THE CELL

A MOLECULAR APPROACH

THIRD EDITION

GEOFFREY M. COOPER • ROBERT E. HAUSMAN

The Cover
This illustration shows a portion of the inside of a cell nucleus, including some of the many proteins that copy, repair, and package DNA. DNA strands are shown in yellow. Running through the center of the illustration, top to bottom, is a replication fork, showing DNA being copied by DNA polymerase. On the right and left sides of the illustration, RNA polymerase is synthesizing messenger RNA. Most of the DNA in the picture is wrapped around nucleosomes. Illustration by David S. Goodsell, The Scripps Research Institute.

Part I opener image
Microtubules and actin filaments are stained with red and green fluorescent dyes, respectively.
(K. G. Murti/Visuals Unlimited)

Part II opener image
High resolution X-ray crystal structures of ribosomal units.
(From N. Ban, P. Nissen, J. Hansen, P. B. Moore and T. A. Steitz, 2000. *Science* 289: 905. Courtesy of Thomas A. Steitz.)

Part III opener image
Probes to repeated sequences on chromosome 4 were hybridized to a human cell. The two copies of chromosome 4, identified by yellow fluorescence, occupy distinct territories in the nucleus.
(From A. I. Lamond and W. C. Earnshaw, 1998. *Science* 280: 547.)

Part IV opener image
Mitosis sequence: Telophase.
(Conly L. Rieder/Biological Photo Service)

The Cell: A Molecular Approach, Third Edition

Address editorial correspondence to ASM Press, c/o The American Society for Microbiology, 1752 N Street NW, Washington, DC 20036 U.S.A.

Address orders and requests for examination copies to Sinauer Associates, Inc.,
P.O. Box 407, 23 Plumtree Road, Sunderland, MA 01375 U.S.A.
Phone: 413-549-4300
FAX: 413-549-1118
email: orders@sinauer.com
www.sinauer.com

Library of Congress Cataloging-in-Publication Data

Cooper, Geoffrey M.
The cell : a molecular approach / Geoffrey M. Cooper, Robert E. Hausman.— 3rd ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 0-87893-214-3 (alk. paper)
1. Cytology. 2. Molecular biology.
[DNLM: 1. Cytology. 2. Molecular Biology. QH 581.2 C776c 2004] I. Hausman, Robert E., 1947- II. Title.
QH581.2.C66 2004
571.6—dc21
2003008953

Printed in U.S.A.

5 4 3 2 1

Chapter 3 Fundamentals of Molecular Biology

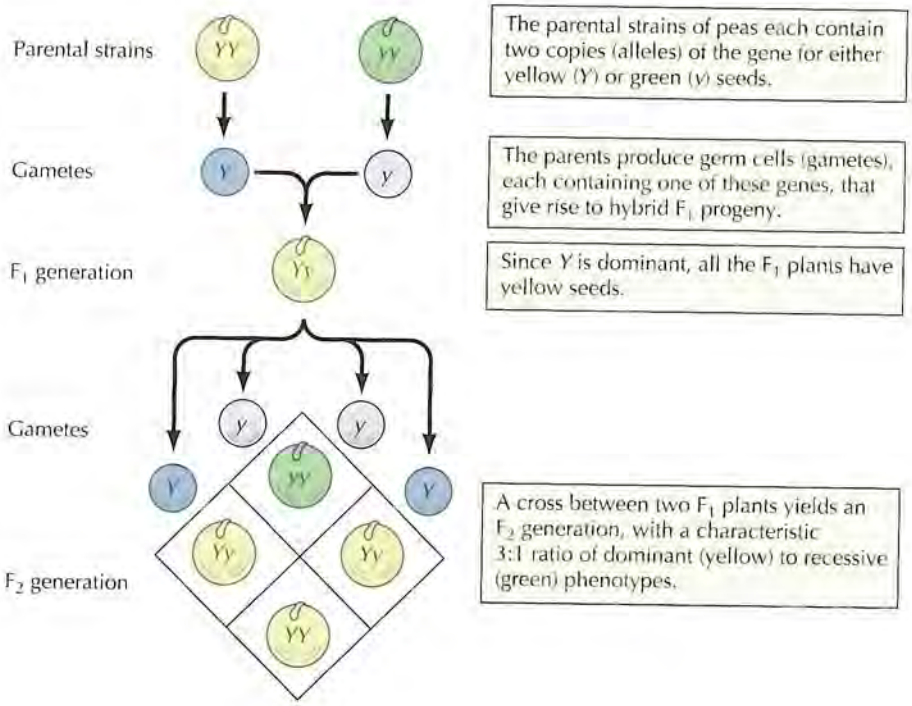
CONTEMPORARY MOLECULAR BIOLOGY is concerned principally with understanding the mechanisms responsible for transmission and expression of the genetic information that governs cell structure and function. As reviewed in Chapter 1, all cells share a number of basic properties, and this underlying unity of cell biology is particularly apparent at the molecular level. Such unity has allowed scientists to choose simple organisms (such as bacteria) as models for many fundamental experiments, with the expectation that similar molecular mechanisms are operative in organisms as diverse as *E. coli* and humans. Numerous experiments have established the validity of this assumption, and it is now clear that the molecular biology of cells provides a unifying theme to understanding diverse aspects of cell behavior.

Initial advances in molecular biology were made by taking advantage of the rapid growth and readily manipulable genetics of simple bacteria, such as *E. coli*, and their viruses. The development of recombinant DNA then allowed both the fundamental principles and many of the experimental approaches first developed in prokaryotes to be extended to eukaryotic cells. The application of recombinant DNA technology has had a tremendous impact, initially allowing individual eukaryotic genes to be isolated and characterized in detail and more recently allowing the determination of the complete genome sequences of complex plants and animals, including humans.

Heredity, Genes, and DNA

Perhaps the most fundamental property of all living things is the ability to reproduce. All organisms inherit the genetic information specifying their structure and function from their parents. Likewise, all cells arise from preexisting cells, so the genetic material must be replicated and passed from parent to progeny cell at each cell division. How genetic information is replicated and transmitted from cell to cell and organism to organism thus represents a question that is central to all of biology. Consequently, elucidation of the mechanisms of genetic transmission and identification of the genetic material as DNA were discoveries that formed the foundation of our current understanding of biology at the molecular level.

Figure 3.1 Inheritance of dominant and recessive genes



Genes and Chromosomes

The classical principles of genetics were deduced by Gregor Mendel in 1865, on the basis of the results of breeding experiments with peas. Mendel studied the inheritance of a number of well-defined traits, such as seed color, and was able to deduce general rules for their transmission. In all cases, he could correctly interpret the observed patterns of inheritance by assuming that each trait is determined by a pair of inherited factors, which are now called **genes**. One gene copy (called an **allele**) specifying each trait is inherited from each parent. For example, breeding two strains of peas—one having yellow seeds, and the other green seeds—yields the following results (Figure 3.1). The parental strains each have two identical copies of the gene specifying yellow (*Y*) or green (*y*) seeds, respectively. The progeny plants are therefore hybrids, having inherited one gene for yellow seeds (*Y*) and one for green seeds (*y*). All these progeny plants (the first filial, or F_1, generation) have yellow seeds, so yellow (*Y*) is said to be **dominant** and green (*y*) **recessive**. The **genotype** (genetic composition) of the F_1 peas is thus *Yy*, and their **phenotype** (physical appearance) is yellow. If one F_1 offspring is bred with another, giving rise to F_2 progeny, the genes for yellow and green seeds segregate in a characteristic manner such that the ratio between F_2 plants with yellow seeds and those with green seeds is 3:1.

Mendel's findings, apparently ahead of their time, were largely ignored until 1900, when Mendel's laws were rediscovered and their importance recognized. Shortly thereafter, the role of **chromosomes** as the carriers of genes was proposed. It was realized that most cells of higher plants and animals are **diploid**—containing two copies of each chromosome. Formation of the germ cells (the sperm and egg), however, involves a unique type of cell division (**meiosis**) in which only one member of each chromosome pair is transmitted to each progeny cell (Figure 3.2). Consequently, the sperm and egg are **haploid**, containing only one copy of each chromosome. The union of these two haploid cells at fertilization creates a new diploid organ-

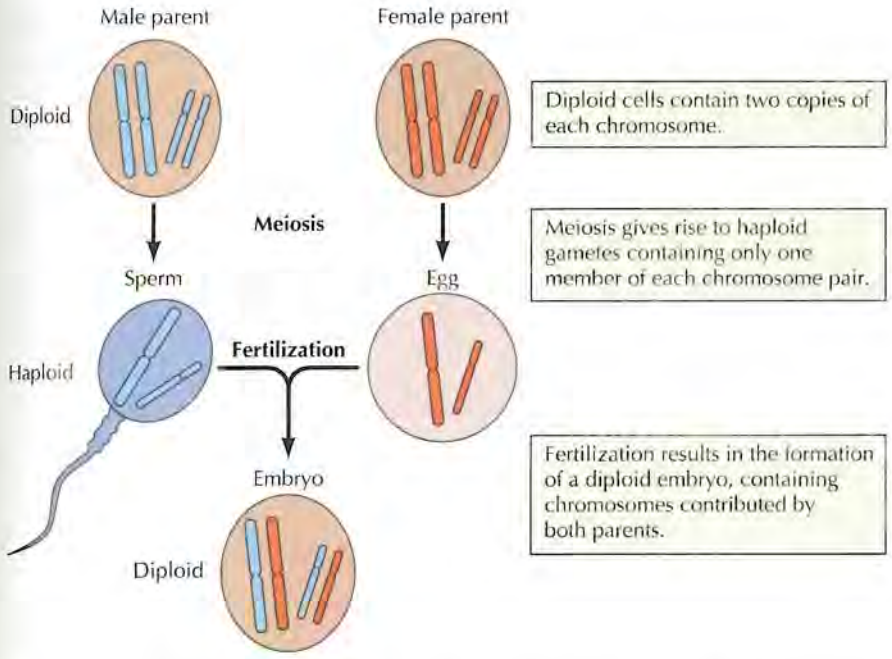


Figure 3.2 Chromosomes at meiosis and fertilization
Two chromosome pairs of a hypothetical organism are illustrated.

ism, now containing one member of each chromosome pair derived from the male and one from the female parent. The behavior of chromosome pairs thus parallels that of genes, leading to the conclusion that genes are carried on chromosomes.

The fundamentals of mutation, genetic linkage, and the relationships between genes and chromosomes were largely established by experiments performed with the fruit fly, *Drosophila melanogaster*. *Drosophila* can be easily maintained in the laboratory, and they reproduce about every two weeks, which is a considerable advantage for genetic experiments. Indeed, these features continue to make *Drosophila* an organism of choice for genetic studies of animals, particularly the genetic analysis of development and differentiation.

In the early 1900s, a number of genetic alterations (**mutations**) were identified in *Drosophila*, usually affecting readily observable characteristics such as eye color or wing shape. Breeding experiments indicated that some of the genes governing these traits are inherited independently of each other, suggesting that these genes are located on different chromosomes that segregate independently during meiosis (Figure 3.3). Other genes, however, are frequently inherited together as paired characteristics. Such genes are said to be linked to each other by virtue of being located on the same chromosome. The number of groups of linked genes is the same as the number of chromosomes (four in *Drosophila*), supporting the idea that chromosomes are carriers of the genes. By 1915, nearly a hundred genes had been defined and mapped onto the four chromosomes of *Drosophila*, leading to general acceptance of the chromosomal basis of heredity.

Genes and Enzymes

Early genetic studies focused on the identification and chromosomal localization of genes that control readily observable characteristics, such as the eye color of *Drosophila*. How these genes lead to the observed phenotypes, however, was unclear. The first insight into the relationship between genes and enzymes came in 1909, when it was realized that the inherited human

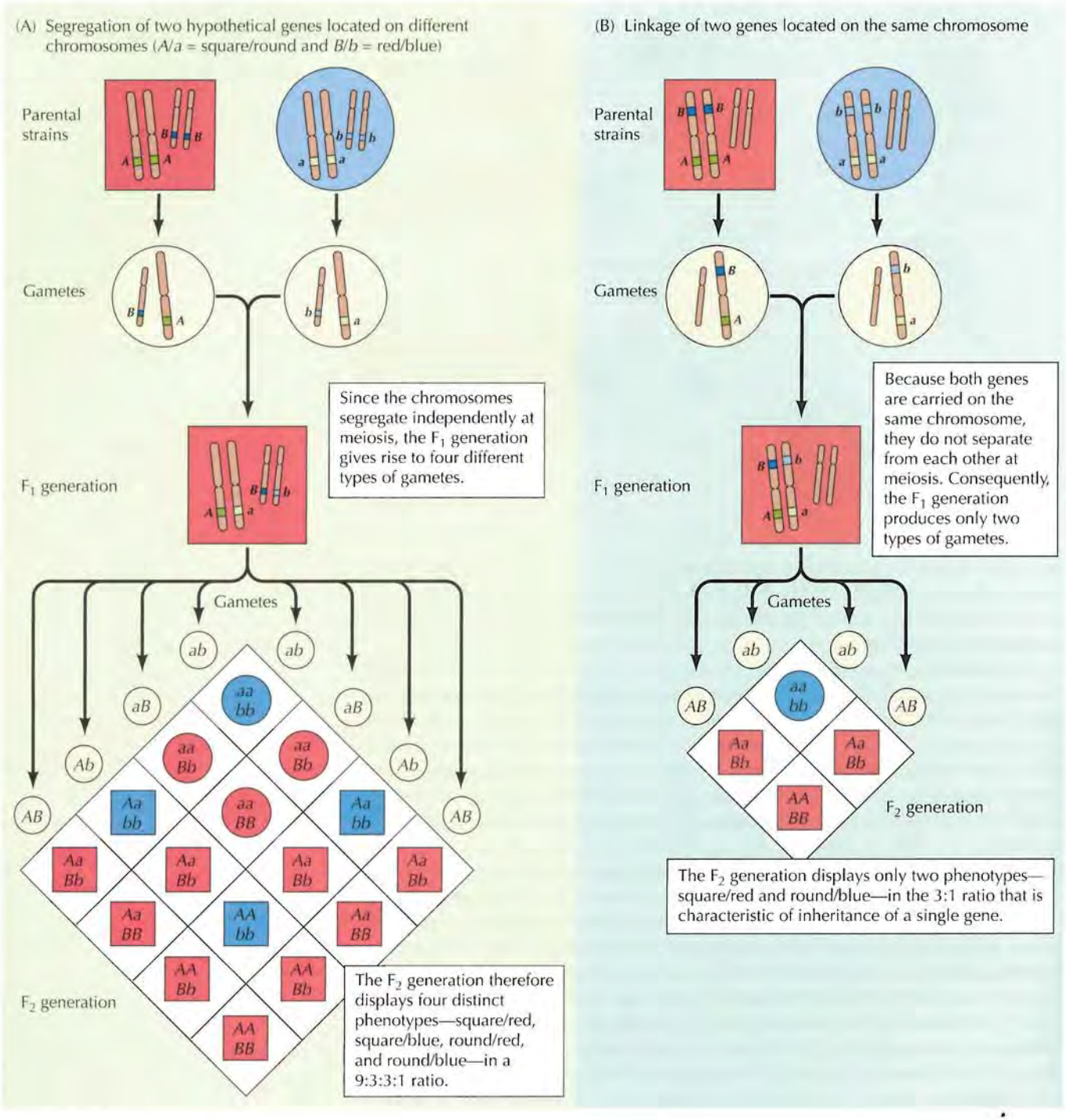


Figure 3.3 Gene segregation and linkage
(A) Segregation of two hypothetical genes for shape (A/a = square/round) and color (B/b = red/blue) located on different chromosomes. (B) Linkage of two genes located on the same chromosome.

disease phenylketonuria (see Molecular Medicine in Chapter 2) results from a genetic defect in metabolism of the amino acid phenylalanine. This defect was hypothesized to result from a deficiency in the enzyme needed to catalyze the relevant metabolic reaction, leading to the general suggestion that genes specify the synthesis of enzymes.

Clearer evidence linking genes with the synthesis of enzymes came from experiments of George Beadle and Edward Tatum, performed in 1941 with

the fungus *Neurospora crassa*. In the laboratory, *Neurospora* can be grown on minimal or rich media similar to those discussed in Chapter 1 for the growth of *E. coli*. For *Neurospora*, minimal media consist only of salts, glucose, and biotin; rich media are supplemented with amino acids, vitamins, purines, and pyrimidines. Beadle and Tatum isolated mutants of *Neurospora* that grew normally on rich media but could not grow on minimal media. Each mutant was found to require a specific nutritional supplement, such as a particular amino acid, for growth. Furthermore, the requirement for a specific nutritional supplement correlated with the failure of the mutant to synthesize that particular compound. Thus, each mutation resulted in a deficiency in a specific metabolic pathway. Since such metabolic pathways were known to be governed by enzymes, the conclusion from these experiments was that each gene specified the structure of a single enzyme—the **one gene–one enzyme hypothesis**. Many enzymes are now known to consist of multiple polypeptides, so the currently accepted statement of this hypothesis is that each gene specifies the structure of a single polypeptide chain.

Identification of DNA as the Genetic Material

Understanding the chromosomal basis of heredity and the relationship between genes and enzymes did not in itself provide a molecular explanation of the gene. Chromosomes contain proteins as well as DNA, and it was initially thought that genes were proteins. The first evidence leading to the identification of DNA as the genetic material came from studies in bacteria. These experiments represent a prototype for current approaches to defining the function of genes by introducing new DNA sequences into cells, as discussed later in this chapter.

The experiments that defined the role of DNA were derived from studies of the bacterium that causes pneumonia (*Pneumococcus*). Virulent strains of *Pneumococcus* are surrounded by a polysaccharide capsule that protects the bacteria from attack by the immune system of the host. Because the capsule gives bacterial colonies a smooth appearance in culture, encapsulated strains are denoted S. Mutant strains that have lost the ability to make a capsule (denoted R) form rough-edged colonies in culture and are no longer lethal when inoculated into mice. In 1928 it was observed that mice inoculated with nonencapsulated (R) bacteria plus heat-killed encapsulated (S) bacteria developed pneumonia and died. Importantly, the bacteria that were then isolated from these mice were of the S type. Subsequent experiments showed that a cell-free extract of S bacteria was similarly capable of converting (or transforming) R bacteria to the S state. Thus, a substance in the S extract (called the transforming principle) was responsible for inducing the genetic **transformation** of R to S bacteria.

In 1944 Oswald Avery, Colin MacLeod, and Maclyn McCarty established that the transforming principle was DNA, both by purifying it from bacterial extracts and by demonstrating that the activity of the transforming principle is abolished by enzymatic digestion of DNA but not by digestion of proteins (Figure 3.4). Although these studies did not immediately lead to the acceptance of DNA as the genetic material, they were extended within a few years by experiments with bacterial viruses. In particular, it was shown that, when a bacterial virus infects a cell, the viral DNA rather than the viral protein must enter the cell in order for the virus to replicate. Moreover, the parental viral DNA (but not the protein) is transmitted to progeny virus particles. The concurrence of these results with continuing studies of the activity of DNA in bacterial transformation led to acceptance of the idea that DNA is the genetic material.

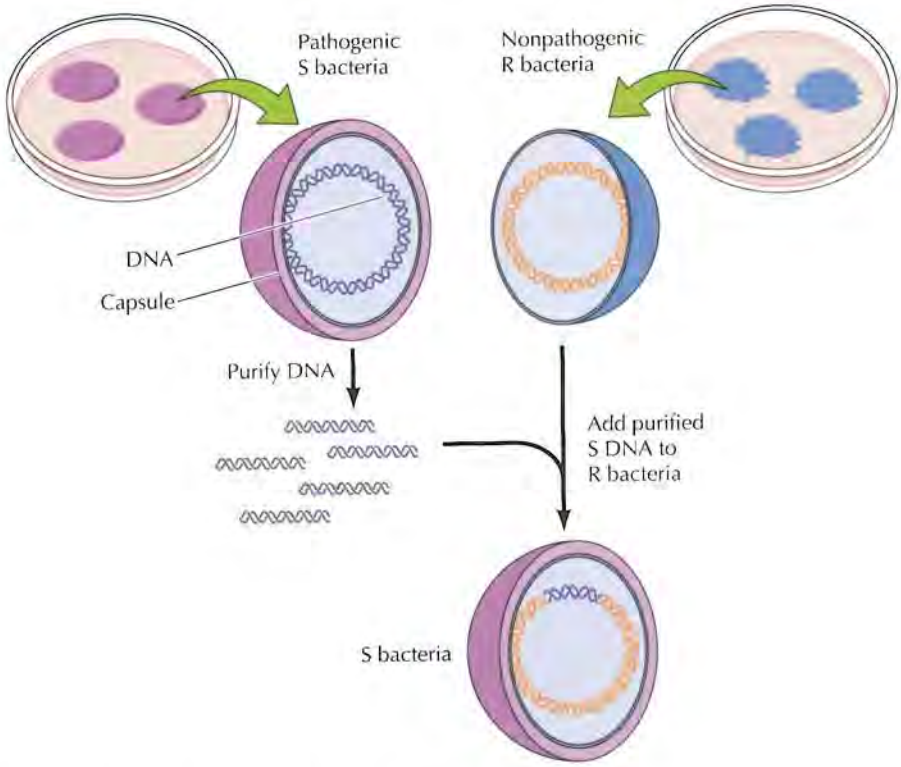


Figure 3.4 Transfer of genetic information by DNA
DNA is extracted from a pathogenic strain of *Pneumococcus*, which is surrounded by a capsule and forms smooth colonies (S). Addition of the purified S DNA to a culture of nonpathogenic, nonencapsulated bacteria (R for "rough" colonies) results in the formation of S colonies. The purified DNA therefore contains the genetic information responsible for transformation of R to S bacteria.

The Structure of DNA

Our understanding of the three-dimensional structure of DNA, deduced in 1953 by James Watson and Francis Crick, has been the basis for present-day molecular biology. At the time of Watson and Crick's work, DNA was known to be a polymer composed of four nucleic acid bases—two purines (adenine [A] and guanine [G]) and two pyrimidines (cytosine [C] and thymine [T])—linked to phosphorylated sugars. Given the central role of DNA as the genetic material, elucidation of its three-dimensional structure appeared critical to understanding its function. Watson and Crick's consideration of the problem was heavily influenced by Linus Pauling's description of hydrogen bonding and the α helix, a common element of the secondary structure of proteins (see Chapter 2). Moreover, experimental data on the structure of DNA were available from X-ray crystallography studies by Maurice Wilkins and Rosalind Franklin. Analysis of these data revealed that the DNA molecule is a helix that turns every 3.4 nm. In addition, the data showed that the distance between adjacent bases is 0.34 nm, so there are ten bases per turn of the helix. An important finding was that the diameter of the helix is approximately 2 nm, suggesting that it is composed of not one but two DNA chains.

From these data, Watson and Crick built their model of DNA (Figure 3.5). The central features of the model are that DNA is a double helix with the sugar–phosphate backbones on the outside of the molecule. The bases are

on the inside, oriented such that hydrogen bonds are formed between purines and pyrimidines on opposite chains. The base pairing is very specific: A always pairs with T and G with C. This specificity accounts for the earlier results of Erwin Chargaff, who had analyzed the base composition of various DNAs and found that the amount of adenine was always equal to that of thymine, and the amount of guanine to that of cytosine. Because of this specific base pairing, the two strands of a DNA molecule are complementary: Each strand contains all the information required to specify the sequences of bases on the other.

Replication of DNA

The discovery of complementary base pairing between DNA strands immediately suggested a molecular solution to the question of how the genetic material could direct its own replication—a process that is required each time a cell divides. It was proposed that the two strands of a DNA molecule could separate and serve as templates for synthesis of new complementary strands, the sequence of which would be dictated by the specificity of base pairing

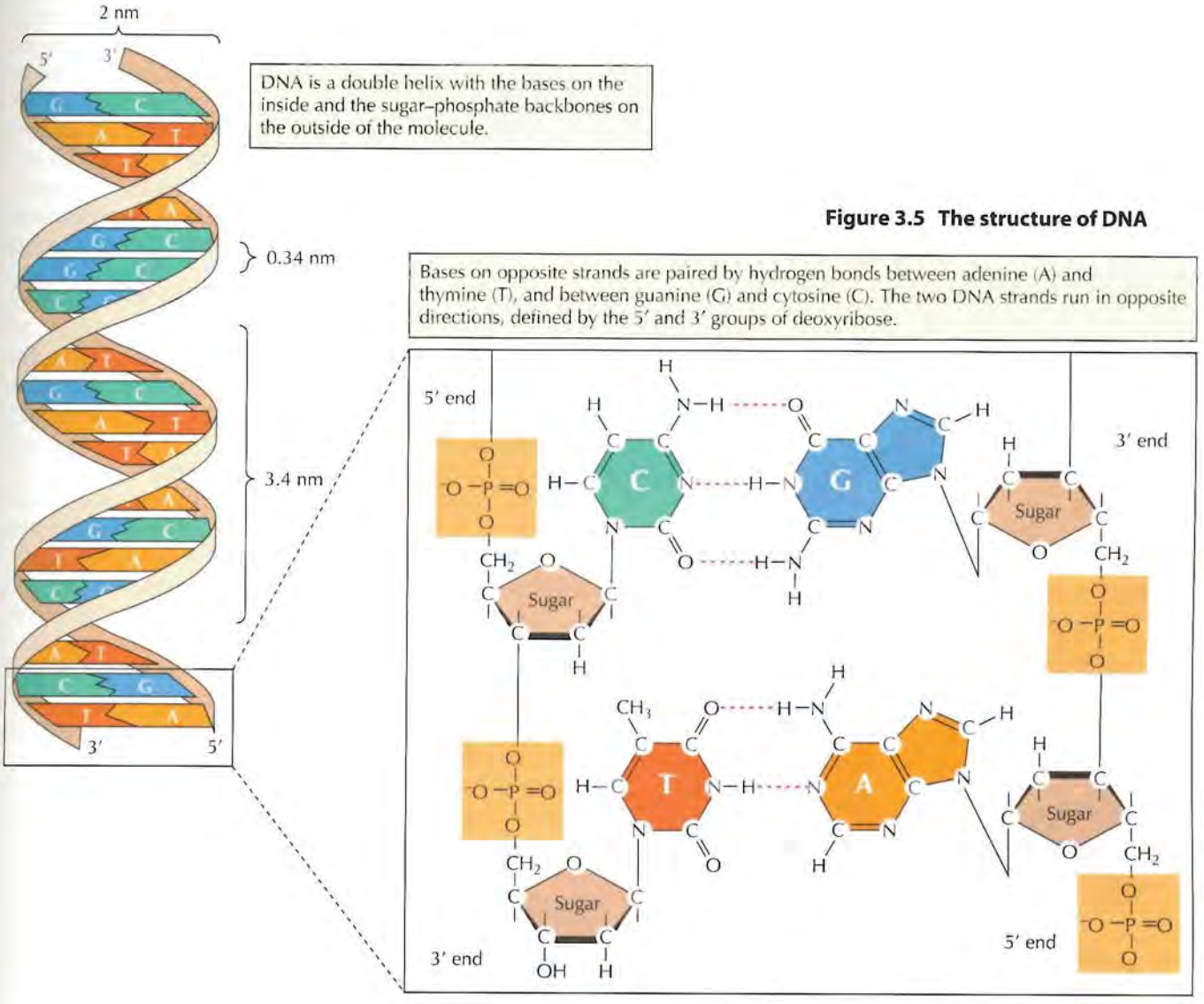


Figure 3.5 The structure of DNA

(Figure 3.6). The process is called **semiconservative replication** because one strand of parental DNA is conserved in each progeny DNA molecule.

Direct support for semiconservative DNA replication was obtained in 1958 as a result of elegant experiments, performed by Matthew Meselson and Frank Stahl, in which DNA was labeled with isotopes that altered its density (Figure 3.7). *E. coli* were first grown in media containing the heavy isotope of nitrogen (^{15}N) in place of the normal light isotope (^{14}N). The DNA of these bacteria consequently contained ^{15}N and was heavier than that of bacteria grown in ^{14}N. Such heavy DNA could be separated from DNA containing ^{14}N by equilibrium centrifugation in a density gradient of CsCl. This ability to separate heavy (^{15}N) DNA from light (^{14}N) DNA enabled the study of DNA synthesis. *E. coli* that had been grown in ^{15}N were transferred to media containing ^{14}N and allowed to replicate one more time. Their DNA was then extracted and analyzed by CsCl density gradient centrifugation. The results of this analysis indicated that all of the heavy DNA had been replaced by newly synthesized DNA with a density intermediate between that of heavy (^{15}N) and that of light (^{14}N) DNA molecules. The implication was that during replication, the two parental strands of heavy DNA separated and served as templates for newly synthesized progeny strands of light DNA, yielding double-stranded molecules of intermediate density. This experiment thus provided direct evidence for semiconservative DNA replication, clearly underscoring the importance of complementary base pairing between strands of the double helix.

The ability of DNA to serve as a template for its own replication was further established with the demonstration that an enzyme purified from *E. coli* (**DNA polymerase**) could catalyze DNA replication *in vitro*. In the presence of DNA to act as a template, DNA polymerase was able to direct the incorporation of nucleotides into a complementary DNA molecule.

Expression of Genetic Information

Genes act by determining the structure of proteins, which are responsible for directing cell metabolism through their activity as enzymes. The identification of DNA as the genetic material and the elucidation of its structure

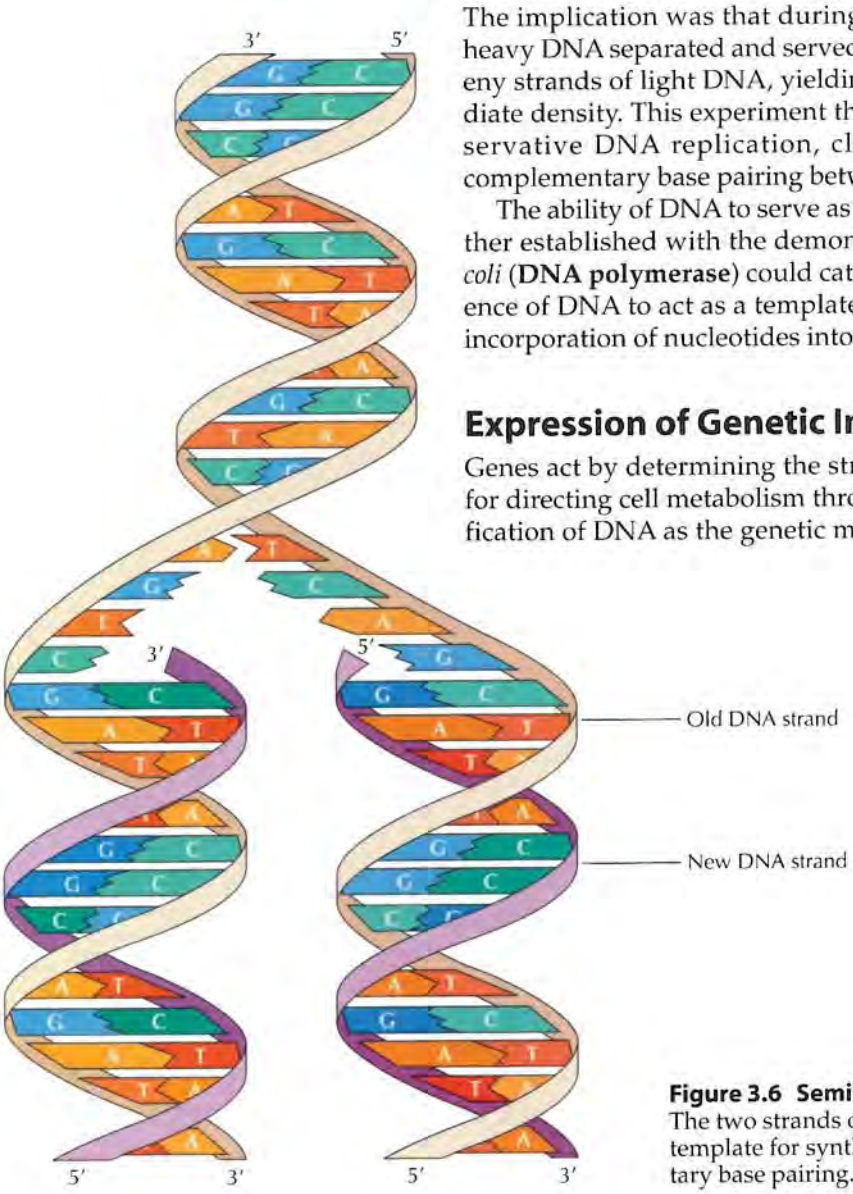


Figure 3.6 Semiconservative replication of DNA
The two strands of parental DNA separate, and each serves as a template for synthesis of a new daughter strand by complementary base pairing.

