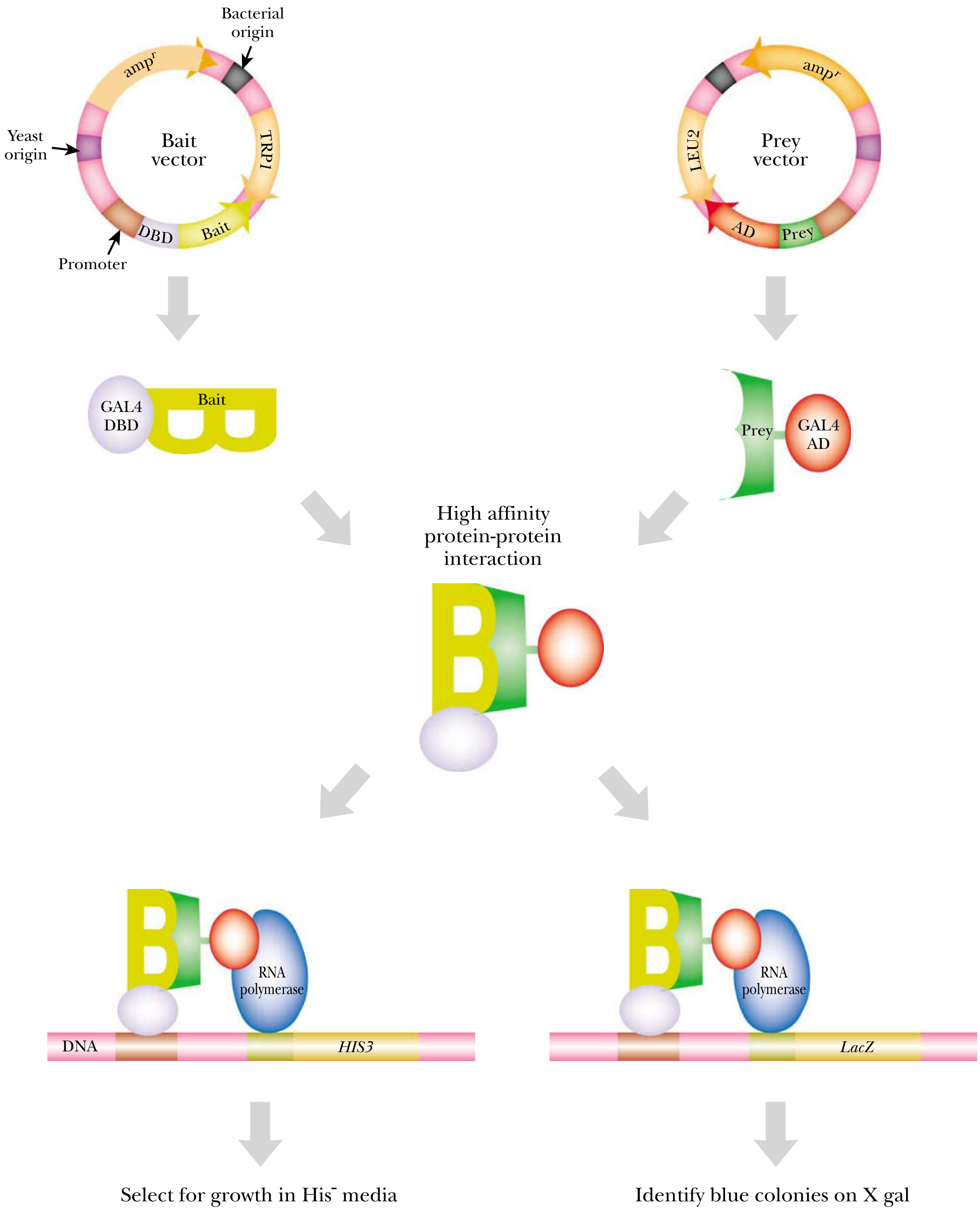


FIGURE 26.18 ***Vectors for Two-Hybrid Analysis***

Two different vectors are necessary for two-hybrid analysis. The bait vector has the coding regions for the DBD and for the Bait protein. The prey vector has the coding regions for the AD and for the Prey protein. These two different constructs are expressed in the same yeast cell. If the Bait and Prey interact, the reporter gene is expressed. Two possible reporter systems are shown here. If the yeast *His3* gene is used, yeast expressing the reporter gene will be able to make histidine and hence to grow in media without histidine provided. If the *lacZ* gene from *E. coli* is used, the yeast cells will turn blue on plates containing X-gal.

and a prey plasmid. If the two fusion proteins X and Y interact, the reporter gene is switched on. In yeast, the *HIS3* or *URA3* genes are usually used. If the reporter gene is not activated, the yeast strain cannot grow unless provided with histidine or uracil respectively. If the reporter gene is turned on the cells can grow on medium without histidine or uracil. Thus the diploid cells from the 6,000 × 6,000 matings are selected on medium lacking the chosen nutrient (Fig. 26.19). Only those combinations where proteins X and Y interact yield viable colonies.

The original two-hybrid system has several limitations. For example, it relies on proteins interacting within the nucleus. Membrane proteins often misfold when localized in the nucleus. Conversely, other proteins are only correctly modified when present in the cytoplasm. Toxic effects and steric problems with very large proteins may also cause some interactions to be missed. Furthermore, many proteins bind RNA and/or rely on small molecules to alter their conformation so promoting protein-protein interactions.

A variety of modified two-hybrid systems have been developed to deal with these issues. One of the most interesting is the RNA three-hybrid system (Fig. 26.20). In this case, the two proteins (DBD-X and Y-AD) are brought together by an intervening RNA molecule that is bound by both X and Y. This can be used to screen for genes encoding RNA-binding proteins.

Libraries of genes from other organisms can also be used for two-hybrid screening provided they are expressed in yeast. In addition, a two-hybrid screening system (BacterioMatch^{TM} from Stratagene Corporation) has recently been devised for use in the bacterium, *E. coli*. This uses two tandem reporter genes, *bla* and *lacZ*, that encode beta-lactamase (ampicillin resistance) and beta-galactosidase respectively.

Protein Interaction by Co-Immunoprecipitation

In mammalian cells, protein interactions can be identified by **co-immunoprecipitation** (Fig. 26.21). The gene for the protein of interest is transfected into mammalian cells and expressed. Then the protein synthesized is isolated from the cytoplasm using antibodies. If no antibody is available for the protein of interest, it can be tagged with the FLAG peptide (or some other convenient tag). Then the antibody to the FLAG tag is used to isolate the protein. The protein is isolated under conditions in which any other proteins that interact with the protein of interest stay associated. Protein A from *Staphylococcus* binds tightly to antibodies and immobilized protein A is therefore used to isolate the antibodies plus any attached proteins. The protein complexes are then separated by electrophoresis on an SDS-PAGE gel to see how many individual proteins are associated in the complex. The identity of the associated proteins can be determined using such techniques as mass spectroscopy or protein sequencing.

If two proteins are associated in the cell and one is precipitated by an antibody, the other should accompany it.

Co-immunoprecipitation can confirm protein interactions that were established using the two-hybrid system (Fig. 26.22). If mammalian proteins are found to interact using the two-hybrid system, their interaction must be confirmed in mammalian cells. First, the two proteins of interest are genetically linked to two different tags, such as the FLAG or His6 tags. The two constructs are co-transfected into cultured mammalian cells. Antibody to one of the tags is added to the cell-free extract from one of the cells and incubated. The antibody complex is isolated by binding to beads coated with Protein A, and the fraction is run on SDS-PAGE. The gel is transferred to nitrocellulose and the membrane is probed with separate antibodies to each of the FLAG and His6 tags. If the two proteins of interest interact in mammalian cells, then both proteins will be present on the Western blot.

co-immunoprecipitation Method of identifying protein-protein interaction by using antibodies to a one of the proteins

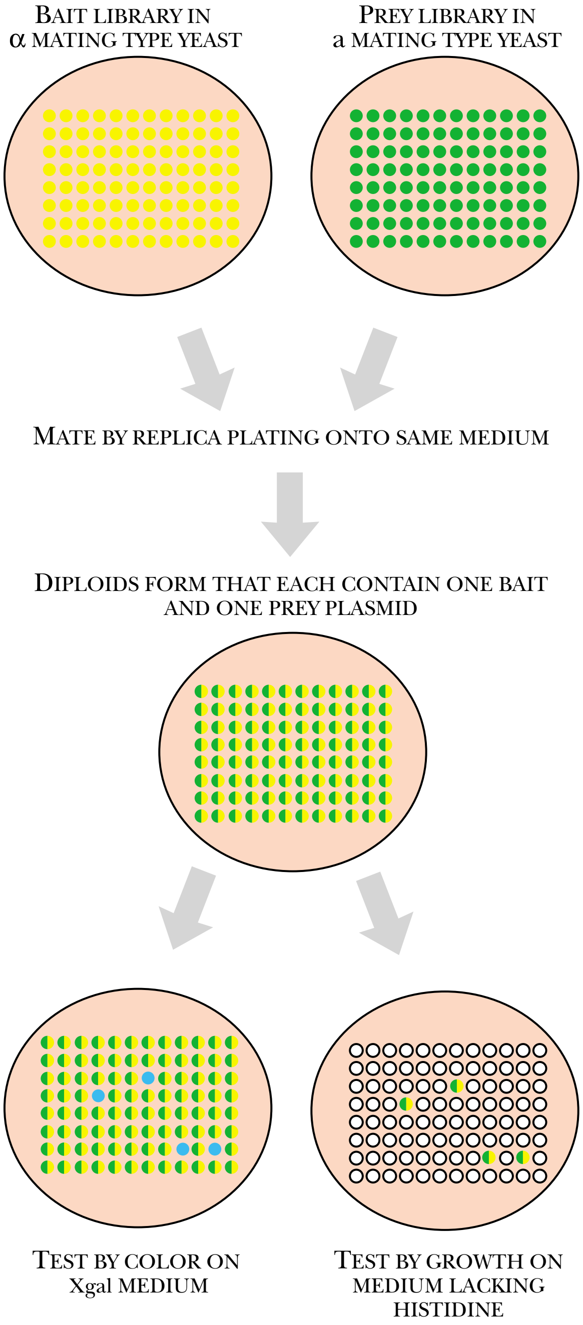

BAIT LIBRARY IN
α MATING TYPE YEAST
PREY LIBRARY IN
a MATING TYPE YEAST
MATE BY REPLICA PLATING ONTO SAME MEDIUM
DIPLOIDS FORM THAT EACH CONTAIN ONE BAIT AND ONE PREY PLASMID
TEST BY COLOR ON Xgal MEDIUM
TEST BY GROWTH ON MEDIUM LACKING HISTIDINE

FIGURE 26.19 ***Two-Hybrid Analysis: Mass Screening by Mating***

To identify all possible protein interactions using the two-hybrid system, haploid α yeast are transformed with the Bait library, and haploid **a** yeast are transformed with the Prey library. When the two yeast types are mated with each other, the diploid cells will each contain a single bait fusion protein and a single prey fusion protein. If the two proteins interact, they activate the reporter gene, which allows the yeast to grow on media lacking histidine (yeast *His3* gene) or turns the yeast cells blue when grown on X-gal media (*lacZ* gene from *E. coli*). This process can be done for all 6,000 yeast proteins using automated techniques.

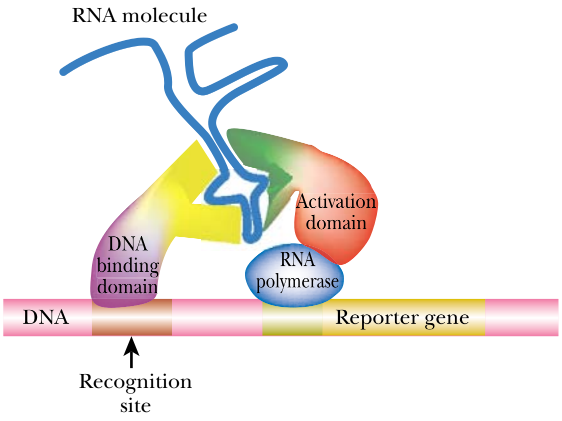


FIGURE 26.20 ***RNA Three-Hybrid System***

The RNA three-hybrid system identifies proteins that interact through an intermediary RNA molecule. Two fusion proteins are used, one (yellow/purple) includes the DBD and the other (green/red) includes the AD of the transcription factor. When these two fusion proteins interact via an RNA molecule, they activate the reporter gene.

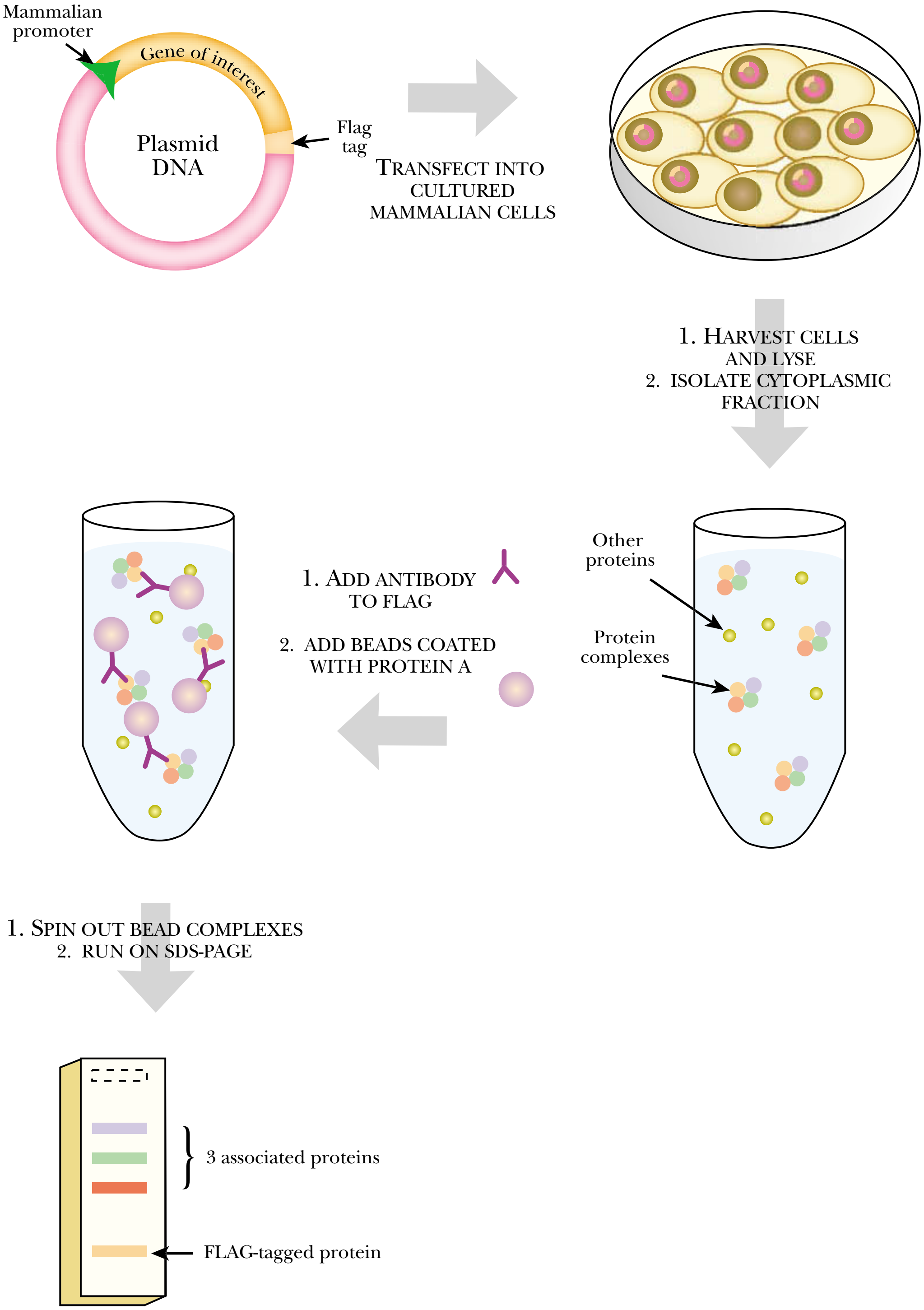


FIGURE 26.21 ***Principle of Co-Immunoprecipitation***

To determine how many other proteins bind to a target protein, they are isolated by precipitating them together using an antibody. An antibody specific to the target protein is needed. If no specific antibody is available a widely used tag (such as FLAG) is added to the coding sequence. The target protein (orange) is cloned behind a mammalian promoter and expressed in mammalian cells where it will bind some other proteins (red, green, purple). The cytoplasmic fraction is isolated. Antibody to the target protein is added and the complexes are isolated by binding the antibody to beads coated with protein A. The beads are spun down. The components are then separated by SDS-PAGE to identify the number and size of the other proteins that bound to the protein of interest.