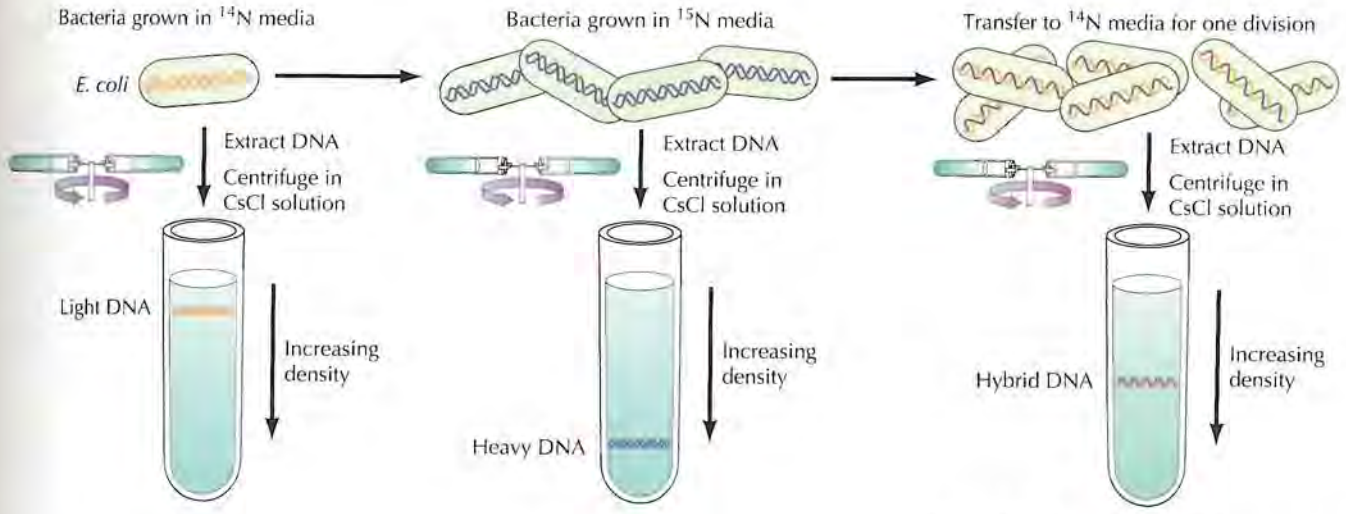


Figure 3.7 Experimental demonstration of semiconservative replication Bacteria grown in medium containing the normal isotope of nitrogen (^{14}N) are transferred into medium containing the heavy isotope (^{15}N) and grown in this medium for several generations. They are then transferred back to medium containing ^{14}N and grown for one additional generation. DNA is extracted from these bacteria and analyzed by equilibrium ultracentrifugation in a CsCl solution. The CsCl sediments to form a density gradient, and the DNA molecules band at a position where their density is equal to that of the CsCl solution. DNA of the bacteria transferred from ^{15}N to ^{14}N medium for a single generation bands at a density intermediate between that of ^{15}N DNA and that of ^{14}N DNA, indicating that it represents a hybrid molecule with one heavy and one light strand.

revealed that genetic information must be specified by the order of the four bases (A, C, G, and T) that make up the DNA molecule. Proteins, in turn, are polymers of 20 amino acids, the sequence of which determines their structure and function. The first direct link between a genetic mutation and an alteration in the amino acid sequence of a protein was made in 1957, when it was found that patients with the inherited disease sickle-cell anemia had hemoglobin molecules that differed from normal ones by a single amino acid substitution. Deeper understanding of the molecular relationship between DNA and proteins came, however, from a series of experiments that took advantage of *E. coli* and its viruses as genetic models.

Colinearity of Genes and Proteins

The simplest hypothesis to account for the relationship between genes and enzymes was that the order of nucleotides in DNA specified the order of amino acids in a protein. Mutations in a gene would correspond to alterations in the sequence of DNA, which might result from the substitution of one nucleotide for another or from the addition or deletion of nucleotides. These changes in the nucleotide sequence of DNA would then lead to corresponding changes in the amino acid sequence of the protein encoded by the gene in question. This hypothesis predicted that different mutations within a single gene could alter different amino acids in the encoded protein, and that the positions of mutations in a gene should reflect the positions of amino acid alterations in its protein product.

The rapid replication and the simplicity of the genetic system of *E. coli* were of major help in addressing these questions. A variety of mutants of *E. coli* could be isolated, including nutritional mutants that (like the *Neurospora* mutants discussed earlier) require particular amino acids for growth. Importantly, the rapid growth of *E. coli* made feasible the isolation and mapping of multiple mutants in a single gene, leading to the first demonstration of the linear relationship between genes and proteins. In these studies, Charles Yanofsky and his colleagues mapped a series of mutations in the gene that encodes an enzyme required for synthesis of the amino acid tryptophan. Analysis of the enzymes encoded by the mutant genes indicated that the relative positions of the amino acid alterations were the same as those of the corresponding mutations (Figure 3.8). Thus, the sequence of amino acids in the protein was colinear with that of mutations in the gene,

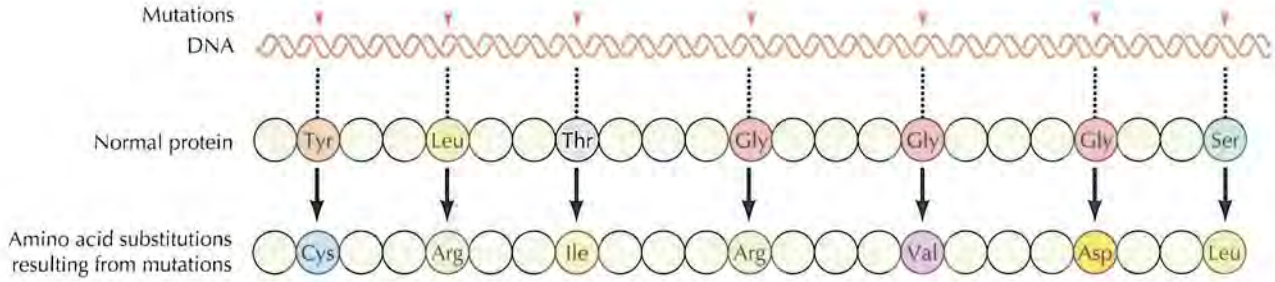


Figure 3.8 Colinearity of genes and proteins
A series of mutations (arrowheads) were mapped in the *E. coli* gene encoding tryptophan synthetase (top line). The amino acid substitutions resulting from each of the mutations were then determined by sequence analysis of the proteins of mutant bacteria (bottom line). These studies revealed that the order of mutations in DNA was the same as the order of amino acid substitutions in the encoded protein.

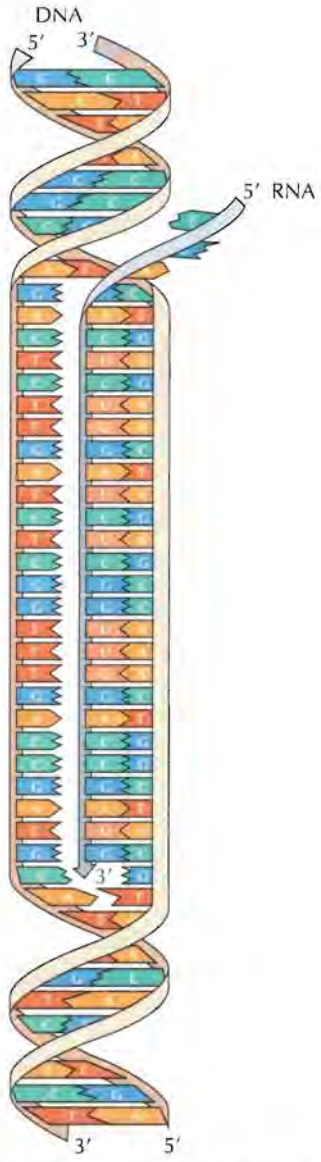


Figure 3.9 Synthesis of RNA from DNA
The two strands of DNA unwind, and one is used as a template for synthesis of a complementary strand of RNA.

as expected if the order of nucleotides in DNA specifies the order of amino acids in proteins.

The Role of Messenger RNA

Although the sequence of nucleotides in DNA appeared to specify the order of amino acids in proteins, it did not necessarily follow that DNA itself directs protein synthesis. Indeed, this appeared not to be the case, since DNA is located in the nucleus of eukaryotic cells, whereas protein synthesis takes place in the cytoplasm. Some other molecule was therefore needed to convey genetic information from DNA to the sites of protein synthesis (the ribosomes).

RNA appeared a likely candidate for such an intermediate because the similarity of its structure to that of DNA suggested that RNA could be synthesized from a DNA template (Figure 3.9). RNA differs from DNA in that it is single-stranded rather than double-stranded, its sugar component is ribose instead of deoxyribose, and it contains the pyrimidine base uracil (U) instead of thymine (T) (see Figure 2.10). However, neither the change in sugar nor the substitution of U for T alters base pairing, so the synthesis of RNA can be readily directed by a DNA template. Moreover, since RNA is located primarily in the cytoplasm, it appeared a logical intermediate to convey information from DNA to the ribosomes. These characteristics of RNA suggested a pathway for the flow of genetic information that is known as the **central dogma** of molecular biology:

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}$$

According to this concept, RNA molecules are synthesized from DNA templates (a process called **transcription**), and proteins are synthesized from RNA templates (a process called **translation**).

Experimental evidence for the RNA intermediates postulated by the central dogma was obtained by Sidney Brenner, Francois Jacob, and Matthew Meselson in studies of *E. coli* infected with the bacteriophage T4. The synthesis of *E. coli* RNA stops following infection by T4, and the only new RNA synthesized in infected bacteria is transcribed from T4 DNA. This T4 RNA becomes associated with bacterial ribosomes, thus conveying the information from DNA to the site of protein synthesis. Because of their role as intermediates in the flow of genetic information, RNA molecules that serve as templates for protein synthesis are called **messenger RNAs (mRNAs)**. They

are transcribed by an enzyme (**RNA polymerase**) that catalyzes the synthesis of RNA from a DNA template.

In addition to mRNA, two other types of RNA molecules are important in protein synthesis. **Ribosomal RNA (rRNA)** is a component of ribosomes, and **transfer RNAs (tRNAs)** serve as adaptor molecules that align amino acids along the mRNA template. The structures and functions of these molecules are discussed in the following section and in more detail in Chapters 6 and 7.

The Genetic Code

How is the nucleotide sequence of mRNA translated into the amino acid sequence of a protein? In this step of gene expression genetic information is transferred between chemically unrelated types of macromolecules—nucleic acids and proteins—raising two new types of problems in understanding the action of genes.

First, since amino acids are structurally unrelated to the nucleic acid bases, direct complementary pairing between mRNA and amino acids during the incorporation of amino acids into proteins seemed impossible. How then could amino acids align on an mRNA template during protein synthesis? This question was solved by the discovery that tRNAs serve as adaptors between amino acids and mRNA during translation (Figure 3.10). Prior to its use in protein synthesis, each amino acid is attached by a specific enzyme to its appropriate tRNA. Base pairing between a recognition sequence on each tRNA and a complementary sequence on the mRNA then directs the attached amino acid to its correct position on the mRNA template.

The second problem in the translation of nucleotide sequence to amino acid sequence was determination of the **genetic code**. How could the information contained in the sequence of four different nucleotides be converted to the sequences of 20 different amino acids in proteins? Because 20 amino acids must be specified by only four nucleotides, at least three nucleotides must be used to encode each amino acid. Used singly, four nucleotides could encode only four amino acids and, used in pairs, four nucleotides could encode only sixteen (4^2) amino acids. Used as triplets, however, four nucleotides could encode 64 (4^3) different amino acids—more than enough to account for the 20 amino acids actually found in proteins.

Direct experimental evidence for the triplet code was obtained by studies of bacteriophage T4 bearing mutations in an extensively studied gene called *rII*. Phages with mutations in this gene form abnormally large plaques, which can be clearly distinguished from those formed by wild-type phages. Hence, isolating and mapping a number of *rII* mutants was easy and led to the establishment of a detailed genetic map of this locus. Study of recombinants between *rII* mutants that had arisen by additions or deletions of nucleotides revealed that phages containing additions or deletions of one or two nucleotides always exhibited the mutant phenotype. Phages containing additions or deletions of three nucleotides, however, were frequently wild-type in function (Figure 3.11). These findings suggested that the gene is read in groups of three nucleotides, starting from a

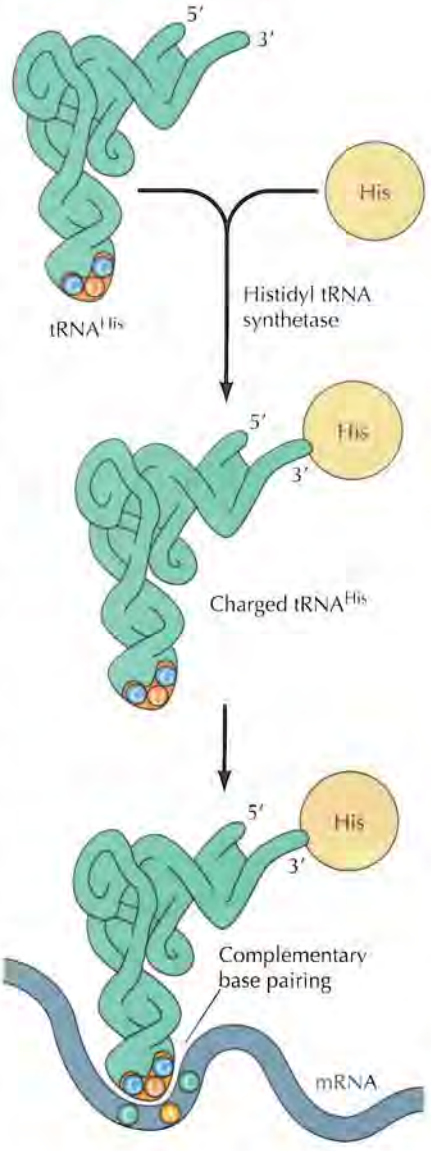


Figure 3.10 Function of transfer RNA
Transfer RNA serves as an adaptor during protein synthesis. Each amino acid (e.g., histidine) is attached to the 3′ end of a specific tRNA by an appropriate enzyme (an aminoacyl tRNA synthetase). The charged tRNAs then align on an mRNA template by complementary base pairing.

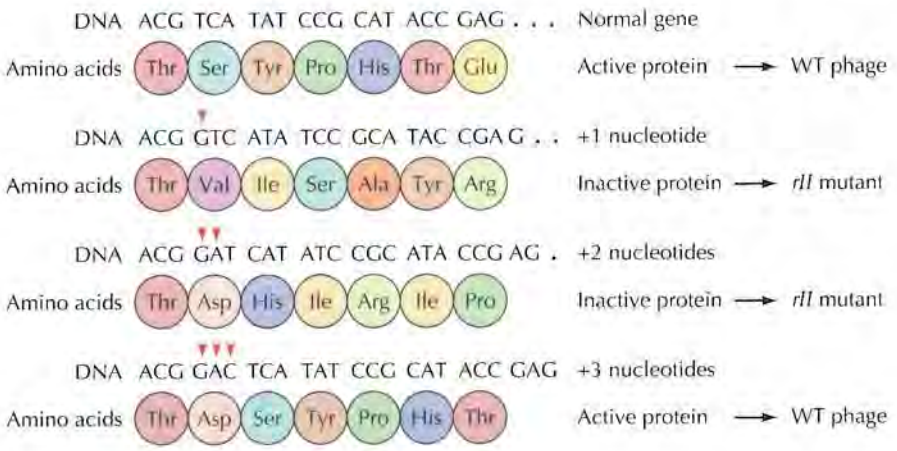


Figure 3.11 Genetic evidence for a triplet code
A series of mutations consisting of additions of one, two, or three nucleotides were studied in the *rII* gene of bacteriophage T4. Additions of one or two nucleotides alter the reading frame of the remainder of the gene. Therefore, all the subsequent amino acids are abnormal, and an inactive protein is produced, giving rise to mutant phage. Additions of three nucleotides, however, alter only a single amino acid. The reading frame of the remainder of the gene is normal, and an active protein giving rise to wild-type (WT) phage is produced.

fixed point. Additions or deletions of one or two nucleotides would then alter the reading frame of the entire gene, leading to the coding of abnormal amino acids throughout the encoded protein. In contrast, additions or deletions of three nucleotides would lead to the addition or deletion of only a single amino acid; the rest of the amino acid sequence would remain unaltered, frequently yielding an active protein.

Deciphering the genetic code thus became a problem of assigning nucleotide triplets to their corresponding amino acids. This problem was approached using *in vitro* systems that could carry out protein synthesis (***in vitro* translation**). Cell extracts containing ribosomes, amino acids, tRNAs, and the enzymes responsible for attaching amino acids to the appropriate tRNAs (aminoacyl-tRNA synthetases) were known to catalyze the incorporation of amino acids into proteins. However, such protein synthesis depends on the presence of mRNA bound to the ribosomes, and can be greatly enhanced by the addition of purified mRNA. Since added mRNA directs protein synthesis in such systems, the genetic code could be deciphered by study of the translation of synthetic mRNAs of known base sequence.

The first such experiment, performed by Marshall Nirenberg and Heinrich Matthaei, involved the *in vitro* translation of a synthetic RNA polymer containing only uracil (Figure 3.12). This poly-U template was found to direct the incorporation of only a single amino acid—phenylalanine—into a polypeptide consisting of repeated phenylalanine residues. Therefore, the triplet UUU encodes the amino acid phenylalanine. Similar experiments with RNA polymers containing only single nucleotides established that AAA encodes lysine and CCC encodes proline. The remainder of the code was deciphered using RNA polymers containing mixtures of nucleotides, leading to the coding assignment of all 64 possible triplets (called **codons**) (Table 3.1). Of the 64 codons, 61 specify an amino acid; the remaining three (UAA, UAG, and UGA) are stop codons that signal the termination of protein synthesis. The code is degenerate; that is, many amino acids are specified by more than one codon. With few exceptions (discussed in Chapter 10), all organisms utilize the same genetic code, providing strong support for the conclusion that all present-day cells evolved from a common ancestor.

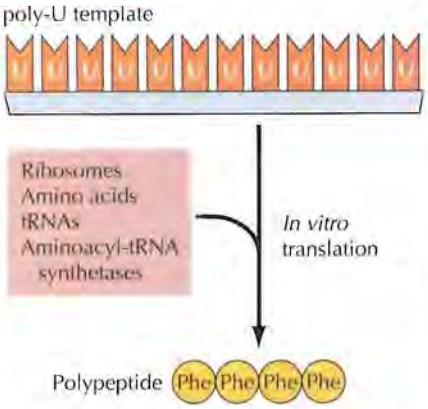


Figure 3.12 The triplet UUU encodes phenylalanine
In vitro translation of a synthetic RNA consisting of repeated uracils (a poly-U template) results in the synthesis of a polypeptide containing only phenylalanine.

RNA Viruses and Reverse Transcription

With the elucidation of the genetic code, the fundamental principles of the molecular biology of cells appeared to have been established. According to

TABLE 3.1 The Genetic Code

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

the central dogma, the genetic material consists of DNA, which is capable of self-replication as well as being transcribed into mRNA, which serves in turn as the template for protein synthesis. However, as noted in Chapter 1, many viruses contain RNA rather than DNA as their genetic material, implying the use of other modes of information transfer.

RNA genomes were first discovered in plant viruses, many of which were found to be composed of only RNA and protein. Direct proof that RNA acts as the genetic material of these viruses was obtained in the 1950s by experiments demonstrating that RNA purified from tobacco mosaic virus could infect new host cells, giving rise to infectious progeny virus. The mode of replication of most viral RNA genomes was subsequently determined by studies of the RNA bacteriophages of *E. coli*. These viruses were found to encode a specific enzyme that could catalyze the synthesis of RNA from an RNA template (RNA-directed RNA synthesis), using the same mechanism of base pairing between complementary strands as is employed during DNA replication or transcription of RNA from DNA.

Although most animal viruses, such as poliovirus or influenza virus, were found to replicate by RNA-directed RNA synthesis, this mechanism did not appear to account for the replication of one family of animal viruses (the RNA tumor viruses), which were of particular interest because of their ability to cause cancer in infected animals. Although these viruses contain genomic RNA in their viral particles, experiments performed by Howard Temin in the early 1960s indicated that their replication requires DNA synthesis in infected cells, leading to the hypothesis that the RNA tumor viruses (now called **retroviruses**) replicate via synthesis of a DNA intermediate, called a DNA provirus (Figure 3.13). This hypothesis was initially met with widespread disbelief because it involves RNA-directed synthesis of DNA—a reversal of the central dogma. In 1970, however, Temin and David Baltimore independently discovered that the RNA tumor viruses contain a

KEY EXPERIMENT

The DNA Provirus Hypothesis

Nature of the Provirus of Rous Sarcoma

Howard M. Temin
McArdle Laboratory, University of Wisconsin, Madison, WI
National Cancer Institute Monographs, Volume 17, 1964, pages 557–570

Howard Temin

The Context

Rous sarcoma virus (RSV), the first cancer-causing virus to be described, was of considerable interest as an experimental system for studying the molecular biology of cancer. Howard Temin began his research in this area when, as a graduate student in 1958, he developed the first assay for the transformation of normal cells to cancer cells in culture following infection with RSV. The availability of such a quantitative *in vitro* assay provided the tool needed for further studies of both cell transformation and virus replication. As Temin proceeded with these studies, he made a series of unexpected observations indicating that the replication of RSV was fundamentally different from that of other RNA viruses. These experiments led to Temin's proposal of the DNA provirus hypothesis, which stated that the viral RNA was copied into DNA in infected cells—a proposal that ran directly counter to the universally accepted central dogma of molecular biology.

The Experiments

The DNA provirus hypothesis was based on several different types of experimental evidence. First, studies of cell transformation using mutants of RSV indicated that important characteristics of transformed cells were determined by genetic information of the virus. This information was regularly transmitted to daughter cells following cell division, even in the absence of virus replication. Temin therefore proposed that the viral genome was present in infected cells in a stably inherited form, which he called a provirus.

Evidence that the provirus was DNA was then derived from experiments with metabolic inhibitors. First, actinomycin D, which inhibits the synthesis of RNA from a DNA template, was found to inhibit virus production by RSV-infected cells (see figure). Second, inhibitors of DNA synthesis inhibited early stages of cell infection by RSV. Thus, DNA synthesis appeared to be required early in infection, and DNA-directed RNA synthesis appeared to be needed subsequently for the production of progeny viruses, leading to the proposal that the provirus was a DNA copy of the viral RNA genome. Temin sought further evidence for this proposal by using nucleic acid hybridization to detect viral sequences in infected cell DNA, but the sensitivity of the available techniques was limited and the data were unconvincing.

The Impact

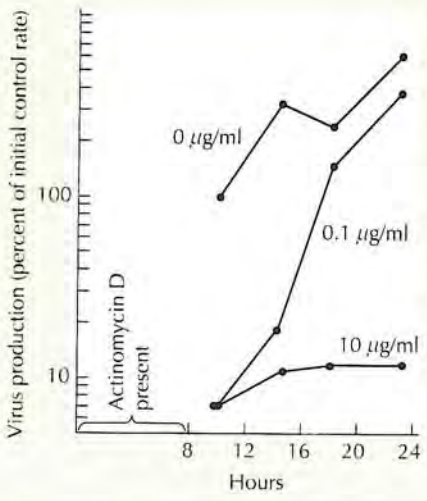
The DNA provirus hypothesis was thus proposed principally on the basis of genetic experiments and the effects of metabolic inhibitors. It was a radical proposal, which contradicted the accepted central dogma of molecular biology. In this setting, Temin's hypothesis that RSV replicated by the transfer of information from RNA to DNA not only failed to win the acceptance of the

novel enzyme that catalyzes the synthesis of DNA from an RNA template. In addition, clear-cut evidence for the existence of viral DNA sequences in infected cells was obtained. The synthesis of DNA from RNA, now called **reverse transcription**, was thus established as a mode of information transfer in biological systems.

Reverse transcription is important not only in the replication of retroviruses, but also in at least two other broad aspects of molecular and cellular biology. First, reverse transcription is not restricted to retroviruses; it also occurs in cells and, as discussed in Chapters 4 and 5, is frequently responsible for the transposition of DNA sequences from one chromosomal location to another. Indeed, the sequence of the human genome has revealed that approximately 40% of human genomic DNA is derived from reverse transcription. Second, enzymes that catalyze RNA-directed DNA synthesis (**reverse transcriptases**) can be used experimentally to generate

scientific community, but was met with general derision. Nonetheless, Temin persevered through the 1960s, continuing with experiments to test his hypothesis and providing increasingly convincing evidence in its support. These efforts culminated in 1970 with the discovery by Temin and Satoshi Mizutani, and at the same time by David Baltimore, of a viral enzyme, now known as reverse transcriptase, that synthesizes DNA from an RNA template—an unambiguous biochemical demonstration that the central dogma could be reversed.

Temin concluded his 1970 paper with the statement that the results "constitute strong evidence that the DNA provirus hypothesis is correct and that RNA tumour viruses have a DNA genome when they are in cells and an RNA genome when they are in virions. This result would have strong implications for theories of viral carcinogenesis and, possibly, for theories of information transfer in other biological systems." As Temin predicted, the discovery of RNA-directed DNA synthesis has led to major advances in our understanding of cancer, human retroviruses, and gene rearrangements. Reverse transcriptase has further provided a critical tool for cDNA cloning, thereby impacting virtually all areas of contemporary cell and molecular biology.



Effect of actinomycin D on RSV replication. RSV-infected cells were cultured with the indicated concentrations of actinomycin D for 8 hours. Actinomycin D was then removed and the amount of virus produced was determined.

DNA copies of any RNA molecule. The use of reverse transcriptase has thus allowed mRNAs of eukaryotic cells to be studied using the molecular approaches that are currently applied to the manipulation of DNA, as discussed in the following section.

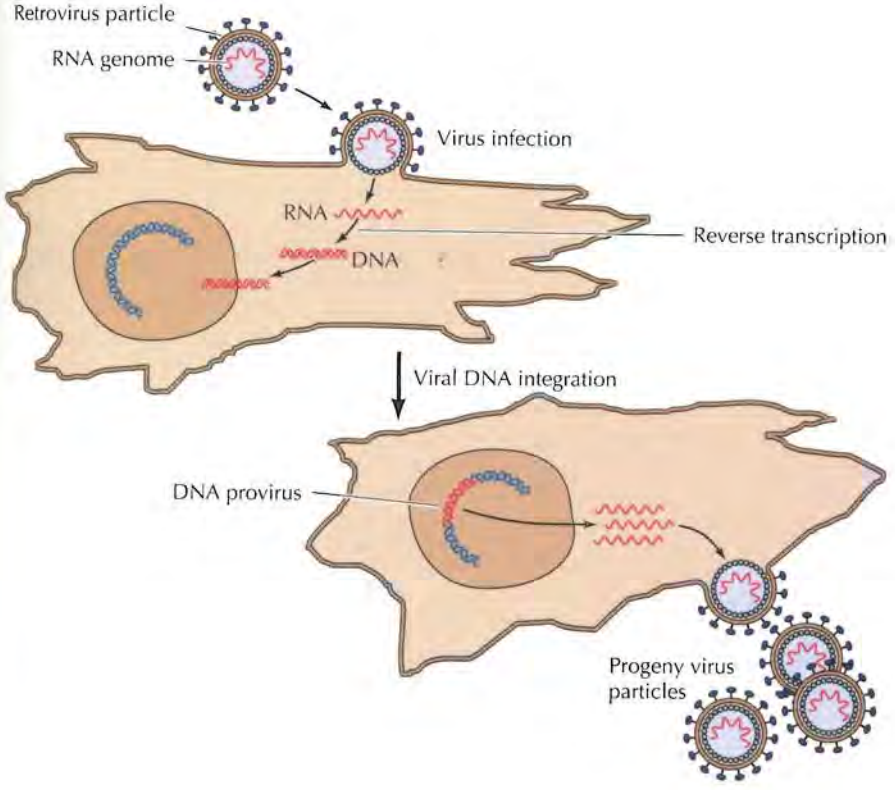


Figure 3.13 Reverse transcription and retrovirus replication
Retroviruses contain RNA genomes in their viral particles. When a retrovirus infects a host cell, however, a DNA copy of the viral RNA is synthesized via reverse transcription. This viral DNA is then integrated into chromosomal DNA of the host to form a DNA provirus, which is transcribed to yield progeny virus RNA.

Recombinant DNA

Classical experiments in molecular biology were strikingly successful in developing our fundamental concepts of the nature and expression of genes. Since these studies were based primarily on genetic analysis, their success depended largely on the choice of simple, rapidly replicating organisms (such as bacteria and viruses) as models. It was not clear, however, how these fundamental principles could be extended to provide a molecular understanding of the complexities of eukaryotic cells, since the genomes of most eukaryotes (e.g., the human genome) are up to a thousand times more complex than that of *E. coli*. In the early 1970s, the possibility of studying such genomes at the molecular level seemed daunting. In particular, there appeared to be no way in which individual genes could be isolated and studied.

This obstacle to the progress of molecular biology was overcome by the development of recombinant DNA technology, which provided scientists with the ability to isolate, sequence, and manipulate individual genes derived from any type of cell. The application of recombinant DNA has thus enabled detailed molecular studies of the structure and function of eukaryotic genes and genomes, thereby revolutionizing our understanding of cell biology.

Restriction Endonucleases

The first step in the development of recombinant DNA technology was the characterization of **restriction endonucleases**—enzymes that cleave DNA at specific sequences. These enzymes were identified in bacteria, where they apparently provide a defense against the entry of foreign DNA (e.g., from a virus) into the cell. Bacteria have a variety of restriction endonucleases that cleave DNA at more than a hundred distinct recognition sites, each of which consists of a specific sequence of four to eight base pairs (examples are given in Table 3.2).

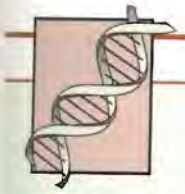
Since restriction endonucleases digest DNA at specific sequences, they can be used to cleave a DNA molecule at unique sites. For example, the restriction endonuclease *Eco*RI recognizes the six-base-pair sequence GAATTC. This sequence is present at five sites in DNA of the bacteriophage

TABLE 3.2 Recognition Sites of Representative Restriction Endonucleases

Enzyme[a]	Source	Recognition site[b]
*Bam*HI	*Bacillus amyloliquefaciens* H	GGATCC
*Eco*RI	*Escherichia coli* RY13	GAATTC
*Hae*III	*Haemophilus aegyptius*	GGCC
*Hind*III	*Haemophilus influenzae* Rd	AAGCTT
*Hpa*I	*Haemophilus parainfluenzae*	GTTAAC
*Hpa*II	*Haemophilus parainfluenzae*	CCGG
*Mbo*I	*Moraxella bovis*	GATC
*Not*I	*Nocardia otitidis-caviarum*	GCGGCCGC
*Sfi*I	*Streptomyces fimbriatus*	GGCCNNNNNGGCC
*Taq*I	*Thermus aquaticus*	TCGA

[a] Enzymes are named according to their species of isolation, followed by a number to distinguish different enzymes isolated from the same organism (e.g., *Hpa*I and *Hpa*II).

[b] Recognition sites show the sequence of only one strand of double-stranded DNA. "N" represents any base.



MOLECULAR MEDICINE

HIV and AIDS

The Disease

Acquired immune deficiency syndrome (AIDS) is a new disease, first described in 1981. It has now become a worldwide pandemic, with more than 50 million people having been infected with HIV and approximately 20 million having died of AIDS. The clinical manifestations of AIDS result principally from failure of the immune system to function normally. In the absence of normal immunity, AIDS patients are sensitive to opportunistic infections by agents (viruses, bacteria, fungi, and protozoans) against which a healthy individual would be resistant. People with AIDS also suffer a high frequency of some types of cancers, particularly lymphomas and Kaposi's sarcoma, although it is the opportunistic infections that are responsible for most deaths.

Molecular and Cellular Basis

AIDS is caused by a retrovirus (human immunodeficiency virus or HIV) that was discovered by the research groups of Robert Gallo and Luc Montagnier in 1983. HIV infects principally a specific type of lymphocyte (the T4 lymphocyte) that is required for a normal immune response. In contrast to many other retroviruses, such as Rous sarcoma virus, HIV does not cause the cells it infects to become cancerous. Instead, HIV eventually kills the cells in which it replicates, ultimately resulting in the depletion of the population of T4 lymphocytes and the failure of the immune system in infected individuals. This failure of the immune system in turn leads to the opportunistic infections and cancers that represent the clinical manifestations of AIDS.

Prevention and Treatment

At present, the only means of preventing AIDS is to avoid HIV infection. HIV is a fragile virus that quickly loses infectivity outside the body, so it cannot be transmitted by casual contact with an infected person. HIV can be transmitted in three ways: through sexual contact, through contaminated blood products, and from mother to child during pregnancy or breast-feeding. Following the isolation of HIV, screening tests were developed to ensure the safety of clotting factors and blood supplies used for transfusions. Prevention of HIV infection by other routes currently depends on individuals minimizing their personal risk of infection by adhering to safe sexual practices and avoiding sources of contaminated blood, such as shared needles used for intravenous drug injection.

Beyond modifying individual behavior to reduce the risk of infection, the identification of HIV as the cause of AIDS opens possibilities for prevention and treatment. A vaccine to prevent HIV infection is being actively pursued, although several features of the biology of HIV pose difficulties to this approach. Alternatively, drugs that inhibit virus replication are now providing effective therapies for HIV-infected individuals. These drugs either act as inhibitors of the HIV reverse transcriptase or of the HIV protease, which is an enzyme required for processing viral proteins. Combinations of such drugs are now prolonging the lives of AIDS patients, although further work is clearly needed to develop drugs that are not only more effective but also less expensive and more practical for use in developing countries.

Scanning electron micrograph of HIV budding from T lymphocytes (Cecil Fox/Photo Researchers, Inc.)

0.1 μm

λ, so *Eco*RI digests λ DNA into six fragments ranging from 3.6 to 21.2 kilobases long (1 kilobase, or kb = 1000 base pairs) (Figure 3.14). These fragments can be separated according to size by **gel electrophoresis**—a common method in which molecules are separated based on the rates of their migration in an electric field. A gel, usually formed from agarose or poly-

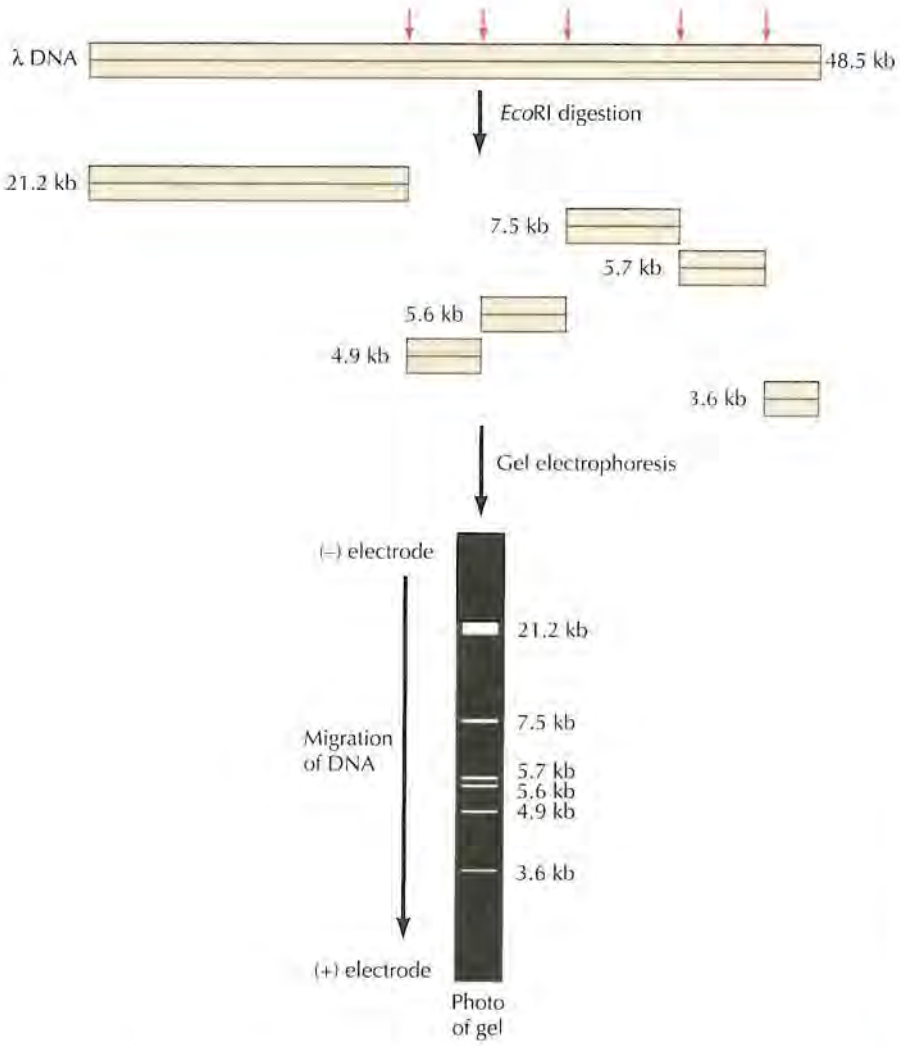


Figure 3.14 *Eco*RI digestion and gel electrophoresis of λ DNA
*Eco*RI cleaves λ DNA at five sites (arrows), yielding six DNA fragments. These fragments are then separated by electrophoresis in an agarose gel. The DNA fragments migrate toward the positive electrode, with smaller fragments moving more rapidly through the gel. Following electrophoresis, the DNA is stained with a fluorescent dye and photographed. The sizes of DNA fragments are indicated.

acrylamide, is placed between two buffer compartments containing electrodes. The sample (e.g., the mixture of DNA fragments to be analyzed) is then pipetted into preformed slots in the gel, and the electric field is turned on. Nucleic acids are negatively charged (because of their phosphate backbone), so they migrate toward the positive electrode. The gel acts like a sieve, selectively retarding the movement of larger molecules. Smaller molecules therefore move through the gel more rapidly, allowing a mixture of nucleic acids to be separated on the basis of size.

In addition to size, the order of restriction fragments can be determined by a variety of methods, yielding (for example) a map of the *Eco*RI sites in λ DNA. The locations of cleavage sites for multiple different restriction endonucleases can be used to generate detailed **restriction maps** of DNA molecules, such as viral genomes (Figure 3.15). In addition, individual DNA fragments produced by restriction endonuclease digestion can be isolated following electrophoresis for further study—including determination of their DNA sequence. The DNAs of many viruses have been characterized by this approach.

Restriction endonuclease digestion alone, however, does not provide sufficient resolution for the analysis of larger DNA molecules, such as cellular genomes. A restriction endonuclease with a six-base-pair recognition site (such as *Eco*RI) cleaves DNA with a statistical frequency of once every 4096

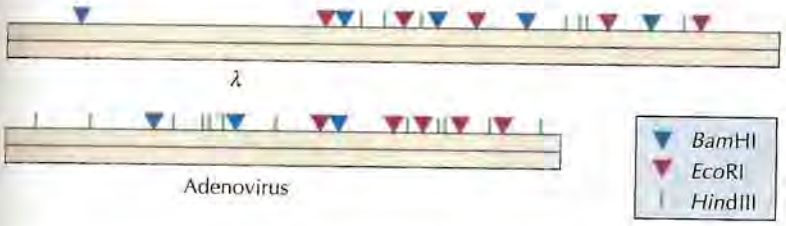


Figure 3.15 Restriction maps of λ and adenovirus DNAs
The locations of cleavage sites for *Bam*HI, *Eco*RI, and *Hind*III are shown in the DNAs of *E. coli* bacteriophage λ (48.5 kb) and human adenovirus-2 (35.9 kb).

base pairs ($1/4^6$). A molecule the size of λ DNA (48.5 kb) would therefore be expected to yield about ten *Eco*RI fragments, consistent with the results illustrated in Figure 3.14. However, restriction endonuclease digestion of larger genomes yields quite different results. For example, the human genome is approximately 3×10^6 kb long and is therefore expected to yield more than 500,000 *Eco*RI fragments. Such a large number of fragments cannot be separated from one another, so agarose gel electrophoresis of *Eco*RI-digested human DNA yields a continuous smear rather than a discrete pattern of DNA fragments. Because it is impossible to isolate single restriction fragments from such digests, restriction endonuclease digestion alone does not yield a source of homogeneous DNA suitable for further analysis. Quantities of such purified DNA fragments, however, can be obtained through molecular cloning.

Generation of Recombinant DNA Molecules

The basic strategy in **molecular cloning** is to insert a DNA fragment of interest (e.g., a segment of human DNA) into a DNA molecule (called a **vector**) that is capable of independent replication in a host cell. The result is a **recombinant molecule** or **molecular clone**, composed of the DNA insert linked to vector DNA sequences. Large quantities of the inserted DNA can be obtained if the recombinant molecule is allowed to replicate in an appropriate host. For example, fragments of human DNA can be cloned in bacteriophage λ vectors (Figure 3.16). These recombinant molecules can then be introduced into *E. coli*, where they replicate efficiently to yield millions of progeny phages containing the human DNA insert. The DNA of these phages can then be isolated, yielding large quantities of recombinant molecules containing a single fragment of human DNA. Whereas this fragment might represent one part in 100,000 of human genomic DNA, it represents approximately one part in 10 after being cloned in the λ vector. Moreover, the fragment can be easily isolated from the rest of the vector DNA by restriction endonuclease digestion and gel electrophoresis, allowing a pure fragment of human DNA to be analyzed and further manipulated.

The DNA fragments used to create recombinant molecules are usually generated by digestion with restriction endonucleases. Many of these enzymes cleave their recognition sequences at staggered sites, leaving overhanging or cohesive single-stranded tails that can associate with each other by complementary base pairing (Figure 3.17). The association between such paired complementary ends can be established permanently by treatment with **DNA ligase**, an enzyme that seals breaks in DNA strands (see Chapter 5). Thus, two different fragments of DNA (e.g., a human DNA insert and a λ DNA vector) prepared by digestion with the same restriction endonuclease can be readily joined to create a recombinant DNA molecule.

The fragments of DNA that can be cloned are not limited to those that terminate in restriction endonuclease cleavage sites. Synthetic DNA "linkers" containing a variety of restriction endonuclease sites can be added to

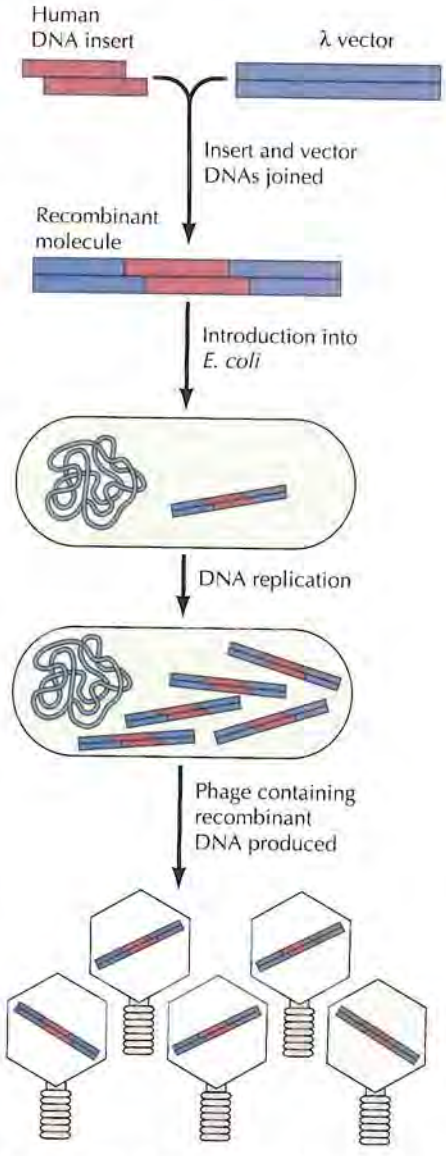


Figure 3.16 Generation of a recombinant DNA molecule
A fragment of human DNA is inserted into a λ DNA vector. The resulting recombinant molecule is then introduced into *E. coli*, where it replicates to yield recombinant progeny phage containing the human DNA insert.

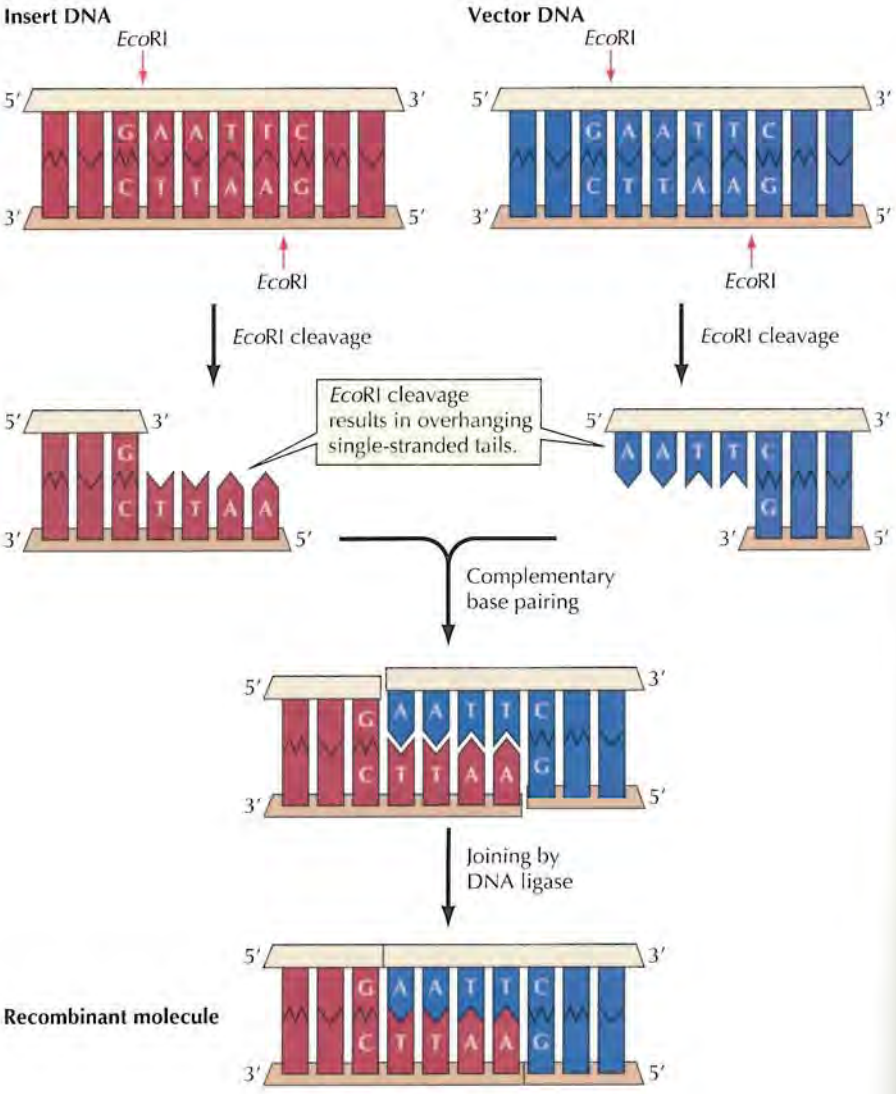


Figure 3.17 Joining of DNA molecules
Vector and insert DNAs are digested with a restriction endonuclease (such as *Eco*RI), which cleaves at staggered sites leaving overhanging single-stranded tails. Vector and insert DNAs can then associate by complementary base pairing, and covalent joining of the DNA strands by DNA ligase yields a recombinant molecule.

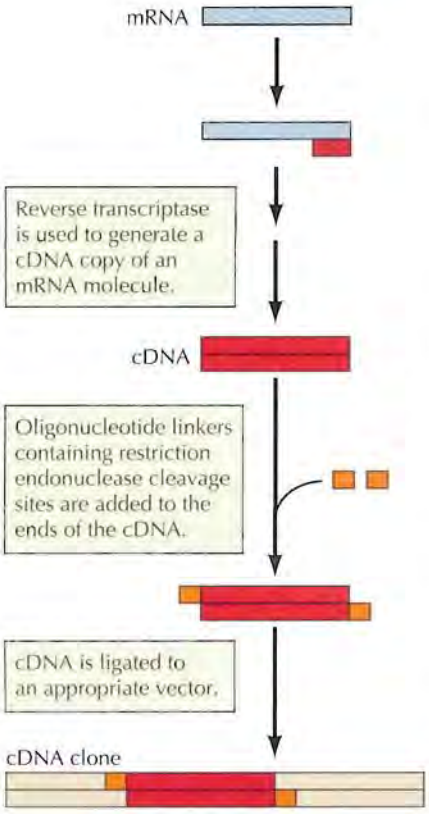


Figure 3.18 cDNA cloning

the blunt ends of any DNA fragment. Linkers are short oligonucleotides that can be readily obtained by chemical synthesis, allowing virtually any fragment of DNA to be prepared for ligation to a vector.

Not only DNA, but also RNA sequences can be cloned (Figure 3.18). The first step is to synthesize a DNA copy of the RNA using the enzyme reverse transcriptase. The DNA product (called a **cDNA** because it is complementary to the RNA used as a template) can then be ligated to vector DNA as already described. Since eukaryotic genes are usually interrupted by noncoding sequences (introns; see Chapter 4), which are removed from mRNA by splicing, the ability to clone cDNA as well as genomic DNA has been critical for understanding gene structure and function.

Vectors for Recombinant DNA

Depending on the size of the insert DNA and the purpose of the experiment, many different types of cloning vectors can be used for the generation of recombinant molecules. The basic vector systems used for the isolation and propagation of cloned DNAs are reviewed here. Other vectors developed for the expression of cloned DNAs and the introduction of

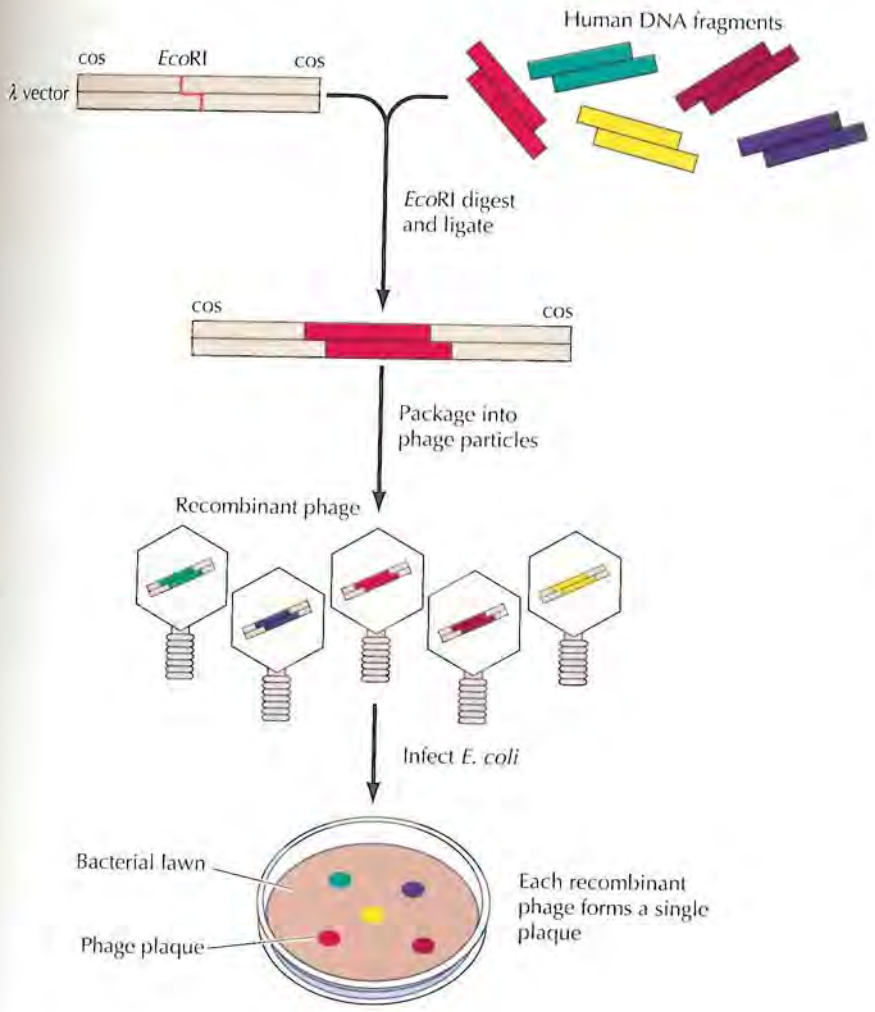


Figure 3.19 Cloning in bacteriophage λ vectors
The vector contains a restriction site (e.g., an *Eco*RI site) for insertion of cloned DNA. In addition, cos sites (cohesive ends), which are required for packaging DNA into phage particles, are present on both ends of the vector DNA. Insert DNA (e.g., human DNA) is ligated to the vector, and the recombinant molecules are packaged into the phage particles by being mixed with phage proteins. The recombinant phages are then used to infect *E. coli*. Each recombinant phage, which carries a unique insert of cloned DNA, forms a single plaque in the infected bacterial culture. Progeny phage carrying unique DNA inserts can then be isolated from individual plaques and grown in large quantities.

recombinant molecules into eukaryotic cells are discussed in subsequent sections.

Bacteriophage λ vectors are frequently used for the initial isolation of either genomic or cDNA clones from eukaryotic cells (Figure 3.19). In λ cloning vectors, sequences of the bacteriophage genome that are dispensable for virus replication have been removed and replaced with unique restriction sites for insertion of cloned DNA. DNA inserts can be as large as about 15 kb and still yield a recombinant genome that can be packaged into phage particles. To isolate genomic clones of human DNA, for example, random fragments of human DNA with an average size of about 15 kb are ligated to λ vector arms. These recombinant DNA molecules can then be efficiently packaged into phage particles by mixing DNA with λ proteins (called packaging extracts) *in vitro*. The phage particles are then used to infect cultures of *E. coli*. Since each recombinant phage forms a single plaque, recombinants carrying unique inserts of human DNA can be isolated. In addition, recombinant phages containing particular genes of interest can be identified by nucleic acid hybridization or other screening methods, as discussed in the next section.

Plasmid vectors (Figure 3.20) allow easier manipulation of cloned DNA sequences than do phage vectors. Plasmids are small circular DNA molecules that can replicate independently—without being associated with

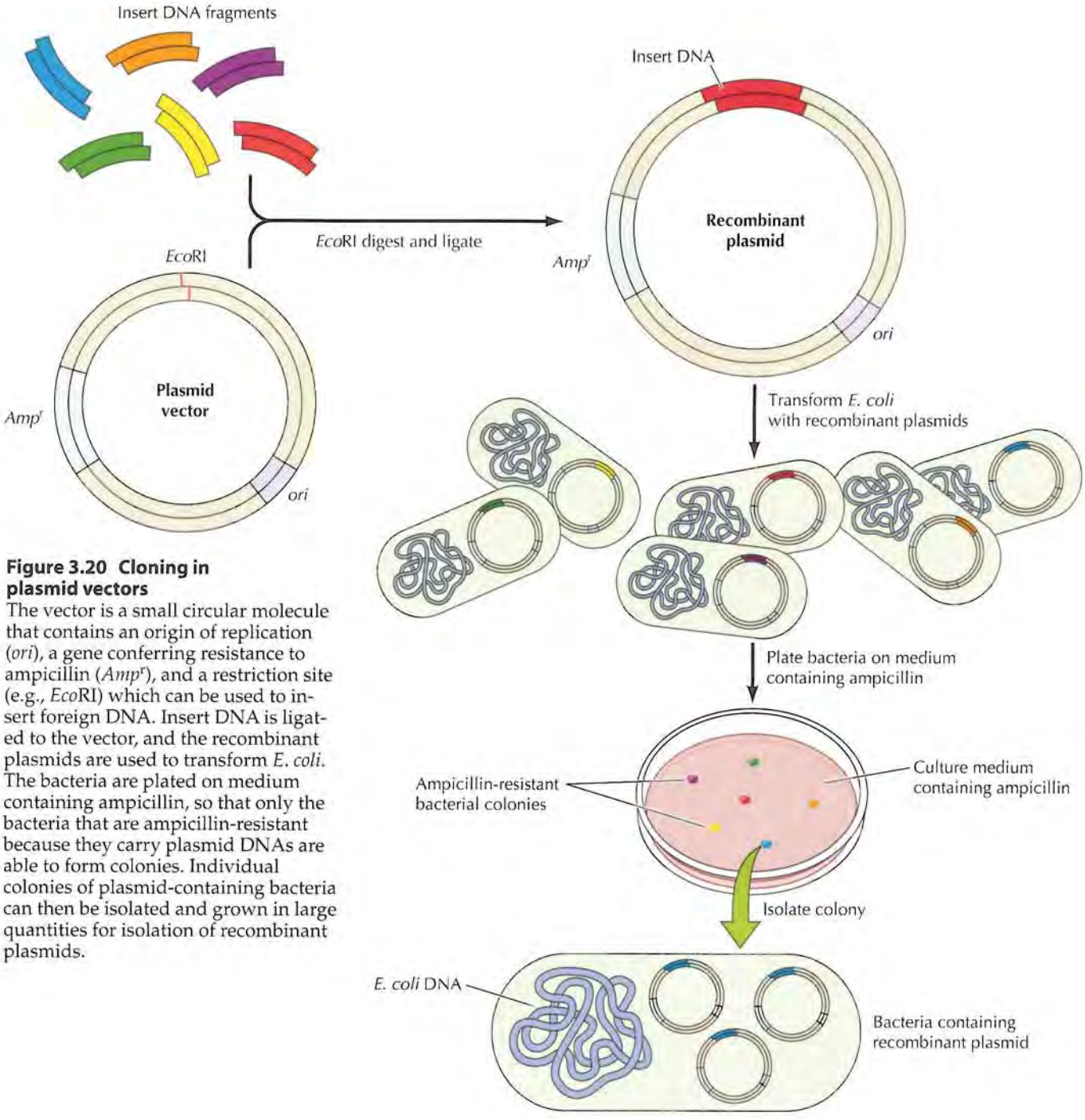


Figure 3.20 Cloning in plasmid vectors
The vector is a small circular molecule that contains an origin of replication (*ori*), a gene conferring resistance to ampicillin (Amp^r), and a restriction site (e.g., *Eco*RI) which can be used to insert foreign DNA. Insert DNA is ligated to the vector, and the recombinant plasmids are used to transform *E. coli*. The bacteria are plated on medium containing ampicillin, so that only the bacteria that are ampicillin-resistant because they carry plasmid DNAs are able to form colonies. Individual colonies of plasmid-containing bacteria can then be isolated and grown in large quantities for isolation of recombinant plasmids.

chromosomal DNA—in bacteria. All that is required on the plasmid DNA is an **origin of replication**—the DNA sequence that signals the host cell DNA polymerase to replicate the DNA molecule. In addition, plasmid vectors carry genes that confer resistance to antibiotics (e.g., ampicillin resistance), so bacteria carrying the plasmids can be selected. Plasmid vectors usually consist of only 2 to 4 kb of DNA, in contrast to the 30 to 45 kb of phage DNA present in λ vectors, facilitating the analysis of an inserted DNA fragment.

To be cloned into a plasmid vector, a fragment of the insert DNA is ligated to an appropriate restriction site in the vector and the recombinant molecule is used to transform *E. coli*. Antibiotic-resistant colonies, which contain plasmid DNA, are selected. Such plasmid-containing bacteria can then be grown in large quantities and their DNA extracted. The small circular plasmid DNA molecules, of which there are often hundreds of copies per cell, can be separated from the bacterial chromosomal DNA; the result is purified plasmid DNA that is suitable for analysis of the cloned insert.

For many studies involving analysis of genomic DNA, it is desirable to clone even larger fragments of DNA than are accommodated by λ vectors. There are five major types of vectors that are used for this purpose (Table 3.3). **Cosmid** vectors accommodate inserts of approximately 45 kb. These vectors contain bacteriophage λ sequences that allow efficient packaging of the cloned DNA into phage particles. In addition, cosmids contain origins of replication and the genes for antibiotic resistance that are characteristic of plasmids, so they are able to replicate as plasmids in bacterial cells. Two other types of vectors are derived from bacteriophage P1, rather than from bacteriophage λ. Bacteriophage P1 vectors, which will accommodate DNA fragments of 70–100 kb, contain sequences that allow recombinant molecules to be packaged *in vitro* into P1 phage particles and then to be replicated as plasmids in *E. coli*. **P1 artificial chromosome (PAC)** vectors also contain sequences of bacteriophage P1, but are introduced directly as plasmids into *E. coli* and will accommodate larger inserts of 130–150 kb. **Bacterial artificial chromosome (BAC)** vectors are derived from a naturally-occurring plasmid of *E. coli* (called the F factor). The replication origin and other F factor sequences allow BACs to replicate as stable plasmids carrying inserts of 120–300 kb. Even larger fragments of DNA (250–400 kb) can be cloned in **yeast artificial chromosome (YAC)** vectors. These vectors contain yeast origins of replication as well as other sequences (centromeres and telomeres, discussed in chapter 4) that allow them to replicate as linear chromosome-like molecules in yeast cells.

DNA Sequencing

Molecular cloning allows the isolation of individual fragments of DNA in quantities suitable for detailed characterization, including the determination of nucleotide sequence. Indeed, determination of the nucleotide sequences of many genes has elucidated not only the structure of their protein products, but also the properties of DNA sequences that regulate gene expression. Furthermore, the coding sequences of novel genes are frequently related to those of previously studied genes, and the functions of newly isolated genes can often be correctly deduced on the basis of such sequence similarities.

Current methods of DNA sequencing are both rapid and accurate, and determining the sequence of several kilobases of DNA is a straightforward

TABLE 3.3 Vectors for Cloning Large Fragments of DNA

Vector	DNA Insert (kb)	Host cell
Cosmids	30–45	*E. coli*
Bacteriophage P1	70–100	*E. coli*
P1 Artificial Chromosomes (PACs)	130–150	*E. coli*
Bacterial Artificial Chromosomes (BACs)	120–300	*E. coli*
Yeast Artificial Chromsomes (YACs)	250–400	Yeast

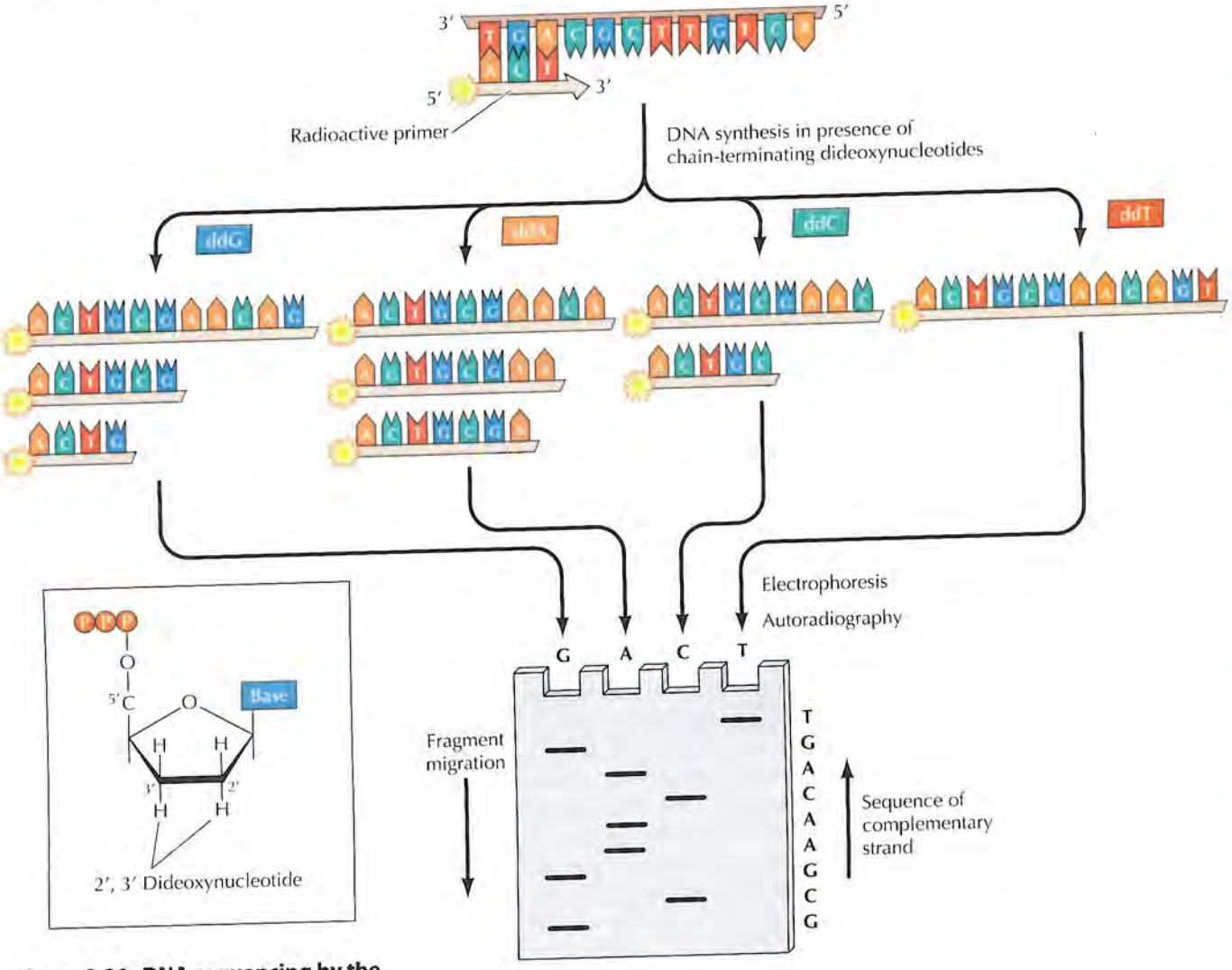


Figure 3.21 DNA sequencing by the Sanger procedure
Dideoxynucleotides, which lack OH groups at the 3′ as well as the 2′ position of deoxyribose, are used to terminate DNA synthesis at specific bases. These molecules are incorporated normally into growing DNA strands. Because they lack a 3′ OH, however, the next nucleotide cannot be added, so synthesis of that DNA strand terminates. DNA synthesis is initiated with a radioactive primer. Four separate reactions are carried out, each containing one dideoxynucleotide mixed with its normal counterpart as well as the three other normal deoxynucleotides. When the dideoxynucleotide is incorporated, DNA synthesis stops, so each reaction yields a series of products extending from the radioactive primer to the base substituted by a dideoxynucleotide. Products of the four reactions are separated by electrophoresis and analyzed by autoradiography to determine the DNA sequence.

task. Thus, it is now far easier to clone and sequence DNA than it is to determine the amino acid sequence of a protein. Since the nucleotide sequence of a gene can be readily translated into the amino acid sequence of its encoded protein, the easiest way of determining protein sequence is the sequencing of a cloned gene or cDNA.