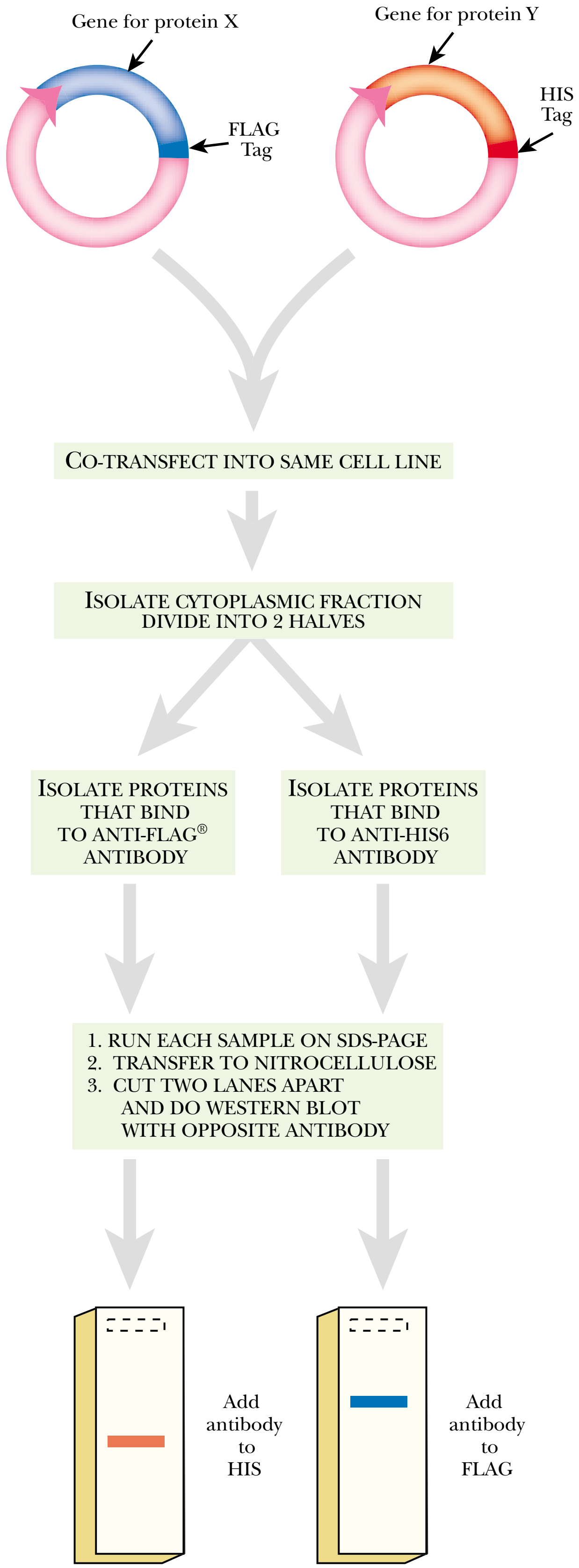

Gene for protein X
Gene for protein Y
FLAG Tag
HIS Tag
CO-TRANSFECT INTO SAME CELL LINE
ISOLATE CYTOPLASMIC FRACTION
DIVIDE INTO 2 HALVES
ISOLATE PROTEINS THAT BIND TO ANTI-FLAG® ANTIBODY
ISOLATE PROTEINS THAT BIND TO ANTI-HIS6 ANTIBODY
1. RUN EACH SAMPLE ON SDS-PAGE
2. TRANSFER TO NITROCELLULOSE
3. CUT TWO LANES APART AND DO WESTERN BLOT WITH OPPOSITE ANTIBODY
Add antibody to HIS
Add antibody to FLAG

FIGURE 26.22
Confirmation of Protein Interaction using Co-Immunoprecipitation

To determine if protein X and Y interact in cultured mammalian cells, they must be fused to two different tag sequences (His6 tag and FLAG tag in this example). The two expression plasmids are transfected into the same mammalian cells. The cytoplasmic protein fraction is isolated and divided into two samples. Antibody to the FLAG tag is added to one sample and antibody to the His6 tag is added to the other. The antibody/protein complexes are isolated using Protein A beads. Each sample is then tested for the presence of the other protein by antibody directed against its tag. That is, the proteins precipitated along with protein X-FLAG are tested for the presence of protein Y. And conversely the proteins precipitated along with protein Y-His6 are tested for protein X.

Protein Arrays

Previous studies of proteins generally examined a single protein at a time. With the recent sequencing of whole genomes, proteome analysis has turned to methods that allow simultaneous monitoring of multiple proteins. Microarrays have been used for DNA for some time, but the variable structures and properties of proteins made such an array approach more difficult. Nonetheless, new technologies have been developed that allow high-throughput analysis of proteins. As a result, **protein microarrays** have recently become available for proteome analysis.

Protein arrays can be constructed if all the proteins carry tags that allow binding to the array support.

Protein microarrays have been used for the biochemical and enzymatic analysis of proteins as well as to survey protein–protein interactions. So far most proteome arrays have used the yeast, *Saccharomyces cerevisiae*, as model organism. A complete proteome analysis needs an array of approximately 6,000 proteins in this case. Such arrays are assembled using proteins that have been tagged with groups allowing binding to solid supports such as 96-well microtiter dishes or glass microscope slides.

Libraries that include nearly 90% of the yeast proteins have been fused to the glutathione-S-transferase (GST) tag, which allows binding to a solid support via glutathione or to the His tag, which allows binding via nickel (Fig. 26.23). These constructs have been expressed under control of the *GAL1* (galactose inducible) or *CUP1* (copper inducible) promoters. Such protein libraries may be pooled or distributed individually into the wells of microtiter dishes. Simpler and less expensive screening is usually done individually, whereas complex or expensive assays are more often run first on pooled protein samples that are subdivided for further analysis if positive results are found.

The functional assays must be designed so that the arrays can be screened conveniently, usually for fluorescence, less often for radioactivity. For example, the yeast proteome has been screened for those proteins that bind **calmodulin** (a small calcium binding protein) or phospholipids, in the laboratory of Michael Snyder at Yale University. The His-tagged proteins were attached to nickel-coated glass slides. Both calmodulin and phospholipid were tagged with biotin. After binding of calmodulin or phospholipids to the proteome array, the biotin was detected by streptavidin carrying a Cy3 fluorescent label (Fig. 26.24). This revealed 39 calmodulin-binding proteins of which six were previously known. Some 150 phospholipid-binding proteins were also found.

Metabolomics

By analogy with genome and proteome, the **metabolome** is the totality of small molecules and metabolic intermediates. NMR of extracts from cells labeled with ^{13}C-glucose has allowed simultaneous measurement of multiple metabolic intermediates. An alternative is the separation of metabolites from cells labeled with ^{14}C-glucose by thin layer chromatography. However, these methods are limited in both sensitivity and in the number and chemical types of compounds that can be readily separated.

Global surveys of low molecular weight metabolites can be done with NMR or high resolution mass spectrometry.

Nearly complete metabolome analysis may be achieved by mass spectrometry. This approach is not limited to particular classes of molecule and is extremely sensitive. Carbon-12 is defined as having a mass of exactly 12 daltons. However, the masses of other atoms, such as ^{14}N or ^{16}O, are not exact integers. Consequently mass spectrometry using extremely high mass resolution (EHMR; that is to 1 ppm or less) allows the unambiguous determination of the molecular formula of any metabolite. Isomers have the same molecular formula, but may be distinguished by the different fragmentation patterns of their molecular ions.

calmodulin A small calcium binding protein of animal cells
metabolome The total complement of small molecules and metabolic intermediates of a cell or organism
protein microarray Microarray of immobilized proteins used for proteome analysis and normally screened by fluorescent or radioactive labeling

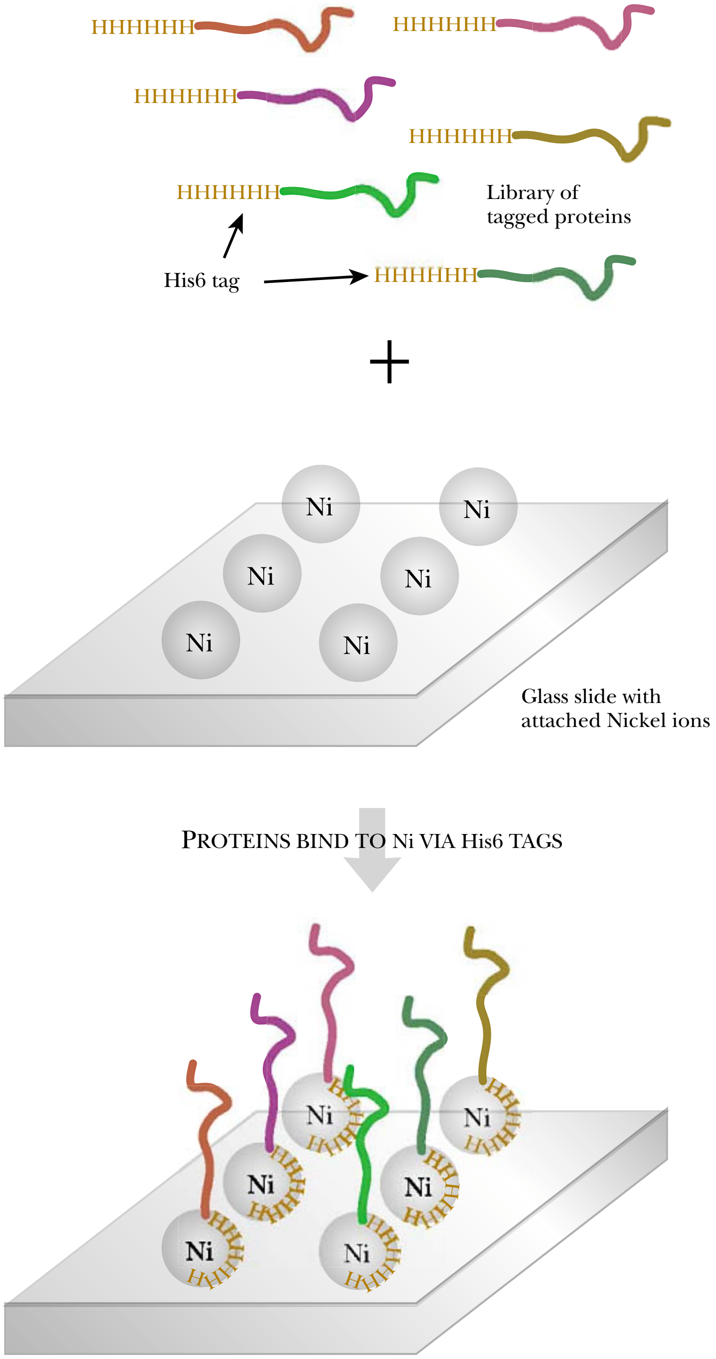

HHHHHH
HHHHHH
HHHHHH
HHHHHH
HHHHHH
Library of tagged proteins
His6 tag
HHHHHH
+
Ni
Ni
Ni
Ni
Ni
Ni
Glass slide with attached Nickel ions
PROTEINS BIND TO Ni VIA His6 TAGS
Ni
Ni
Ni
Ni
Ni
Ni

FIGURE 26.23 ***Protein Microarray—Principle***

To assemble a protein microarray, a library of His-tagged proteins is incubated with a nickel coated glass slide. The proteins adhere to the slide wherever a nickel ion is attached.

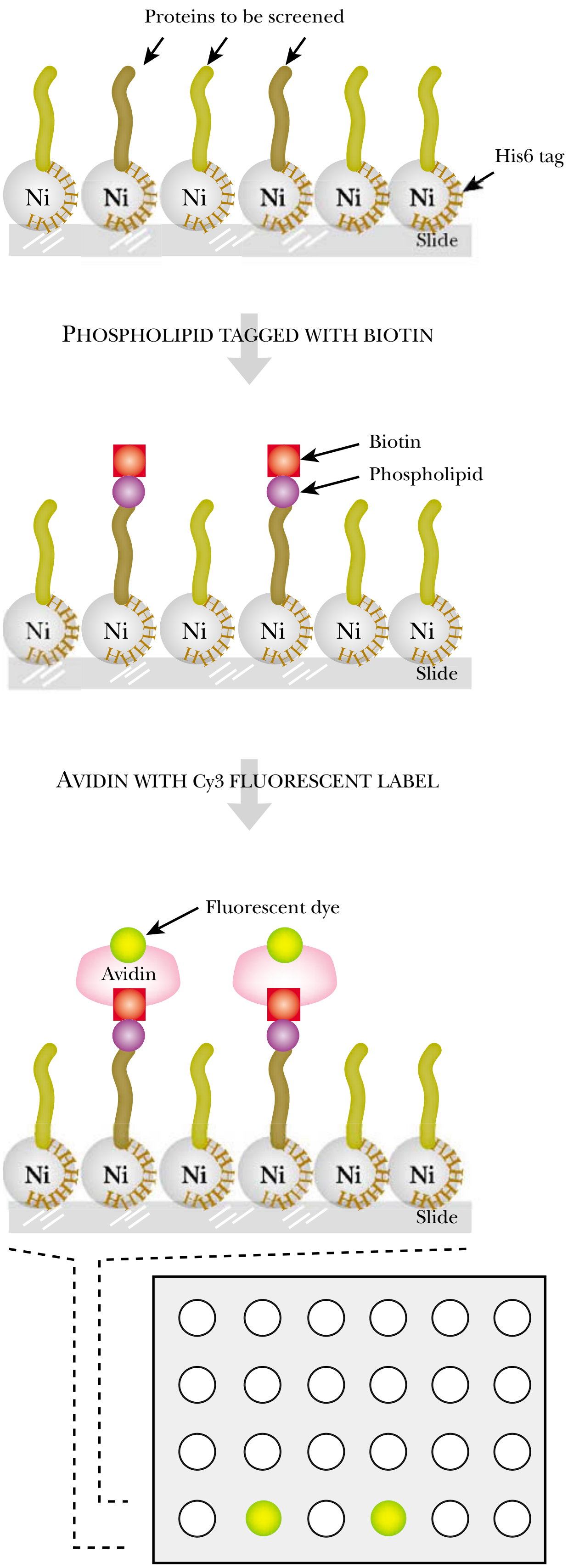


FIGURE 26.24 ***Screening Protein Microarray using Biotin/Streptavidin***

Protein microarrays can be screened to find proteins that bind to phospholipids, for example. The protein microarray is incubated with phospholipid bound to biotin. Then the bound phospholipid is visualized by adding avidin conjugated to a fluorescent dye. Spots that fluoresce represent specific proteins that bind phospholipids.

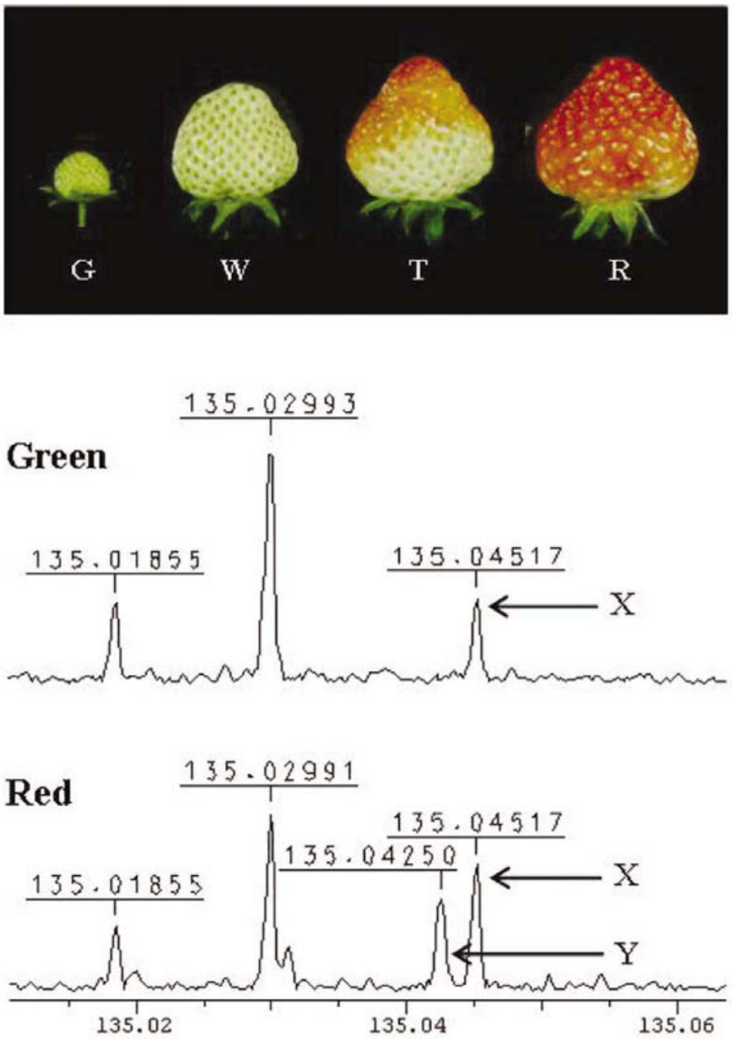


FIGURE 26.25
Metabolome Analysis of Strawberries

Non-targeted metabolic analysis in strawberry. (A) Four consecutive stages of strawberry fruit development (G, green; W, white; T, turning; R, red) were subjected to metabolic analysis using FTMS. Similar fruit samples were used earlier to perform gene expression analysis using cDNA microarrays. (B) An example of high resolution (>100,000) separation of very close mass peaks in data obtained from the analysis of green and red stages of fruit development. Peaks marked with an X have the same mass, while peak Y is different by a mere 3ppm. Courtesy of Phenomenome Discoveries Inc., Saskatoon, Canada.

Metabolome analysis is especially useful for analysis of plants, which typically make many secondary metabolites including pigments, scents, flavors, alkaloids, and other commercially important products. For example, using both aqueous and organic extracts from strawberries has allowed the measurement of nearly 7,000 different metabolites. If printed out, the complete EHMR mass spectra for such mixtures would be a couple of miles long. Comparison of white mutants with wild type red strawberries directly revealed differences in the levels of several intermediates in the pathway for pigment synthesis as well as of the red pigment itself (Fig. 26.25).

Glossary

A (acceptor) site Binding site on the ribosome for the tRNA that brings in the next amino acid

Ac element Intact and active version of a transposon found in maize

acceptor stem Base paired stem of tRNA to which the amino acid is attached

acetylation Addition of an acetyl (CH_3CO) group

aconitase An enzyme of the Krebs cycle that, in animals, also acts as an iron regulatory protein

acridine orange A mutagenic agent that acts by intercalation

activator protein Protein that switches a gene on

active site Special site or pocket on a protein where other molecules are bound and the chemical reaction occurs

acyl phosphate Phosphate derivative in which the phosphate is attached to a carboxyl group

adenine (A) A purine base that pairs with thymine, found in DNA or RNA

adenosine The nucleoside consisting of adenine plus (deoxy)ribose

adhesin Protein that enables bacteria to attach themselves to the surface of animal cells. Same as colonization factor

A-DNA A rare alternative form of double stranded helical DNA

A-form An alternative form of the double helix, with 11 base pairs per turn, often found for double stranded RNA, but rarely for DNA

African Eve Hypothetical female human ancestor thought to have lived in Africa around 100,000–200,000 years ago

agarose A polysaccharide from seaweed that is used to form gels for separating nucleic acids by electrophoresis

agarose gel electrophoresis Technique for separation of nucleic acid molecules by passing an electric current through a gel made of agarose

AIDS (aquired immunodeficiency syndrome) Disease caused by human immunodeficiency virus (HIV) that damages the immune system

alkaline phosphatase An enzyme that cleaves phosphate groups from a wide range of molecules

allele One particular version of a gene, or more broadly, a particular version of any locus on a molecule of DNA

***allo*-lactose** An isomer of lactose that is the true inducer of the *lac* operon

allosteric enzyme Enzyme that changes shape and activity when it binds a small molecule

allosteric protein Protein that changes shape when it binds a small molecule

alpha- (α-) carbon Central carbon atom of an amino acid that carries both the amino group and the carboxyl group

alpha- (α-) helix A helical secondary structure found in proteins

alpha complementation Assembly of functional β-galactosidase from N-terminal alpha fragment plus rest of protein

alpha fragment N-terminal fragment of β-galactosidase

alternative sigma factor A nonstandard sigma factor needed to recognize a specialized subset of genes

alternative splicing Alternative ways to make two or more different final mRNA molecules by using different segments from the same original gene

Alu element An example of a SINE, a particular short DNA sequence found in many copies on the chromosomes of humans and other primates

Ames test Test for mutagenic activity that makes use of bacteria

amethopterin Another name for the anti-cancer drug methotrexate

amino acid Monomer from which polypeptide chains are built

amino- or N-terminus The end of a polypeptide chain that is made first and that has a free amino group

amino-acyl tRNA synthetase Enzyme that attaches an amino acid to tRNA

aminoglycosides Family of antibiotics that inhibit protein synthesis by binding to the small subunit of the ribosome; includes streptomycin, kanamycin, neomycin, tobramycin, gentamycin and many others

***amp* gene** Gene conveying resistance to ampicillin and related antibiotics and encoding beta-lactamase. Same as *bla* gene

ampicillin A widely used antibiotic of the penicillin family

anaerobic respiration Respiration using other oxidizing agents (e.g. nitrate) instead of oxygen

analog A chemical substance that mimics another well enough to be mistaken for it by biological macromolecules, in particular enzymes, receptor proteins, or regulatory proteins

anchor sequence Sequence added to primers or probes that may be used for binding to a support or may incorporate convenient restriction sites, primer binding sites for future manipulations, or primer bindings sites for subsequent PCR reactions

aneuploid Having irregular numbers of different chromosomes

annealing The re-pairing of separated single strands of DNA to form a double helix

anti-anti-sigma factor Protein that binds to an anti-sigma factor and so prevents the anti-sigma factor from binding to and inhibiting a sigma factor

antibiotics Chemical substances that inhibit specific biochemical processes and thereby stop bacterial growth selectively; that is, without killing the patient too.

antibody Protein made by the immune system to recognize and bind to foreign proteins or other macromolecules

anticodon loop Loop of tRNA molecule that contains the anticodon

anticodon Group of three complementary bases on tRNA that recognize and bind to a codon on the mRNA

antifreeze protein Protein that prevents freezing of blood, tissue fluids or cells of organisms living at subzero temperatures

antiparallel Parallel, but running in opposite directions

antisense RNA An RNA molecule that is complementary to messenger RNA or another functional RNA molecule

anti-Shine-Dalgarno sequence Sequence on 16S rRNA that is complementary to the Shine-Dalgarno sequence of mRNA

anti-sigma factor Protein that binds to a sigma factor and blocks its role in the initiation of transcription

anti-termination factor Same as anti-termination protein

anti-terminator protein Protein that allows transcription to continue through a transcription terminator

AP endonuclease Endonuclease that nicks DNA next to an AP-site

Apicomplexa Phylum of parasitic single-celled eukaryotes that contain both mitochondria and degenerate non-photosynthetic chloroplasts

apicoplast Degenerate non-photosynthetic chloroplast found in members of the Apicomplexa, including the malaria parasite

apoprotein That portion of a protein consisting only of the polypeptide chains without any extra cofactors or prosthetic groups

apoptosis Programmed suicide of unwanted cells during development or to fight infection

AP-site A site in DNA where a base is missing (AP-site = apurinic site or apyrimidinic site depending on the nature of the missing base)

***araBAD* operon** Operon that encodes proteins involved in metabolism of the sugar arabinose

arabinose A five-carbon sugar often found in plant cell wall material that can be used as a carbon source by many bacteria

Archaebacteria (or Archaea) Type of bacteria forming a genetically distinct domain of life. Includes many bacteria growing under extreme conditions

archaebacteria One of the three domains of life comprising the "ancient" bacteria

archea New name for archaebacteria, one of the three domains of life

ascomycete Type of fungus that produces four (or sometimes eight) spores in a structure known as an ascus

ascospore Type of spore made inside an ascus by fungi of the ascomycete group, including yeasts and molds

ascus Specialized spore forming structure of ascomycete fungus

asexual or vegetative reproduction Form of reproduction in which there is no reshuffling of the genes between two individuals

asymmetric center Carbon atom with four different groups attached. This results in optical isomerism

***att*λ** Lambda attachment site—site where lambda inserts its DNA into the bacterial chromosome

attenuation Type of transcriptional regulation that works by premature termination and depends on alternative stem and loop structures in the leader region of the mRNA

attenuation protein Regulatory protein involved in attenuation and that binds to the leader region of mRNA

autogenous regulation Self regulation, i.e. when a DNA-binding protein regulates the expression of its own gene

autoradiography Allowing radioactive materials to take pictures of themselves by laying them flat on photographic film

auxin Plant hormone that induces plant cells to grow bigger

avidin A protein from egg white that binds biotin very tightly

azidothymidine (AZT) Nucleoside analog that acts as a DNA chain terminator during reverse transcription and is used against AIDS. Also known as zidovudine

B1 element An example of a SINE found in mice; the precursor sequence from which the human Alu element evolved

bacteria Primitive, relatively simple, single-celled organisms that lack a cell nucleus

bacterial (70S) ribosome Type of ribosome found in bacterial cells

bacterial artificial chromosome (BAC) Single copy vector based on the F-plasmid of *E. coli* that can carry very long inserts of DNA. Widely used in the human genome project

bacteriocin A toxic protein made by bacteria to kill other, closely related, bacteria

bacterioferritin The bacterial analog of ferritin, an iron storage protein

bacteriophage (phage) Virus that infects bacteria

bacteriophage ΦX174 A small spherical virus that contains circular single-stranded DNA and infects *Escherichia coli*

bacteriophage lambda Virus of *E. coli* with both lytic and lysogenic alternatives to its life cycle, which is widely used as a cloning vector

bacteriophage M13 A small male-specific filamentous virus that contains circular single-stranded DNA and infects *Escherichia coli*

bacteriophage Mu A bacterial virus that replicates by transposition and causes mutations by insertion within host cell genes

bacteriophage T7 A bacteriophage that infects *E. coli* whose promoters are only recognized by its own RNA polymerase

bait The fusion between the DNA binding domain of a transcriptional activator protein and another protein as used in two-hybrid screening

bandshift assay Method for testing binding of a protein to DNA by measuring the change in mobility of DNA during gel electrophoresis. Same as gel retardation or mobility shift assay

Barr body Inactive and highly condensed X-chromosome as seen in the light microscope

base analog Chemical mutagen that mimics a DNA base

base pair A pair of two complementary bases (A with T or G with C) held together by hydrogen bonds

base substitution Mutation in which one base is replaced by another

base Alkaline chemical substance, in molecular biology especially refers to the cyclic nitrogen compounds found in DNA and RNA

basic HLH (b/HLH) protein DNA-binding protein with a positively charged (basic) region next to a HLH-motif

bent DNA Double helical DNA that is bent due to several runs of As

beta- (β-) sheet A flat sheet-like secondary structure found in proteins

beta-galactosidase or β-galactosidase (LacZ) Enzyme that splits lactose and related molecules to release galactose

beta-lactamase or β-lactamase Enzyme that inactivates β-lactam antibiotics such as ampicillin by cleaving the lactam ring

beta-lactams or β-lactams Family of antibiotics that inhibit cross-linking of the peptidoglycan of the bacterial cell wall; includes penicillins and cephalosporins

B-form or B-DNA The normal form of the DNA double helix, as originally described by Watson and Crick

bi-directional replication Replication that proceeds in two directions from a common origin

binary fission Simple form of cell division in which the cell elongates and divides down the middle after replication of the DNA

binding protein Protein whose role is to bind another molecule

bioinformatics The computerized analysis of large amounts of biological sequence data

biopanning Method of screening a phage display library for a desired displayed protein by binding to a bait molecule attached to a solid support

biotin Vitamin that is widely used to label or tag nucleic acids in molecular biology because it may be bound very tightly by avidin or streptavidin

***bla* gene** Gene conveying resistance to ampicillin and related antibiotics and encoding beta-lactamase. Same as *amp* gene

blue/white screening Screening procedure based on insertional inactivation of the gene for β-galactosidase

blunt ends Ends of a double-stranded DNA molecule that are fully base paired and have no unpaired single-stranded overhang

bovine spongiform encephalopathy (BSE) Mad cow disease, a prion disease of cattle

bp Abbreviation for base pair(s)

branch site Site in the middle of an intron where branching occurs during splicing

Buchnera Genus of gram-negative bacterial symbionts found in insects that supply their host insect with essential amino acids

budding Type of cell division seen in yeasts in which a new cell forms as a bulge on the mother cell, enlarges, and finally separates

CAAT box A sequence often found in the upstream region of eukaryotic promoters and which binds transcription factors

calmodulin A small calcium binding protein of animal cells

cap Structure at the 5′-end of eukaryotic mRNA consisting of a methylated guanosine attached in reverse orientation

capsid Shell or protective layer that surrounds the DNA or RNA of a virus particle

carboxy- or C-terminus The end of a polypeptide chain that is made last and has a free carboxy-group

carrier protein Protein that carries other molecules around the body or within the cell

CAT Chloramphenicol acetyl transferase

catenane Structure in which two or more circles of DNA are interlocked

cauliflower mosaic virus A small spherical virus of plants with circular DNA. Some of its promoters are used in plant genetic engineering

CD4 protein A protein found on the surface of T-cells that acts as a receptor during the immune response

cDNA (complementary DNA) DNA copy of a gene that lacks introns and therefore consists solely of the coding sequence. Made by reverse transcription of mRNA

cDNA library Collection of genes in their cDNA form, lacking introns

cell cycle Series of stages that a cell goes through from one cell division to the next

cell The cell is the basic unit of life. Each cell is surrounded by a membrane and usually has a full set of genes that provide it with the genetic information necessary to operate

Cen sequence See centromere sequence

central dogma Basic plan of genetic information flow in living cells which relates genes (DNA), message (RNA) and proteins

centrifugation Process in which samples are spun at high speed and the centrifugal force causes the larger or heavier components to sediment to the bottom

centriole Organelle involved in organizing chromosome partition during the division of eukaryotic cells

centromere (Cen) sequence Sequence at centromere of eukaryotic chromosome that is needed for correct partition of chromosomes during cell division

centromere Region of eukaryotic chromosome, usually more or less central, where the microtubules attach during mitosis and meiosis

cephalosporins Group of antibiotics of the β-lactam type that inhibit cross-linking of the peptidoglycan of the bacterial cell wall

CG-islands Region of DNA in eukaryotes that contains many clustered CG sequences that are used as targets for cytosine methylation

chain termination mutation Same as nonsense mutation

chain termination sequencing Method of sequencing DNA by using dideoxynucleotides to terminate synthesis of DNA chains. Same as dideoxy sequencing

chain terminator Agent that prevents continued elongation of a strand of DNA

chaotropic agent Chemical compound that disrupts water structure and so helps hydrophobic groups to dissolve

chaperone Sometimes "molecular chaperone"; same as chaperonin

chaperonin Protein that oversees the correct folding of other proteins

charged tRNA tRNA with an amino acid attached

chemiluminescence Production of light by a chemical reaction

chi sites Specific sequences on the DNA of eukaryotes where crossovers form

chimera Hybrid molecule that includes DNA from more than one source

chiral center Same as asymmetric center

chloramphenicol Antibiotic that binds to 23S rRNA and inhibits protein synthesis

chloramphenicol acetyl transferase (CAT) Enzyme that inactivates chloramphenicol by adding acetyl groups

chlorophyll Green pigment that absorbs light during photosynthesis

choleratoxin Type of toxin made by *Vibrio cholerae* the cholera bacterium

chromatid Single double-helical DNA molecule making up whole or half of a chromosome. A chromatid also contains histones and other DNA-associated proteins.

chromatin remodeling complex A protein assembly that rearranges the histones of chromatin in order to allow transcription

chromatin Complex of DNA plus protein which constitutes eukaryotic chromosomes

chromogenic substrate Colorless or pale substrate that is converted to a strongly colored product by an enzyme

chromosome banding technique Visualization of chromosome bands by using specific stains that emphasize regions lacking genes

chromosome walking Method for cloning neighboring regions of a chromosome by successive cycles of hybridization using overlapping probes

chromosome Structure containing the genes of a cell and made of a single molecule of DNA

***cI* gene** Gene encoding the lambda repressor or cI protein

cI protein Lambda repressor protein responsible for maintaining bacteriophage lambda in the lysogenic state

ciliates Group of free-living protozoans that move by means of cilia attached to the cell surface

ciprofloxacin A fluoroquinolone antibiotic that inhibits DNA gyrase

2μ circle Same as 2μ plasmid

cistron Segment of DNA (or RNA) that encodes a single polypeptide chain

clamp-loading complex Group of proteins that loads the sliding clamp of DNA polymerase onto the DNA

clavulanic acid Beta-lactam compound that reacts with and inactivates beta-lactamases

cloning vector Any molecule of DNA that can replicate itself inside a cell and is used for carrying cloned genes or segments of DNA. Usually a small multicopy plasmid or a modified virus

cloverleaf structure 2-D structure showing base pairing in a tRNA molecule

coding strand The strand of DNA equivalent in sequence to the messenger RNA (same as plus strand)

co-dominance When two different alleles both contribute to the observed properties

codon Group of three RNA or DNA bases that encodes a single amino acid

cofactor Extra chemical group bound (often temporarily) to a protein but which is not part of the polypeptide chain

co-immunoprecipitation Method of identifying protein-protein interaction by using antibodies to a one of the proteins

cointegrate A temporary structure formed by linking the strands of two molecules of DNA during transposition, recombination or similar processes

ColE plasmid Small multicopy plasmid that carries genes for colicins of the E group.

ColEI plasmid Small multicopy plasmid of *Escherichia coli* that forms the basis of many cloning vectors widely used in molecular biology

"cold" Slang for non-radioactive

colicin Toxic protein or bacteriocin made by *Escherichia coli* to kill closely related bacteria

colonization factor Protein that enables bacteria to attach themselves to the surface of animal cells. Same as adhesin

competent cell Cell that is capable of taking up DNA from the surrounding medium

competitive inhibitor Chemical substance that inhibits an enzyme by mimicking the true substrate well enough to be mistaken for it

complementary DNA (cDNA) Version of a gene that lacks the introns and is made from the corresponding mRNA by using reverse transcriptase

complementary sequences Two nucleic acid sequences whose bases pair with each other because A, T, G, C in one sequence correspond to T, A, C, G, respectively, in the other

complex transposon A transposon that moves by replicative transposition

composite transposon A transposon that consists of two insertion sequences surrounding a central block of genes

conditional mutation Mutation whose phenotypic effects depend on environmental conditions such as temperature or pH

conjugated protein Complex of protein plus another molecule

conjugation bridge Junction that forms between two cells and provides a channel for DNA to move from donor to recipient during conjugation

conjugation Process in which genes are transferred by cell to cell contact

conjugative transposon A transposon that is also capable of transferring itself from one bacterial cell to another by conjugation

consensus sequence Idealized base sequence consisting of the bases most often found at each position

conservative substitution Replacement of an amino acid with another that has similar chemical and physical properties

conservative transposition Same as cut-and-paste transposition

constitutive gene Gene that is expressed all the time

contig A stretch of known DNA sequence that is contiguous and lacks gaps

contractile protein Protein that uses energy (usually ATP) to contract

controlled pore glass (CPG) Glass with pores of uniform sizes that is used as a solid support for chemical reactions such as artificial DNA synthesis

Coomassie Blue A blue dye used to stain proteins

copy number The number of copies of a gene or plasmid found within a single host cell

core enzyme The part of DNA or RNA polymerase that synthesizes new DNA or RNA (i.e. lacking the recognition and/or attachment subunits)

co-repressor In prokaryotes—a small signal molecule needed for some repressor proteins to bind to DNA; in eukaryotes—an accessory protein, often a histone deacetylase, involved in gene repression

***cos* sequences (lambda cohesive ends)** Complementary 12 bp long overhangs found at each end of the linear form of the lambda genome

cosmid Small multicopy plasmid that carries lambda *cos* sites and can carry around 45 kb of cloned DNA

cotransfer frequency Frequency with which two genes remain associated during transfer of DNA between cells

cotranslational export Export of a protein across a membrane while it is still being synthesized by a ribosome

coupled transcription-translation When ribosomes of bacteria start translating an mRNA molecule that is still being transcribed from the DNA

covalently closed circular DNA (cccDNA) Circular DNA with no nicks in either strand

cPABP (chloroplast polyadenylate binding protein) A translational activator protein that controls expression of chloroplast mRNA

crista (plural cristae) Infolding of the photosynthetic membrane in chloroplast

crossing over When two different strands of DNA are broken and are then joined to one another

crossover Structure formed when the strands of two DNA molecules are broken and joined to each other

crossover resolvase Bacterial enzyme that separates covalently fused chromosomes

crown gall Type of tumor formed on plants due to infection by *Agrobacterium* carrying a Ti-plasmid