The most common method of DNA sequencing is based on premature termination of DNA synthesis resulting from the inclusion of chain-terminating **dideoxynucleotides** (which do not contain the deoxyribose 3′ hydroxyl group) in DNA polymerase reactions (Figure 3.21). DNA synthesis is initiated from a primer that has been labeled at one end with a radioisotope. Four separate reactions are run, each including one dideoxynucleotide (either A, C, G, or T) in addition to its normal counterpart. Incorporation of a dideoxynucleotide stops further DNA synthesis because no 3′ hydroxyl group is available for addition of the next nucleotide. Thus, a series of labeled DNA molecules is generated, each terminating at the base represented by the dideoxynucleotide in each reaction. These fragments of DNA are then separated according to size by gel electrophoresis and detected by exposure of the gel to X-ray film (**autoradiography**). The size of each fragment is determined by its terminal dideoxynu-

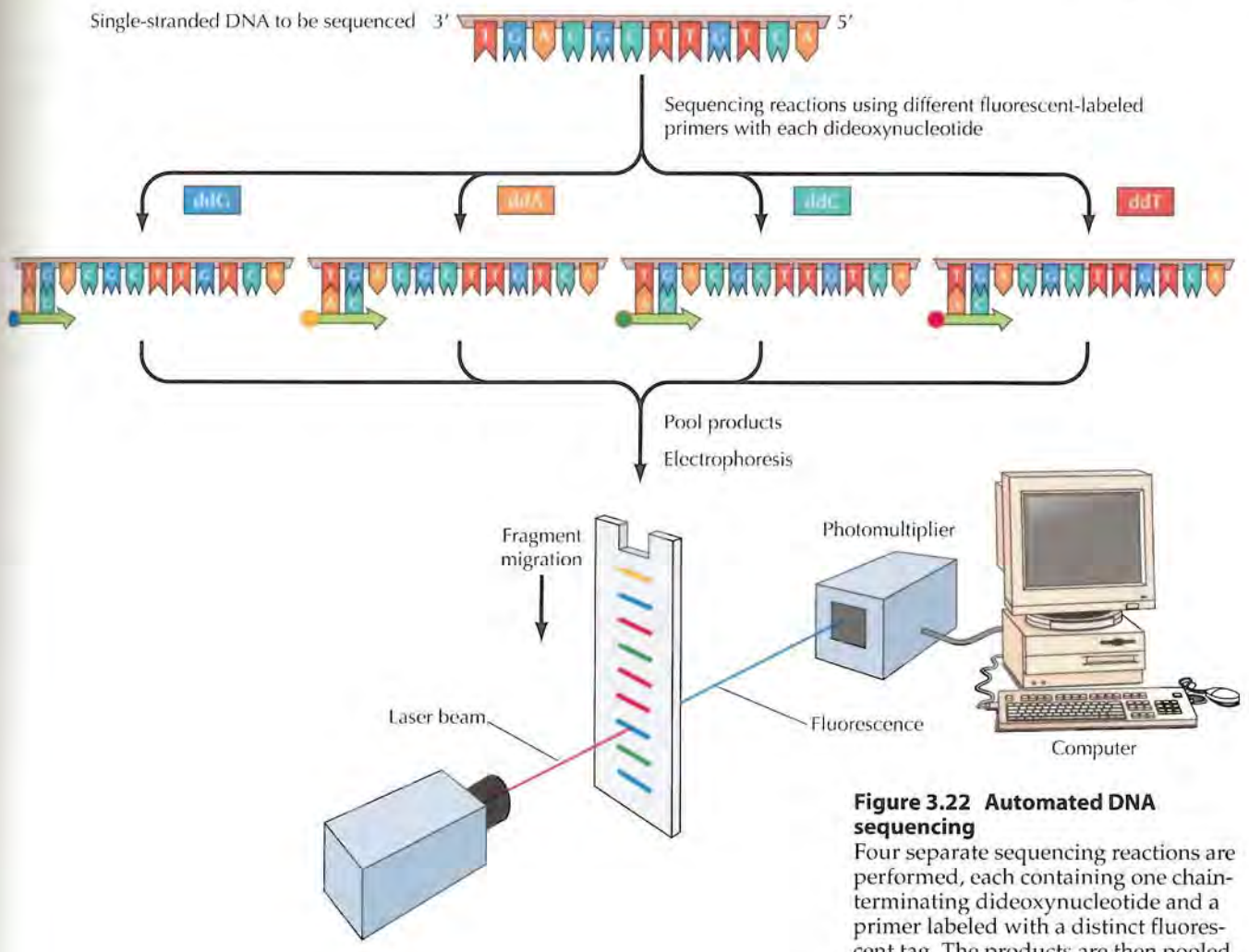


Figure 3.22 Automated DNA sequencing
Four separate sequencing reactions are performed, each containing one chain-terminating dideoxynucleotide and a primer labeled with a distinct fluorescent tag. The products are then pooled and subjected to gel electrophoresis. As the DNA strands migrate through the gel, they pass through a laser beam that excites the fluorescent label. The emitted light is detected by a photomultiplier, which is connected to a computer that collects and analyzes the data.

cleotide, so the DNA sequence corresponds to the order of fragments read from the gel.

Large-scale DNA sequencing is frequently performed using automated systems, which use fluorescent-labeled primers in dideoxynucleotide sequencing reactions (Figure 3.22). As the newly synthesized DNA strands are electrophoresed through a gel, they pass through a laser beam that excites the fluorescent label. The resulting emitted light is then detected by a photomultiplier, and a computer collects and analyzes the data. This type of automated DNA sequencing has enabled the large-scale analysis required for determination of the complete sequence of the human genome, as well as the genome sequences of a number of species of bacteria, yeast, *Arabidopsis*, *C. elegans*, *Drosophila*, and the mouse.

Expression of Cloned Genes

In addition to enabling determination of the nucleotide sequences of genes—and hence the amino acid sequences of their protein products—molecular cloning has provided new approaches to obtaining large amounts of proteins for structural and functional characterization. Many proteins of interest are present at only low levels in eukaryotic cells and therefore cannot be purified in significant amounts by conventional biochemical techniques. Given a cloned gene, however, this problem can be

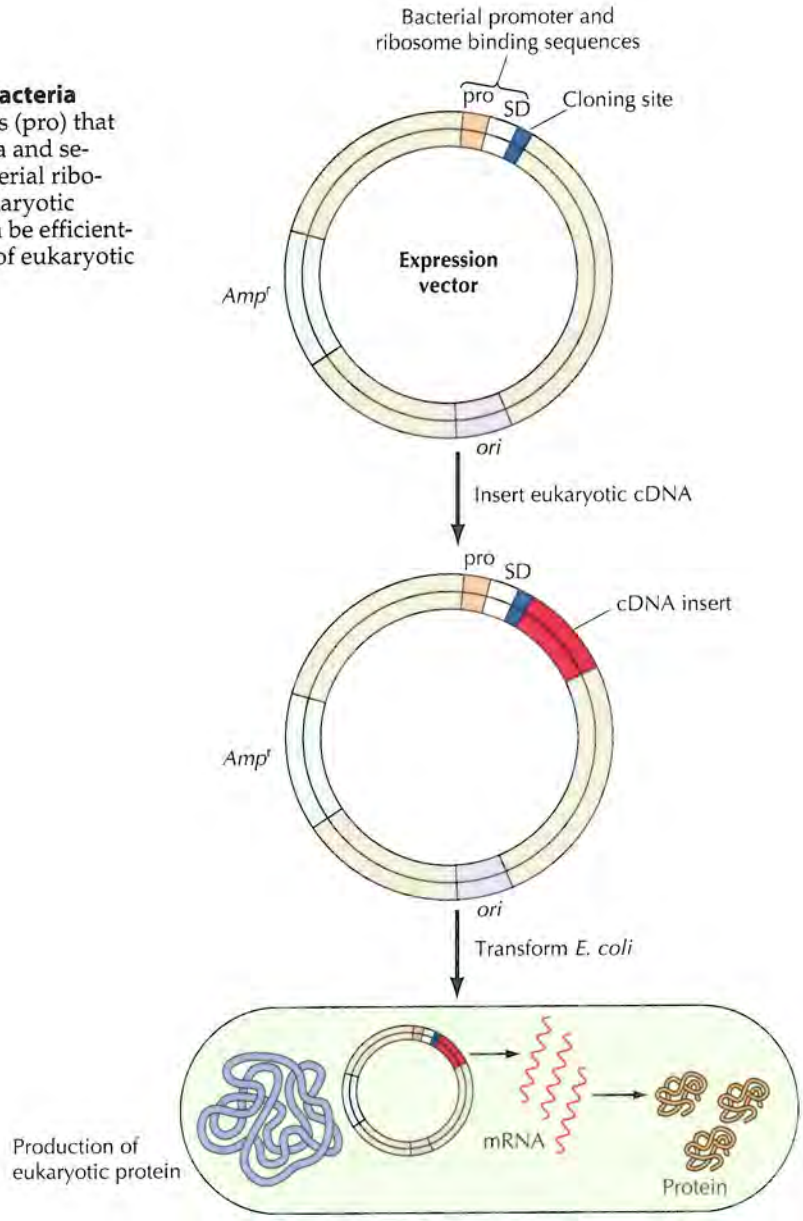


Figure 3.23 Expression of cloned genes in bacteria
Expression vectors contain promoter sequences (pro) that direct transcription of inserted DNA in bacteria and sequences required for binding of mRNA to bacterial ribosomes (Shine-Delgarno [SD] sequences). A eukaryotic cDNA inserted adjacent to these sequences can be efficiently expressed in *E. coli*, resulting in production of eukaryotic proteins in transformed bacteria.

solved by the engineering of vectors that lead to high levels of gene expression in either bacteria or eukaryotic cells.

To express a eukaryotic gene in *E. coli*, the cDNA of interest is cloned into a plasmid or phage vector (called an **expression vector**) that contains sequences that drive transcription and translation of the inserted gene in bacterial cells (Figure 3.23). Inserted genes often can be expressed at levels high enough that the protein encoded by the cloned gene corresponds to as much as 10% of the total bacterial protein. Purifying the protein encoded by the cloned gene in quantities suitable for detailed biochemical or structural studies is then a straightforward matter.

It is frequently useful to express high levels of a cloned gene in eukaryotic cells, rather than in bacteria. This mode of expression may be important, for example, to ensure that posttranslational modifications of the protein (such as addition of carbohydrates or lipids) occur normally. Such protein expression in eukaryotic cells can be achieved, as in *E. coli*, by insertion of the cloned gene into a vector (usually derived from a virus) that

directs high-level gene expression. One system frequently used for protein expression in eukaryotic cells is infection of insect cells by **baculovirus** vectors, which direct very high levels of expression of genes inserted in place of a viral structural protein. Alternatively, high levels of protein expression can be achieved using appropriate vectors in mammalian cells. Expression of cloned genes in yeast is particularly useful because simple methods of yeast genetics can be employed to identify proteins that interact with other cloned proteins or with specific DNA sequences.

Amplification of DNA by the Polymerase Chain Reaction

Molecular cloning allows individual DNA fragments to be propagated in bacteria and isolated in large amounts. An alternative method to isolating large amounts of a single DNA molecule is the **polymerase chain reaction (PCR)**, which was developed by Kary Mullis in 1988. Provided that some sequence of the DNA molecule is known, PCR can achieve a striking amplification of DNA via reactions carried out entirely *in vitro*. Essentially, DNA polymerase is used for repeated replication of a defined segment of DNA. The number of DNA molecules increases exponentially, doubling with each round of replication, so a substantial quantity of DNA can be obtained from a small number of initial template copies. For example, a single DNA molecule amplified through 30 cycles of replication would theoretically yield 2^{30} (approximately 1 billion) progeny molecules. Single DNA molecules can thus be amplified to yield readily detectable quantities of DNA that can be isolated by molecular cloning or further analyzed directly by restriction endonuclease digestion or nucleotide sequencing.

The general procedure for PCR amplification of DNA is illustrated in Figure 3.24. The starting material can be either a cloned DNA fragment or a mixture of DNA molecules—for example, total DNA from human cells. A specific region of DNA can be amplified from such a mixture, provided that the nucleotide sequence surrounding the region is known so that primers can be designed to initiate DNA synthesis at the desired point. Such primers are usually chemically synthesized oligonucleotides containing 15 to 20 bases of DNA. Two primers are used to initiate DNA synthesis in opposite directions from complementary DNA strands. The reaction is started by heating the template DNA to a high temperature (e.g., 95°C) so that the two strands separate. The temperature is then lowered to allow the primers to pair with their complementary sequences on the template strands. DNA polymerase then uses the primers to synthesize a new strand complementary to each template. Thus in one cycle of amplification, two new DNA molecules are synthesized from one template molecule. The process can be repeated multiple times, with a twofold increase in DNA molecules resulting from each round of replication.

The multiple cycles of heating and cooling involved in PCR are performed by programmable heating blocks called thermocyclers. The DNA polymerases used in these reactions are heat-stable enzymes from bacteria such as *Thermus aquaticus*, which lives in hot springs at temperatures of about 75°C. These polymerases are stable even at the high temperatures used to separate the strands of double-stranded DNA, so PCR amplification can be performed rapidly and automatically. RNA sequences can also be amplified by this method if reverse transcriptase is used to synthesize a cDNA copy prior to PCR amplification.

If enough of the sequence of a gene is known that primers can be specified, PCR amplification provides an extremely powerful method of obtaining readily detectable and manipulable amounts of DNA from starting

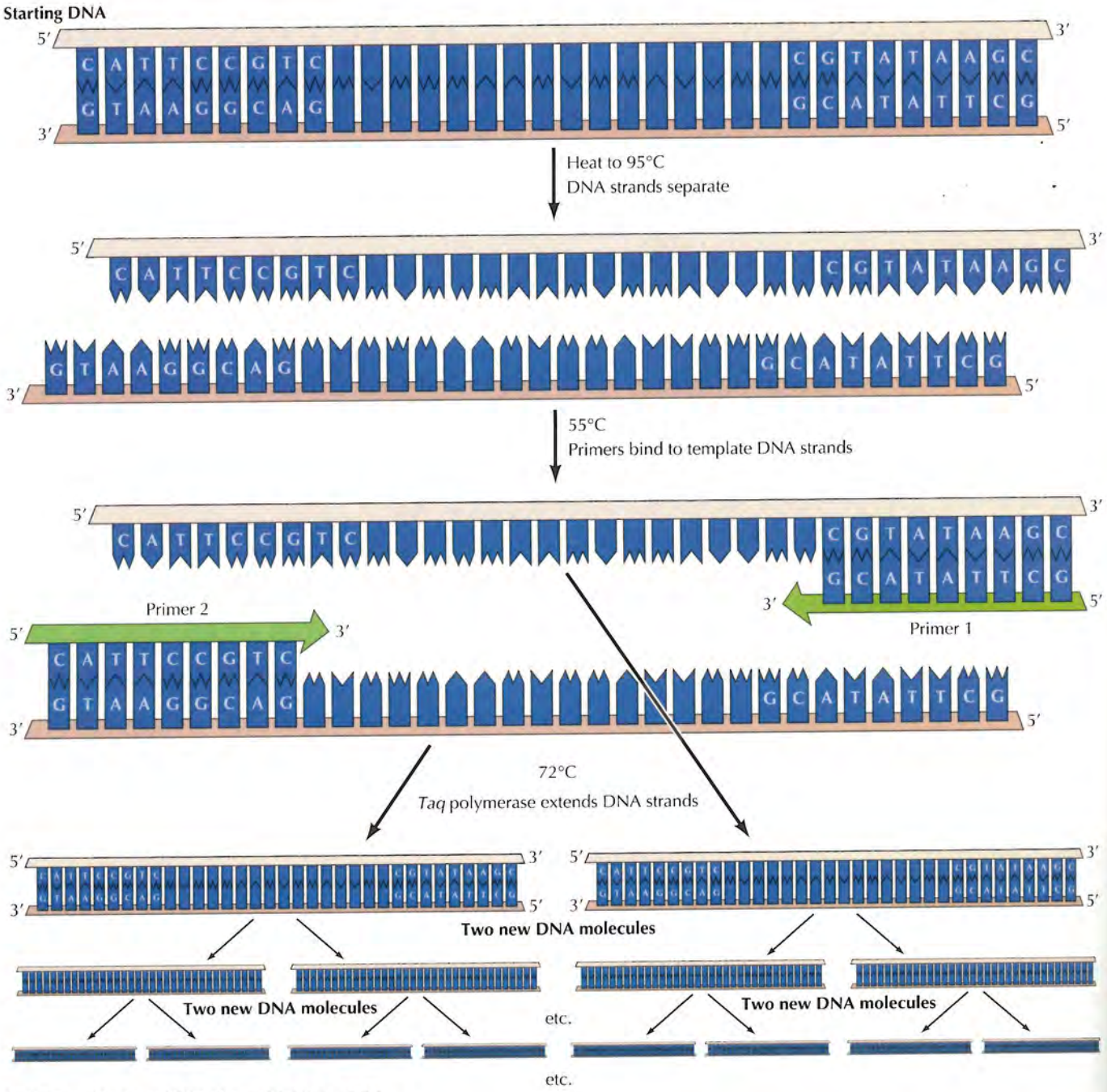


Figure 3.24 Amplification of DNA by PCR
The region of DNA to be amplified is flanked by two sequences used to prime DNA synthesis. The starting double-stranded DNA is heated to separate the strands and then cooled to allow primers (usually oligonucleotides of 15 to 20 bases) to bind to each strand of DNA. DNA polymerase from *Thermus aquaticus* (*Taq* polymerase) is used to synthesize new DNA strands starting from the primers, resulting in the formation of two new DNA molecules. The process can be repeated for multiple cycles, each resulting in a twofold amplification of DNA.

material that may contain only a few molecules of the desired DNA sequence in a complex mixture of other molecules. For example, defined DNA sequences of up to several kilobases can be readily amplified from total genomic DNA, or a single cDNA can be amplified from total cell RNA. These amplified DNA segments can then be further manipulated or analyzed, for example, to detect mutations within a gene of interest. PCR is thus a powerful addition to the repertoire of recombinant DNA techniques. Its power is particularly apparent in applications such as the diagnosis of inherited diseases, studies of gene expression during development, and forensic medicine.

Detection of Nucleic Acids and Proteins

The advent of molecular cloning has enabled the isolation and characterization of individual genes from eukaryotic cells. Understanding the role of genes within cells, however, requires analysis of the intracellular organization and expression of individual genes and their encoded proteins. In this section, the basic procedures currently available for detection of specific nucleic acids and proteins are discussed. These approaches are important for a wide variety of studies, including the mapping of genes to chromosomes, the analysis of gene expression, and the localization of proteins to subcellular organelles. The same general procedures are also used to isolate specific genes as molecular clones.

Nucleic Acid Hybridization

The key to detection of specific nucleic acid sequences is base pairing between complementary strands of RNA or DNA. At high temperatures (e.g., 90 to 100°C), the complementary strands of DNA separate (denature), yielding single-stranded molecules. If such denatured DNA strands are then incubated under appropriate conditions (e.g., 65°), they will renature to form double-stranded molecules as dictated by complementary base pairing—a process called **nucleic acid hybridization**. Nucleic acid hybrids can be formed between two strands of DNA, two strands of RNA, or one strand of DNA and one of RNA.

Nucleic acid hybridization provides a means for detecting DNA or RNA sequences that are complementary to any isolated nucleic acid, such as a viral genome or a cloned DNA sequence (Figure 3.25). The cloned DNA is radiolabeled, usually by being synthesized in the presence of radioactive nucleotides. This radioactive DNA is then used as a **probe** for hybridization to complementary DNA or RNA sequences, which are detected by virtue of the radioactivity of the resulting double-stranded hybrids.

Southern blotting (a technique developed by E. M. Southern) is widely used for detection of specific genes in cellular DNA (Figure 3.26). The DNA to be analyzed is digested with a restriction endonuclease, and the digested DNA fragments are separated by gel electrophoresis. The gel is then overlaid with a nitrocellulose filter or nylon membrane, to which the DNA fragments are transferred (blotted) to yield a replica of the gel. The filter is then incubated with a radiolabeled probe, which hybridizes to the DNA fragments that contain the complementary sequence. These fragments are then visualized by exposure of the filter to X-ray film.

Northern blotting is a variation of the Southern blotting technique (hence its name) that is used for detection of RNA instead of DNA. In this method, total cellular RNAs are extracted and fractionated according to size by gel electrophoresis. As in Southern blotting, the RNAs are transferred to a filter and detected by hybridization with a radioactive probe. Northern blotting is frequently used in studies of gene expression—for example, to determine whether specific mRNAs are present in different types of cells.

Rather than analyzing one gene at a time, as in Southern or Northern blotting, hybridization to **DNA microarrays** allows tens of thousands of genes to be analyzed simultaneously. As the complete sequences of eukaryotic genomes have become available, hybridization to DNA microarrays has enabled researchers to undertake global analyses of sequences present in either cellular DNA or RNA samples. A DNA microarray consists of a glass slide or membrane filter onto which oligonucleotides or fragments of cDNAs are printed by a robotic system in small spots at a high density (Fig-

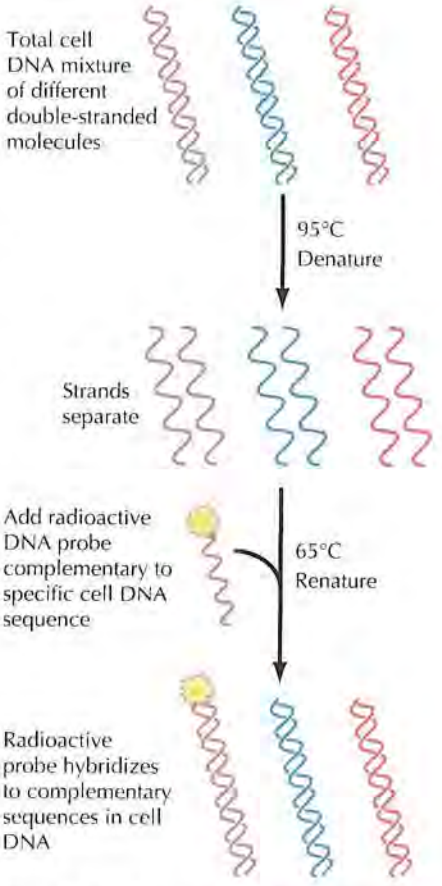


Figure 3.25 Detection of DNA by nucleic acid hybridization
A specific sequence can be detected in total cell DNA by hybridization with a radiolabeled DNA probe. The DNA is denatured by heating to 95°C, yielding single-stranded molecules. The radiolabeled probe is then added and the temperature is lowered to 65°C, allowing complementary DNA strands to renature by pairing with each other. The radioactive probe hybridizes to complementary sequences in cell DNA, which can then be detected as radioactive double-stranded molecules.

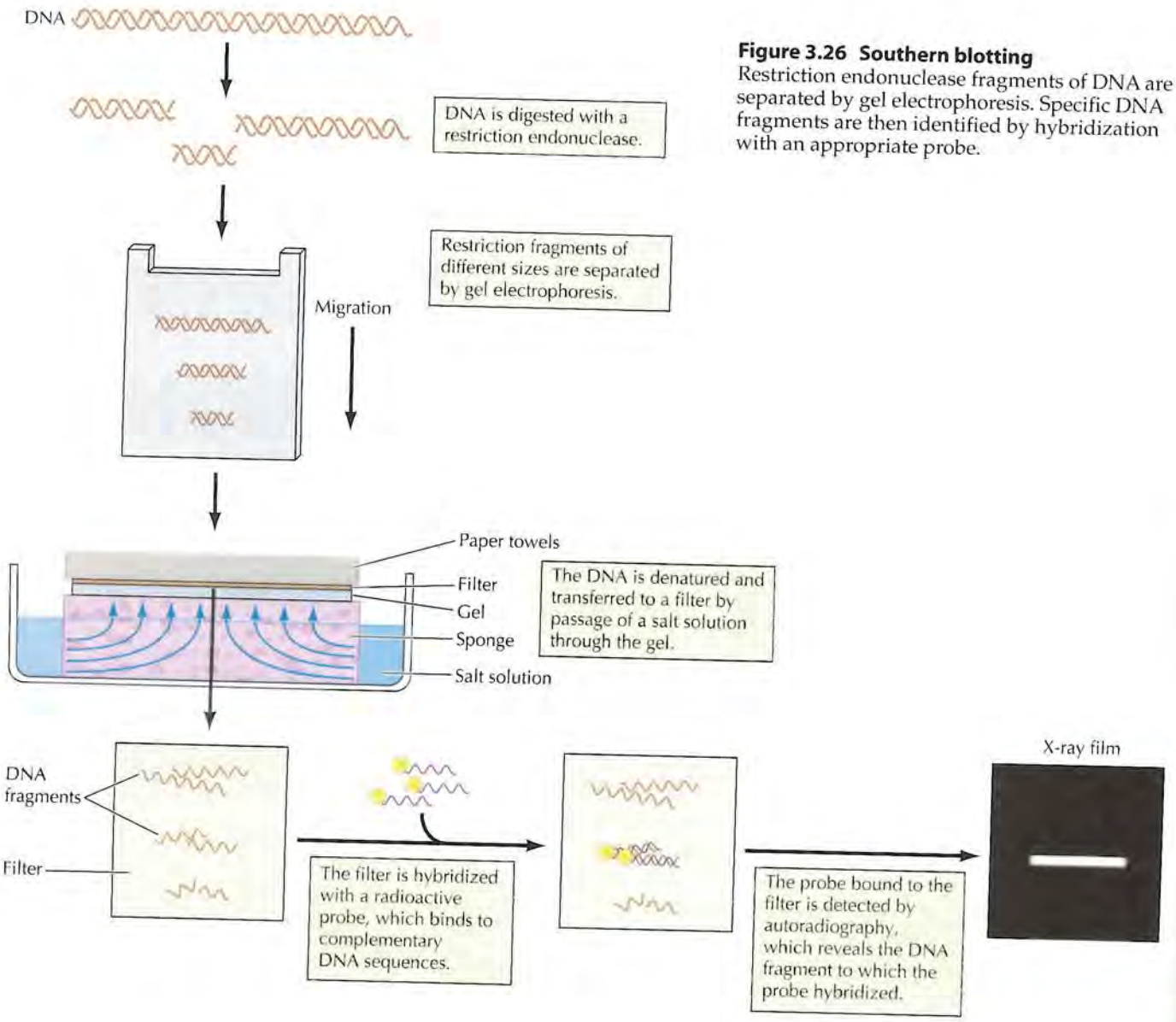


Figure 3.26 Southern blotting
Restriction endonuclease fragments of DNA are separated by gel electrophoresis. Specific DNA fragments are then identified by hybridization with an appropriate probe.

ure 3.27). Each spot on the array consists of a single oligonucleotide or cDNA. More than 10,000 unique DNA sequences can be printed onto a typical glass microscope slide, so it is readily possible to produce DNA microarrays containing sequences representing all of the genes in cellular genomes. As illustrated in Figure 3.27, one widespread application of DNA microarrays is in studies of gene expression, for example a comparison of the genes expressed by two different types of cells. In an experiment of this type, cDNA probes are synthesized from the mRNAs expressed in each of the two cell types (e.g., cancer cells and normal cells). The two cDNAs are labeled with different fluorescent dyes (typically red and green), and a mixture of the cDNAs is hybridized to a DNA microarray in which 10,000 or more human genes are represented as single spots. The array is then analyzed using a high-resolution laser scanner, and the relative extent of transcription of each gene in the cancer cells compared to the normal cells is indicated by the ratio of red to green fluorescence at the appropriate spot on the array.

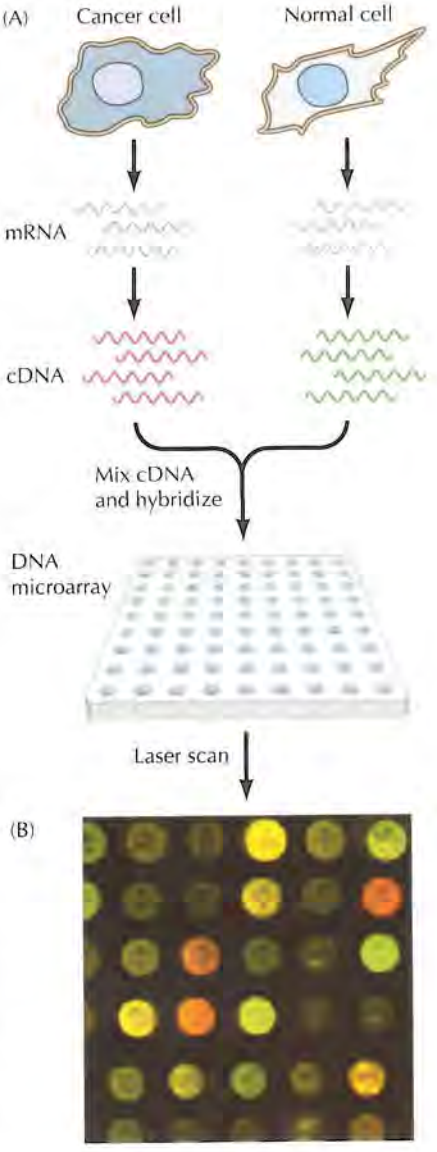


Figure 3.27 DNA microarrays
(A) An example of comparative analysis of gene expression in cancer cells and normal cells. mRNAs extracted from cancer cells and normal cells are used as templates for synthesis of cDNA probes labeled with different fluorescent dyes (e.g., a red fluorescent label for cancer cell cDNA and green for normal cell cDNA). The two cDNA probes are mixed and hybridized to a DNA microarray containing spots of oligonucleotides corresponding to 10,000 or more distinct human genes. The relative level of expression of each gene in cancer cells compared to normal cells is indicated by the ratio of red to green fluorescence at each position on the microarray. (B) Photograph of a portion of a microarray.

Nucleic acid hybridization can be used to detect homologous DNA or RNA sequences not only in cell extracts, but also in chromosomes or intact cells—a procedure called ***in situ* hybridization** (Figure 3.28). In this case, the hybridization of radioactive or fluorescent probes to specific cells or subcellular structures is analyzed by microscopic examination. For example, labeled probes can be hybridized to intact chromosomes in order to identify the chromosomal regions that contain a gene of interest. *In situ* hybridization can also be used to detect specific mRNAs in different types of cells within a tissue.

Detection of Small Amounts of DNA or RNA by PCR

Amplification of DNA by the polymerase chain reaction is a much more sensitive technique for detecting cellular DNA or RNA sequences than is Southern or Northern blotting. Approximately 100,000 copies of a DNA or RNA sequence are required for detection by blot hybridization. In contrast, PCR can amplify single copies of DNA (or RNA after reverse transcription) to readily detectable levels.

As discussed earlier, the specificity of amplification in PCR is provided by the use of oligonucleotide primers that hybridize to complementary sequences on the template molecule. Therefore, PCR can selectively amplify a specific DNA molecule from complex mixtures, such as total cell DNA or RNA. Consequently, PCR amplification can be used to detect specific DNA or RNA molecules in very small amounts of starting material, such as extracts of single cells. This extraordinary sensitivity has made PCR an important method for a variety of applications, including the analysis of gene expression in cells available in only limited quantities.

Antibodies as Probes for Proteins

Studies of gene expression and function require the detection not only of DNA and RNA, but also of specific proteins. For these studies, **antibodies** take the place of nucleic acid probes as reagents that can selectively react with unique protein molecules. Antibodies are proteins produced by cells of the immune system (B lymphocytes) that react against molecules (**antigens**) that the host organism recognizes as foreign substances—for example, the protein coat of a virus. The immune systems of vertebrates are capable of producing millions of different antibodies, each of which specifically recognizes a unique antigen, which may be a protein, a carbohydrate, or a nonbiological molecule. An individual lymphocyte produces only a single type of

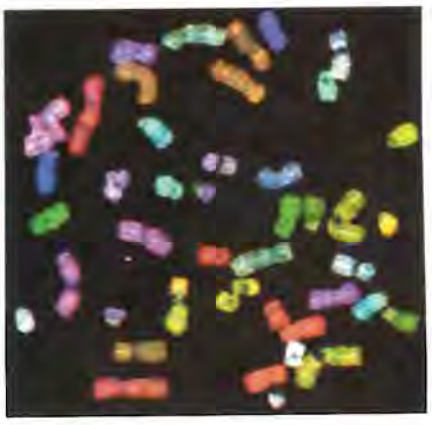
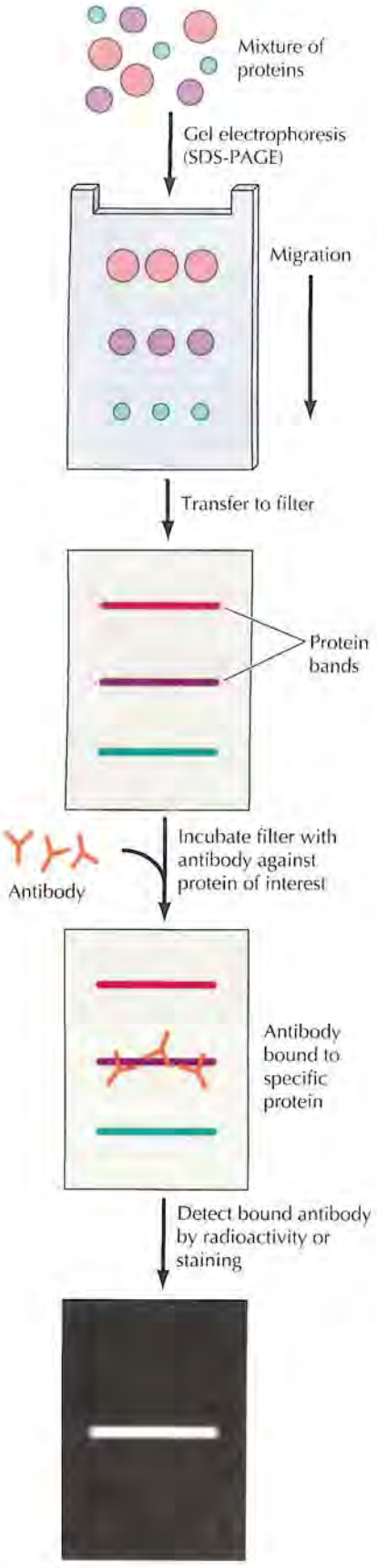


Figure 3.28 Fluorescence *in situ* hybridization
Hybridization of human chromosomes with chromosome-specific fluorescent probes that label each of the 24 chromosomes a different color. (Courtesy of Thomas Reid and Hesed Padilla-Nash, National Cancer Institute.)



antibody, but the antibody genes of different lymphocytes vary as a result of programmed gene rearrangements during development of the immune system (see Chapter 5). This variation gives rise to an array of lymphocytes with distinct antibody genes, programmed to respond to different antigens.

Antibodies can be generated by inoculation of an animal with any foreign protein. For example, antibodies against human proteins are frequently raised in rabbits. The sera of such immunized animals contain a mixture of antibodies (produced by different lymphocytes) that react against multiple sites on the immunizing antigen. However, single species of antibodies (**monoclonal antibodies**) can also be produced by the culturing of clonal lines of B lymphocytes from immunized animals (usually mice). Because each lymphocyte is programmed to produce only a single antibody, a clonal line of lymphocytes produces a monoclonal antibody that recognizes only a single antigenic determinant, thereby providing a highly specific immunological reagent.

Although antibodies can be raised against proteins purified from cells, other materials may also be used for immunization. For example, animals may be immunized with intact cells to raise antibodies against unknown proteins expressed by a specific cell type (e.g., a cancer cell). Such antibodies may then be used to identify proteins specifically expressed by the cell type used for immunization. In addition, antibodies are frequently raised against proteins expressed in bacteria as recombinant clones. In this way, molecular cloning allows the production of antibodies against proteins that may be difficult to isolate from eukaryotic cells. Moreover, antibodies can be raised against synthetic peptides that consist of only 10 to 15 amino acids, rather than against intact proteins. Therefore, once the sequence of a gene is known, antibodies against peptides synthesized from part of its predicted protein sequence can be produced. Because antibodies against these synthetic peptides frequently react with the intact protein as well, it is possible to produce antibodies against a protein starting with only the sequence of a cloned gene.

Antibodies can be used in a variety of ways to detect proteins in cell extracts. Two common methods are **immunoblotting** (also called **Western blotting**) and **immunoprecipitation**. Western blotting (Figure 3.29) is another variation of Southern blotting. Proteins in cell extracts are first separated according to size by gel electrophoresis. Because proteins have different shapes and charges, however, this process requires a modification of the methods used for electrophoresis of nucleic acids. Proteins are separated by a method known as **SDS-polyacrylamide gel electrophoresis (SDS-PAGE)**, in which they are dissolved in a solution containing the negatively charged detergent sodium dodecyl sulfate (SDS). Each protein binds many detergent molecules, which denature the protein and give the protein an overall negative charge. Under these conditions, all proteins migrate toward the positive electrode—their rates of migration determined (like those of nucleic acids) only by size. Following electrophoresis, the proteins are transferred to a filter, which is then allowed to react with antibodies against the protein of interest. The antibody bound to the filter can be

Figure 3.29 Western blotting
Proteins are separated according to size by SDS-polyacrylamide gel electrophoresis and transferred from the gel to a filter. The filter is incubated with an antibody directed against a protein of interest. The antibody bound to the filter can then be detected by reaction with various reagents, such as a radioactive probe that binds to the antibody.

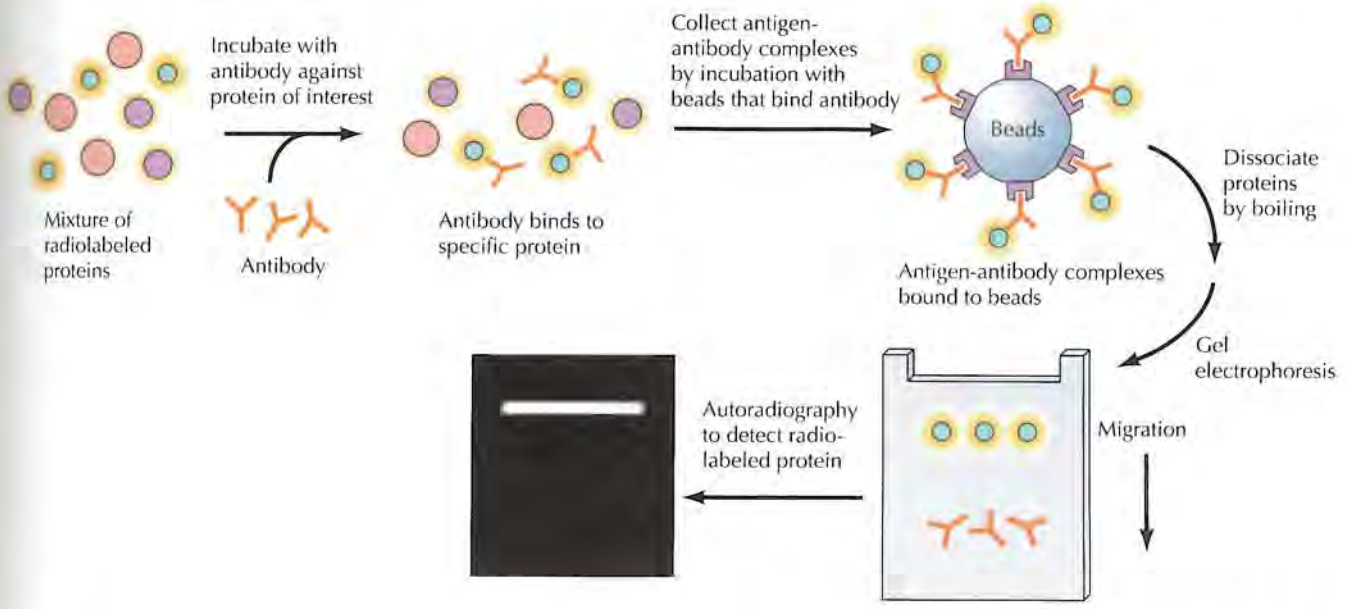


Figure 3.30 Immunoprecipitation Radiolabeled proteins are incubated with an antibody, which forms complexes with the protein against which it is directed (the antigen). These antigen-antibody complexes are collected on beads that bind the antibody. The beads are then boiled to dissociate the antigen-antibody complexes, and the recovered proteins are analyzed by SDS-polyacrylamide gel electrophoresis. The radioactive protein that was immunoprecipitated is detected by autoradiography.

detected by various methods, thereby identifying the protein against which the antibody is targeted.

In immunoprecipitation, antibodies are used to isolate the proteins against which they are directed (Figure 3.30). Typically, cells are incubated with radioactive amino acids to label their proteins. Such a radiolabeled cell extract is then incubated with an antibody, which binds to its antigenic target protein. The resulting antigen-antibody complexes are isolated and subjected to electrophoresis, allowing detection of the radioactive antigen by autoradiography.

As discussed in Chapter 1, antibodies can also be used to visualize proteins within cells, as well as in cell lysates. For example, cells can be stained with antibodies labeled with fluorescent dyes, and the subcellular localization of the antigenic proteins can be visualized by fluorescence microscopy (see Figure 1.28). Antibodies can also be labeled with tags that are visible in the electron microscope, such as heavy metals, allowing visualization of antigenic proteins at the ultrastructural level.

Probes for Screening Recombinant DNA Libraries

The same basic methods for detecting nucleic acids and proteins in cell extracts are used for identifying molecular clones that contain specific cellular DNA inserts. For example, nucleic acid hybridization can be used to identify genomic or cDNA clones that contain DNA sequences for which a probe is available. Cloned cDNAs in expression vectors can also be identified by the use of antibodies against their encoded proteins.

The first step in isolation of either genomic or cDNA clones is frequently the preparation of **recombinant DNA libraries**—collections of clones that contain all the genomic or mRNA sequences of a particular cell type (Figure 3.31). A **genomic library** of human DNA, for example, might be prepared by the cloning of random DNA fragments with average sizes of about 15 kb in a λ vector. Since the size of the human genome is about 3×10^6 kb, the DNA equivalent of one genome would be represented by 200,000 such λ clones. Because of statistical fluctuations in sampling, however, many genes will not be represented in a library of 200,000 recombinants, while other

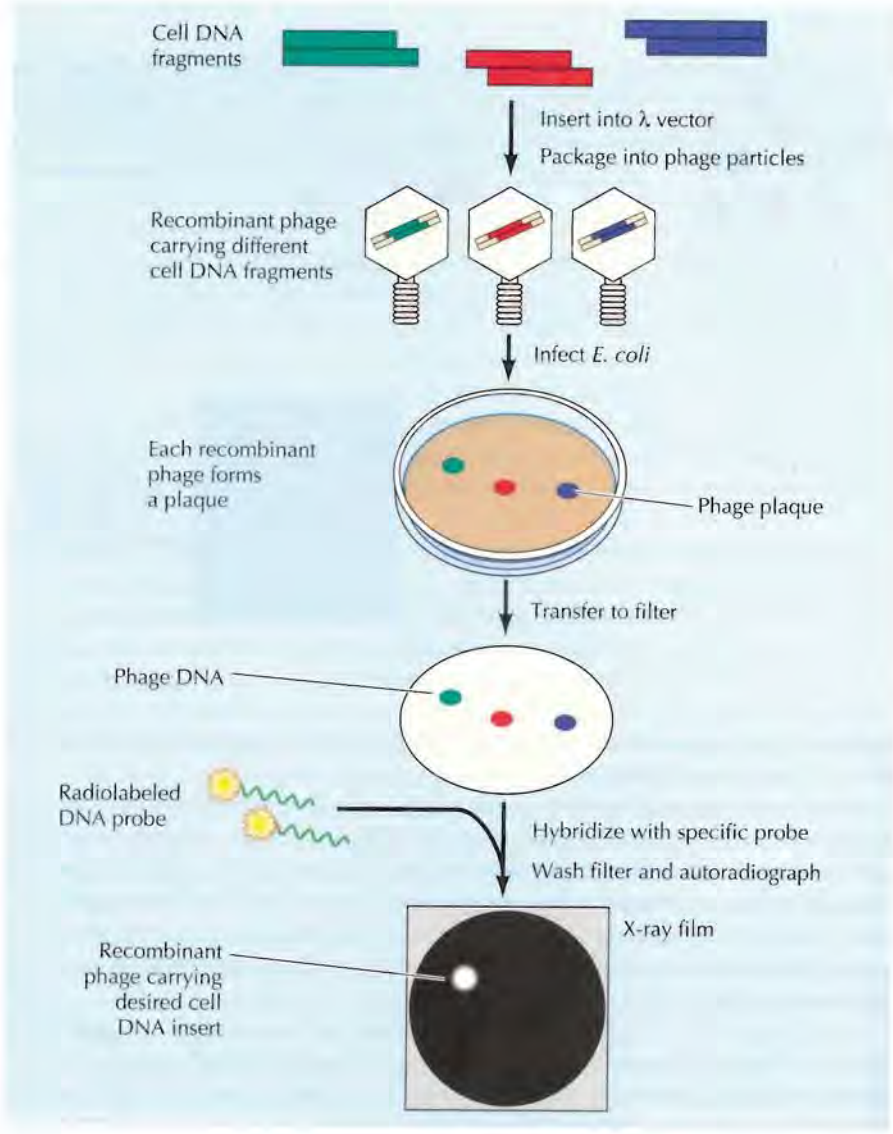


Figure 3.31 Screening a recombinant library by hybridization
Fragments of cell DNA are cloned in a bacteriophage λ vector and packaged into phage particles, yielding a collection of recombinant phage carrying different cell inserts. The phages are used to infect bacteria, and the culture is overlaid with a filter. Some of the phages in each plaque are transferred to the filter, which is then hybridized with a radiolabeled probe to identify the phage plaque containing the desired gene. The appropriate phage plaque can then be isolated from the original culture plate.

genes will be represented by multiple clones. Therefore, larger libraries, consisting of approximately 1 million recombinant phages are usually prepared to ensure a high likelihood that any gene of interest is represented in the collection.

Any gene for which a probe is available can readily be isolated from such a recombinant library. The recombinant phages are plated on *E. coli*, and each phage replicates to produce a plaque on the lawn of bacteria. The plaques are then blotted onto filters in a process similar to the transfer of DNA from a gel to a filter during Southern blotting, and the filters are hybridized with a radiolabeled probe to identify the phage plaques that contain the gene of interest. The appropriate plaque can then be isolated from the original plate in order to propagate the recombinant phage that carries the desired cell insert. Similar procedures can be used to screen bacterial colonies carrying plasmid DNA clones, so specific clones can be isolated by hybridization from either phage or plasmid libraries.

A variety of probes can be used for screening recombinant libraries. For example, a cDNA clone can be used as a probe to isolate the corresponding genomic clone, or a gene cloned from one species (e.g., mouse) can be used

to isolate a related gene from a different species (e.g., human). In addition to isolated DNA fragments, synthetic oligonucleotides can be used as probes, enabling the isolation of genes on the basis of partial amino acid sequences of their encoded proteins. For example, oligonucleotides consisting of 15 to 20 bases can be synthesized on the basis of the partial amino acid sequence of a protein of interest. These oligonucleotides can then be used as probes to isolate cDNA clones, which (as already discussed) are much easier to sequence and to characterize than is the protein itself. It is thus possible to proceed experimentally from a partial peptide sequence of a protein to the isolation of a cloned gene.

An alternative approach to isolating a gene on the basis of its protein product is the use of antibodies as probes to screen expression libraries. In this case a **cDNA library** is generated in an expression vector that drives protein synthesis in *E. coli*. Phage plaques or bacterial colonies are then transferred to a filter as already described, but the filter is then reacted with an antibody (as in Western blotting) to identify clones that are producing the protein of interest.

These procedures for identifying molecular clones and detecting genes and gene products in cells illustrate the flexibility of recombinant DNA technology. Starting with any cloned gene, it is possible not only to determine the nucleotide sequence of that gene and use it as a probe for studies of gene organization and transcription, but also to express and raise antisera against its encoded protein. Conversely, genes can be cloned on the basis of limited characterization of a protein of interest, using either oligonucleotide or antibody probes. Thus, recombinant DNA has allowed experimental analyses to proceed either from DNA to protein or from protein to DNA, providing great versatility to the strategies currently employed for studies of eukaryotic genes and their encoded proteins.

Gene Function in Eukaryotes

The recombinant DNA techniques discussed in the preceding sections provide powerful approaches to the isolation and detailed characterization of the genes of eukaryotic cells. Understanding the function of a gene, however, requires analysis of the gene within cells or intact organisms—not simply as a molecular clone in bacteria. In classical genetics, the function of genes has generally been revealed by the altered phenotypes of mutant organisms. The advent of recombinant DNA has added a new dimension to studies of gene function. Namely, it has become possible to investigate the function of a cloned gene directly by reintroducing the cloned DNA into eukaryotic cells. In simpler eukaryotes, such as yeasts, this technique has made possible the isolation of molecular clones corresponding to virtually any mutant gene. In addition, there are several methods by which cloned genes can be introduced into cultured animal and plant cells, as well as intact organisms, for functional analysis. These approaches can be coupled with the ability to introduce mutations in cloned DNA *in vitro*, extending the power of recombinant DNA to allow functional studies of the genes of more complex eukaryotes.

Genetic Analysis in Yeasts

Yeasts are particularly advantageous for studies of eukaryotic molecular biology (see Chapter 1). The genome of *Saccharomyces cerevisiae*, which consists of approximately 1.2×10^7 base pairs, is 200 times smaller than the human genome. Moreover, yeasts can easily be grown in culture, reproduc-

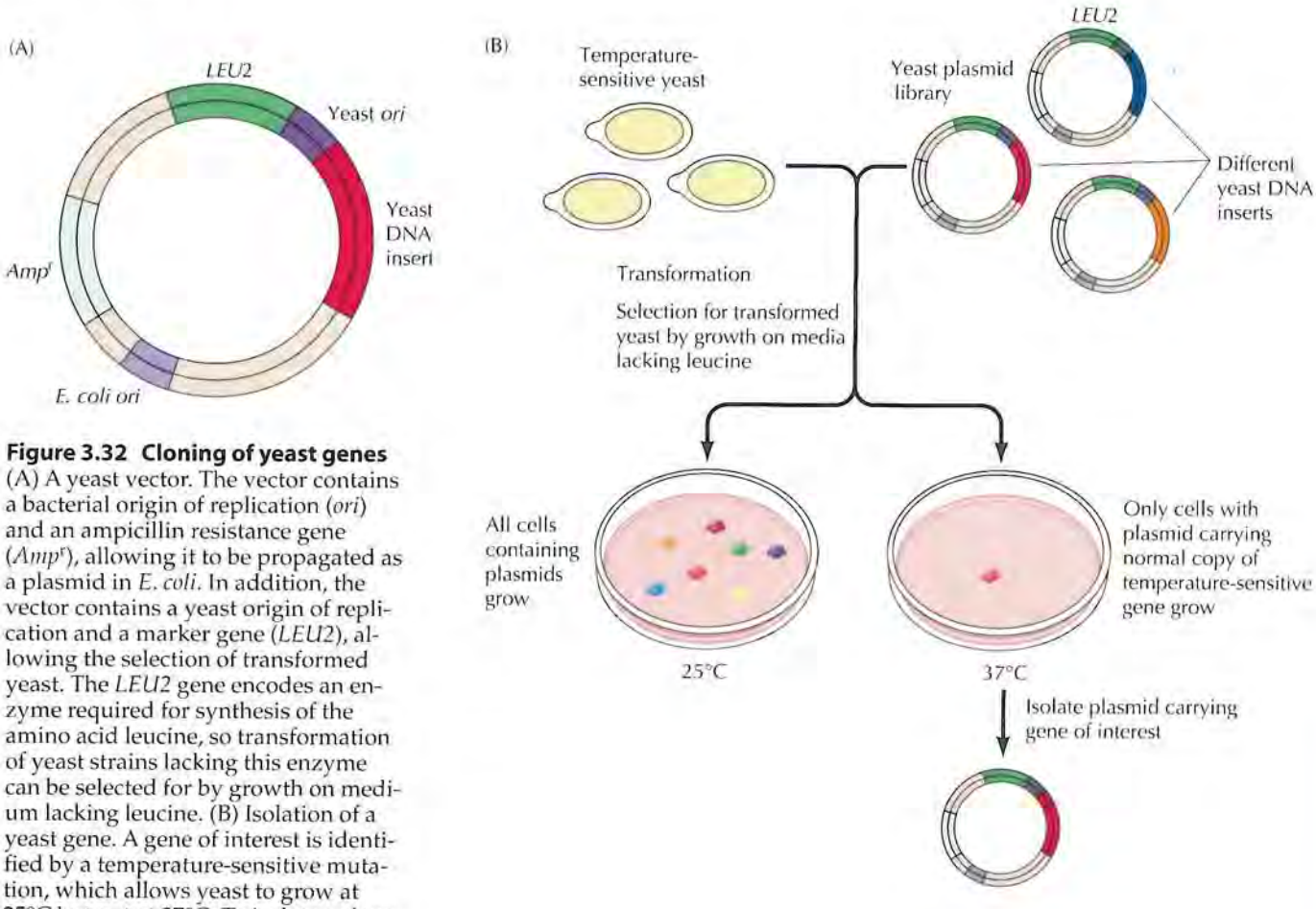


Figure 3.32 Cloning of yeast genes
(A) A yeast vector. The vector contains a bacterial origin of replication (*ori*) and an ampicillin resistance gene (Amp^r), allowing it to be propagated as a plasmid in *E. coli*. In addition, the vector contains a yeast origin of replication and a marker gene (*LEU2*), allowing the selection of transformed yeast. The *LEU2* gene encodes an enzyme required for synthesis of the amino acid leucine, so transformation of yeast strains lacking this enzyme can be selected for by growth on medium lacking leucine. (B) Isolation of a yeast gene. A gene of interest is identified by a temperature-sensitive mutation, which allows yeast to grow at 25°C but not at 37°C. To isolate a clone of the gene, the temperature-sensitive yeasts are transformed with a plasmid library containing a collection of genes encompassing the entire yeast genome. All yeasts transformed by plasmid DNAs are able to grow on media lacking leucine at 25°C, but only those yeasts transformed by a plasmid carrying a normal copy of the gene of interest are able to grow at 37°C. The desired plasmid can be isolated from transformed yeasts that form colonies at the nonpermissive temperature.

ing with a division time of about 2 hours. Thus, yeasts offer the same basic advantages—a small genome and rapid reproduction—that are afforded by bacteria.

Mutations in yeasts can be identified as readily as in *E. coli*. For example, yeast mutants that require a particular amino acid or other nutrient for growth can easily be isolated. In addition, yeasts with defects in genes required for fundamental cell processes (in contrast to metabolic defects) can be isolated as **temperature-sensitive mutants**. Such mutants encode proteins that are functional at one temperature (the permissive temperature) but not another (the nonpermissive temperature), whereas normal proteins are functional at both. A yeast with a temperature-sensitive mutation in an essential gene can be identified by its ability to grow only at the permissive temperature. The ability to isolate such temperature-sensitive mutants has allowed the identification of yeast genes controlling many fundamental cell processes, such as RNA synthesis and processing, progression through the cell cycle, and transport of proteins between cellular compartments.

The relatively simple genetics of yeast also enables a gene corresponding to any yeast mutation to be cloned, simply on the basis of its functional activity (Figure 3.32). First, a genomic library of normal yeast DNA is prepared in vectors that replicate as plasmids in yeasts as well as in *E. coli*. The small size of the yeast genome means that a complete library consists of only a few thousand plasmids. A mixture of such plasmids is then used to transform a temperature-sensitive yeast mutant, and transformants that are able to grow at the nonpermissive temperature are selected. Such transfor-

mants have acquired a normal copy of the gene of interest on plasmid DNA, which can then be easily isolated from the transformed yeast cells for further characterization.

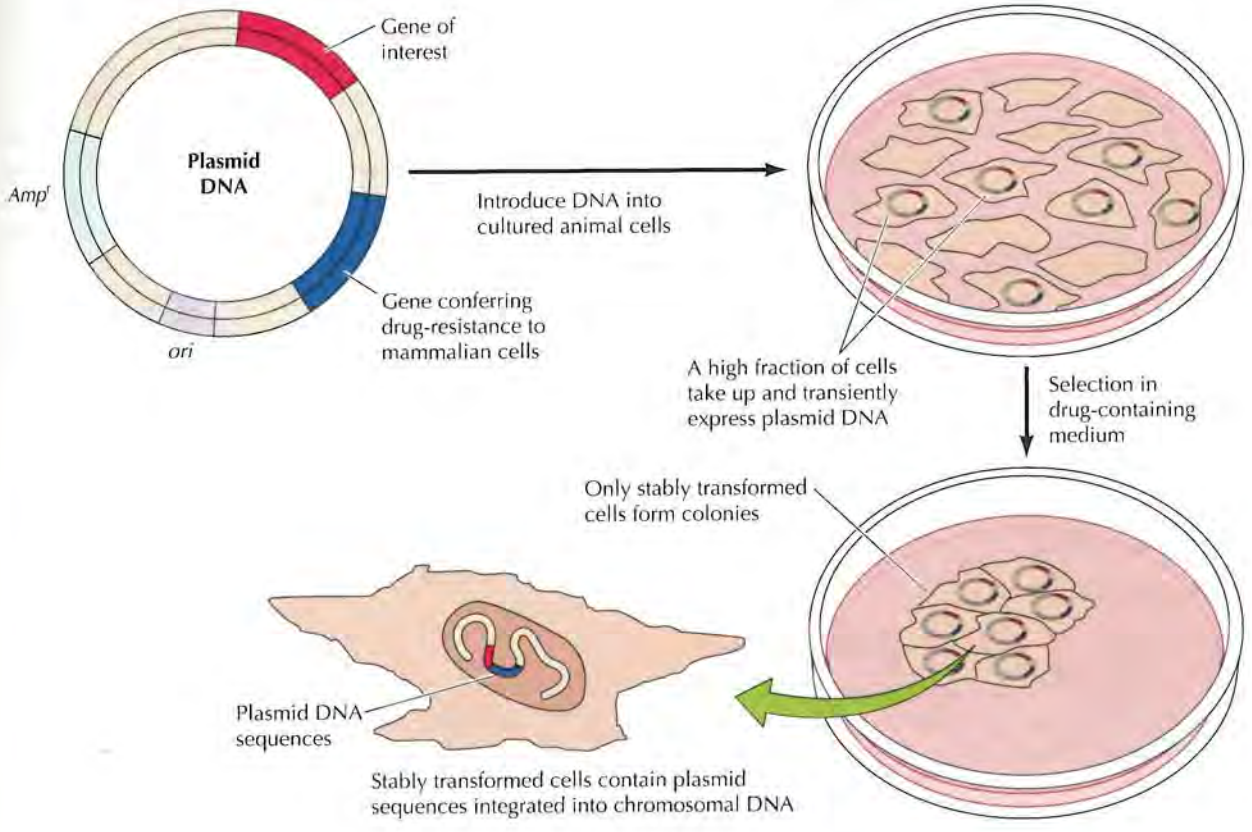
Yeast genes encoding a wide variety of essential proteins have been identified in this manner. In many cases, such genes isolated from yeasts have also been useful in identifying and cloning related genes from mammalian cells. Thus, the simple genetics of yeast has not only provided an important model for eukaryotic cells, but has also led directly to the cloning of related genes from more complex eukaryotes.

Gene Transfer in Plants and Animals

Although the cells of complex eukaryotes are not amenable to the simple genetic manipulations possible in yeasts, gene function can still be assayed by the introduction of cloned DNA into plant and animal cells. Such experiments (generally called **gene transfer**) have proven critical to addressing a wide variety of questions, including studies of the mechanisms that regulate gene expression and protein processing. In addition, as discussed later in the book, gene transfer has enabled the identification and characterization of genes that control animal cell growth and differentiation, including a variety of genes responsible for the abnormal growth of human cancer cells.

The methodology for introduction of DNA into animal cells was initially developed for infectious viral DNAs and is therefore frequently called **transfection** (a word derived from *trans*formation + in*fection*) (Figure 3.33). DNA can be introduced into animal cells in culture by a variety of methods, including direct microinjection into the cell nucleus, coprecipitation of DNA with calcium phosphate to form small particles that are taken up by the

Figure 3.33 Introduction of DNA into animal cells
A eukaryotic gene of interest is cloned in a plasmid containing a drug resistance marker that can be selected for in cultured animal cells. The plasmid DNA is taken up and expressed by a high fraction of the cells for a few days (transient expression). Stably transformed cells, in which the plasmid DNA becomes integrated into chromosomal DNA, can be selected for their ability to grow in drug-containing medium.



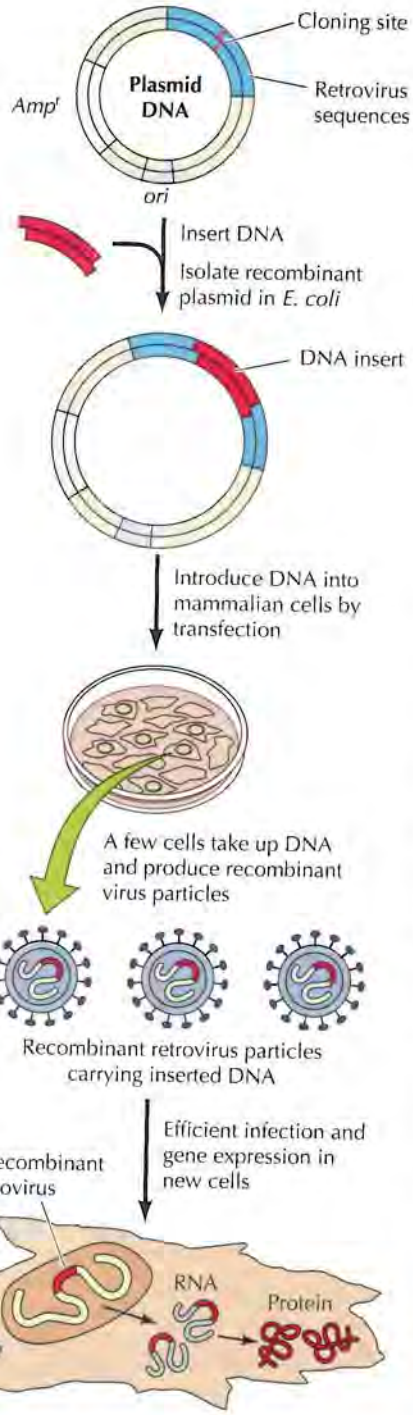


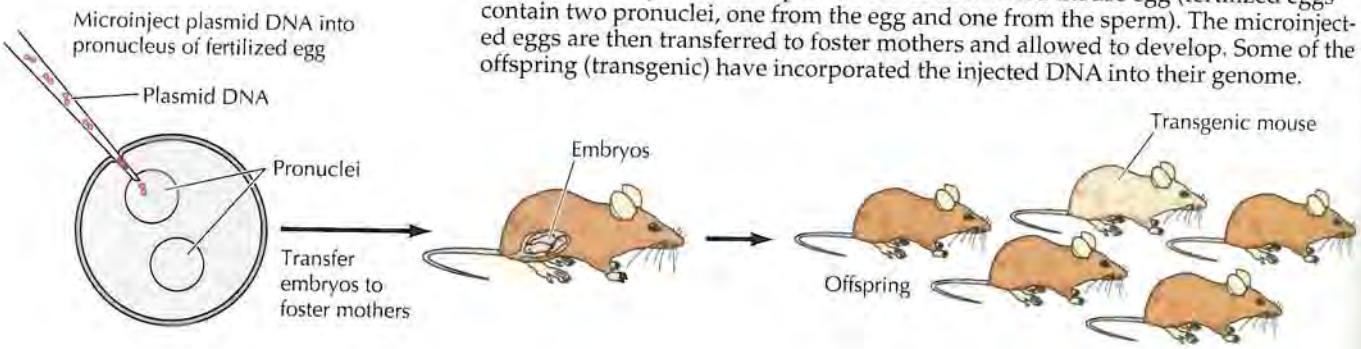
Figure 3.34 Retroviral vectors
The vector consists of retroviral sequences cloned in a plasmid that can be propagated in *E. coli*. Foreign DNA is inserted into the viral sequences, and recombinant plasmids are isolated in bacteria. Animal cells in culture are then transfected with the recombinant DNA. The DNA is taken up by a small fraction of the cells, which produce recombinant retroviruses carrying the inserted DNA. These recombinant retroviruses can be used to efficiently infect new cells, where the viral genome carrying the inserted gene integrates into chromosomal DNA as a provirus.

cells, incorporation of DNA into lipid vesicles (**liposomes**) that fuse with the plasma membrane, and exposure of cells to a brief electric pulse that transiently opens pores in the plasma membrane (**electroporation**). The DNA taken up by a high fraction of cells is transported to the nucleus, where it can be transcribed for several days—a phenomenon called **transient expression**. In a smaller fraction of cells (usually 1% or less), the foreign DNA becomes stably integrated into the cell genome and is transferred to progeny cells at cell division just as any other cell gene is. These stably transformed cells can be isolated if the transfected DNA contains a selectable marker, such as resistance to a drug that inhibits the growth of normal cells. Thus, any cloned gene can be introduced into mammalian cells by being transferred together with a drug resistance marker that can be used to isolate stable transformants. The effects of such cloned genes on cell behavior—for example, cell growth or differentiation—can then be analyzed.