CRP (cyclic AMP receptor protein) Bacterial protein that binds cyclic AMP and then binds to DNA

cruciform structure Cross shaped structure in double stranded DNA (or RNA) formed from an inverted repeat

cryptic plasmid A plasmid that confers no identified characteristics or phenotypic properties

CTD (carboxy-terminal domain) Repetitive region at the C-terminus of RNA polymerase II that may be phosphorylated

C-terminus Carboxy-terminus. The end of a polypeptide chain that is made last and has a free carboxy-group

cut-and-paste transposition Type of transposition in which a transposon is completely excised from its original location and moves as a whole unit to another site

cycloheximide An antibiotic that inhibits eukaryotic protein synthesis

cytidine The nucleoside consisting of cytosine plus (deoxy)ribose

cytokinesis Cell division

cytokinin Plant hormone that induces plant cells to divide

cytoplasm The portion of a cell that is inside the cell membrane but outside the nucleus

cytosine (C) One of the pyrimidine bases found in DNA or RNA and which pairs with guanine

D- and L-forms The two isomeric forms of an optically active substance; also called D- and L-isomers

data mining The use of computer analysis to find useful information by filtering or sifting through large amounts of data

***de novo* methylase** An enzyme that adds methyl groups to wholly nonmethylated sites

deaminase An enzyme that removes an amino group

deamination Loss of an amino group

defective interfering RNA (DI-RNA) RNA molecule derived from viral RNA by deletions that remove essential functions. DI-RNA therefore depends on the parental virus for replication

defective phage Mutant phage that lacks genes for making virus particles

degenerate primer Primer with several alternative bases at certain positions

deletion Mutation in which one or more bases is lost from the DNA sequence

demethylase An enzyme that removes methyl groups

denaturant Chemical compound that destroys the 3D structure of proteins, especially by breaking hydrogen bonds

denaturation When describing proteins or other biological polymers, refers to the loss of correct 3-D structure

denaturing gradient gel electrophoresis (DGGE) Combination of gel electrophoresis with DNA denaturation that allows separation of DNA molecules differing in sequence by only a single base

deoxynucleoside A nucleoside containing deoxyribose as the sugar

deoxynucleotide A nucleotide containing deoxyribose as the sugar

deoxyribonuclease (DNase) Enzyme that cuts or degrades DNA

deoxyribonuclease I (DNase I) Non-specific nuclease that cuts DNA between any two nucleotides. Often used in footprint analysis

deoxyribonucleic acid (DNA) The nucleic acid polymer of which the genes are made

deoxyribonucleoside 5′-triphosphate (deoxyNTP) Precursor for DNA synthesis consisting of a base, deoxyribose and three phosphate groups

deoxyribose The sugar with five carbon atoms that is found in DNA

detergent Molecule that is hydrophobic at one end and highly hydrophilic at the other and which is used to dissolve lipids or grease

deuterostomes Group of animal phyla including echinoderms and vertebrates

Dicer Ribonuclease that cleaves double-stranded RNA into segments of 21–23 bp

dideoxy sequencing Method of sequencing DNA by using dideoxynucleotides to terminate synthesis of DNA chains. Same as chain termination sequencing

dideoxynucleotide Nucleotide whose sugar is dideoxyribose instead of ribose or deoxyribose

dideoxyribose Derivative of ribose that lacks the oxygen of both the 2′ and the 3′ hydroxyl groups

***dif* site** Site on bacterial chromosome used by crossover resolvase to separate covalently fused chromosomes

differential display PCR Variant of RT-PCR that specifically amplifies messenger RNA from eukaryotic cells using oligo(dT) primers

differentiation Progressive changes in the structure and gene expression of cells belonging to a single organism that leads to the formation of different types of cell

digoxigenin A steroid from foxglove plant widely used for chemical labeling of DNA molecules

dihydrofolate (DHF) Cofactor with a variety of roles including making precursors for DNA and RNA synthesis

dihydrofolate reductase Enzyme that converts dihydrofolate back to tetrahydrofolate

dimethoxytrityl (DMT) group Group used for blocking the 5′-hydroxyl of nucleotides during artificial DNA synthesis

diphthamide Modified amino acid found only in eukaryotic elongation factor eEF2 that is the target for diphtheria toxin

diploid Having two copies of each chromosome and hence of each gene

dipolar ion Same as zwitterion; a molecule with both a positive and a negative charge

directed mutagenesis Deliberate alteration of the DNA sequence of a gene by any of a variety of artificial techniques

D-isomer That one of a pair of optical isomers that rotates light in a clockwise direction

disulfide bond A sulfur to sulfur bond formed between two sulfhydryl groups, in particular between those of cysteine, and which binds together two protein chains

***Dmd* gene** Gene responsible for Duchenne muscular dystrophy

DNA Deoxyribonucleic acid, nucleic acid polymer of which the genes are made

DNA adenine methylase (Dam) A bacterial enzyme that methylates adenine in the sequence GATC

DNA array DNA chip used to simultaneously detect and identify many short RNA or DNA fragments by hybridization. Also known as DNA chip or oligonucleotide array detector

DNA chip Chip used to simultaneously detect and identify many short DNA fragments by DNA-DNA hybridization. Also known as DNA array or oligonucleotide array detector

DNA cytosine methylase (Dcm) A bacterial enzyme that methylates cytosine in the sequences CCAGG and CCTGG

DNA fingerprint Individually unique pattern due to multiple bands of DNA produced using restriction enzymes, separated by electrophoresis and usually visualized by Southern blotting

DNA glycosylase Enzyme that breaks the bond between a base and the deoxyribose of the DNA backbone

DNA gyrase An enzyme that introduces negative supercoils into DNA, a member of the type II topoisomerase family

DNA helicase Enzyme that unwinds double helical DNA

DNA library Collection of cloned segments of DNA that is big enough to contain at least one copy of every gene from a particular organism. Same as gene library

DNA ligase Enzyme that joins DNA fragments covalently, end to end

DNA microarray same as DNA array or DNA chip

DNA polymerase An enzyme that elongates strands of DNA, especially when chromosomes are being replicated

DNA polymerase α Enzyme that makes short segment of initiator DNA during replication of animal chromosomes

DNA polymerase δ Enzyme that makes most of the DNA when animal chromosomes are replicated

DNA polymerase eta (η) A repair polymerase in animals that can replicate past thymine dimers

DNA polymerase I (Pol I) Bacterial enzyme that makes small stretches of DNA to fill in gaps between Okazaki fragments or during repair of damaged DNA

DNA polymerase III (Pol III) Enzyme that makes most of the DNA when bacterial chromosomes are replicated

DNA polymerase V A repair polymerase in bacteria that can replicate past pyrimidine dimers and AP-sites

DNA virus A virus whose genome consists of DNA

DnaA protein Protein that binds to the origin of bacterial chromosomes and helps initiate replication

domain (of life) Highest ranking group into which living creatures are divided, based on the most fundamental genetic properties

domain (of protein) A region of a polypeptide chain that folds up more or less independently to give a local 3D-structure

dominant allele Allele whose properties are expressed in the phenotype whether present as a single or double copy

donor cell Cell that donates DNA to another cell

double helix Structure formed by twisting two strands of DNA spirally around each other

Ds elements Defective version of a transposon found in maize; cannot move alone but needs the Ac element to provide transposase

dsRNA Double-stranded RNA

Duchenne muscular dystrophy One of several inherited diseases affecting mucle function

duplication Mutation in which a segment of DNA is duplicated

E (exit) site Site on the ribosome that a tRNA occupies just before leaving the ribosome

early genes Genes expressed early during virus infection and that mainly encode enzymes involved in virus DNA (or RNA) replication

EDTA (ethylene diamine tetraacetate) A widely used chelating agent that binds di-positive ions such as Ca^{2+} and Mg^{2+}

effective genome The portion of the genome that consists of useful genetic information and ignores the intervening and non-coding DNA. Only applicable to eukaryotic organisms

electrophoresis Movement of charged molecules due to an electric field. Used to separate and purify nucleic acids and proteins

electroporator Device that uses a high voltage discharge to make cells competent to take up DNA

electrospray ionization (ESI) Type of mass spectrometry in which gas-phase ions are generated from ions in solution

elongation factors Proteins that are required for the elongation of a growing polypeptide chain

enantiomers A pair of mirror-image optical isomers (i.e., D- and L-isomers)

endonuclease A nuclease that cuts a nucleic acid in the middle

endoplasmic reticulum Internal system of membranes found in eukaryotic cells

endosymbiosis Form of symbiosis where one organism lives inside the other

enhancer Regulatory sequence outside, and often far away from, the promoter region that binds transcription factors

Entamoeba A very primitive single-celled eukaryote that lacks mitochondria

enterotoxins Types of toxin made by enteric bacteria including some pathogenic strains of *E. coli*

enzyme A protein or RNA molecule that catalyses a chemical reaction

epistasis When a mutation in one gene masks the effect of alterations in another gene

error-prone repair Type of DNA repair process that introduces mutations

erythromycin An antibiotic that inhibits bacterial protein synthesis

***Escherichia coli* (*E. coli*)** A species of bacterium commonly used in genetics and molecular biology

essential amino acids Those amino acids that animals are unable to synthesize for themselves (Arg, Val, Ile, Leu, Phe, Trp, Thr, Met, Lys, His)

ethidium bromide A stain that specifically binds to DNA or RNA and appears orange if viewed under ultraviolet light

Eubacteria One of the three domains of life comprising bacteria of the "normal" kind as opposed to the genetically distinct Archaea. Includes the plastids and mitochondria.

euchromatin Normal chromatin, as opposed to heterochromatin

eukaryote Higher organism with advanced cells, which have more than one chromosome within a compartment called the nucleus

eukaryotic (80S) ribosome Type of ribosome found in cytoplasm of eukaryotic cell and encoded by genes in the nucleus

excision repair system Also known as "cut and patch" repair. A DNA repair system that recognizes bulges in the DNA double helix, removes the damaged strand and replaces it

excisionase Enzyme that reverses DNA integration by removing a segment of dsDNA and resealing the gap. In particular, lambda excisionase removes integrated lambda DNA

exon Segment of a gene that codes for protein and that is still present in the messenger RNA after processing is complete

exon cassette selection Type of alternative splicing that makes different mRNA molecules by choosing different selections of exons from the primary transcript

exon trapping Experimental procedure for isolating exons by using their flanking splice recognition sites

exonuclease Enzyme that cleaves nucleic acid molecules at the end and usually removes just a single nucleotide at a time

3′-exonuclease An enzyme that degrades nucleic acids from the 3′-end

expressed sequence tag (EST) A special type of STS derived from a region of DNA that is expressed by transcription into mRNA

expression site Special location on chromosome where the chosen copy of a gene present in multiple copies may be expressed

expression vector Vector specifically designed to place a cloned gene under control of a plasmid-borne promoter

extein A segment of a protein that remains after the splicing out of any inteins

extragenic suppression Reversion of a mutation by a second change that is within another distinct gene

Fe_4S_4 cluster A group of inorganic iron and sulfur atoms found as a cofactor in several proteins

fermentation A biochemical process that releases energy without oxygen or light

ferritin An iron storage protein

fertility plasmid Plasmid that enables a cell to donate DNA by conjugation

filial generations Successive generations of descendants from a genetic cross which are numbered F1, F2, F3, etc., to keep track of them

filoviruses A family of filamentous negative single-stranded RNA viruses that consist of an inner helical capsid covered by an outer envelope

FISH See Fluorescence in Situ Hybridization

FLAG® tag A short peptide tag (AspTyrLysAspAspAspAspLys) that is bound by a specific anti-FLAG® antibody that may be attached to a resin for use in column purification of proteins

flippase Same as Flp recombinase

Flp recombinase (or flippase) Enzyme encoded by the 2μ plasmid of yeast that catalyzes recombination between inverted repeats (FRT sites)

Flp recombination target (or FRT site) Recognition site for Flp recombinase

fluorescence Process in which a molecule absorbs light of one wavelength and then emits light of another, longer, lower energy wavelength

fluorescence activated cell sorter (FACS) Instrument that sorts cells (or chromosomes) based on fluorescent labeling

fluorescence in situ hybridization (FISH) Using a fluorescent probe to visualize a molecule of DNA or RNA in its natural location

fluorescence resonance energy transfer (FRET) Transfer of energy from short-wavelength fluorophore to long-wavelength fluorophore so quenching the short wave emission

fluorophore A fluorescent chemical group

fluoroquinolones Subgroup of the quinolone family of antibiotics that inhibits DNA gyrase and certain other topoisomerases

folate Cofactor involved in carrying one carbon groups in DNA synthesis

footprint Method for testing binding of a protein to DNA by its protection of DNA from chemical degradation

F-plasmid Fertility plasmid that allows *E. coli* to donate DNA by conjugation

frameshift Alteration in the reading frame during polypeptide synthesis

frameshift mutation Mutation in which the reading frame of a structural gene is altered by insertion or deletion of one or a few bases

FRT site Flp recombination target, the recognition site for Flp recombinase

functional genomics The study of the whole genome and its expression

Fur (ferric uptake regulator) Global regulatory protein that senses iron levels in bacteria

fusidic acid An antibiotic that inhibits protein synthesis

G1 phase Stage of the eukaryotic cell cycle following cell division; cell growth occurs here

G2 phase Stage of the eukaryotic cell cycle between DNA synthesis and mitosis: preparation for division

galactoside Compound of galactose, such as lactose, ONPG or X-gal

β-galactosidase Enzyme that splits lactose and related compounds

gametes Cells specialized for sexual reproduction that are haploid (have one set of genes)

gametophyte Haploid phase of a plant, especially of lower plants such as mosses and liverworts, where it forms a distinct multicellular body

gap A break in a strand of DNA or RNA where bases are missing

GC ratio The amount of G plus C relative to all four bases in a sample of DNA. The GC ratio is usually expressed as a percentage

gel electrophoresis Electrophoresis of charged molecules through a gel meshwork in order to sort them by size

gel retardation Method for testing binding of a protein to DNA by measuring the change in mobility of DNA during gel electrophoresis. Same as bandshift assay or mobility shift assay

gene A unit of genetic information

gene cassette Deliberately designed segment of DNA that is flanked by convenient restriction sites and usually carries a gene for resistance to an antibiotic or some other easily observed character

gene conversion Recombination and repair of DNA during meiosis that leads to replacement of one allele by another. This may result in a non-Mendelian ratio among the progeny of a genetic cross

gene creature Genetic entitiy that consists primarily of genetic information, sometimes with a protective covering, but without its own machinery to generate energy or replicate macromolecules

gene family Group of closely related genes that arose by successive duplication and perform similar roles

gene fusion Structure in which parts of two genes are joined together, in particular when the regulatory region of one gene is joined to the coding region of a reporter gene

gene library Collection of cloned segments of DNA that is big enough to contain at least one copy of every gene from a particular organism. Same as DNA library

gene product End product of gene expression; usually a protein but includes various untranslated RNAs such as rRNA, tRNA, and snRNA

gene superfamily Group of related genes that arose by several stages of successive duplication. Members of a superfamily have often diverged so far that their ancestry may be difficult to recognize

GeneChip® array The first brand of DNA chip, made by Affymetrix Corporation

generalized transduction Type of transduction where fragments of bacterial DNA are packaged at random and all genes have roughly the same chance of being transferred

generation time The time from the start of one cell division to the start of the next

genetic code The code for converting the base sequence in nucleic acids, read in groups of three, into the sequence of a polypeptide chain

genetic element Any molecule or segment of DNA or RNA that carries genetic information and acts as a heritable unit

genome mining The use of computer analysis to find useful information by filtering or sifting through large amounts of biological sequence data

genome The entire genetic information of an individual organism

genomics Study of genomes as a whole rather than one gene at a time

genotype The genetic make-up of an organism

genus A group of closely related species

germ line cells Reproductive cells producing eggs or sperm that take part in forming the next generation

germline cell Cell capable of forming gametes and so contributing to the next generation of animals

Giardia A very primitive single-celled eukaryote that lacks mitochondria

gigabase pair (Gbp) 10^9 base pairs

global regulation Regulation of a large group of genes in response to the same stimulus

global regulator A regulator that controls a large group of genes, generally in response to some stimulus or developmental stage

globins Family of related proteins, including hemoglobin and myoglobin, that carry oxygen in the blood and tissues of animals

β-D-glucosyl hydroxymethyluracil See J-base

glutathione-S-transferase (GST) Enzyme that binds to the tripeptide, glutathione. GST is often used in making fusion proteins

glycine The simplest amino acid

glycogen Storage carbohydrate found both in bacteria and in the livers of animals

glycoprotein Complex of protein plus carbohydrate

Golgi apparatus A membrane bound organelle that takes part in export of materials from eukaryotic cells

gp120 A protein (i.e. a "gene product") of 120 kd found in the outer envelope of HIV that binds to the CD4 receptor

gram-negative bacteria Major division of Eubacteria that possess an extra outer membrane lying outside the cell wall

gram-positive bacteria Major division of Eubacteria that lack an extra outer membrane lying outside the cell wall

gratuitous inducer A molecule (usually artificial) that induces a gene but is not metabolized like the natural substrate; the best known example is the induction of the *lac* operon by IPTG

green fluorescent protein (GFP) A jellyfish protein that emits green fluorescence and is widely used in genetic analysis

Gregor Mendel Discovered the basic laws of genetics by crossing pea plants

guanidine Non-ionized form of guanidinium

guanidinium chloride A widely used denaturant of proteins

guanine (G) A purine base found in DNA or RNA that pairs with cytosine

guanosine The nucleoside consisting of guanine plus (deoxy)ribose

guide RNA Small RNA used to locate sequences on a longer mRNA during RNA editing

hairpin A double stranded base-paired structure formed by folding a single strand of DNA or RNA back upon itself

haploid Possessing only a single copy of each chromosome or gene

haploid genome A complete set containing a single copy of all the genes (generally used of organisms that have two or more sets of each gene)

H-DNA A form of DNA consisting of a triple helix. Its formation is promoted by acid conditions and by runs of purine bases

headful packaging Type of virus packaging mechanism that depends on the amount of DNA the head of the virus particle can hold (as opposed to using specific recognition sequences)

heat shock protein (HSP) Protein induced in response to high temperature. Many heat shock proteins are chaperonins

heat shock response Response to high temperature by expressing a set of genes that encode heat shock proteins

helicase Enzyme that unwinds double helical DNA

helix-loop-helix (HLH) One type of DNA-binding motif common in proteins

helix-turn-helix (HTH) One type of DNA-binding motif common in proteins

helper phage Phage that provides the necessary genes so allowing a defective phage to make virus particles

helper virus A virus that provides essential functions for defective viruses, satellite viruses and satellite RNA

hemi-methylated Methylated on only one strand

hemolysin Type of toxin that lyses red blood cells

herpes viruses A family of spherical animal DNA viruses with an outer envelope of material stolen from the nuclear membrane of the host cell

heterochromatin A highly condensed form of chromatin that cannot be transcribed because it cannot be accessed by RNA polymerase

heterodimer Dimer composed of two different subunits

heteroduplex A DNA double helix composed of single strands from two different DNA molecules

heterozygous Having two different alleles of the same gene

Hfr-strain Bacterial strain that transfers chromosomal genes at high frequency due to an integrated fertility plasmid

highly repetitive DNA DNA sequences that exist in hundreds of thousands of copies

His tag Six tandem histidine residues that are fused to proteins so allowing purification by binding to nickel ions that are attached to a solid support. Also known as polyhistidine tag

histone Special positively charged protein that binds to DNA and helps to maintain the structure of chromosomes in eukaryotes

histone acetyl transferase (HAT) Enzyme that adds acetyl groups to histones

histone deacetylase (HDAC) Enzyme that removes acetyl groups from histones

histone-like protein Bacterial protein that binds nonspecifically to DNA and participates in maintaining the structure of the nucleoid; they do not actually have much in common with true histones

H-NS protein (histone-like nucleoid structuring protein) A bacterial protein that binds nonspecifically to DNA and helps maintain the higher level structure of the nucleoid

Holliday junction DNA structure formed during recombination and found at the crossover point where the two molecules of DNA are joined

homeobox genes Genes encoding transcription factors containing a homeodomain that help specify the body plan of multicellular organisms

homeodomain Conserved region of about 60 amino acid residues found in homeobox proteins that that binds DNA by a helix-turn-helix motif

homing intron A mobile intron that encodes a protein enabling it to insert itself into a recognition sequence within a target gene

homologous Related in sequence to an extent that implies common genetic ancestry

homologous chromosomes Two chromosomes are homologous when they carry the same sequence of genes in the same linear order

homologous recombination Recombination between two lengths of DNA that are identical, or nearly so, in sequence

homozygous Having two identical alleles of the same gene

Hoogsteen base pair A type of nonstandard base pair found in triplex DNA, in which a pyrimidine is bound sideways on to a purine

horizontal gene transfer Movement of genes sideways between unrelated organisms. Same as lateral gene transfer

hormone Messenger molecule that circulates inside multicellular organisms such as animals and plants

"hot" Slang for radioactive

hot spots Site in DNA or RNA where mutations are unusually frequent

housekeeping genes Genes that are switched on all the time because they are needed for essential life functions

***Hox* genes** Family of homeobox genes that control overall body layout by regulating the expression of many other regulatory genes, including those for other transcription factors

HU protein (heat-unstable nucleoid protein) A bacterial protein that binds to DNA with low specificity and is involved in bending of DNA

Human Genome Project Program to sequence all the DNA making up the human genome

human immunodeficiency virus (HIV) The retrovirus that causes AIDS

hybrid DNA Artificial double stranded DNA molecule made by pairing two single strands from two different sources

hybridization Pairing of single strands of DNA or RNA from two different (but related) sources to give a hybrid double helix

hydrogen bond Bond resulting from the attraction of a positive hydrogen atom to both of two other atoms with negative charges

hydrophilic Water-loving; readily dissolves in water

hydrophobic Water-hating; repelled by water and dissolves in water only with great difficulty

ice nucleation factor Protein found on surface of certain bacteria that promotes the formation of ice crystals

IHF (integration host factor) A bacterial protein that bends DNA so helping the initiation of transcription of certain genes; named after its role in helping the integration of bacteriophage lambda into the chromosome of *E. coli*

immunity protein Protein that provides immunity. In particular bacteriocin immunity proteins bind to the corresponding bacteriocins and render them harmless

immunization Process of preparing the immune system for future infection by treating the patient with weak or killed versions of an infectious agent

immunological screening Screening procedure that relies on the specific binding of antibodies to the target protein

imprinting When the expression of a particular allele depends on whether it originally came from the father or the mother; (imprinting is a rare exception to the normal rules of genetic dominance)

***in vitro* packaging** Procedure in which virus proteins are mixed with DNA *in vitro* to assemble infectious virus particles. Often used for packaging recombinant DNA into bacteriophage lambda

incompatibility The inability of two plasmids of the same family to co-exist in the same host cell

induced fit When the binding of the substrate induces a change in enzyme conformation so that the two fit together better

induced mutation Mutation caused by external agents such as mutagenic chemicals or radiation

inducer A signal molecule that turns on a gene by binding to a regulatory protein

influenza virus A member of the orthomyxovirus family whose segmented genome consists of eight molecules of negative single-stranded RNA inside a nucleocapsid surrounded by an outer envelope

inherited disease Disease due to a genetic defect that is passed on from one generation to the next

30S initiation complex Initiation complex for translation that contains only the small subunit of the bacterial ribosome

70S initiation complex Initiation complex for translation that contains both subunits of the bacterial ribosome

initiation complex (for replication) Assemblage of proteins that binds to the origin and initiates replication of DNA

initiation factors Proteins that are required for the initiation of a new polypeptide chain

initiator box Sequence at the start of transcription of a eukaryotic gene

initiator DNA (iDNA) Short segment of DNA made just after the RNA primer during replication of animal chromosomes

initiator tRNA The tRNA that brings the first amino acid to the ribosome when starting a new polypeptide chain

inosine A purine nucleoside, found most often in transfer RNA, that contains the unusual base hypoxanthine

insertion Mutation in which one or more extra bases are inserted into the DNA sequence

insertion sequence A simple transposon consisting only of inverted repeats surrounding a gene encoding transposase

insertional inactivation Inactivation of a gene by inserting a foreign segment of DNA into the middle of the coding sequence

insulator A DNA sequence that shields promoters from the action of enhancers and also prevents the spread of heterochromatin

insulator binding protein (IBP) Protein that binds to insulator sequence and is necessary for the insulator to function

Int protein Same as integrase

integrase Enzyme that inserts a segment of dsDNA into another DNA molecule at a specific recognition sequence. In particular, lambda integrase inserts lambda DNA into the chromosome of *E. coli*

integration Insertion of a segment of dsDNA into another DNA molecule at a specific recognition sequence

integron Genetic element consisting of an integration site plus a gene encoding an integrase

intein Self-splicing intervening sequence that is found in a protein

intercalation Insertion of a flat chemical molecule between the bases of DNA, often leading to mutagenesis

intergenic DNA Non-coding DNA that lies between genes

intergenic region DNA sequence between genes

internal eliminated segment (IES) Extra sequences in the DNA of the ciliate micronucleus that are eliminated during conversion of a micronucleus to a macronucleus

internal resolution site (IRS) Site within a complex transposon where resolvase cuts the DNA to release two separate molecules of DNA from the cointegrate during replicative transposition

interphase Part of the eukaryotic cell cycle between two cell divisions and consisting of G1-, S- and G2- phases

intervening sequence An alternative name for an intron

intracellular parasite Parasite that lives inside the cells of its host organism

intragenic suppression Reversion of a mutation by a second change at a different site but within the same gene

intron Segment of a gene that does not code for protein but is transcribed and forms part of the primary transcript

inverse PCR Method for using PCR to amplify unknown sequences by circularizing the template molecule

inversion Mutation in which a segment of DNA has its orientation reversed, but remains at the same location

invertase (Strictly, DNA invertase) An enzyme that recognizes specific sequences at the two ends of an invertible segment and inverts the DNA between them

inverted repeat Sequence of DNA that is the same when read forwards as when read backwards on the complementary strand. One type of palindrome.

inverted repeats Sequences of DNA that are identical when read in opposite orientations on complementary strands.

ionizing radiation Radiation that ionizes molecules that it strikes

IPTG (*iso*-propyl-thiogalactoside) A gratuitous inducer of the *lac* operon

iron regulatory protein (IRP) Translational regulator that controls expression of mRNA in animals in response to the level of iron

iron sulfur cluster Group of iron and sulfur atoms found in proteins and involved in oxidation/reduction reactions

iron-responsive element (IRE) Site on mRNA where the IRP binds

irreversible inhibition Type of inhibition in which an enzyme is permanently inactivated by a chemical change

isoelectric focusing Technique for separating proteins according to their charge by means of electrophoresis through a pH gradient

isoschizomers Restriction enzymes from different species that share the same recognition sequence

ISWI ("imitation switch") complex Smaller type of chromatin remodeling complex

J-base β-D-glucosyl hydroxymethyluracil, an unusual base found in regions of trypanosome DNA that are silenced and which is made by modification of thymine

jumping gene Popular name for a transposable element

junk DNA Defective selfish DNA that is of no use to the host cell it inhabits and which can no longer move or express its genes

kanamycin Antibiotic of the aminoglycoside family that inhibits protein synthesis

kappa particle Endosymbiotic bacteria (*Caedibacter*) that grow and divide inside killer *Paramecium*

karyotype The complete set of chromosomes found in the cells of a particular individual

killer *Paramecium* *Paramecium* that contains kappa particles in the cytoplasm, which it uses to kill other strains of *Paramecium*

kilobase ladder A set of DNA fragments that are exact multiples of 1000 bp and are often used as standards in gel electrophoresis

kinase Enzyme that attaches a phosphate group to another molecule

kinetic proofreading Proofreading of DNA that occurs during the process of DNA synthesis

kinetochore Protein structure that attaches to the DNA of the centromere during cell division and also binds the microtubules

kinetoplast Single giant mitochondrion found inside the cells of protozoans such as trypanosomes

kingdom Major subdivision of eukaryotic organisms, in particular the plant, fungus and animal kingdoms

Klenow polymerase DNA polymerase I from *E. coli* that lacks the 5′ to 3′ exonuclease domain

K_m See Michaelis constant

kuru A prion disease of cannibals

L- and D-forms The two isomeric forms of an optically active substance; also called L- and D-isomers

LacI protein Repressor that controls the *lac* operon

lactose permease (LacY) The transport protein for lactose

β-lactamase Enzyme that destroys antibiotics of the β-lactam class that includes penicillins and cephalosporins

***lacZ* gene** Gene encoding β-galactosidase; widely used as a reporter gene

lagging strand The new strand of DNA which is synthesized in short pieces during replication and then joined later

lambda (or λ) Specialized transducing phage of *Escherichia coli* that may insert its DNA into the bacterial chromosome

lambda attachment site (*att*λ) Recognition site on DNA used during integration of lambda DNA into *E. coli* chromosome

lambda left promoter (P_L) One of the promoters repressed by binding of the lambda repressor or cI protein

lambda repressor (cI protein) Repressor protein responsible for maintaining bacteriophage lambda in the lysogenic state

large subunit The larger of the two ribosomal subunits, 50S in bacteria, 60S in eukaryotes

lariat structure Branched, lariat-shaped segment of RNA generated by splicing out an intron

late genes Genes expressed later in virus infection and that mainly encode enzymes involved in virus particle assembly

latency State in which a virus replicates its genome in step with the host cell without making virus particles or destroying the host cell. Same as lysogeny, but generally used to describe animal viruses

lateral gene transfer Movement of genes sideways between unrelated organisms. Same as horizontal gene transfer

leader peptidase Enzyme that removes the leader sequence after protein export

leader peptide A short peptide coded by the leader region of certain genes controlled by the attenuation mechanism

leader region The region of an mRNA molecule in front of the structural genes, especially when involved in regulation by the attenuation mechanism

leading strand The new strand of DNA that is synthesized continuously during replication

leaky mutation Mutation where partial activity remains

leucine zipper One type of DNA-binding motif common in proteins

ligase Enzyme that joins up DNA fragments end to end

LINE Long interspersed element

LINE-1 (or L1) element A particular LINE found in many copies in the genome of humans and other mammals

linkage When two alleles are inherited together more often than would be expected by chance. Usually this is because they reside on the same DNA molecule (typically, on the same chromosome).

linkage group A group of alleles carried on the same DNA molecule (that is, on the same chromosome)

linking number (L) The sum of the superhelical turns (the writhe, W) plus the double helical turns (the twist, T)

lipoprotein Complex of protein plus lipid

L-isomer That one of a pair of optical isomers that rotates light in an anticlockwise direction

living cell A unit of life that possesses a genome made of DNA and sends genetic messages (RNA) from its genes (DNA) to its own ribosomes to make its own proteins with energy it generates itself

lock and key model Model of enzyme action in which the active site of an enzyme fits the substrate precisely

locus (plural, loci) A place or location on a chromosome; it may be a genuine gene or just any site with variations in the DNA sequence that can be detected, like RFLPs or VNTRs

long interspersed element (LINE) Long sequence found in multiple copies that makes up much of the moderately repetitive DNA of mammals

long terminal repeats (LTRs) Direct repeats of several hundred base pairs found at the ends of retroviruses and some other retro-elements

***luc* gene** Gene encoding luciferase from eukaryotes

luciferase Enzyme that emits light when provided with a substrate known as luciferin

luciferin Chemical substrate used by luciferase to emit light

Lumi-phos An artificial substrate that is split by alkaline phosphatase, releasing an unstable molecule that emits light

***lux* gene** Gene encoding luciferase from bacteria

Lyme disease Infection caused by *Borrelia burgdorferii* and transmitted by ticks

lysogen A cell containing a lysogenic virus

lysogeny Type of virus infection in which the virus becomes largely quiescent, makes no new virus particles and duplicates its genome in step with the host cell. Same as latency but used of bacterial viruses

lysosome Membrane bound organelle of eukaryotic cells that contains degradative enzymes

lysozyme An enzyme found in many bodily fluids that degrades the peptidoglycan of bacterial cell walls

lytic growth Type of infection in which a virus generates many virus particles and destroys the cell

M phase (or mitosis) Stage of the eukaryotic cell cycle in which cell division occurs

M13 Rod-shaped bacteriophage that infects *E. coli*, contains a circle of single stranded DNA, and is used to manufacture DNA for sequencing

macromolecule Large polymeric molecule; in living cells especially DNA, RNA, protein or polysaccharide

macronucleus Large somatic nucleus of ciliates that contains multiple copies of genes that are expressed

macronucleus-destined segment (MDS) Segments of DNA that remain during conversion of a ciliate micronucleus to macronucleus and are spliced to form uninterrupted genes that can be expressed

mad cow disease Same as **bovine spongiform encephalopathy (BSE)**, a prion disease of cattle

maintenance methylase Enzyme that adds a second methyl group to the other DNA strand of half-methylated sites

MALDI Matrix-assisted laser desorption-ionization. Type of mass spectrometry in which gas-phase ions are generated from a solid sample by a pulsed laser

male-specific phage Virus that only infects "male" bacteria, i.e. those bacteria carrying the F-plasmid

maltose-binding protein (MBP) Protein of *E. coli* that binds maltose during transport. MBP is often used in making fusion proteins

map unit A subdivision that is one hundredth of the length of the bacterial chromosome

Mariner elements A widespread family of conservative DNA-based transposons first found in *Drosophila*

mass spectrometry Technique for measuring the mass of molecular ions derived from volatilized molecules

***MAT* locus** Chromosomal locus in yeast that controls the mating type and exists as two alternative forms, *MATa* or *MATα*

mating factor Chemical messenger or pheromone that indicates the mating type and promotes sexual conjugation

mating types Equivalent of different sexes found in lower eukaryotes. They are structurally identical but biochemically distinct

matrix attachment region (MAR) Site on eukaryotic DNA that binds to proteins of the nuclear matrix or of the chromosomal scaffold—same as SAR sites

matrix-assisted laser desorption-ionization (MALDI) Type of mass spectrometry in which gas-phase ions are generated from a solid sample by a pulsed laser

maximum velocity (V_m or V_{max}) Velocity reached when all the active sites of an enzyme are filled with substrate

Mbp Megabase pairs or million base pairs

mechanical protein Protein that uses chemical energy to perform physical work

mediator A protein complex that transmits the signal from transcription factors to the RNA polymerase in eukaryotic cells

meiosis Formation of haploid gametes from diploid parent cells

melting When used of DNA, refers to its separation into two strands as a result of heating

melting temperature (Tm) The temperature at which the two strands of a DNA molecule are half unpaired

membrane A thin flexible structural layer made of protein and phospholipid that is found surrounding all living cells

membrane-bound organelle Organelle that is separated from the rest of the cytoplasm by membranes

Mendelian character Trait that is clear cut and discrete and can be unambiguously assigned to one category or another

Mendelian ratios Whole number ratios of inherited characters found as the result of a genetic cross

β-mercaptoethanol (BME) A small molecule with free sulfhydryl groups often used to break disulfide bonds in proteins

messenger RNA (mRNA) The class of RNA molecule that carries genetic information from the genes to the rest of the cell

metabolism The processes by which nutrient molecules are transported and transformed within the cell to release energy and to provide new cell material

metabolome The total complement of small molecules and metabolic intermediates of a cell or organism

metallothionein Protein that protects animal cells by binding toxic metals

methotrexate (or amethopterin) Anti-cancer drug that inhibits dihydrofolate reductase of animals

methylcytosine binding protein (MeCP) Type of protein in eukaryotes that recognizes methylated CG-islands