Animal viruses can also be used as vectors for more efficient introduction of cloned DNAs into cells. Retroviruses are particularly useful in this respect, since their life cycle involves the stable integration of viral DNA into the genome of infected cells (Figure 3.34). Consequently, retroviral vectors can be used to efficiently introduce cloned genes into a wide variety of cell types, making them an important vehicle for a broad range of applications.

Cloned genes can also be introduced into the germ line of multicellular organisms, allowing them to be studied in the context of the intact animal rather than in cultured cells. One method used to produce mice that carry such foreign genes (**transgenic mice**) is the direct microinjection of cloned DNA into the pronucleus of a fertilized egg (Figure 3.35). The injected eggs are then transferred to foster mothers and allowed to develop to term. In a fraction of the progeny mice (often about 10%), the foreign DNA will have integrated into the genome of the fertilized egg and is therefore present in all cells of the animal. Since the foreign DNA is present in germ cells as well as in somatic cells, it is transferred by breeding to new progeny mice just as any other cell gene would be.

Figure 3.35 Production of transgenic mice
DNA is microinjected into a pronucleus of a fertilized mouse egg (fertilized eggs contain two pronuclei, one from the egg and one from the sperm). The microinjected eggs are then transferred to foster mothers and allowed to develop. Some of the offspring (transgenic) have incorporated the injected DNA into their genome.



The properties of **embryonic stem (ES) cells** provide an alternative means of introducing cloned genes into mice (Figure 3.36). ES cells can be established in culture from early mouse embryos. They can also be reintroduced into early embryos, where they participate normally in development and can give rise to cells in all tissues of the mouse—including germ cells. It is thus possible to introduce cloned DNA into ES cells in culture, select stably transformed cells, and then introduce these cells back into mouse embryos. Such embryos give rise to chimeric offspring in which some cells are derived from the normal embryo cells and some from the transfected ES cells. In some such mice the transfected ES cells are incorporated into the germ line. Breeding these mice therefore leads to the direct inheritance of the transfected gene by their progeny.

Cloned DNAs can also be introduced into plant cells. One approach is to remove the plant cell wall, forming protoplasts that are surrounded only by a plasma membrane. DNA can then be introduced into such protoplasts by electroporation, as was described for animal cells. Alternatively, DNA can be introduced into intact plant cells by bombardment of the cells with DNA-coated microprojectiles, such as small particles of tungsten. The DNA-coated particles are shot directly into the plant cells; some of the cells are killed, but others survive and become stably transformed.

Vectors for more efficient introduction of recombinant DNA into plant cells have been developed from plant viruses. In addition, a plasmid from the bacterium *Agrobacterium tumefaciens* (the **Ti plasmid**) provides a novel vehicle for the introduction of cloned DNAs into various plants (Figure 3.37). In nature, *Agrobacterium* attaches to the leaves of plants and the Ti plasmid is trans-

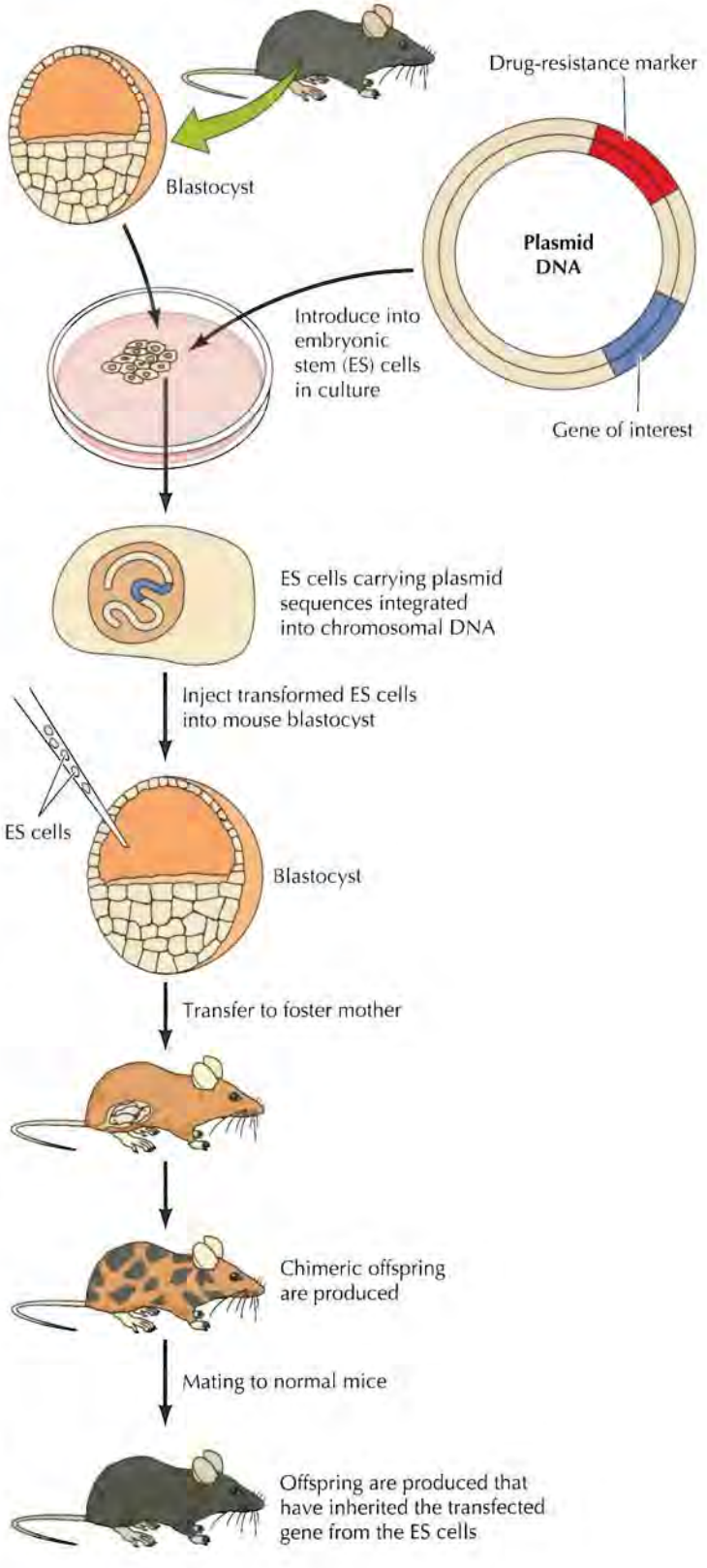


Figure 3.36 Introduction of genes into mice via embryonic stem cells
Embryonic stem (ES) cells are cultured cells derived from early mouse embryos (blastocysts). DNA can be introduced into these cells in culture, and stably transformed ES cells can be isolated. These transformed ES cells can then be injected into a recipient blastocyst, where they are able to participate in normal development of the embryo. Some of the progeny mice that develop after transfer of injected embryos to foster mothers therefore contain cells derived from transformed ES cells, as well as from the normal cells of the blastocyst. Since these mice are a mixture of two different cell types, they are referred to as chimeric. Offspring carrying the transfected gene can then be produced by the breeding of chimeric mice in which descendants of the transformed ES cells have been incorporated into the germ line.

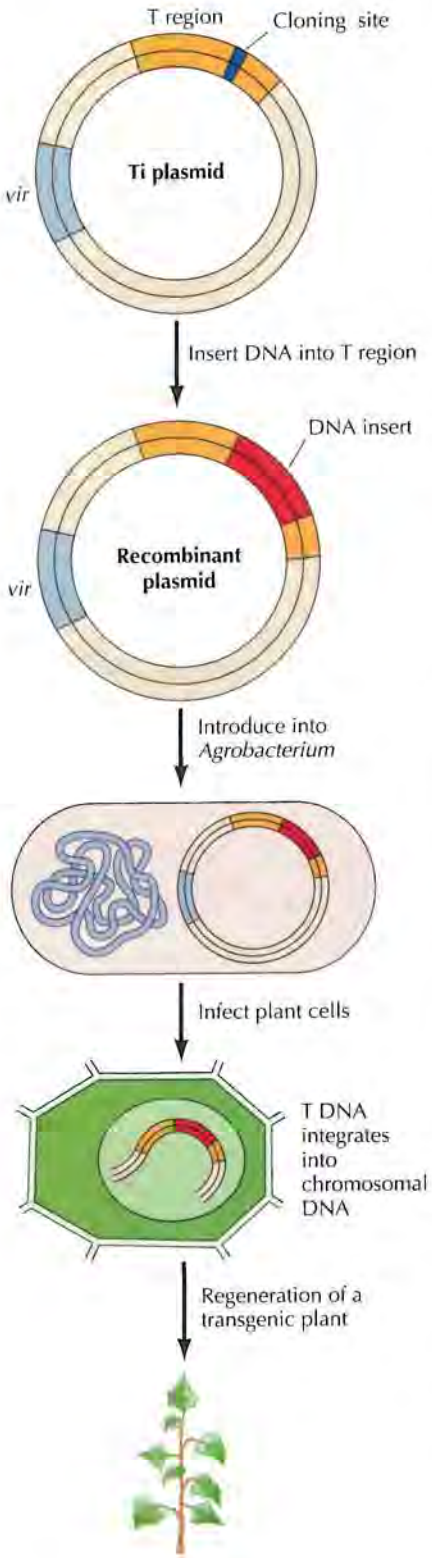


Figure 3.37 Introduction of genes into plant cells via the Ti plasmid
The Ti plasmid contains the T region, which is transferred to infected plant cells, and virulence (*vir*) genes, which function in T DNA transfer. In Ti plasmid vectors, foreign DNA is inserted into the T region. The recombinant plasmid is introduced into *Agrobacterium tumefaciens*, which is then used to infect cultured cells. The T region of the plasmid (carrying the inserted DNA) is transferred to the plant cells and integrates into chromosomal DNA. A transgenic plant can then be generated from the transformed cells.

ferred into plant cells, where it becomes incorporated into chromosomal DNA. Vectors developed from the Ti plasmid therefore provide an efficient means of introducing recombinant DNA into sensitive plant cells. Since many plants can be regenerated from single cultured cells (see Chapter 1), transgenic plants can be established directly from cells into which recombinant DNA has been introduced in culture—a much simpler procedure than the production of transgenic animals.

Mutagenesis of Cloned DNAs

In classical genetic studies (e.g., in bacteria or yeasts), mutants are the key to identifying genes and understanding their function by observing the altered phenotype of mutant organisms. In such studies, mutant genes are detected because they result in observable phenotypic changes—for example, temperature-sensitive growth or a specific nutritional requirement. The isolation of genes by recombinant DNA, however, has opened a different approach to mutagenesis. It is now possible to introduce any desired alteration into a cloned gene and to determine the effect of the mutation on gene function. Such procedures have been called **reverse genetics**, since a mutation is introduced into a gene first and its functional consequence is determined second. The ability to introduce specific mutations into cloned DNAs (***in vitro* mutagenesis**) has proven to be a powerful tool in studying the expression and function of eukaryotic genes.

Cloned genes can be altered by many *in vitro* mutagenesis procedures, which can lead to the introduction of deletions, insertions, or single nucleotide alterations. One common method of mutagenesis is the use of synthetic oligonucleotides to generate nucleotide changes in a DNA sequence (Figure 3.38). In this procedure a synthetic oligonucleotide bearing the mutant base is used as a primer for DNA synthesis. Newly synthesized DNA molecules containing the mutation can then be isolated and characterized. For example, specific amino acids of a protein can be altered in order to characterize their role in protein function.

Variations of this approach, combined with the versatility of other methods for manipulating recombinant DNA molecules, can be used to introduce virtually any desired alteration in a cloned gene. The effects of such mutations on gene expression and function can then be determined by introduction of the gene into an appropriate cell type. *In vitro* mutagenesis has thus allowed detailed characterization of the functional roles of both the regulatory and protein-coding sequences of cloned genes.

Introducing Mutations into Cellular Genes

Although the transfer of cloned genes into cells (particularly in combination with *in vitro* mutagenesis) provides a powerful approach to studying gene structure and function, such experiments fall short of defining the role of an unknown gene in a cell or intact organism. The cells used as recipients for transfer of cloned genes usually already have normal copies of the gene in

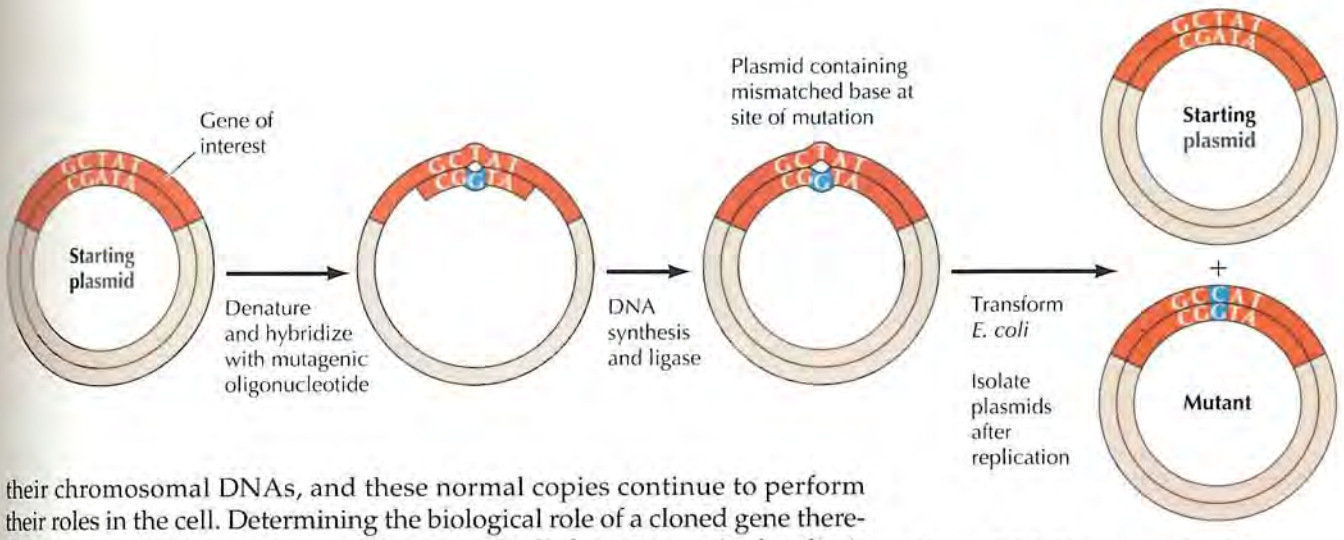


Figure 3.38 Mutagenesis with synthetic oligonucleotides
An oligonucleotide carrying the desired base alteration is used to prime DNA synthesis from plasmid DNA, and the DNA is circularized by incubation with DNA ligase. This process yields plasmids in which one strand contains the normal base and the other strand the mutant base. Replication of the plasmid DNAs after transformation of *E. coli* therefore yields a mixture of both starting and mutant plasmids.

their chromosomal DNAs, and these normal copies continue to perform their roles in the cell. Determining the biological role of a cloned gene therefore requires that the activity of the normal cellular gene copies be eliminated. As discussed in the following section, this can be readily accomplished in yeasts. In animal cells, this task is considerably more difficult, although several approaches to either inactivating the chromosomal copies of a cloned gene or inhibiting normal gene function are now widely used.

Mutating the chromosomal genes in yeasts is relatively easy because DNA introduced into yeast cells frequently undergoes **homologous recombination** with its chromosomal copy (Figure 3.39). In homologous recombination, the cloned yeast gene replaces the normal allele, so mutations introduced into the cloned gene *in vitro* become incorporated into the

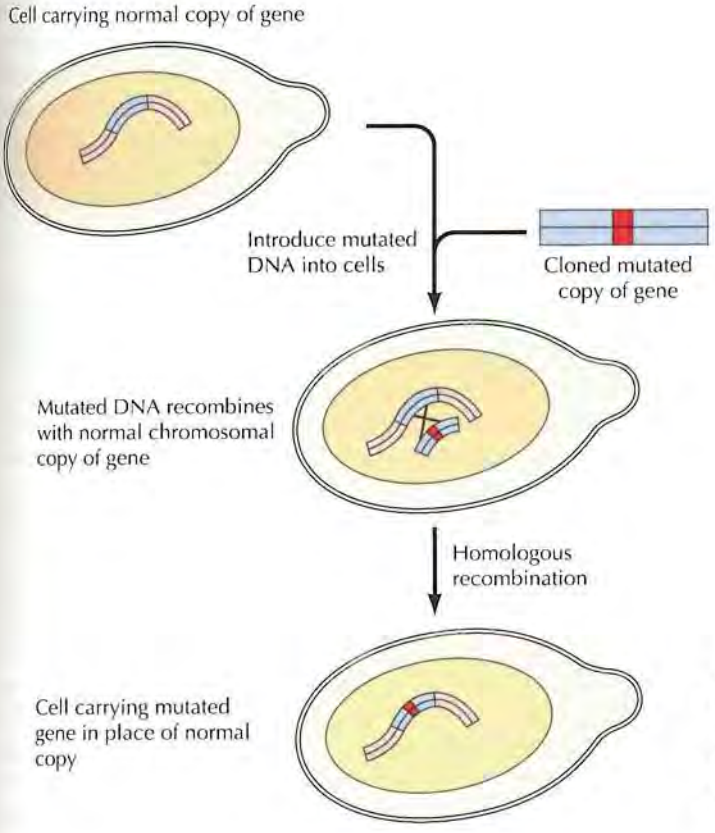


Figure 3.39 Gene inactivation by homologous recombination
A mutated copy of the cloned gene is introduced into cells. The cloned gene may then replace the normal gene copy by homologous recombination, yielding a cell carrying the desired mutation in its chromosomal DNA.

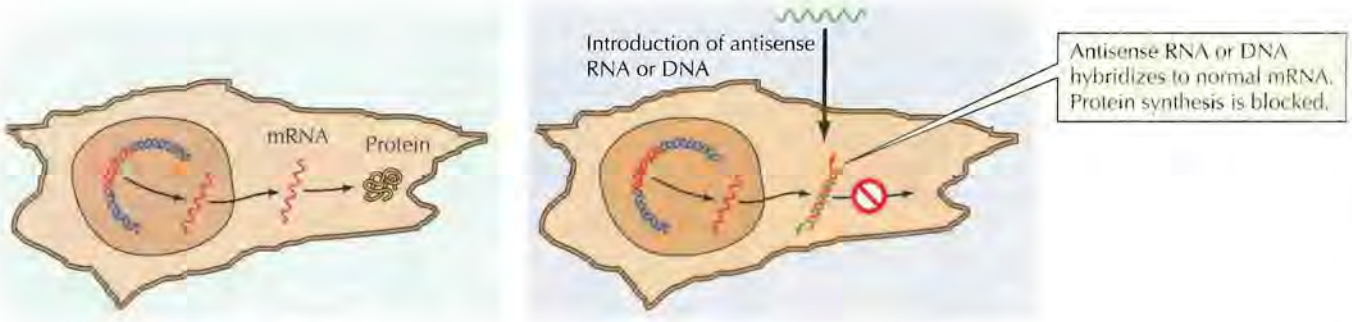


Figure 3.40 Inhibition of gene expression by antisense RNA or DNA Antisense RNA or single-stranded DNA is complementary to the mRNA of a gene of interest. Antisense nucleic acids therefore form hybrids with their target mRNAs, blocking the translation of mRNA into protein and leading to mRNA degradation.

chromosomal copy of the yeast gene. In the simplest case, mutations that inactivate the cloned gene can be introduced in place of the normal gene copy in order to determine its role in cellular processes. Since yeasts have both haploid and diploid stages of their life cycle, even genes that are required for cell growth can be inactivated and studied. An inactive gene copy is introduced into diploid cells, which then contain one functional and one inactive copy of the target gene. The cells are induced to undergo meiosis, and the effect of gene inactivation on the progeny haploid cells can be observed.

Recombination between transferred DNA and the homologous chromosomal gene is a rare event in mammalian cells, so gene inactivation by this approach is more difficult than it is in yeasts. Possibly because the genomes of mammalian cells are so much larger than that of yeasts, most transfected DNA that integrates into the recipient cell genome does so at random sites by recombination with unrelated sequences. However, various procedures have been developed both to increase the frequency of homologous recombination and to select and isolate the transformed cells in which homologous recombination has occurred, so genes in mammalian cells can be inactivated by this approach. Importantly, genes can be inactivated or disrupted not only in somatic cell lines but also in embryonic stem cells in culture. As already described, ES cells can be used to generate transgenic mice, so the effects of inactivation of a gene can be investigated in the context of the intact animal. The functions of hundreds of mouse genes have been investigated in this way, and such studies have been particularly important in revealing the roles of specific genes in mouse development.

An alternative to gene inactivation by homologous recombination is the use of **antisense nucleic acids** to block gene expression (Figure 3.40). In this approach, RNA or single-stranded DNA complementary to the mRNA of the gene of interest (antisense) is introduced into a cell. The antisense RNA or DNA hybridizes with the mRNA and blocks its translation into protein. Moreover, the RNA-DNA hybrids resulting from the introduction of antisense DNA molecules are usually degraded within the cell. Antisense RNAs can be introduced into cells directly by microinjection. Alternatively, cells can be transfected with vectors that have been engineered to express antisense RNA. Antisense DNA is usually in the form of short oligonucleotides (about 20 bases long) that are microinjected into cells. Alternatively, because cells are able to take up such oligonucleotides from the culture medium, antisense oligonucleotides can simply be added to the cell culture.

RNA interference (RNAi) provides an additional, and very effective, method for interfering with gene expression at the mRNA level (Figure 3.41). In RNA interference, short double-stranded RNA molecules (21 to 23 nucleotides) induce the degradation of complementary mRNAs. Although

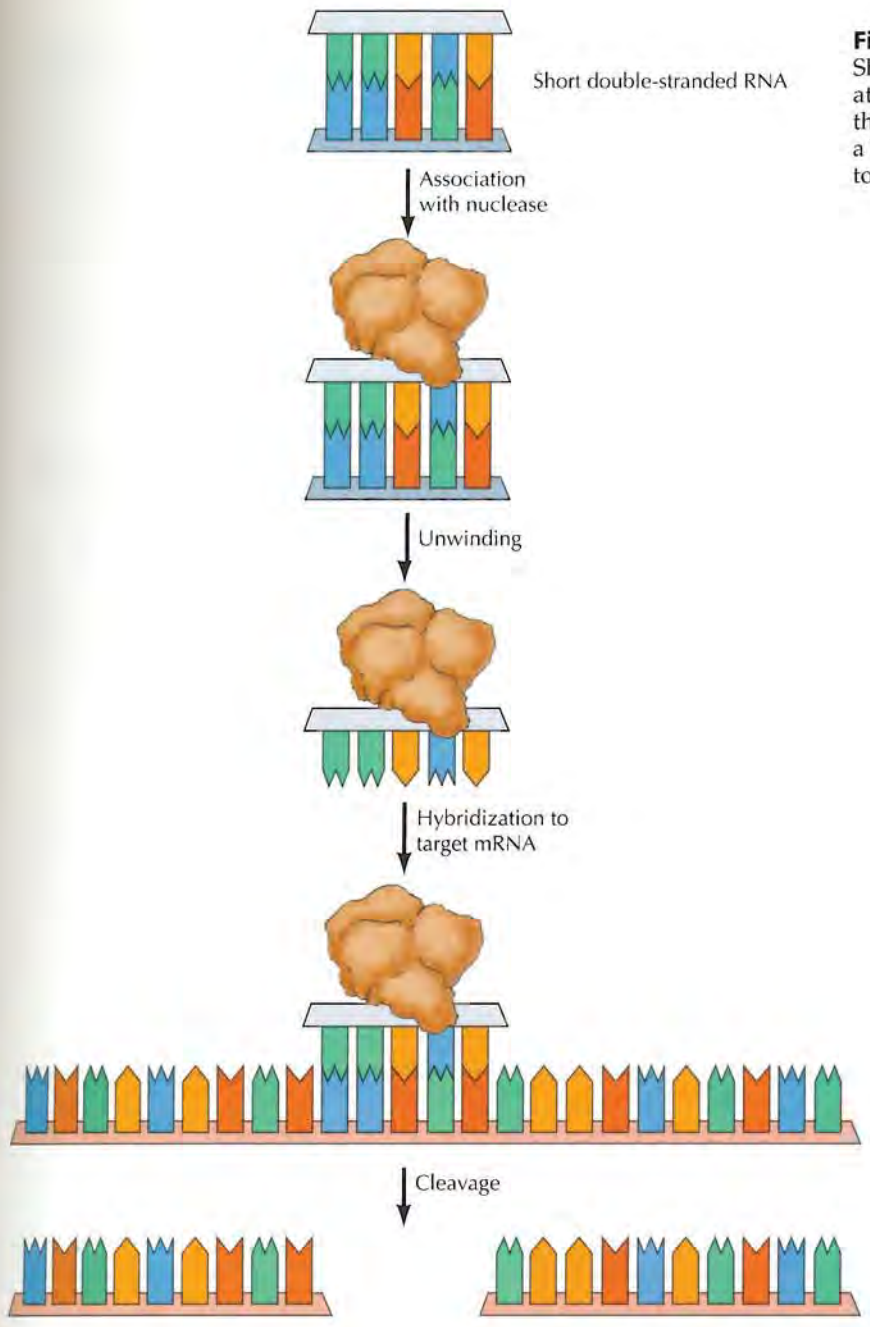


Figure 3.41 RNA interference
Short double-stranded RNA molecules associate with a protein ribonuclease. Unwinding of the double-stranded RNA and hybridization to a homologous mRNA then targets the nuclease to the mRNA, leading to mRNA cleavage.

the mechanism by which short double-stranded RNAs induce the degradation of their target mRNAs is not yet fully understood, it appears to involve the action of a ribonuclease that associates with the double-stranded RNAs and is then guided to the target mRNA by complementary base pairing. RNAi was first found to effectively induce mRNA degradation in *C. elegans*, and its use has been extended to *Drosophila*, *Arabidopsis*, and most recently mammalian cells.

In addition to inactivating a gene or inducing degradation of an mRNA, it is sometimes possible to interfere directly with the function of proteins within cells (Figure 3.42). One approach is to microinject antibodies that block the activity of the protein against which they are directed. Alterna-

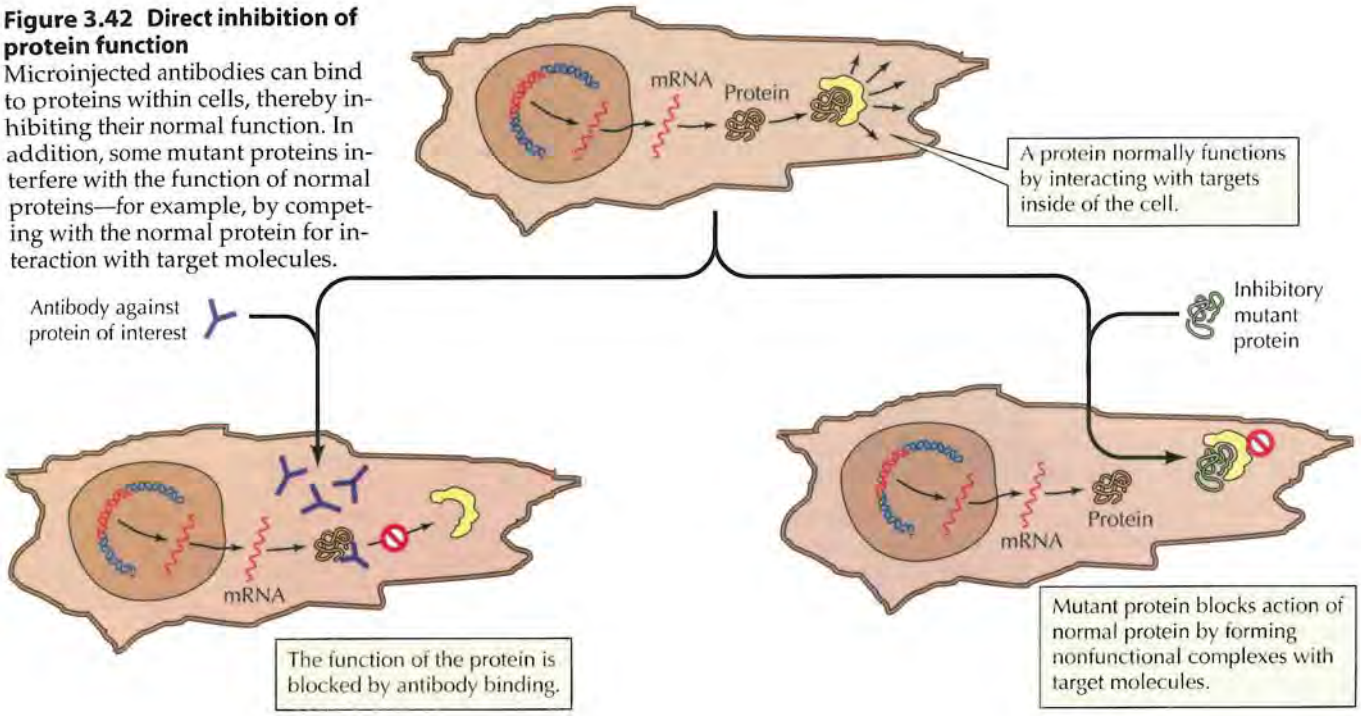


Figure 3.42 Direct inhibition of protein function
Microinjected antibodies can bind to proteins within cells, thereby inhibiting their normal function. In addition, some mutant proteins interfere with the function of normal proteins—for example, by competing with the normal protein for interaction with target molecules.

tively, some mutant proteins interfere with the function of their normal counterparts when they are expressed within the same cell—for example, by competing with the normal protein for binding to its target molecule. Cloned DNAs encoding such mutant proteins (called **dominant inhibitory mutants**) can be introduced into cells by gene transfer and used to study the effects of blocking normal gene function.

KEY TERMS

gene, allele, dominant, recessive, genotype, phenotype, chromosome, diploid, meiosis, haploid, mutation

one gene–one enzyme hypothesis

transformation

semiconservative replication, DNA polymerase

SUMMARY

HEREDITY, GENES, AND DNA

Genes and Chromosomes: Chromosomes are the carriers of genes.

Genes and Enzymes: A gene specifies the amino acid sequence of a polypeptide chain.

Identification of DNA as the Genetic Material: DNA was identified as the genetic material by bacterial transformation experiments.

The Structure of DNA: DNA is a double helix in which hydrogen bonds form between purines and pyrimidines on opposite strands. Because of specific base pairing—A with T and G with C—the two strands of a DNA molecule are complementary in sequence.

Replication of DNA: DNA replicates by semiconservative replication, in which the two strands separate and each serves as a template for synthesis of a new progeny strand.

EXPRESSION OF GENETIC INFORMATION

Colinearity of Genes and Proteins: The order of nucleotides in DNA specifies the order of amino acids in proteins.