4-methylumbelliferyl phosphate An artificial substrate that is cleaved by alkaline phosphatase, releasing a fluorescent molecule

Michaelis constant (K_m) The substrate concentration that gives half maximal velocity in an enzyme reaction. It is an inverse measure of the affinity of the substrate for the active site

Michaelis-Menten equation Equation describing relationship between substrate concentration and the rate of an enzyme reaction

micro RNA (miRNA) Small regulatory RNA molecules of eukaryotic cells

2 micron plasmid See 2μ plasmid

micronucleus Small germline nucleus of ciliates whose genes are not expressed

minichromosomes Miniature chromosomes of 50–150 kbp found in trypanosomes that carry silent copies of the *VSG* gene

mini-satellite Another term for a VNTR (variable number tandem repeats)

mirrorlike palindrome Sequence of DNA that is the same when read forwards and backwards on the same strand. One type of palindrome

mismatch Wrong pairing of two bases in a double helix of DNA

mismatch repair system DNA repair system that recognizes mispaired bases and cuts out part of the DNA strand containing the wrong base

missense mutation Mutation in which a single codon is altered so that one amino acid in a protein is replaced with a different amino acid

mistranslation Errors made during translation

mitochondrion Membrane-bound organelle found in eukaryotic cells that produces energy by respiration

mitosis Division of eukaryotic cell into two daughter cells with identical sets of chromosomes

mobile DNA Segment of DNA that moves from site to site within or between other molecules of DNA

mobile genetic element A discrete segment of DNA that is able to change its location within larger DNA molecules by transposition or integration and excision

mobility shift assay Method for testing binding of a protein to DNA by measuring the change in mobility of DNA during gel electrophoresis. Same as bandshift assay or gel retardation

mobilizability Ability of a non-transferable plasmid to be moved from one host cell to another by a transferable plasmid

moderately repetitive sequence DNA sequences that exist in thousands of copies (but less than a hundred thousand)

modification enzyme Enzyme that binds to the DNA at the same recognition site as the corresponding restriction enzyme but methylates the DNA

modified base Nucleic acid base that is chemically altered after the nucleic acid has been synthesized

modifier gene Gene that modifies the expression of another gene

molecular beacon A fluorescent probe molecule that contains both a fluorophore and a quenching group and that fluoresces only when it binds to a specific DNA target sequence

molecular sewing Creation of a hybrid gene by joining segments from multiple sources using PCR

monocistronic mRNA mRNA carrying the information of a single cistron, that is a coding sequence for only a single protein

multimeric Formed of multiple subunits

multiple cloning site (MCS) A stretch of artificially synthesized DNA that contains cut sites for seven or eight widely used restriction enzymes. Same as polylinker

mutagen Any agent, including chemicals and radiation, that can cause mutations

mutant Organism carrying a mutated gene

mutation An alteration in the DNA (or RNA) that comprises the genetic information

mutator gene Gene whose mutation alters the mutation frequency of the organism, usually because it codes for a protein involved in DNA synthesis or repair

mutualism Form of symbiosis where both partners benefit

MyoD A eukaryotic transcription factor that takes part in muscle cell differentiation

N protein An anti-terminator protein used by bacteriophage lambda

nalidixic acid A quinolone antibiotic that inhibits DNA gyrase

nanopore detector Detector that allows a single strand of DNA through a molecular pore and records its characteristics as it passes through

30 nanometer fiber Chain of nucleosomes that is arranged helically, approximately 30 nm in diameter

negative control See negative regulation

negative feedback Form of negative regulation where the final product of a pathway inhibits the first enzyme in the pathway

negative or "minus" strand The non-coding strand of RNA or DNA

negative regulation Regulatory mode in which a repressor keeps a gene switched off until it is removed

negative supercoiling Supercoiling with a left handed or counterclockwise twist

neomycin phosphotransferase Enzyme that inactivates the antibiotics kanamyin and neomycin by adding a phosphate group

neomycin Antibiotic of the aminoglycoside family that inhibits protein synthesis

N-formyl-methionine or fMet Modified methionine used as the first amino acid during protein synthesis in bacteria

nick A break in the backbone of a DNA or RNA molecule (but where no bases are missing)

nick translation The removal of a short stretch of DNA or RNA, starting from a nick, and its replacement by newly made DNA

non-coding DNA DNA that does not code for proteins or functional RNA molecules

non-coding RNA RNA molecule that functions without being translated into protein; includes tRNA, rRNA, snRNA, snoRNA, scRNA, tmRNA and some regulatory RNA molecules

non-homologous end joining DNA repair system found in eukaryotes that mends double-stranded breaks

non-homologous recombination Recombination between two lengths of DNA that are largely unrelated. It involves specific proteins, that recognize particular sequences and form crossovers between them. Same as site-specific recombination

nonsense mutation Mutation due to changing the codon for an amino acid to a stop codon

norfloxacin A fluoroquinolone antibiotic that inhibits DNA gyrase

Northern blotting Hybridization technique in which a DNA probe binds to an RNA target molecule

novobiocin An antibiotic that inhibits type II topoisomerases, especially DNA gyrase, by binding to the B-subunit

***npt* gene** Gene for neomycin phosphotransferase. Provides resistance against the antibiotics kanamyin and neomycin

N-terminus The end of a polypeptide chain that is made first and has a free amino group; same as amino terminus

nuclear envelope Envelope consisting of two concentric membranes that surrounds the nucleus of eukaryotic cells

nuclear matrix A mesh of filamentous proteins found on the inside of the nuclear membrane and used in anchoring DNA

nuclear membranes Pair of concentric membranes comprising the nuclear envelope that separates off the nucleus from the rest of a eukaryotic cell

nuclear pore Pore in nuclear membrane that allows proteins, RNA and other molecules into and out of the nucleus

nuclease Enzyme that cuts or degrades nucleic acids

nucleic acid Polymer made of nucleotides that carries genetic information

nucleocapsid Inner protein shell of a virus particle that contains the nucleic acid

nucleolar organizer Chromosomal region associated with the nucleolus; actually a cluster of rRNA genes

nucleolus Region of the nucleus where ribosomal RNA is made and processed

nucleomorph Degenerate remains of the nucleus of a symbiotic eukaryote that was incorporated by secondary endosymbiosis into another eukaryotic cell

nucleoprotein Complex of protein plus nucleic acid

nucleoside The union of a purine or pyrimidine base with a pentose sugar

nucleosome Subunit of a eukaryotic chromosome consisting of DNA coiled around histone proteins

nucleotide Monomer or subunit of a nucleic acid, consisting of a pentose sugar plus a base plus a phosphate group

nucleus An internal compartment surrounded by the nuclear membrane and containing the chromosomes. Only the cells of higher organisms have nuclei.

null allele Mutant version of a gene which completely lacks any activity

null mutation Mutation that totally inactivates a gene

Nus proteins A family of bacterial proteins involved in termination of transcription and/or in anti-termination

NusA protein A bacterial protein involved in termination of transcription

***nut* site** The site on DNA where the anti-terminator N-protein binds

oil drop model Model of protein structure in which the hydrophobic groups cluster together on the inside away from the water

Okazaki fragments The short pieces of DNA that make up the lagging strand

oligo(dT) Stretch of single-stranded DNA consisting solely of dT or deoxythymidine residues

oligo(U) Stretch of single-stranded RNA consisting solely of U or uridine residues

oligonucleotide array detector Chip used to simultaneously detect and identify many short DNA fragments by DNA-DNA hybridization. Also known as DNA array or DNA chip

***o*-nitrophenyl phosphate** Artificial substrate that is split by alkaline phosphatase, releasing yellow *o*-nitrophenol

ONPG (*o*-nitrophenyl galactoside) Artificial substrate that is split by β-galactosidase, releasing yellow *o*-nitrophenol

open circle Circular DNA with one strand nicked and hence with no supercoiling

open reading frame (ORF) Sequence of bases (either in DNA or RNA) that can be translated (at least in theory) to give a protein

operator Site on DNA to which a repressor protein binds

operon A cluster of prokaryotic genes that are transcribed together to give a single mRNA (i.e. polycistronic mRNA)

opportunistic infection Any disease caused by an infectious agent that is only capable of invading a host with an impaired immune system

optical isomers Isomers where the molecules differ only in their 3D arrangement and consequently affect the rotation of polarized light

ORF See open reading frame

organelle Subcellular structure that carries out a specific task. Membrane-bound organelles are separated from the rest of the cytoplasm by membranes but other organelles, such as the ribosome, are not.

origin of chromosome (*oriC*) Origin of replication of a chromosome

origin of replication (*ori*) Site on a chromosome or any other DNA molecule where replication begins

orthologous genes Homologous genes that are found in separate species and which diverged when the organisms containing them diverged

***ortho*-nitrophenyl galactoside (ONPG)** Artificial substrate for β-galactosidase that yields a yellow color upon cleavage

outer membrane Extra membrane lying outside the cell wall in gram-negative but not gram-positive bacteria

overlap primer PCR primer that matches small regions of two different gene segments and is used in joining segments of DNA from different sources

P (peptide) site Binding site on the ribosome for the tRNA that is holding the growing polypeptide chain

P1 artificial chromosome (PAC) Single copy vector based on the P1-phage/plasmid of *E. coli* that can carry very long inserts of DNA

P1 Generalized transducing phage of *Escherichia coli*

P22 Generalized transducing phage of *Salmonella*

PAGE Polyacrylamide gel electrophoresis. Technique for separating proteins by electrophoresis on a gel made from polyacrylamide

palindrome A sequence that reads the same backwards as forwards

palindromic Reading the same backwards as forwards

paralogous genes Homologous genes that are located within the same organism due to gene duplication

Paramecium A type of free-living protozoan that feeds on bacteria

parasite An organism or genetic entity that replicates at the expense of another creature

parasitism Form of symbiosis where one organism lives at the expense of the other

partial diploidy Situation in which a cell is diploid for only some of its genes

partial dominance When a functional allele only partly masks a defective allele

patch recombinant DNA double helix with a short patch of heteroduplex due to transient formation of a crossover

pathogen Parasite that seriously incapacitates or kills its host

pathogenic Disease causing

pathogenicity island Region of bacterial chromosome containing clustered genes for virulence

PCNA protein The sliding clamp for the DNA polymerase of eukaryotic cells (PCNA = proliferating cell nuclear antigen)

PCR See polymerase chain reaction

PCR machine See thermocycler

PCR primers Short pieces of single-stranded DNA that match the sequences at either end of the target DNA segment and which are needed to initiate DNA synthesis in PCR

penetrance Variability in the phenotypic expression of an allele

penicillin Group of antibiotics of the β-lactam type that inhibit cross-linking of the peptidoglycan of the bacterial cell wall. Originally isolated from a mold called *Penicillium*, which grows on bread producing a blue layer of fungus

pentose A five carbon sugar, such as ribose or deoxyribose

peptide bond Type of chemical linkage holding amino acids together in a polypeptide chain

peptide nucleic acid (PNA) Artificial analog of nucleic acids with a polypeptide backbone

peptidoglycan Polymer that makes up eubacterial cell walls; consists of long chains of sugar derivatives, cross-linked at intervals with short chains of amino acids

peptidyl transferase Enzyme activity on the ribosome that makes peptide bonds; actually 23S rRNA (bacterial) or 28S rRNA (eukaryotic)

permease A protein that transports nutrients or other molecules across a membrane

phage display library Collection of a large number of modified phages displaying different peptide or protein sequences

phage display Fusion of a protein or peptide to the coat protein of a bacteriophage whose genome also carries the cloned gene encoding the protein. The protein is displayed on the outside of the virus particle and the corresponding gene is carried on the inside

phage Short for bacteriophage, a virus that infects bacteria

phase variation Reversible inversion of a segment of DNA leading to differences in gene expression

phenol extraction Technique for removing protein from nucleic acids by dissolving the protein in phenol

phenotype The visible or measurable effect of the genotype

pheromone Hormone or messenger molecule that travels between organisms, rather than circulating within the same organism

***phoA* gene** Gene encoding alkaline phosphatase; widely used as a reporter gene

phosphatase An enzyme that removes phosphate groups

phosphate group Group of four oxygen atoms surrounding a central phosphorus atom found in the backbone of DNA and RNA

phosphodiester The linkage between nucleotides in a nucleic acid that consists of a central phosphate group esterified to sugar hydroxyl groups on either side

phospholipid A hydrophobic molecule found making up cell membranes and consisting of a soluble head group and two fatty acids both linked to glycerol phosphate

phosphoramidate Phosphate derivative in which the phosphate group is attached to an amino group

phosphoramidite method Method for artificial synthesis of DNA that utilizes the reactive phosphoramidite group to make linkages between nucleotides

phosphorothioate A phosphate group in which one of the four oxygen atoms around the central phosphorus is replaced by sulfur

phylum (plural phyla) Major groups into which animals are divided, roughly equivalent in rank to the divisions of plants or bacteria

plaque (When referring to viruses) A clear zone caused by virus destruction in a layer of cultured cells or a lawn of bacteria

plasmid Self-replicating genetic elements that are sometimes found in both prokaryotic and eukaryotic cells. They are not chromosomes nor part of the host cell's permanent genome. Most plasmids are circular molecules of double stranded DNA although rare linear plasmids and RNA plasmids are known

2μ plasmid (or 2μ circle) A multicopy plasmid found in the yeast, *Saccharomyces cerevisiae*, whose derivatives are widely used as vectors

Plasmodium The malaria parasite, a protozoan belonging to the Apicomplexa

plastid Any organelle that is genetically equivalent to a chloroplast, whether functional in photosynthesis or not

ploidy The number of sets of chromosomes possessed by an organism

PNA clamp Two identical PNA strands that are joined by a flexible linker and are intended to form a triple helix with a complementary strand of DNA or RNA

point mutation Mutation that affects a single base pair

polarity When the insertion of a segment of DNA affects the expression of downstream genes, usually by preventing their transcription

poly(A) polymerase Enzyme that adds the poly(A) tail to the end of mRNA

poly(A) tail A stretch of multiple adenosine residues found at the 3′-end of mRNA

poly(A)-binding protein (PABP) Protein that binds to mRNA via its poly(A) tail

polyacrylamide gel electrophoresis (PAGE) Technique for separating proteins by electrophoresis on a gel made from polyacrylamide

polyacrylamide Polymer used in separation of proteins or very small nucleic acid molecules by gel electrophoresis

polyadenylation complex Protein complex that adds the poly(A) tail to eukaryotic mRNA

polycistronic mRNA mRNA carrying multiple coding sequences (cistrons) that may be translated to give several different protein molecules; only found in prokaryotic (bacterial) cells

polyhistidine tag (His tag) Six tandem histidine residues that are fused to proteins so allowing purification by binding to nickel ions that are attached to a solid support

polylinker A stretch of artificially synthesized DNA that contains cut sites for seven or eight widely used restriction enzymes. Same as multiple cloning site (MCS)

polymerase Enzyme that synthesizes nucleic acids

polymerase chain reaction (PCR) Amplification of a DNA sequence by repeated cycles of strand separation and replication

polymerase eta (η) A repair DNA polymerase in animals that can replicate past thymine dimers

polymerase I (Pol I) See DNA polymerase I

polymerase III (Pol III) See DNA polymerase III

polymorphism A difference in DNA sequence between two related individual organisms

polypeptide chain A polymer that consists of amino acids

polyphosphate Compound consisting of multiple phosphate groups linked by high energy phosphate bonds

polyprotein A long polypeptide that is cut up to generate several smaller proteins

polysome Group of ribosomes bound to and translating the same mRNA

positive control See positive regulation

positive or "plus" strand The coding strand of RNA or DNA

positive regulation Control by an activator that promotes gene expression when it binds

post-transcriptional gene silencing (PTGS) Plant version of the RNA interference response to double-stranded RNA that results in the degradation of mRNA or other RNA transcripts homologous to the inducing dsRNA

post-translational modification Modification of a protein or its constituent amino acids after translation is finished

potential intrastrand triplex (PIT) Stretch of DNA that might be expected from its sequence to form H-type triplex DNA

poxviruses A family of large and complex dsDNA animal viruses with 150 to 200 genes

prey The fusion between the activator domain of a transcriptional activator protein and another protein as used in two-hybrid screening

PriA Protein of the primosome that helps primase bind

Pribnow box Another name for the –10 region of the bacterial promoter

primary atmosphere The original atmosphere of the earth consisting mostly of hydrogen and helium

primary endosymbiosis Original uptake of prokaryotes by the ancestral eukaryotic cell, giving rise to mitochondria and chloroplasts

primary structure The linear order in which the subunits of a polymer are arranged

primary transcript The original RNA molecule obtained by transcription from a DNA template, before any processing or modification has occurred

primase Enzyme that starts a new strand of DNA by making an RNA primer

primer Short segment of nucleic acid that binds to the template strand and allows synthesis of a new chain of DNA to get started. RNA primers are used in cells and DNA primers are used in PCR

primer extension Method to locate the 5′ start site of transcription by using reverse transcriptase to extend a primer bound to mRNA so locating the 5′-end of the transcript

primer walking Approach to sequencing a long cloned DNA molecule by using successive primers located at stages along the molecule

primitive soup Mixture of random molecules, including amino acids, sugars, and nucleic acid bases, found in solution on the primeval earth

primosome Cluster of proteins (including PriA and primase) that synthesizes a new RNA primer during DNA replication

prion A protein that can mis-fold into an alternative pathological form that then promotes its own formation auto-catalytically. Misfolded prion proteins are responsible for the neurodegenerative diseases known as spongiform encephalopathies that include scrapie, kuru and BSE

prion protein (PrP) The prion protein found in the nervous tissue of mammals and whose misfolded form is responsible for prion diseases

***Prn-p* gene** Gene encoding the prion protein (PrP)

probe molecule Molecule that is tagged in some way (usually radioactive or fluorescent) and is used to bind to and detect another molecule

processed pseudogene Pseudogene lacking introns because it was reverse transcribed from messenger RNA by reverse transcriptase

prokaryote Lower organism, such as a bacterium, with a primitive type of cell containing a single chromosome and having no nucleus

promoter Region of DNA in front of a gene that binds RNA polymerase and so promotes gene expression

proofreading Process that checks whether the correct nucleotide has been inserted into new DNA. Usually refers to DNA polymerase checking whether it has inserted the correct base

prophage Bacteriophage genome that is integrated into the DNA of the bacterial host cell

prophase Stage of mitosis during which condensed chromosomes become visible, the centrioles divide and the nuclear membrane dissolves

prosthetic group Extra chemical group, which is not part of the polypeptide chain, covalently attached to a protein

protease Same as proteinase; an enzyme that degrades proteins

proteasome Protein assembly found in eukaryotic cells that degrades proteins

protein Polymer made from amino acids; may consist of several polypeptide chains

protein A Antibody binding protein from *Staphylococcus* that is often used in making fusion proteins

protein interactome The total of all the protein-protein interactions in a particular cell or organism

protein kinase An enzyme that adds phosphate groups to another protein

protein microarray Microarray of immobilized proteins used for proteome analysis and normally screened by fluorescent or radioactive labeling

protein primer Protein used instead of RNA as a primer for DNA synthesis in some bacteria and viruses

proteinase Same as protease; an enzyme that degrades proteins

proteinoid Artificially synthesized polypeptide containing randomly linked amino acids

proteolipid A type of lipoprotein that is extremely hydrophobic and found in the interior of membranes

proteome The total set of proteins encoded by a genome or the total protein complement of an organism

protomer A single polymer chain that is itself a subunit for a higher level of assembly

protostomes Grouping of animal phyla that includes arthropods, annelids, molluscs, flatworms etc

provirus Virus genome that is integrated into the host cell DNA

PrP (prion protein) The prion protein found in the nervous tissue of mammals and whose misfolded form is responsible for prion diseases

PrP^{C} (cellular PrP) The healthy, normal form of the prion protein

PrP^{Sc} (scrapie PrP) The pathological form of the prion protein, sometimes known as the scrapie agent

pseudogene Defective copy of a genuine gene

pseudouridine An isomer of uridine that is introduced into some RNA molecules by post-transcriptional modification

pulsed field gel electrophoresis (PFGE) Type of gel electrophoresis used for analysis of very large DNA molecules and which uses an electric field of "pulses" delivered from a hexagonal array of electrodes

purine Type of nitrogenous base with a double ring found in DNA and RNA

pyrimidine Type of nitrogenous base with a single ring found in DNA and RNA

quasi-species A set of closely related sequences (especially virus genomes) whose individual members vary from consensus by frequent errors or mutations

quaternary structure Aggregation of more than one polymer chain to form a final structure

quencher Molecule that prevents fluorescence by binding to the fluorophore and absorbing its activation energy

quinolone antibiotics A family of antibiotics, including nalidixic acid, norfloxacin and ciprofloxacin, that inhibit DNA gyrase and other type II topoisomerases by binding to the A-subunit

RACE See rapid amplification of cDNA ends

racemase An enzyme that interconverts the D- and L-isomers of an optically active substance

racemic mixture Mixture of equal amounts of both D- and L-isomers of an optically active substance

Rad proteins Group of proteins involved in recombination and repair of DNA damage in yeast and animal cells. Rad51 corresponds to the prokaryotic RecA protein

radiation hybrid A cell (usually from a rodent) that contains fragments of chromosomes (generated by irradiation) from another species

radical replacement Replacement of an amino acid with another that has different chemical and physical properties

radioisotope Radioactive form of an element

random coil Region of polypeptide chain lacking secondary structure

randomly amplified polymorphic DNA (RAPD) Method for testing genetic relatedness using PCR to amplify arbitrarily chosen sequences

rapid amplification of cDNA ends (RACE) RT-PCR-based technique that gener-ates the complete 5′ or 3′ end of a cDNA starting from a partial sequence

R-bodies Toxic proteins that form crystals inside the kappa particles of killer *Paramecium*

reading frame One of three alternative ways of dividing up a sequence of bases in DNA or RNA into codons

RecA protein Protein involved in recombination and repair of DNA in *E. coli* that binds single-stranded DNA

recessive allele The allele whose properties are not observed because they are masked by the dominant allele

recipient cell Cell that receives DNA from another cell

recombination Exchange of genetic information between chromosomes or other molecules of DNA

–10 region Region of bacterial promoter 10 bases back from the start of transcription that is recognized by RNA polymerase

–35 region Region of bacterial promoter 35 bases back from the start of transcription that is recognized by RNA polymerase

regulatory protein A protein that regulates the expression of a gene or the activity of another protein

regulatory region DNA sequence in front of a gene, used for regulation rather than to encode a protein

regulon A set of genes or operons that are regulated by the same regulatory protein even though they are at different locations on the chromosome

release factor Protein that recognizes a stop codon and brings about the release of a finished polypeptide chain from the ribosome

renaturation Re-annealing of single-stranded DNA or refolding of a denatured protein to give the original natural 3-D structure

repeated sequences DNA sequences that exist in multiple copies

repetitive sequences Same as repeated sequences

replication Duplication of the genomic DNA (or RNA), typically prior to cell division or virus replication

replication bubble (replication eye) Bulge where DNA is in the process of replication

replication factor C (RFC) Eukaryotic protein that binds to initiator DNA and loads DNA polymerase δ plus its sliding clamp onto the DNA

replication fork Region where the enzymes replicating a DNA molecule are bound to untwisted, single stranded DNA

replicative form (RF) Double-stranded form of the genome of a single-stranded DNA (or RNA) virus. The RF first replicates itself and is then used to generate the ssDNA (or ssRNA) to pack into the virus particles

replicative transposition Type of transposition in which two copies of the transposon are generated, one in the original site and another at a new location

replicon Molecule of DNA or RNA that contains an origin of replication and can self-replicate

replisome Assemblage of proteins (including primase, DNA polymerase, helicase, SSB protein) that replicates DNA

reporter gene Gene that is used in genetic analysis because its product is convenient to assay or easy to detect

reporter protein A protein that is easy to detect and gives a signal that can be used to reveal its location and/or indicate levels of gene expression

repressor Regulatory protein that prevents a gene from being transcribed

resolution Cleavage of the junction where two DNA molecules are fused together so releasing two separate DNA molecules. Refers to the breakdown both of crossovers formed during recombination and of cointegrates formed by transposition

resolvase An enzyme that cuts apart (i.e., "resolves") a cointegrate releasing two separate molecules of DNA

restriction enzyme Type of endonuclease that cuts double stranded DNA at a specific sequence of bases, the recognition site

restriction fragment length polymorphism (RFLP) A difference in restriction sites between two related DNA molecules that results in production of restriction fragments of different lengths

restriction map A diagram showing the location of restriction enzyme cut sites on a segment of DNA

retro-element A genetic element that uses reverse transcriptase to convert the RNA form of its genome to a DNA copy

retron Genetic element found in bacteria that encodes reverse transcriptase and uses it to make a bizarre RNA/DNA hybrid molecule

retroposon Short for retrotransposon

retro-pseudogene Another name for a processed pseudogene

retrotransposon A transposable element that uses reverse transcriptase to convert the RNA form of its genome to a DNA copy

retroviruses A family of animal viruses with single-stranded RNA inside two protein shells surrounded by an outer envelope. Once inside the host cell they use reverse transcriptase to convert the RNA version of the genome to a DNA copy

reverse transcriptase An enzyme that uses single-stranded RNA as a template for making double-stranded DNA

reverse transcriptase PCR (RT-PCR) Variant of PCR that allows genes to be amplified and cloned as intron-free DNA copies by starting with mRNA and using reverse transcriptase

reverse transcription The process in which single-stranded RNA is used as a template for making double-stranded DNA

reverse turn Region of polypeptide chain that turns around and goes back in the same direction

reversion Alteration of DNA that reverses the effects of a prior mutation

R-group Any unspecified chemical group; in particular the side chain of an amino acid

Rho (ρ) protein Protein factor needed for successful termination at certain transcriptional terminators

Rho-dependent terminator Transcriptional terminator that depends on Rho protein

Rho-independent terminator Transcriptional terminator that does not need Rho protein

ribonuclease A nuclease that cuts RNA

ribonuclease (RNase) Enzyme that cuts or degrades RNA

ribonuclease III A ribonuclease of bacteria whose main function is processing rRNA and tRNA precursors

ribonuclease H (RNase H) Enzyme that degrades the RNA strand of DNA:RNA hybrid double helixes. In bacteria it removes the major portion of RNA primers used to initiate DNA synthesis.

ribonuclease P A ribonuclease involved in processing tRNA in bacteria that consists of an RNA ribozyme plus an accessory protein

ribonucleic acid (RNA) Nucleic acid that differs from DNA in having ribose in place of deoxyribose and having uracil in place of thymine

ribonucleoside A nucleoside whose sugar is ribose (not deoxyribose)

ribonucleotide reductase Enzyme that reduces ribonucleotides to deoxyribonucleotides

ribonucleotide A nucleotide whose sugar is ribose (not deoxyribose)

ribose The 5-carbon sugar found in RNA

ribosomal RNA (rRNA) Class of RNA molecule that makes up part of the structure of a ribosome

70S ribosome Type of ribosome found in bacterial cells

80S ribosome Type of ribosome found in cytoplasm of eukaryotic cells

ribosome The cell's machinery for making proteins

ribosome binding site (RBS) Same as Shine-Dalgarno sequence; sequence close to the front of mRNA that is recognized by the ribosome; only found in prokaryotic cells

ribosome modulation factor (RMF) Protein that inactivates surplus ribosomes during slow growth or stationary phase in bacteria

ribosome recycling factor (RRF) Protein that dissociates the ribosomal subunits after a polypeptide chain has been finished and released

riboswitch Domain of messenger RNA that directly senses a signal and controls translation by alternating between two structures

ribozyme An RNA enzyme, that is an RNA molecule with catalytic activity

rickettsia Type of degenerate bacterium that is an obligate parasite and infects the cells of higher organisms

right-handed helix In a right-handed helix, as the observer looks down the helix axis (in either direction), each strand turns clockwise as it moves away from the observer

R-loop analysis Hybridization of the DNA copy of a gene to the corresponding mRNA that results in the appearance of loops, which represent the intervening sequences in the DNA that have no partners in the mRNA

7SL RNA Non-coding RNA that forms part of the machinery for protein export across intracellular membranes in eukaryotic cells

RNA editing Changing the coding sequence of an RNA molecule after transcription by altering, adding or removing bases

RNA interference Response that is triggered by the presence of double-stranded RNA and results in the degradation of mRNA or other RNA transcripts homologous to the inducing dsRNA

RNA or ribonucleic acid Nucleic acid that differs from DNA in having ribose in place of deoxyribose

RNA polymerase Enzyme that synthesizes RNA

RNA polymerase I Eukaryotic RNA polymerase that transcribes the genes for the large ribosomal RNAs

RNA polymerase II Eukaryotic RNA polymerase that transcribes the genes encoding proteins

RNA polymerase III Eukaryotic RNA polymerase that transcribes the genes for 5S ribosomal RNA and transfer RNA

RNA primer Short segment of RNA used to initiate synthesis of a new strand of DNA during replication

RNA replicase Special RNA polymerase used by RNA viruses to replicate their RNA genomes

RNA virus A virus whose genome consists of RNA

RNA world The hypothetical stage of early life in which RNA encoded genetic information and carried out enzyme reactions without the need for either DNA or protein

RNA-dependent RNA polymerase (RdRP) RNA polymerase that uses RNA as a template and is involved in the amplification of the RNAi response

RNA-induced silencing complex (RISC) Protein complex induced by siRNA that degrades single-stranded RNA corresponding in sequence to the siRNA

rolling circle amplification technology (RCAT) Method based on rolling circle replication that uses DNA polymerase to amplify target DNA at normal temperatures

rolling circle replication Mechanism of replicating double stranded circular DNA that starts by nicking and unrolling one strand and using the other, still circular, strand as a template for DNA synthesis. Used by some plasmids and viruses

R-plasmid or R-factor Plasmid that carries genes for antibiotic resistance

Rubisco (ribulose bisphosphate carboxylase) A critical enzyme in the fixation of carbon dioxide during photosynthesis

S1 nuclease Endonuclease from *Aspergillus oryzae* that cleaves single-stranded RNA or DNA but does not cut double-stranded nucleic acids

S1 nuclease mapping Method using S1 nuclease to locate the 5′-end or 3′-end of a transcript

satellite DNA Highly repetitive DNA of eukaryotic cells that is found as long clusters of tandem repeats and is permanently coiled tightly into heterochromatin

satellite RNA Parasitic RNA molecule that requires a helper virus for replication and capsid formation

satellite virus A defective virus that needs an unrelated helper virus to infect the same host cell in order to provide essential functions

saturated (Referring to enzymes) When all the active sites are filled with substrate and the enzyme cannot work any faster

scaffold attachment region (SAR) Site on eukaryotic DNA that binds to proteins of the chromosomal scaffold or of the nuclear matrix—same as MAR sites

scintillant Molecule that emits pulses of light when hit by a particle of radioactivity

scintillation counter Machine that detects and counts pulses of light

scintillation counting Detection and counting of individual microscopic pulses of light

Scorpion primer DNA primer joined to a molecular beacon by an inert linker. When the probe sequence binds target DNA, the quencher and fluorophore are separated allowing fluorescence

scrapie An infectious disease of sheep that causes degeneration of the brain and is caused by mis-folded prion proteins

SECIS element See selenocyteine insertion sequence

second messenger Chemical messenger that is made inside a cell in response to a message recieved from outside the cell

secondary atmosphere The atmosphere of the earth after the light gases were lost and resulting mostly from volcanic out-gassing. It contained reduced gases but no oxygen

secondary endosymbiosis Uptake by an ancestral eukaryotic cell of another single-celled eukaryote, usually an alga, thus providing chloroplasts at second hand

secondary structure Initial folding up of a polymer due to hydrogen bonding

second-site revertant Revertant in which the change in the DNA, which suppresses the effect of the mutation, is at a different site to the original mutation

segregation Replication of a hybrid DNA molecule (whose two strands differ in sequence) to give two separate DNA molecules, each with a different sequence

selenocysteine (Sec) Amino acid resembling cysteine but containing selenium instead of sulfur

selenocyteine insertion sequence (SECIS element) Recognition sequence that signals for insertion of selenocysteine at a UGA stop codon

self-assembly The spontaneous assembly of a biological structure from its subunits

selfish DNA A sequence of DNA that manages to replicate but which is of no use to the host cell it inhabits

self-splicing Splicing out of an intron by the ribozyme activity of the RNA molecule itself without the requirement for a separate protein enzyme

semi-conservative replication Mode of DNA replication in which each daughter molecule gets one of the two original strands and one new complementary strand

sense RNA RNA (such as mRNA) that has been made using the non-coding strand of DNA as a template. Equivalent to plus or coding strand RNA

sensor kinase A protein that phosphorylates itself when it senses a specific signal (often an environmental stimulus, but sometimes an internal signal)

septum Cross-wall that separates two new bacterial cells after division

Sequenase® Genetically modified DNA polymerase from bacteriophage T7 used for sequencing DNA

sequence tagged site (STS) A short sequence (usually 100–500 bp) that is unique within the genome and can be easily detected, usually by PCR

sequestration protein (SeqA) Protein that binds the origin of replication, thereby delaying its methylation

serial analysis of gene expression (SAGE) Method to monitor level of multiple mRNA molecules by sequencing a DNA concatemer that contains many serially linked sequence tags derived from the mRNAs

sex chromosome A chromosome involved in determining the sex of an individual

sex pilus Protein filament made by donor bacteria that binds to a suitable recipient and draws the two cells together

sex-linked A gene is sex-linked when it is carried on one of the sex chromosomes

sexual reproduction Form of reproduction that involves reshuffling of the genes between two individuals

Shine-Dalgarno (S-D) sequence Same as RBS; sequence close to the front of mRNA that is recognized by the ribosome; only found in prokaryotic cells

short interfering RNA (siRNA) Double-stranded RNA molecules of 21–22 nucleotides involved in triggering RNA interference in eukaryotes

short interspersed element (SINE) Short repeated sequence that makes up a significant fraction of the moderately or highly repetitive DNA of mammals

shotgun sequencing Approach in which the genome is broken into many random short fragments for sequencing. The complete genome sequence is then assembled by computerized searching for overlaps between individual sequences

shuttle vector A vector that can survive in and be moved between more than one type of host cell

sigma subunit Subunit of bacterial RNA polymerase that recognizes and binds to the promoter sequence

signal molecule Molecule that exerts a regulatory effect by binding to a regulatory protein

signal sequence Short, largely hydrophobic sequence of amino acids at the front of a protein that label it for export

silencing In genetic terminology, refers to switching off genes in a relatively nonspecific manner

silent mutation An alteration in the DNA sequence that has no effect on the phenotype

simian virus 40 (SV40) A small, spherical dsDNA virus that causes cancer in monkeys by inserting its DNA into the host chromosome

simple sequence length polymorphism (SSLP) Any DNA region consisting of tandem repeats that vary in number from individual to individual, including VNTRs, microsatellites, and other tandem repeats

SINE Short interspersed element

single nucleotide polymorphism (SNP) A difference in DNA sequence of a single base change between two individuals

single strand binding protein (SSB protein) A protein that keeps separated strands of DNA apart

site-directed mutagenesis Deliberate alteration of a specific DNA sequence by any artificial technique

site-specific recombination Recombination between two lengths of DNA that are largely unrelated. It involves specific proteins that recognize particular sequences and form crossovers between them. Same as non-homologous recombination