The Role of Messenger RNA: Messenger RNA functions as an intermediate to convey information from DNA to the ribosomes, where it serves as a template for protein synthesis.

central dogma, transcription, translation, messenger RNA (mRNA), RNA polymerase, ribosomal RNA (rRNA), transfer RNA (tRNA)

The Genetic Code: Transfer RNAs serve as adaptors between amino acids and mRNA during translation. Each amino acid is specified by a codon consisting of three nucleotides.

genetic code, *in vitro* translation, codon

RNA Viruses and Reverse Transcription: DNA can be synthesized from RNA templates, as first discovered in retroviruses.

retrovirus, reverse transcription, reverse transcriptase

RECOMBINANT DNA

Restriction Endonucleases: Restriction endonucleases cleave specific DNA sequences, yielding defined fragments of DNA molecules.

restriction endonuclease, gel electrophoresis, restriction map

Generation of Recombinant DNA Molecules: Recombinant DNA molecules consist of a DNA fragment of interest ligated to a vector that is able to replicate independently in an appropriate host cell.

molecular cloning, vector, recombinant molecule, molecular clone, DNA ligase, cDNA

Vectors for Recombinant DNA: A variety of vectors are used to clone different sizes of DNA fragments.

plasmid, origin of replication, cosmid, P1 artificial chromosome (PAC), bacterial artificial chromosome (BAC), yeast artificial chromosome (YAC)

DNA Sequencing: The nucleotide sequences of cloned DNA fragments can be readily determined.

dideoxynucleotide, autoradiography

Expression of Cloned Genes: The proteins encoded by cloned genes can be expressed at high levels in either bacteria or eukaryotic cells.

expression vector, baculovirus

Amplification of DNA by the Polymerase Chain Reaction: PCR allows the amplification and isolation of specific fragments of DNA *in vitro*.

polymerase chain reaction (PCR)

DETECTION OF NUCLEIC ACIDS AND PROTEINS

Nucleic Acid Hybridization: Nucleic acid hybridization allows the detection of specific DNA or RNA sequences.

nucleic acid hybridization, probe, Southern blotting, Northern blotting, DNA microarray, *in situ* hybridization

Detection of Small Amounts of DNA or RNA by PCR: PCR provides a sensitive method for detecting small amounts of specific DNA or RNA molecules.

Antibodies as Probes for Proteins: Antibodies are used to detect specific proteins in cells or cell extracts.

antibody, antigen, monoclonal antibody, immunoblotting, Western blotting, immunoprecipitation, SDS-polyacrylamide gel electrophoresis (SDS-PAGE)

Probes for Screening Recombinant DNA Libraries: Specific DNA inserts can be detected in recombinant DNA libraries by the use of either nucleic acid hybridization or antibody probes.

recombinant DNA library, genomic library, cDNA library

GENE FUNCTION IN EUKARYOTES

temperature-sensitive mutant

***Genetic Analysis in Yeasts*:** The simple genetics and rapid replication of yeasts facilitate the molecular cloning of a gene corresponding to any yeast mutation.

gene transfer, transfection, transient expression, liposome, electroporation, transgenic mouse, embryonic stem (ES) cell, Ti plasmid

***Gene Transfer in Plants and Animals*:** Cloned genes can be introduced into complex eukaryotic cells and multicellular organisms for functional analysis.

reverse genetics, *in vitro* mutagenesis

***Mutagenesis of Cloned DNAs*:** *In vitro* mutagenesis of cloned DNAs is used to study the effect of engineered mutations on gene function.

homologous recombination, antisense nucleic acids, RNA interference (RNA), dominant inhibitory mutant

***Introducing Mutations into Cellular Genes*:** Mutations can be introduced into chromosomal gene copies by homologous recombination with cloned DNA sequences. In addition, the expression or function of specific gene products can be blocked by antisense nucleic acids, RNA interference, or dominant inhibitory mutants.

Questions

1. Define translation in context of molecular biology.

2. What components must be present in order to do *in vitro* protein synthesis?

3. How was the first codon assignment for an amino acid discovered?

4. What does it mean to say that the genetic code is degenerate?

5. Addition or deletion of one or two nucleotides in the coding part of a gene produces a non-functioning protein, whereas addition or deletion of three nucleotides often produces a protein with nearly normal function. Explain.

6. Describe the features a yeast artificial chromosome must have in order to be used to clone a piece of human DNA cut with *Eco*R1 in yeast.

7. Why is the polymerase chain reaction (PCR) so useful?

8. What is the difference between a genomic library and a cDNA library?

9. You are studying an enzyme in which an active-site cysteine residue is encoded by the triplet UGU. How would mutating the third base to a C affect enzyme function? How about mutating it to an A?

10. Digestion of a 4-kb DNA molecule with *Eco*RI yields two fragments of 1 kb and 3 kb each. Digestion of the same molecule with *Hin*dIII yields fragments of 1.5 kb and 2.5 kb. Finally, digestion with *Eco*RI and *Hin*dIII in combination yields fragments of 0.5 kb, 1 kb, and 2.5 kb. Draw a restriction map indicating the positions of the *Eco*RI and *Hin*dIII cleavage sites.

11. Starting with DNA from a single sperm, how many copies of a specific gene sequence will be obtained after 10 cycles of PCR amplification? After 30 cycles?

12. You have cloned a cDNA of unknown function. How could you experimentally determine the subcellular localization of the protein it encodes?

13. You are interested in identifying the amino acid residues that are important for the catalytic activity of an enzyme. Assuming you have a cDNA clone available, what experimental strategies could you use?

References and Further Reading

General References

Lewin, B. 2000. *Genes VII*. Oxford: Oxford Univ. Press.

Weaver, R. F. 2002. *Molecular Biology*. 2nd ed. New York: McGraw-Hill.

Heredity, Genes, and DNA

Avery, O. T., C. M. MacLeod and M. McCarty. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.* 79: 137–158. [P]

Franklin, R. E. and R. G. Gosling. 1953. Molecular configuration in sodium thymonucleate. *Nature* 171: 740–741. [P]

Kornberg, A. 1960. Biologic synthesis of deoxyribonucleic acid. *Science* 131: 1503–1508. [P]

Meselson, M. and F. W. Stahl. 1958. The replication of DNA in *Escherichia coli*. *Proc. Natl. Acad. Sci.* USA 44: 671–682. [P]

Watson, J.D. and F. H. C. Crick. 1953. Genetical implications of the structure of deoxyribonucleic acid. *Nature* 171: 964–967. [P]

Watson, J.D. and F. H. C. Crick. 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* 171: 737–738. [P]

Wilkins, M. H. F., A. R. Stokes and H. R. Wilson. 1953. Molecular structure of deoxypentose nucleic acids. *Nature* 171: 738–740. [P]

Expression of Genetic Information

Baltimore, D. 1970. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* 226: 1209–1211. [P]

Brenner, S., F. Jacob and M. Meselson. 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190: 576–581. [P]

Crick, F. H. C., L. Barnett, S. Brenner and R. J. Watts-Tobin. 1961. General nature of the genetic code for proteins. *Nature* 192: 1227–1232. [P]

Ingram, V. M. 1957. Gene mutations in human hemoglobin: The chemical difference between normal and sickle cell hemoglobin. *Nature* 180: 326–328. [P]

Nirenberg, M. and P. Leder. 1964. RNA codewords and protein synthesis. *Science* 145: 1399–1407. [P]

Nirenberg, M. W. and J. H. Matthaei. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA* 47: 1588–1602. [P]

Temin, H. M. and S. Mizutani. 1970. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* 226: 1211–1213. [P]

Yanofsky, C., B. C. Carlton, J. R. Guest, D. R. Helinski and U. Henning. 1964. On the colinearity of gene structure and protein structure. *Proc. Natl. Acad. Sci. USA* 51: 266–272. [P]

Recombinant DNA

Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith and K. Struhl. eds. 1989. *Current Protocols in Molecular Biology*. New York: Greene Publishing and Wiley Interscience. [R]

Burke, D. T., G. F. Carle and M. V. Olson. 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* 236: 806–812. [P]

Cohen, S. N., A. C. Y. Chang, H. W. Boyer and R. B. Helling. 1973. Construction of biologically functional bacterial plasmids *in vitro*. *Proc. Natl. Acad. Sci. USA* 70: 3240–3244. [P]

Nathans, D. and H. O. Smith. 1975. Restriction endonucleases in the analysis and restructuring of DNA molecules. *Ann. Rev. Biochem.* 44: 273–293. [R]

Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis and H. A. Erlich. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239: 487–491. [P]

Sambrook, J., and D. Russell. 2001. *Molecular Cloning: A Laboratory Manual*. 3rd ed. Plainview, N.Y.: Cold Spring Harbor Laboratory Press.

Sanger, F., S. Nicklen and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74: 5463–5467. [P]

Detection of Nucleic Acids and Proteins

Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith and K. Struhl. eds. 1989. *Current Protocols in Molecular Biology*. New York: Greene Publishing and Wiley Interscience.

Benton, W. D. and R. W. Davis. 1977. Screening λgt recombinant clones by hybridization to single plaques *in situ*. *Science* 196: 180–182. [P]

Broome, S. and W. Gilbert. 1978. Immunological screening method to detect specific translation products. *Proc. Natl. Acad. Sci. USA* 75: 2746–2749. [P]

Brown, P. O. and D. Botstein. 1999. Exploring the new world of the genome with DNA microarrays. *Nature Genetics* 21: 33-37. [R]

Caruthers, M. H. 1985. Gene synthesis machines: DNA chemistry and its uses. *Science* 230: 281–285. [R]

Gerhold, D., T. Rushmore and C. T. Caskey. 1999. DNA chips: promising toys have become powerful tools. *Trends Biochem. Sci.* 24: 168–173. [R]

Grunstein, M. and D. S. Hogness. 1975. Colony hybridization: A method for the isolation of cloned DNAs that contain a specific gene. *Proc. Natl. Acad. Sci. USA* 72: 3961–3965. [P]

Harlow, E. and D. Lane. 1999. *Using Antibodies: A Laboratory Manual*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Kohler, G. and C. Milstein. 1975. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* 256: 495–497. [P]

Maniatis, T., R. C. Hardison, E. Lacy, J. Lauer, C. O'Connell, D. Quon, G. K. Sim and A. Efstratiadis. 1978. The isolation of structural genes from libraries of eucaryotic DNA. *Cell* 15: 687–701. [P]

Sambrook, J., and D. Russell. 2001. *Molecular Cloning: A Laboratory Manual*. 3rd ed. Plainview, N.Y.: Cold Spring Harbor Laboratory Press.

Schildkraut, C. L., J. Marmur and P. Doty. 1961. The formation of hybrid DNA molecules, and their use in studies of DNA homologies. *J. Mol. Biol.* 3: 595–617. [P]

Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98: 503–517. [P]

Gene Function in Eukaryotes

Botstein, D. and D. Shortle. 1985. Strategies and applications of *in vitro* mutagenesis. *Science* 229: 1193–1201. [R]

Bronson, S. K. and O. Smithies. 1994. Altering mice by homologous recombination using embryonic stem cells. *J. Biol. Chem.* 269: 27155–27158. [R]

Capecchi, M. R. 1989. Altering the genome by homologous recombination. *Science* 244: 1288–1292. [R]

Gasser, C. S. and R. T. Fraley. 1989. Genetically engineering plants for crop improvement. *Science* 244: 1293–1299. [R]

Gordon, J. W., G. A. Scangos, D. J. Plotkin, J. A. Barbosa and F. H. Ruddle. 1980. Genetic transformation of mouse embryos by microinjection of purified DNA. *Proc. Natl. Acad. Sci. USA* 77: 7380–7384. [P]

Herskowitz, I. 1987. Functional inactivation of genes by dominant negative mutations. *Nature* 329: 219–222. [R]

Horsch, R. B., J. E. Fry, N. L. Hoffmann, D. Eichholtz, S. G. Rogers and R. T. Fraley. 1985. A simple and general method for transferring genes into plants. *Science* 227: 1229–1231. [P]

Hutvagner, G. and P. D. Zamore. 2002. RNAi: nature abhors a double-strand. *Curr. Opin. Genet. Dev.* 12: 225-232. [R]

Izant, J. G. and H. Weintraub. 1984. Inhibition of thymidine kinase gene expression by antisense RNA: A molecular approach to genetic analysis. *Cell* 36: 1007–1015. [P]

Jaenisch, R. 1988. Transgenic animals. *Science* 240: 1468–1474. [R]

Joyner, A. L., ed. 1993. *Gene Targeting. A Practical Approach.* Oxford, England: IRL Press.

Kuhn, R., F. Schwenk, M. Aguet and K. Rajewsky. 1995. Inducible gene targeting in mice. *Science* 269: 1427–1429. [P]

Maliga, P., D. F. Klessig, A. R. Cashmore, W. Gruissem and J. E. Varner, eds. 1994. *Methods in Plant Molecular Biology: A Laboratory Course Manual.* Plainview, N.Y.: Cold Spring Harbor Laboratory Press.

Palmiter, R. D. and R. L. Brinster. 1986. Germline transformation of mice. *Ann. Rev. Genet.* 20: 465–499. [R]

Robertson, E., A. Bradley, M. Kuehn and M. Evans. 1986. Germline transmission of genes introduced into cultured pluripotential cells by retroviral vector. *Nature* 323: 445–448. [P]

Sedivy, J. M. 1999. Gene targeting and somatic cell genetics: A rebirth or a coming of age? *Trends Genet.* 15: 88–90. [R]

Sharp, P. A. 2001. RNA interference—2001. *Genes Dev.* 15: 485–490. [R]

Smith, M. 1985. *In vitro* mutagenesis. *Ann. Rev. Genet.* 19: 423–462. [R]

Struhl, K. 1983. The new yeast genetics. *Nature* 305: 391–397. [R]

Thomas, K. R. and M. R. Capecchi, 1987. Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell* 51: 503–512. [P]

Wagner, R. W. 1994. Gene inhibition using antisense oligonucleotides. *Nature* 372: 333–335. [R]

Wigler, M., R. Sweet, G. K. Sim, B. Wold, A. Pellicer, E. Lacy, T. Maniatis, S. Silverstein and R. Axel. 1979. Transformation of mammalian cells with genes from prokaryotes and eukaryotes. *Cell* 16: 777–785. [P]

Chapter 5

Replication, Maintenance, and Rearrangements of Genomic DNA

THE FUNDAMENTAL BIOLOGICAL PROCESS OF REPRODUCTION requires the faithful transmission of genetic information from parent to offspring. Thus, the accurate replication of genomic DNA is essential to the lives of all cells and organisms. Each time a cell divides, its entire genome must be duplicated, and complex enzymatic machinery is required to copy the large DNA molecules that make up both prokaryotic and eukaryotic chromosomes. In addition, cells have evolved mechanisms to correct mistakes that sometimes occur during DNA replication and to repair DNA damage that can result from the action of environmental agents, such as radiation. Abnormalities in these processes result in a failure of accurate replication and maintenance of genomic DNA—a failure that can have disastrous consequences, such as the development of cancer.

Despite the importance of accurate DNA replication and maintenance, cell genomes are far from static. In order for species to evolve, mutations and gene rearrangements are needed to maintain genetic variation between individuals. Recombination between homologous chromosomes during meiosis plays an important role in this process by allowing parental genes to be rearranged into new combinations in the next generation. Rearrangements of DNA sequences within the genome are also thought to contribute to evolution by creating novel combinations of genetic information. In addition, some DNA rearrangements are programmed to regulate gene expression during the differentiation and development of individual cells and organisms. In humans, a prominent example is the rearrangement of antibody genes during development of the immune system. A careful balance between maintenance and variation of genetic information is thus critical to the development of individual organisms as well as to evolution of the species.

DNA Replication

As discussed in Chapter 3, DNA replication is a semiconservative process in which each parental strand serves as a template for the synthesis of a new complementary daughter strand. The central enzyme involved is DNA polymerase, which catalyzes the joining of deoxyribonucleoside 5′-triphosphates (dNTPs) to form the growing DNA chain. However, DNA replication is much more com-

plex than a single enzymatic reaction. Other proteins are involved, and proofreading mechanisms are required to ensure that the accuracy of replication is compatible with the low frequency of errors that is needed for cell reproduction. Additional proteins and specific DNA sequences are also needed both to initiate replication and to copy the ends of eukaryotic chromosomes.

DNA Polymerases

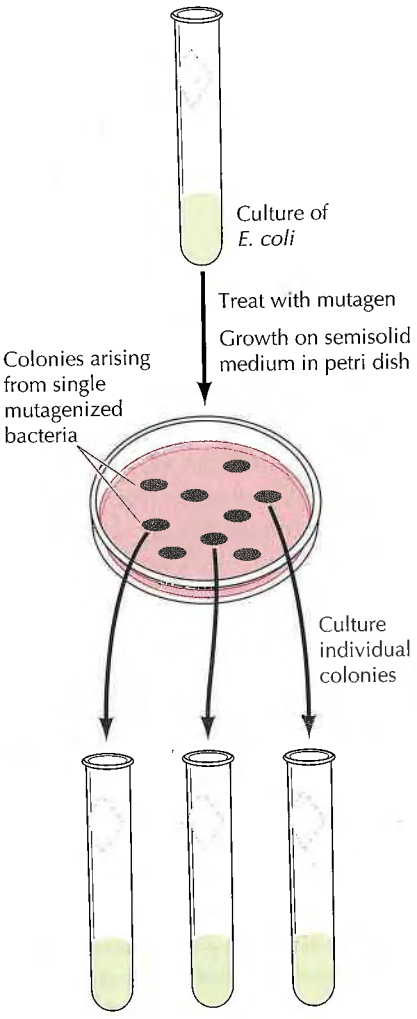
DNA polymerase was first identified in lysates of *E. coli* by Arthur Kornberg in 1956. The ability of this enzyme to accurately copy a DNA template provided a biochemical basis for the mode of DNA replication that was initially proposed by Watson and Crick, so its isolation represented a landmark discovery in molecular biology. Ironically, however, this first DNA polymerase to be identified (now called DNA polymerase I) is not the major enzyme responsible for *E. coli* DNA replication. Instead, it is now clear that both prokaryotic and eukaryotic cells contain several different DNA polymerases that play distinct roles in the replication and repair of DNA.

The multiplicity of DNA polymerases was first revealed by the isolation of a mutant strain of *E. coli* that was deficient in polymerase I (Figure 5.1). Cultures of *E. coli* were treated with a chemical (a **mutagen**) that induces a high frequency of mutations, and individual bacterial colonies were isolated and screened to identify a mutant strain lacking polymerase I. Analysis of a few thousand colonies led to the isolation of the desired mutant, which was almost totally defective in polymerase I activity. Surprisingly, the mutant bacteria grew normally, leading to the conclusion that polymerase I is not required for DNA replication. On the other hand, the mutant bacteria were extremely sensitive to agents that damage DNA (e.g., ultraviolet light), suggesting that polymerase I is involved primarily in the repair of DNA damage rather than in DNA replication per se.

The conclusion that polymerase I is not required for replication implied that *E. coli* must contain other DNA polymerases, and subsequent experiments led to the identification of two such enzymes, now called DNA polymerases II and III. The potential roles of these enzymes were investigated by the isolation of appropriate mutants. Strains of *E. coli* with mutations in polymerase II were found to grow normally, and the role of this enzyme in a specialized form of error-prone DNA repair (discussed in the section "DNA Repair") has only recently been established. Temperature-sensitive polymerase III mutants, however, were unable to replicate their DNA at high temperature, and subsequent studies have confirmed that polymerase III is the major replicative enzyme in *E. coli*.

It is now known that, in addition to polymerase III, polymerase I is also required for replication of *E. coli* DNA. The original polymerase I mutant was not completely defective in that enzyme, and later experiments showed that the residual polymerase I activity in this strain plays a key role in the replication process. The replication of *E. coli* DNA thus involves two distinct DNA polymerases, the specific roles of which are discussed below.

Eukaryotic cells contain five classical DNA polymerases: α, β, γ, δ, and ε. Polymerase γ is located in mitochondria and is responsible for replication of mitochondrial DNA. The other four enzymes are located in the nucleus and are therefore candidates for involvement in nuclear DNA replication. Polymerases α, δ, and ε are most active in dividing cells, suggesting that they function in replication. In contrast, polymerase β is active in nondividing as well as dividing cells, consistent with its function in the repair of DNA damage.



Screen for DNA polymerase activity to identify mutant lacking polymerase I

Figure 5.1 Isolation of a mutant deficient in polymerase I
A culture of *E. coli* was treated with a chemical mutagen, and individual bacterial colonies were isolated by growth on semisolid medium. Several thousand colonies were then cultured and screened to identify a mutant lacking polymerase I.

Two types of experiments have provided further evidence addressing the roles of polymerases α, δ, and ε in DNA replication. First, replication of the DNAs of some animal viruses, such as SV40, can be studied in cell-free extracts. The ability to study replication *in vitro* has allowed direct identification of the enzymes involved, and analysis of such cell-free systems has shown that polymerases α and δ are required for SV40 DNA replication. Second, polymerases α, δ, and ε are found in yeasts as well as in mammalian cells, enabling the use of the powerful approaches of yeast genetics (see Chapter 3) to test their biological roles directly. Such studies indicate that yeast mutants lacking any of these three DNA polymerases are unable to proliferate, implying a critical role for polymerase ε as well as for α and δ. However, further studies have shown that the essential function of polymerase ε in yeast does not require its enzymatic activity as a DNA polymerase. The role of polymerase ε in DNA replication thus remains unclear, although it probably functions similarly to polymerase δ, which is sufficient to catalyze DNA replication in the absence of polymerase ε both in cell-free systems and in yeast.

All known DNA polymerases share two fundamental properties that carry critical implications for DNA replication (Figure 5.2). First, all polymerases synthesize DNA only in the 5′ to 3′ direction, adding a dNTP to the 3′ hydroxyl group of a growing chain. Second, DNA polymerases can add a new deoxyribonucleotide only to a preformed primer strand that is hydrogen-bonded to the template; they are not able to initiate DNA synthesis *de novo* by catalyzing the polymerization of free dNTPs. In this respect, DNA

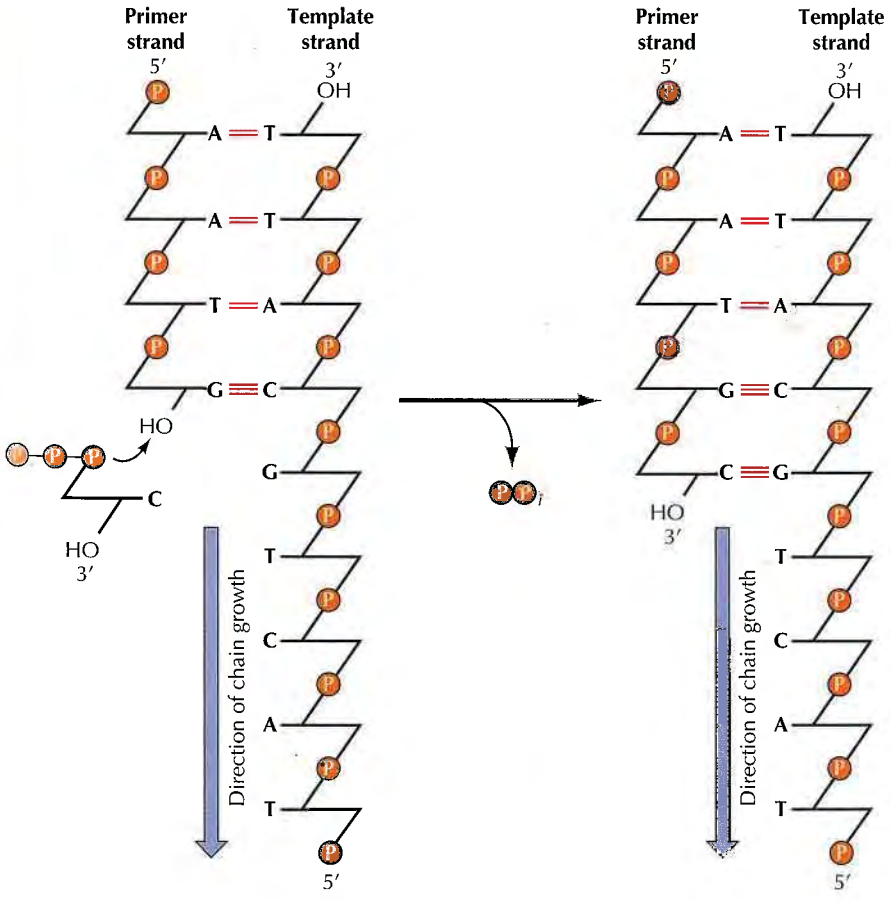


Figure 5.2 The reaction catalyzed by DNA polymerase
All known DNA polymerases add a deoxyribonucleoside 5′-triphosphate to the 3′ hydroxyl group of a growing DNA chain (the primer strand).

polymerases differ from RNA polymerases, which can initiate the synthesis of a new strand of RNA in the absence of a primer. As discussed later in this chapter, these properties of DNA polymerases appear critical for maintaining the high fidelity of DNA replication that is required for cell reproduction.

The Replication Fork

DNA molecules in the process of replication were first analyzed by John Cairns in experiments in which *E. coli* were grown in the presence of radioactive thymidine, which allowed subsequent visualization of newly replicated DNA by autoradiography (Figure 5.3). In some cases, complete circular molecules in the process of replicating could be observed. These DNA molecules contained two **replication forks**, representing the regions of active DNA synthesis. At each fork the parental strands of DNA separated and two new daughter strands were synthesized.

The synthesis of new DNA strands complementary to both strands of the parental molecule posed an important problem to understanding the biochemistry of DNA replication. Since the two strands of double-helical DNA run in opposite (antiparallel) directions, continuous synthesis of two new strands at the replication fork would require that one strand be synthesized in the 5′ to 3′ direction while the other is synthesized in the opposite (3′ to 5′) direction. But DNA polymerase catalyzes the polymerization of dNTPs only in the 5′ to 3′ direction. How, then, can the other progeny strand of DNA be synthesized?

This enigma was resolved by experiments showing that only one strand of DNA is synthesized in a continuous manner in the direction of overall DNA replication; the other is formed from short (1–3 kb), discontinuous pieces of DNA that are synthesized backward with respect to the direction of movement of the replication fork (Figure 5.4). These small pieces of newly synthesized DNA (called **Okazaki fragments** after their discoverer) are joined by the action of **DNA ligase**, forming an intact new DNA strand. The continuously synthesized strand is called the **leading strand**, since its elongation in the direction of replication fork movement exposes the template used for the synthesis of Okazaki fragments (the **lagging strand**).

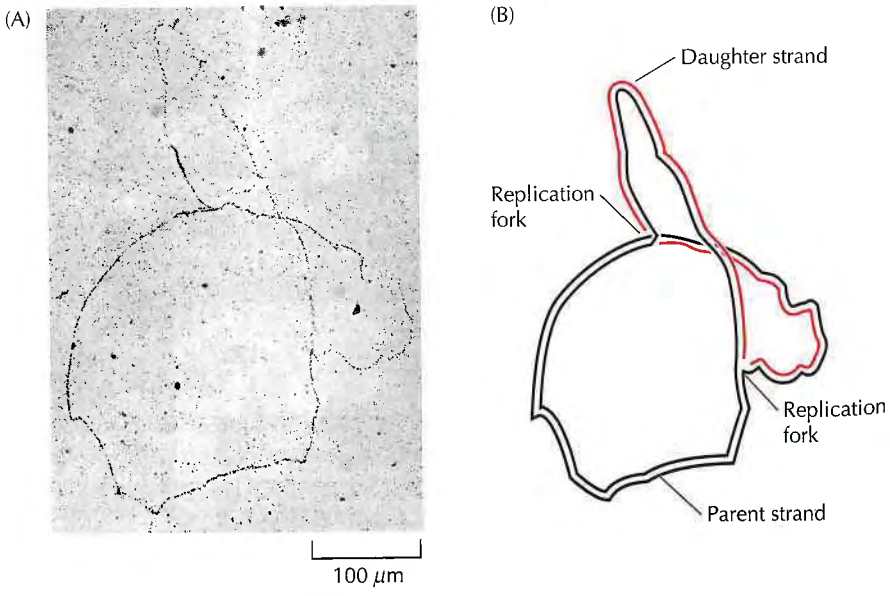


Figure 5.3 Replication of *E. coli* DNA (A) An autoradiograph showing bacteria that were grown in [^{3}H]thymidine for two generations to label the DNA, which was then extracted and visualized by exposure to photographic film. (B) This schematic illustrates the two replication forks shown in (A). (From J. Cairns, 1963. *Cold Spring Harbor Symp. Quant. Biol.* 28: 43.)

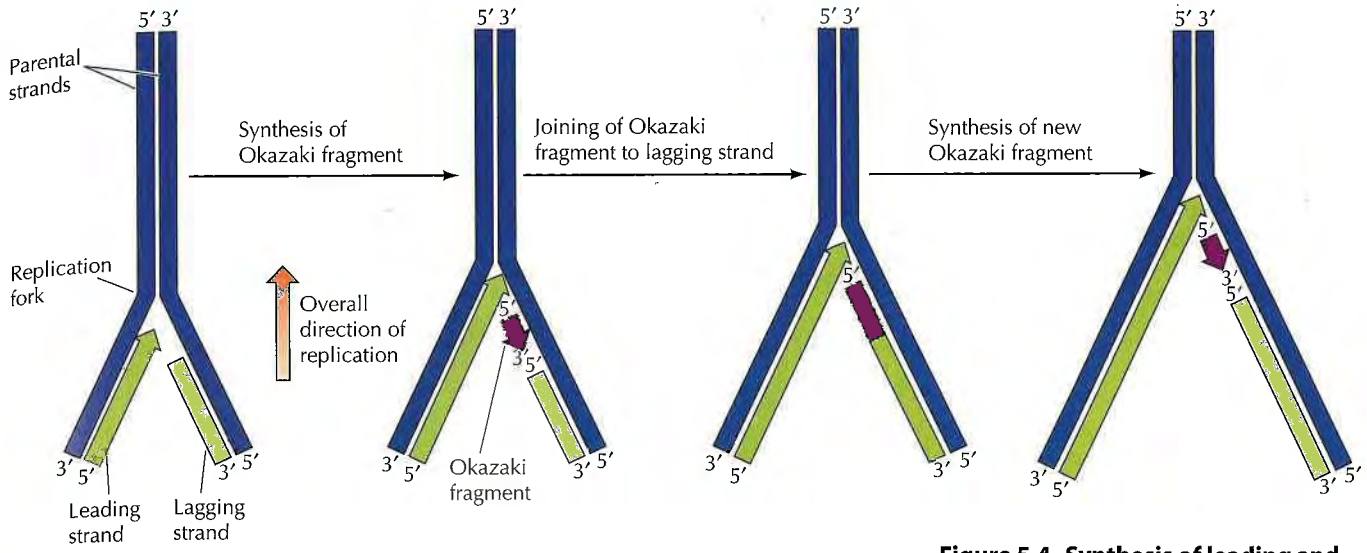
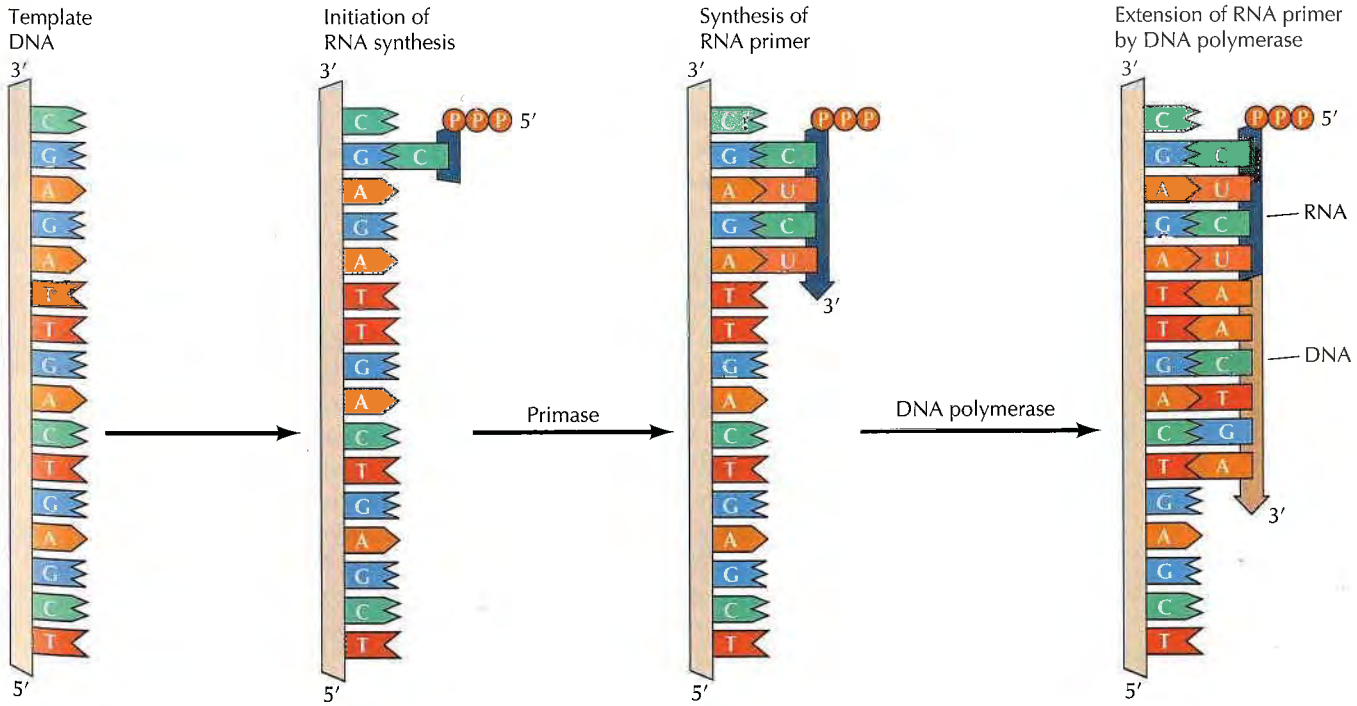


Figure 5.4 Synthesis of leading and lagging strands of DNA
The leading strand is synthesized continuously in the direction of replication fork movement. The lagging strand is synthesized in small pieces (Okazaki fragments) backward from the overall direction of replication. The Okazaki fragments are then joined by the action of DNA ligase.