Slicer Ribonuclease activity of the RISC complex

sliding clamp Subunit of DNA polymerase that encircles the DNA, thereby holding the core enzyme onto the DNA

small cytoplasmic RNA (scRNA) Small RNA molecules of varied function found in the cytoplasm of eukaryotic cells

small nuclear ribonucleoprotein (snRNP) Complex of snRNA plus protein

small nuclear RNA (snRNA) Small RNA molecule that is found only in the nucleus of eukaryotes where it oversees the splicing of mRNA

small nucleolar RNA (snoRNA) Small RNA molecules that are involved in ribosomal RNA base modification in the nucleolus of eukaryotic cells

small subunit The smaller of the two ribosomal subunits, 30S in bacteria, 40S in eukaryotes

snurp snRNP or small nuclear ribonucleoprotein

sodium dodecyl sulfate (SDS) A detergent widely used to denature and solubilize proteins before separation by electrophoresis

somatic cell Cell making up the body, as opposed to the germline

SOS system An error-prone repair system of bacteria that responds to severe DNA damage

Southern blotting A method to detect single stranded DNA that has been transferred to nylon paper by using a probe that binds DNA

South-Western blotting Detection technique in which a DNA probe binds to a protein target molecule

specialized transduction Type of transduction where certain regions of the bacterial DNA are carried preferentially

species A group of closely related organisms with a relatively recent common ancestor. Among animals, species are populations that breed among themselves but not with individuals of populations. No satisfactory definition exists for bacteria or other organisms that do not practice sexual reproduction.

specific regulation Regulation that applies to a single gene or operon or to a very small number of related genes

S-phase Stage in the eukaryotic cell cycle in which chromosomes are duplicated

3′ splice site Recognition site for splicing at the downstream or 3′-end of the intron

5′ splice site Recognition site for splicing at the upstream or 5′-end of the intron

spliceosome Complex of proteins and small nuclear RNA molecules that removes introns during the processing of messenger RNA

splicing Removal of intervening sequences and rejoining the ends of a molecule; usually refers to removal of introns from RNA

spontaneous mutation Mutation that occurs "naturally" without the help of mutagenic chemicals or radiation

spore A cell specialized for survival under adverse conditions and/or designed for distribution

start codon The special AUG codon that signals the start of a protein

stem and loop Structure made by folding an inverted repeat sequence

steroid receptor Protein that binds steroid hormones

sticky ends Ends of a double-stranded DNA molecule that have unpaired single-stranded overhangs, generated by a staggered cut

stop codon Codon that signals the end of a protein

streptavidin Protein from *Streptomyces* that binds biotin extremely tightly and specifically. Used in detection procedures for molecules labeled with biotin

streptomycin Antibiotic of the aminoglycoside family that inhibits protein synthesis

structural gene Sequence of DNA (or RNA) that codes for a protein or for an untranslated RNA molecule

structural protein A protein that forms part of a cellular structure

substrate Molecule that binds to an enzyme and is the target of enzyme action

subtractive hybridization Technique used to remove unwanted DNA or RNA by hybridization so leaving behind the DNA or RNA molecule of interest

30S subunit Small subunit of a 70S ribosome

40S subunit Small subunit of an 80S ribosome

50S subunit Large subunit of a 70S ribosome

60S subunit Large subunit of an 80S ribosome

subviral agent Infectious agents that are more primitive than viruses and encode fewer of their own functions

sulfhydryl group -SH; Chemical group of sulfur and hydrogen

sulfonamide Antibiotic that inhibits the synthesis of the folate cofactor

sulfonamides Synthetic antibiotics that are analogs of *p*-aminobenzoic acid, a precursor of the vitamin folic acid. Sulfonamides inhibit dihydropteroate synthetase

supercoiling Higher level coiling of DNA that is already a double helix

suppressor mutation A mutation that restores function to a defective gene by suppressing the effect of a previous mutation

suppressor tRNA A mutant tRNA that recognizes a stop codon and can insert an amino acid when it reads a stop codon on the mRNA

S-value The sedimentation coefficient is the velocity of sedimentation divided by the centrifugal field. It is dependent on mass and is measured in Svedberg units

Swi/Snf ("switch sniff") complex Larger type of chromatin remodeling complex

SYBR® Green I A DNA-binding fluorescent dye that binds only to double-stranded DNA and becomes fluorescent only when bound

symbiosis Association of two living organisms that interact

symbiotic theory Theory that the organelles of eukaryotic cells are derived from symbiotic prokaryotes

T4 ligase Type of DNA ligase from bacteriophage T4 and which is capable of ligating blunt ends

TA cloning Procedure that uses Taq polymerase to generate single 3′-A overhangs on the ends of DNA segments that are used to clone DNA into a vector with matching 3′-T overhangs

TA cloning vector Vector with single 3′-T overhangs (in its linearized form) that is used to clone DNA segments with single 3′-A overhangs generated by Taq polymerase

tail specific protease Enzyme that destroys mis-made proteins by degrading them tail first, i.e., from the carboxyl end

tandem duplication Mutation in which a segment of DNA is duplicated and the second copy remains next to the first

tandem mass spectrometry (MS/MS) Two successive rounds of mass spectrometry in which a parent ion is first isolated and then fragmented into daughter ions for more detailed analysis

tandem repeats Repeated sequences of DNA (or RNA) that lie next to each other

Taq polymerase Heat resistant DNA polymerase from *Thermus aquaticus* that is used for PCR

TaqMan® probe Fluorescent probe consisting of two fluorophores linked by a DNA probe sequence. Fluorescence increases only after the fluorophores are separated by degradation of the linking DNA

target DNA DNA that is the target for binding by a probe during hybridization or the target for amplification by PCR

target sequence a) Sequence on host DNA molecule into which a transposon inserts itself; b) Sequence within the original DNA template that is amplified in a PCR reaction

TATA binding protein (TBP) Transcription factor that recognizes the TATA box

TATA box Binding site for a transcription factor that guides RNA polymerase II to the promoter in eukaryotes

TATA box factor Another name for TATA binding protein

tautomerization Alternation of a molecule, in particular a base of a nucleic acid, between two different isomeric structures

Tc1 element Transposon *Caenorhabditis* 1. A transposon of the mariner family found in the nematode *Caenorhabditis*

T-cell Type of white blood cell belonging to the immune system

T-DNA (tumor-DNA) Region of the Ti-plasmid that is transferred into the plant cell nucleus

telomerase Enzyme that adds DNA to the telomere of a eukaryotic chromosome

telomere Specific repetitive sequence of DNA found at the end of linear eukaryotic chromosomes

temperature-sensitive (ts) mutation Mutation whose phenotypic effects depend on temperature

template strand Strand of DNA used as a guide for synthesizing a new strand by complementary base pairing

***Ter* site** Site in the terminus region that blocks movement of a replication fork

teratogen An agent that causes abnormal embryo development leading to gross structural defects or monstrosities

5′-terminal oligopyrimidine tract (5′-TOP) Long pyrimidine-rich tracts located between the 5′-end of mRNA and the start codon.

terminator DNA sequence at end of a gene that tells RNA polymerase to stop transcribing

terminus of chromosome (*terC*) The place on a chromosome where replication ends

terminus of replication (*ter*) The place on any DNA molecule where replication ends

tertiary atmosphere The present atmosphere of the earth resulting from biological activity

tertiary structure Final 3-D folding of a polymer chain

tetracycline Antibiotic that binds to 16S ribosomal RNA and inhibits protein synthesis

tetrahydrofolate (THF) Reduced form of dihydrofolate cofactor that is needed for making precursors for DNA and RNA synthesis

tetraploid Having four copies of each gene

thermocycler Machine used to rapidly shift samples between several temperatures in a pre-set order (for PCR)

Thermus aquaticus Thermophilic bacterium found in hot springs and used as a source of thermostable DNA polymerase

theta-replication Mode of replication in which two replication forks go in opposite directions around a circular molecule of DNA

third base redundancy Situation where a set of four codons all code for the same amino acid and thus the identity of the third codon base makes no difference during translation

thymidine The nucleoside consisting of thymine plus deoxyribose

thymidylate synthetase Enzyme that adds a methyl group, so converting the uracil of dUMP to thymine

thymine (T) A pyrimidine base found in DNA that pairs with adenine

tight mutation Mutation whose phenotype is clear-cut due to the complete loss of function of a particular gene product

time-of-flight (TOF) Type of mass spectrometry detector that measures the time for an ion to fly from the ion source to the detector

Ti-plasmid Tumor-inducing plasmid. Plasmid that is carried by soil bacteria of the *Agrobacterium* group and confers the ability to infect plants and produce tumors

tmRNA Specialized RNA used to terminate protein synthesis when a ribosome is stalled by a damaged mRNA

tobacco mosaic virus A filamentous single-stranded RNA virus that infects a wide range of plants

topoisomerase Enzyme that alters the level of supercoiling or catenation of DNA (i.e. changes the topological conformation)

topoisomerase IV A particular topoisomerase involved in DNA replication in bacteria

topoisomers Isomeric forms that differ in topology—i.e. their level of supercoiling or catenation

totipotency Ability of a single cell to give rise to the whole multicellular organism from which it is derived

totipotent Capable of giving rise to a complete multicellular organism

***tra* genes** Genes needed for plasmid transfer

Tra$^+$ Transfer positive (refers to a plasmid capable of self-transfer)

transcription Process by which information from DNA is converted into its RNA equivalent

transcription bubble Region where DNA double helix is temporarily opened up so allowing transcription to occur

transcription factor Protein that regulates gene expression by binding to DNA in the control region of the gene

transcription-coupled repair Preferential repair of the template strand of DNA that may be transcribed

transcriptome The total sum of the RNA transcripts found in a cell, under any particular set of conditions

transduction Process in which genes are transferred inside virus particles

transfection Process in which purified viral DNA enters a cell by transformation. Often used to refer to entry of any DNA, even if not of viral origin, into an animal cell

transfer RNA (tRNA) RNA molecules that carry amino acids to the ribosome

transferability Ability of certain plasmids to move themselves from one bacterial cell to another

transformation (As used in bacterial genetics) Process in which genes are transferred into a cell as free molecules of DNA

transformation (As used of cancer) Changing a normal cell into a cancer cell, even if no extra DNA enters the cell

transition Mutation in which a pyrimidine is replaced by another pyrimidine or a purine is replaced by another purine

transition state Another term for the activated intermediate in a chemical reaction

transition state analog Enzyme inhibitor that mimics the reaction intermediate or transition state, rather than the substrate

transition state energy Energy difference between the reactants and the activated reaction intermediate or transition state

translation Making a protein using the information provided by messenger RNA

translational activator A protein that binds to mRNA and promotes its translation

translational repression Form of control in which the translation of a messenger RNA is prevented

translational repressor A protein that binds to mRNA and prevents its translation

translatome The total set of proteins that have actually been translated and are present in a cell under any particular set of conditions

translocase Enzyme complex that transports proteins across membranes

translocation a) Transport of a newly made protein across a membrane by means of a translocase; b) Sideways movement of the ribosome on mRNA during translation and c) Removal of a segment of DNA from its original location and its reinsertion in a different place

transport protein A protein that carries other molecules across membranes or around the body

transposable element A mobile segment of DNA that is always inserted in another, host molecule, of DNA. It has no origin of replication of its own and relies on the host DNA molecule for replication. Includes both DNA-based transposons and retrotransposons

transposase Enzyme responsible for moving a transposon

transposition The process by which a transposon moves from one host DNA molecule to another

transposon Same as transposable element, although the term is usually restricted to DNA-based elements that do not use reverse transcriptase

trans-splicing Splicing of a segment from one RNA molecule into another distinct RNA molecule

transversion Mutation in which a pyrimidine is replaced by a purine or vice versa

trimethoprim Antibiotic that is an analog of the pterin ring portion of the folate cofactor. It inhibits dihydrofolate reductase

triploid Having three copies of each chromosome

trisomy Having three copies of a particular chromosome

true revertant Revertant in which the original base sequence is exactly restored

trypanosomes Group of parasitic single-celled eukaryotes that cause sleeping sickness and other tropical diseases

Tus protein Bacterial protein that binds to *Ter* sites and blocks movement of replication forks

twist, T The number of double helical turns in a molecule of DNA (or double-stranded RNA)

two-component regulatory system A regulatory system consisting of two proteins, a sensor kinase and a DNA-binding regulator

two-hybrid system Method of screening for protein-protein interactions that uses fusions of the proteins being investigated to the two separate domains of a transcriptional activator protein

Ty1 element Transposon yeast 1. A retrotransposon of yeast that moves via an RNA intermediate

type I restriction enzyme Type of restriction enzyme that cuts the DNA a thousand or more base pairs away from the recognition site

type I topoisomerase Topoisomerase that cuts a single strand of DNA and therefore changes the linking number by one

type II restriction enzyme Type of restriction enzyme that cuts the DNA in the middle of the recognition site

type II restriction enzyme Type or restriction enzyme that cuts a fixed number of bases away from its recognition site

type II topoisomerase Topoisomerase that cuts both strands of DNA and therefore changes the linking number by two

U1 Snurp (snRNP) that recognizes the upstream slice site

U2 Snurp (snRNP) that recognizes the branch site

U2AF (U2 accessory factor) Protein involved in splicing of introns that recognizes the down stream splice site

ubiquitin Small protein attached to other proteins as a signal that they should be degraded; used by eukaryotic cells, not bacteria

uncharged tRNA tRNA without an amino acid attached

unequal crossing over Crossing over in which the two segments that cross over are of different lengths; often due to misalignment during pairing of DNA strands

Ung protein Same as uracil-N-glycosylase

universal genetic code Version of the genetic code used by almost all organisms

3′-untranslated region (3′-UTR) Sequence at the 3′-end of mRNA, downstream of the final stop codon, that is not translated into protein

5′-untranslated region (5′-UTR) Region of an mRNA between the 5′-end and the translation start site

upstream element DNA sequence upstream of the TATA box in eukaryotic promoters that is recognized by specific proteins

upstream region Region of DNA in front (i.e. beyond the 5′-end) of a structural gene; its bases are numbered negatively counting backwards from the start of transcription

uracil (U) A pyrimidine base found in RNA that may pair with adenine

uracil-N-glycosylase Enzyme that removes uracil from DNA

urea A nitrogen waste product of animals; also widely used as a denaturant of proteins

uridine The nucleoside consisting of uracil plus ribose

urkaryote Hypothetical ancestor that provided the genetic information of the eukaryotic nucleus

U-RNA Uracil-rich small RNA (includes snRNA and snoRNA)

vaccination Artificial induction of the immune response by injecting foreign proteins or other antigens

variable number tandem repeats (VNTR) Cluster of tandemly repeated sequences in the DNA, whose number of repeats differs from one individual to another

variant surface glycoprotein (VSG) Glycoprotein found on surface of trypanosomes that is encoded by multiple gene copies and varied to avoid recognition by the animal immune system

vector (a) In molecular biology a vector is molecule of DNA which can replicate and is used to carry cloned genes or DNA fragments; (b) in general biology a vector is an organism (such as a mosquito) that carries and distributes a disease-causing microorganisms (such as yellow fever or malaria)

vegetative reproduction Form of reproduction in which there is no reshuffling of the genes between two individuals (same as asexual reproduction)

vertical gene transfer Transfer of genetic information from an organism to its descendents

viral genome Molecule of DNA or RNA that carries the genes of a virus

virion Virus particle

viroid Naked single-stranded circular RNA that forms a stable highly base-paired rod-like structure and replicates inside infected plant cells. Viroids do not encode any proteins but possess self-cleaving ribozyme activity

virulence factors Proteins that promote virulence in infectious bacteria. Include toxins, adhesins and proteins protecting bacteria from the immune system

virulence plasmid Plasmid that carries genes for virulence factors that play a role in bacterial infection

virus Subcellular parasite with genes of DNA or RNA which replicates inside the host cell upon which it relies for energy and protein synthesis. In addition, it has an extracellular form, in which the virus genes are contained inside a protective coat

virusoid Parasitic RNA molecule that does not encode any proteins but depends on a helper virus for replication and capsid formation.

V_{max} Maximum velocity of an enzyme

VNTR See variable number tandem repeats

Western blotting Detection technique in which a probe, usually an antibody, binds to a protein target molecule

wild-type The original or “natural” version of a gene or organism

wobble rules Rules allowing less rigid base pairing but only for codon/anticodon pairing

writhe Same as writhing number, W

writhing number, W The number of supercoils in a molecule of DNA (or double-stranded RNA)

X-chromosome Female sex chromosome; possession of two X-chromosomes causes female gender in mammals

X-gal (5-bromo-4-chloro-3-indolyl β-D-galactoside) Artificial substrate that is split by β-galactosidase, releasing a blue dye

X-inactivation The condensation and complete shutting down of gene expression of one of the two X-chromosomes in cells of female mammals

Xis protein Enzyme that reverses DNA integration by removing a segment of dsDNA and resealing the gap leaving behind an intact recognition sequence. Same as excisionase. Not to be confused with Xist RNA involved in X chromosome silencing

Xist gene A gene that causes the inactivation of the X-chromosome that carries it

X-phos 5-bromo-4-chloro-3-indolyl phosphate, an artificial substrate that is split by alkaline phosphatase, releasing a blue dye

Y-chromosome Male sex chromosome; possession of a Y-chromosome plus an X-chromosome causes male gender in mammals

yeast artificial chromosome (YAC) Single copy vector based on yeast chromosome that can carry very long inserts of DNA. Widely used in the human genome project

Y-guy Hypothetical male human ancestor thought to have lived in Africa around 100,000–200,000 years ago

Z-DNA An alternative form of DNA double helix with left-handed turns and 12 base pairs per turn

Z-form An alternative form of double helix with left-handed turns and 12 base pairs per turn. Both DNA and dsRNA may be found in the Z-form

zinc finger One type of DNA-binding motif common in proteins

Zoo blotting Comparative Southern blotting using DNA target molecules from several different animals to test whether the probe DNA is from a coding region

zwitterion Same as dipolar ion; a molecule with both a positive and a negative charge

zygote Cell formed by union of sperm and egg which develops into a new individual

Index

Note: Italicized page numbers refer to illustrations, tables, and notes