Although the discovery of discontinuous synthesis of the lagging strand provided a mechanism for the elongation of both strands of DNA at the replication fork, it raised another question: Since DNA polymerase requires a primer and cannot initiate synthesis *de novo*, how is the synthesis of Okazaki fragments initiated? The answer is that short fragments of RNA serve as primers for DNA replication (Figure 5.5). In contrast to DNA synthesis, the synthesis of RNA can initiate *de novo*, and an enzyme called **primase** synthesizes short fragments of RNA (e.g., three to ten nucleotides long) complementary to the lagging strand template at the replication fork. Okazaki fragments are then synthesized via extension of these RNA primers by DNA polymerase. An important consequence of such RNA priming is that newly synthesized Okazaki fragments contain an

Figure 5.5 Initiation of Okazaki fragments with RNA primers
Short fragments of RNA serve as primers that can be extended by DNA polymerase.



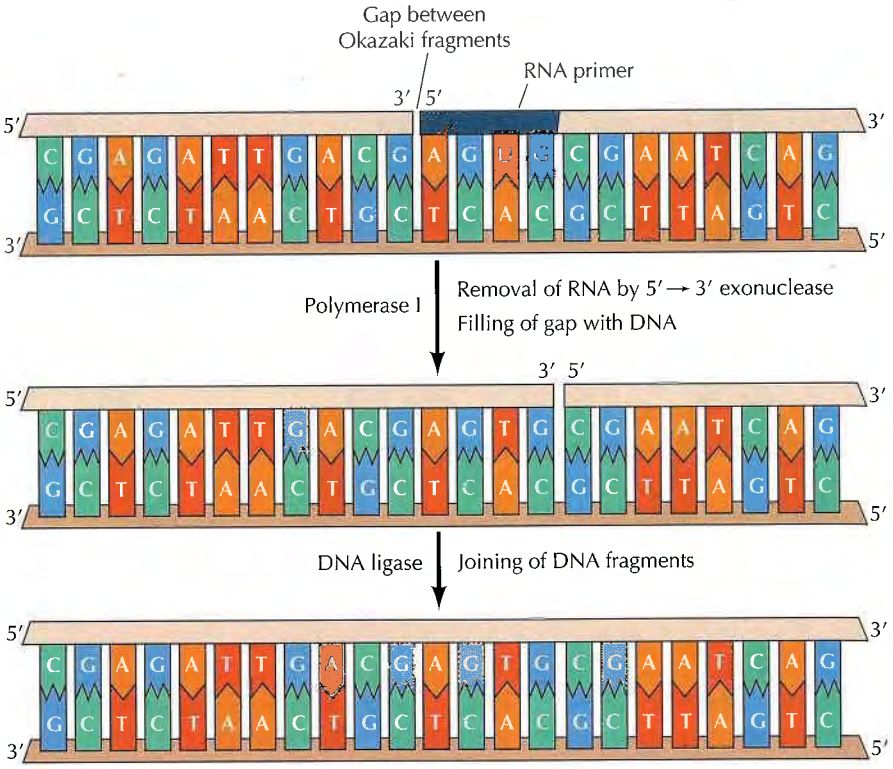


Figure 5.6 Removal of RNA primers and joining of Okazaki fragments
Because of its 5′ to 3′ exonuclease activity, DNA polymerase I removes RNA primers and fills the gaps between Okazaki fragments with DNA. The resultant DNA fragments can then be joined by DNA ligase.

RNA-DNA joint, the discovery of which provided critical evidence for the role of RNA primers in DNA replication.

To form a continuous lagging strand of DNA, the RNA primers must eventually be removed from the Okazaki fragments and replaced with DNA. In *E. coli*, RNA primers are removed by the combined action of **RNase H**, an enzyme that degrades the RNA strand of RNA-DNA hybrids, and polymerase I. This is the aspect of *E. coli* DNA replication in which polymerase I plays a critical role. In addition to its DNA polymerase activity, polymerase I acts as an **exonuclease** that can hydrolyze DNA (or RNA) in either the 3′ to 5′ or 5′ to 3′ direction. The action of polymerase I as a 5′ to 3′ exonuclease removes ribonucleotides from the 5′ ends of Okazaki fragments, allowing them to be replaced with deoxyribonucleotides to yield fragments consisting entirely of DNA (Figure 5.6). In eukaryotic cells, other exonucleases take the place of *E. coli* polymerase I in removing primers, and the gaps between Okazaki fragments are filled by the action of polymerase δ. As in prokaryotes, these DNA fragments can then be joined by DNA ligase.

The different DNA polymerases thus play distinct roles at the replication fork (Figure 5.7). In prokaryotic cells, polymerase III is the major replicative polymerase, functioning in the synthesis both of the leading strand of DNA and of Okazaki fragments by the extension of RNA primers. Polymerase I then removes RNA primers and fills the gaps between Okazaki fragments. In eukaryotic cells, polymerase α is found in a complex with primase, and it appears to function in conjunction with primase to synthesize short RNA-DNA fragments during lagging strand synthesis. Polymerase δ can then synthesize both the leading and lagging strands, acting to extend the RNA-DNA primers initially synthesized by the polymerase α-primase complex. In addition, polymerase δ can take the place of *E. coli* polymerase I in filling the gaps between Okazaki fragments following primer removal. Although the roles of polymerase ε remain to be fully understood, its activities seem to be similar to those of polymerase δ.

Not only polymerases and primase but also a number of other proteins act at the replication fork. These additional proteins have been identified both by

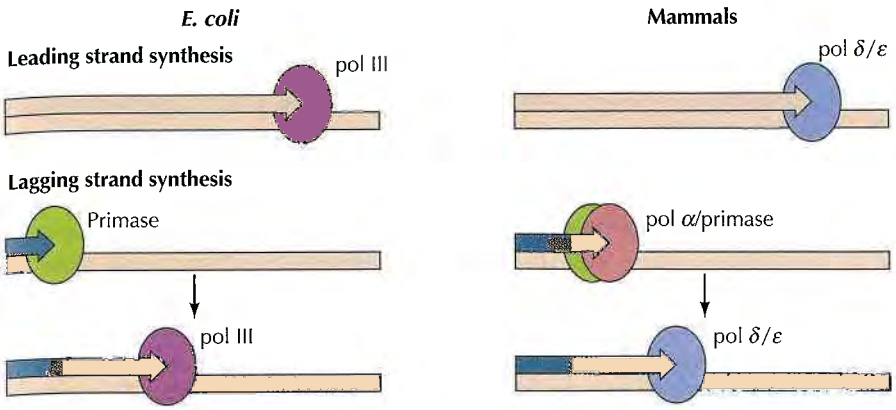


Figure 5.7 Roles of DNA polymerases in *E. coli* and mammalian cells
The leading strand is synthesized by polymerase III (pol III) in *E. coli* and by polymerases δ and ε (pol δ/ε) in mammalian cells. In *E. coli*, lagging strand synthesis is initiated by primase, and RNA primers are extended by polymerase III. In mammalian cells, lagging strand synthesis is initiated by a complex of primase with polymerase α (pol α). The short RNA-DNA fragments synthesized by this complex are then extended by polymerases δ and ε.

the analysis of *E. coli* mutants defective in DNA replication and by the purification of the mammalian proteins required for *in vitro* replication of SV40 DNA. One class of proteins required for replication binds to DNA polymerases, increasing the activity of the polymerases and causing them to remain bound to the template DNA so that they continue synthesis of a new DNA strand. Both *E. coli* polymerase III and eukaryotic polymerases δ and ε are associated with two types of accessory proteins (sliding-clamp proteins and clamp-loading proteins) that load the polymerase onto the primer and maintain its stable association with the template (Figure 5.8). The clamp-

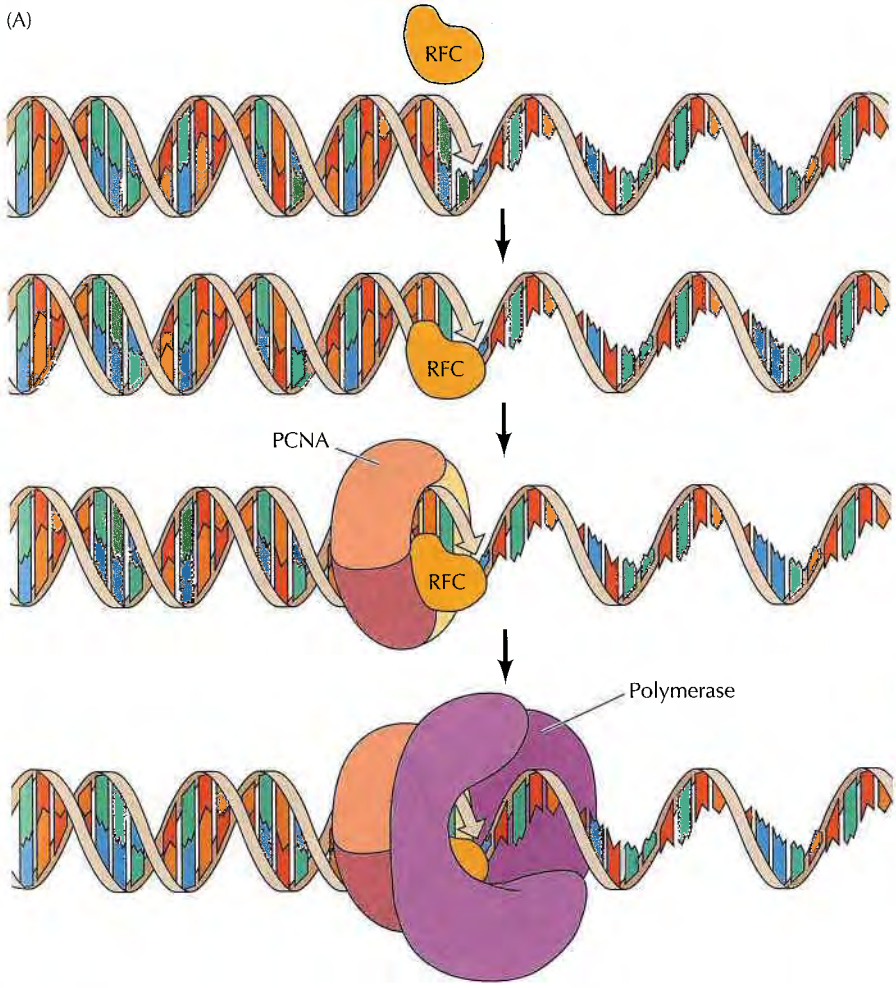
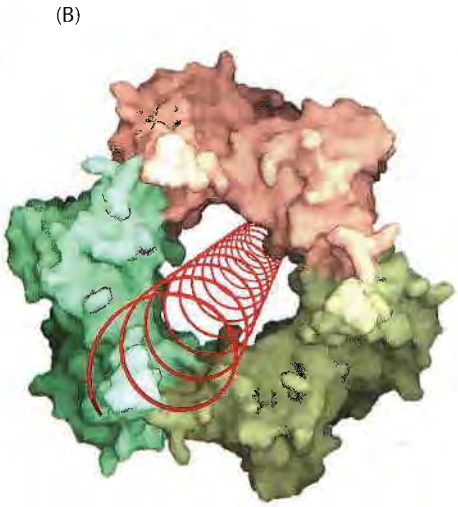


Figure 5.8 Polymerase accessory proteins
(A) The clamp-loading protein (RFC in mammalian cells) binds DNA at the junction between primer and template. The sliding-clamp protein (PCNA in mammalian cells) binds adjacent to the RFC, and DNA polymerase then binds to PCNA. (B) Model of PCNA bound to DNA. (B, from T. S. Krishna, X. P. Kong, S. Gary, P. M. Burgers and J. Kuriyan, 1994. *Cell* 79: 1233.)



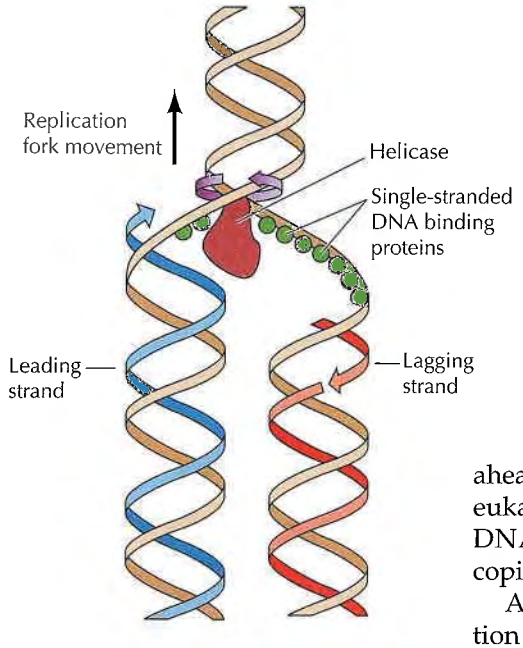


Figure 5.9 Action of helicases and single-stranded DNA-binding proteins
Helicases unwind the two strands of parental DNA ahead of the replication fork. The unwound DNA strands are then stabilized by single-stranded DNA-binding proteins so that they can serve as templates for new DNA synthesis.

loading proteins (called the γ complex in *E. coli* and replication factor C [RFC] in eukaryotes) specifically recognize and bind DNA at the junction between the primer and template. The sliding-clamp proteins (β protein in *E. coli* and proliferating cell nuclear antigen [PCNA] in eukaryotes) bind adjacent to the clamp-loading proteins, forming a ring around the template DNA. The sliding-clamp proteins then load the DNA polymerase onto DNA at the primer-template junction. The ring formed by the sliding clamp maintains the association of the polymerase with its template as replication proceeds, allowing the uninterrupted synthesis of many thousands of nucleotides of DNA.

Other proteins unwind the template DNA and stabilize single-stranded regions (Figure 5.9). **Helicases** are enzymes that catalyze the unwinding of parental DNA, coupled to the hydrolysis of ATP, ahead of the replication fork. **Single-stranded DNA-binding proteins** (e.g., eukaryotic replication factor A [RFA]) then stabilize the unwound template DNA, keeping it in an extended single-stranded state so that it can be copied by the polymerase.

As the strands of parental DNA unwind, the DNA ahead of the replication fork is forced to rotate. Unchecked, this rotation would cause circular DNA molecules (such as SV40 DNA or the *E. coli* chromosome) to become twisted around themselves, eventually blocking replication (Figure 5.10). This problem is solved by **topoisomerases,** enzymes that catalyze the reversible breakage and rejoining of DNA strands. There are two types of these enzymes: Type I topoisomerases break just one strand of DNA; type II topoisomerases introduce simultaneous breaks in both strands. The breaks introduced by type I and type II topoisomerases serve as "swivels" that allow the two strands of template DNA to rotate freely around each other so that replication can proceed without twisting the DNA ahead of the fork (see Figure 5.10). Although eukaryotic chromosomes are composed of linear rather than circular DNA molecules, their replication also requires topoisomerases; otherwise, the complete chromosomes would have to rotate continually during DNA synthesis.

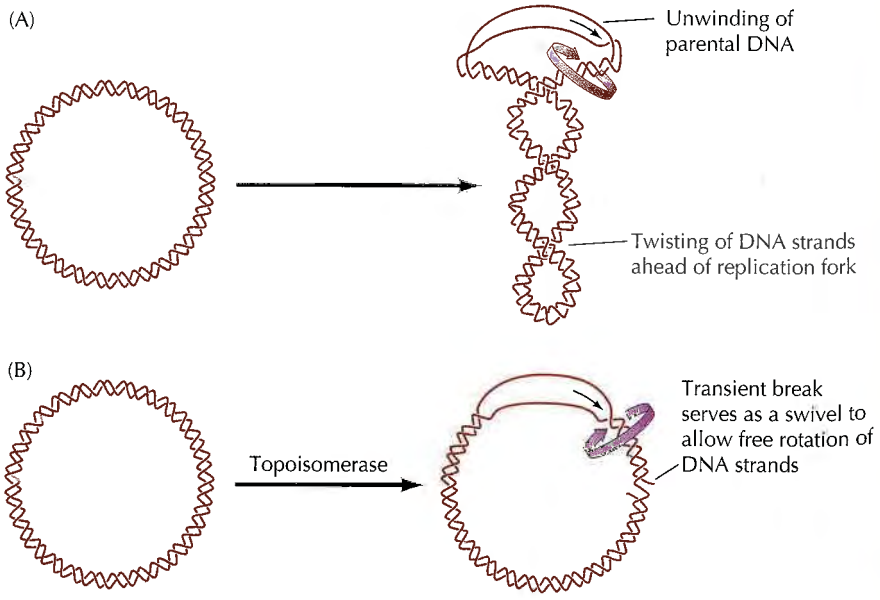


Figure 5.10 Action of topoisomerases during DNA replication
(A) As the two strands of template DNA unwind, the DNA ahead of the replication fork is forced to rotate in the opposite direction, causing circular molecules to become twisted around themselves. (B) This problem is solved by topoisomerases, which catalyze the reversible breakage and joining of DNA strands. The transient breaks introduced by these enzymes serve as swivels that allow the two strands of DNA to rotate freely around each other.

Type II topoisomerase is needed not only to unwind DNA but also to unravel newly replicated circular DNA molecules that become interwined with each other. In eukaryotic cells, topoisomerase II appears to be involved in mitotic chromosome condensation. In addition, studies of yeast mutants, as well as experiments in *Drosophila* and mammalian cells, indicate that topoisomerase II is required for the separation of daughter chromatids at mitosis, suggesting that it functions to untangle newly replicated loops of DNA in the chromosomes of eukaryotes.

The enzymes involved in DNA replication act in a coordinated manner to synthesize both leading and lagging strands of DNA simultaneously at the replication fork (Figure 5.11). This task is accomplished by the formation of dimers of the replicative DNA polymerases (polymerase III in *E. coli* or polymerase δ/ε in eukaryotes), each with its appropriate accessory proteins. One molecule of polymerase then acts in synthesis of the leading strand while the other acts in synthesis of the lagging strand. The lagging strand template is thought to form a loop at the replication fork so that the polymerase subunit engaged in lagging strand synthesis moves in the same overall direction as the other subunit, which is synthesizing the leading strand.

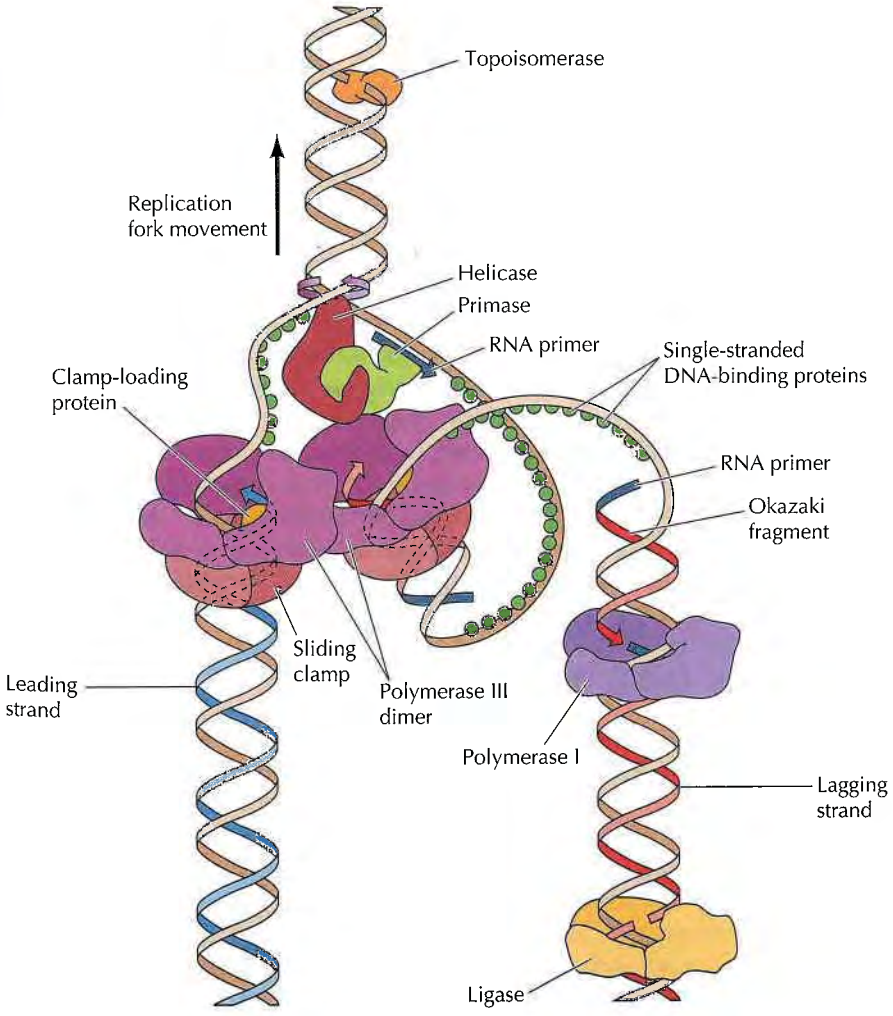


Figure 5.11 Model of the *E. coli* replication fork
Helicase, primase, and two molecules of DNA polymerase III carry out coordinated synthesis of both the leading and lagging strands of DNA. The lagging strand template is folded so that the polymerase responsible for lagging strand synthesis moves in the same direction as overall movement of the fork. Topoisomerase acts as a swivel ahead of the fork, and DNA polymerase I and ligase remove RNA primers and join Okazaki fragments behind the fork.

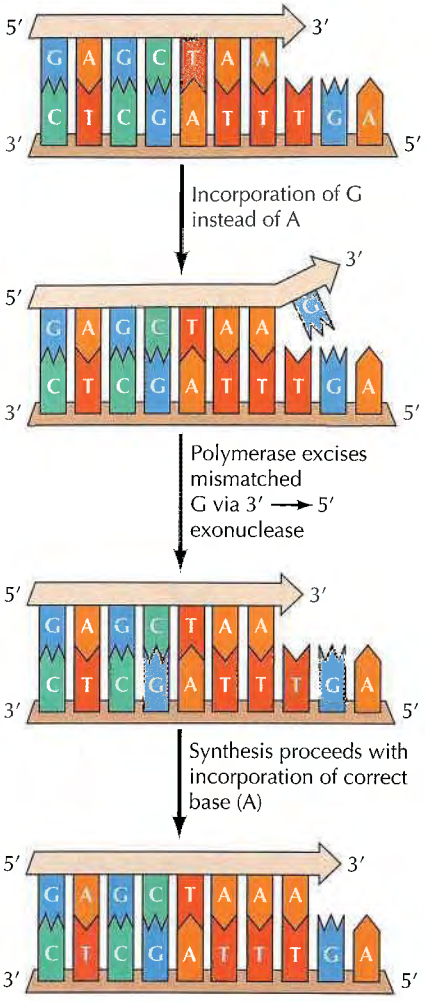


Figure 5.12 Proofreading by DNA polymerase
G is incorporated in place of A as a result of mispairing with T on the template strand. Because it is mispaired, the 3′ terminal G is not hydrogen-bonded to the template strand. This mismatch at the 3′ terminus of the growing chain is recognized and excised by the 3′ to 5′ exonuclease activity of DNA polymerase, which requires a primer hydrogen-bonded to the template strand in order to continue synthesis. Following excision of the mismatched G, DNA synthesis can proceed with incorporation of the correct nucleotide (A).

The Fidelity of Replication

The accuracy of DNA replication is critical to cell reproduction, and estimates of mutation rates for a variety of genes indicate that the frequency of errors during replication corresponds to only one incorrect base per 10^9 to 10^{10} nucleotides incorporated. This error frequency is much lower than would be predicted simply on the basis of complementary base pairing. In particular, the free energy differences resulting from the changes in hydrogen bonding between correctly matched and mismatched bases are only large enough to favor the formation of correctly matched base pairs by about 1,000 fold. Consequently, base selection determined simply by hydrogen bonding between complementary bases would result in an error frequency corresponding to the incorporation of about one incorrect base per 10^3. The much higher degree of fidelity actually achieved results largely from the activities of DNA polymerase.

One mechanism by which DNA polymerase increases the fidelity of replication is by helping to select the correct base for insertion into newly synthesized DNA. The polymerase does not simply catalyze incorporation of whatever nucleotide is hydrogen-bonded to the template strand. Instead, it actively discriminates against incorporation of a mismatched base by adapting to the conformation of a correct base pair. In particular, recent structural studies of several DNA polymerases indicate that the binding of correctly matched dNTPs induces conformational changes in DNA polymerase that lead to the incorporation of the nucleotide into DNA. This ability of DNA polymerase to select for incorporation of matched nucleotides appears to increase the accuracy of replication about a thousandfold, reducing the expected error frequency from 10^{-3} to approximately 10^{-6}.

The other major mechanism responsible for the accuracy of DNA replication is the **proofreading** activity of DNA polymerase. As already noted, *E. coli* polymerase I has 3′ to 5′ as well as 5′ to 3′ exonuclease activity. The 5′ to 3′ exonuclease operates in the direction of DNA synthesis and helps remove RNA primers from Okazaki fragments. The 3′ to 5′ exonuclease operates in the reverse direction of DNA synthesis, and participates in proofreading newly synthesized DNA (Figure 5.12). Proofreading is effective because DNA polymerase requires a primer and is not able to initiate synthesis *de novo*. Primers that are hydrogen-bonded to the template are preferentially used, so when an incorrect base is incorporated, it is likely to be removed by the 3′ to 5′ exonuclease activity rather than being used to continue synthesis. Such 3′ to 5′ exonuclease activities are also associated with *E. coli* polymerase III and eukaryotic polymerases δ and ε. The 3′ to 5′ exonucleases of these polymerases selectively excise mismatched bases that have been incorporated at the end of a growing DNA chain, thereby increasing the accuracy of replication by a hundred- to a thousandfold.

The importance of proofreading may explain the fact that DNA polymerases require primers and catalyze the growth of DNA strands only in the 5′ to 3′ direction. When DNA is synthesized in the 5′ to 3′ direction, the energy required for polymerization is derived from hydrolysis of the 5′ triphosphate group of a free dNTP as it is added to the 3′ hydroxyl group of the growing chain (see Figure 5.2). If DNA were to be extended in the 3′ to 5′ direction, the energy of polymerization would instead have to be derived from hydrolysis of the 5′ triphosphate group of the terminal nucleotide already incorporated into DNA. This would eliminate the possibility of proofreading, because removal of a mismatched terminal nucleotide would also remove the 5′ triphosphate group needed as an energy source for further chain elongation. Thus, although the ability of DNA polymerase to extend a primer only in the

5′ to 3′ direction appears to make replication a complicated process, it is necessary for ensuring accurate duplication of the genetic material.

Combined with the ability to discriminate against the insertion of mismatched bases, the proofreading activity of DNA polymerases is sufficient to reduce the error frequency of replication to about one mismatched base per 10^9. Additional mechanisms (discussed in the section "DNA Repair") act to remove mismatched bases that have been incorporated into newly synthesized DNA, further ensuring correct replication of the genetic information.

Origins and the Initiation of Replication

The replication of both prokaryotic and eukaryotic DNAs starts at a unique sequence called the **origin of replication**, which serves as a specific binding site for proteins that initiate the replication process. The first origin to be defined was that of *E. coli*, in which genetic analysis indicated that replication always begins at a unique site on the bacterial chromosome. The *E. coli* origin has since been studied in detail and found to consist of 245 base pairs of DNA, elements of which serve as binding sites for proteins required to initiate DNA replication (Figure 5.13). The key step is the binding of an initiator protein to specific DNA sequences within the origin. The initiator protein begins to unwind the origin DNA and recruits the other proteins involved in DNA synthesis. Helicase and single-stranded DNA-binding proteins then act to continue unwinding and exposing the template DNA, and primase initiates the synthesis of leading strands. Two replication forks are formed and move in opposite directions along the circular *E. coli* chromosome.

The origins of replication of several animal viruses, such as SV40, have been studied as models for the initiation of DNA synthesis in eukaryotes. SV40 has a single origin of replication (consisting of 64 base pairs) that functions both in infected cells and in cell-free systems. Replication is initiated by a virus-encoded protein (called T antigen) that binds to the origin and also acts as a helicase. A single-stranded DNA-binding protein is required to stabilize the unwound template, and the DNA polymerase α-primase complex then initiates DNA synthesis.

Although single origins are sufficient to direct the replication of bacterial and viral genomes, multiple origins are needed to replicate the much larger genomes of eukaryotic cells within a reasonable period of time. For example, the entire genome of *E. coli* (4×10^6 base pairs) is replicated from a single origin in approximately 30 minutes. If mammalian genomes (3×10^9 base pairs) were replicated from a single origin at the same rate, DNA replication would require about 3 weeks (30,000 minutes). The problem is further exacerbated by the fact that the rate of DNA replication in mammalian cells is actually about tenfold lower than in *E. coli*, possibly as a result of the

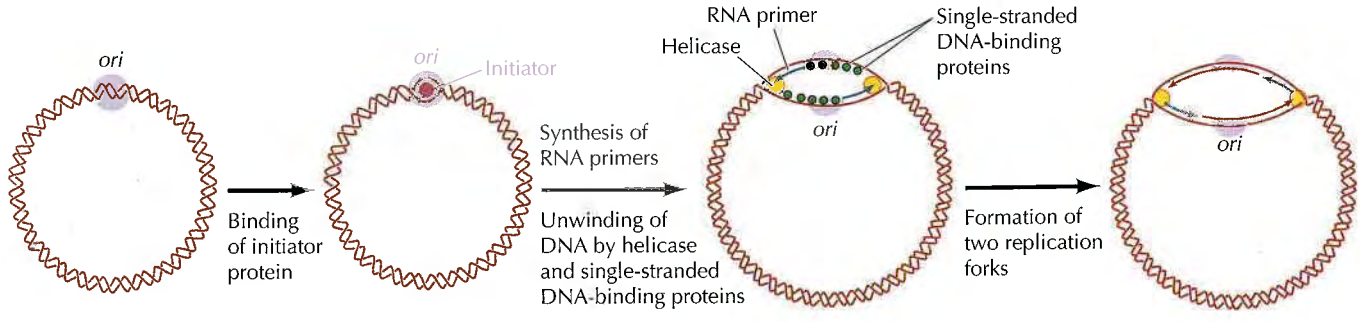


Figure 5.13 Origin of replication in *E. coli*
Replication initiates at a unique site on the *E. coli* chromosome, designated the origin (*ori*). The first event is the binding of an initiator protein to *ori* DNA, which leads to partial unwinding of the template. The DNA continues to unwind by the actions of helicase and single-stranded DNA-binding proteins, and RNA primers are synthesized by primase. The two replication forks formed at the origin then move in opposite directions along the circular DNA molecule.

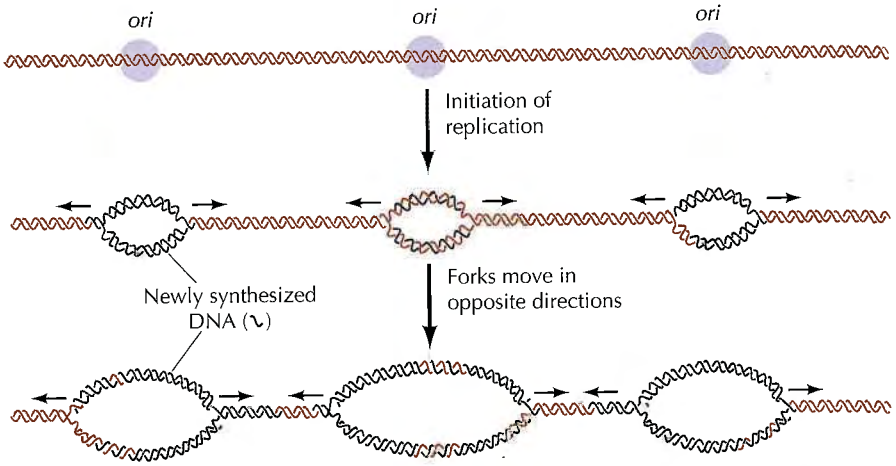


Figure 5.14 Replication origins in eukaryotic chromosomes
Replication initiates at multiple origins (*ori*), each of which produces two replication forks.

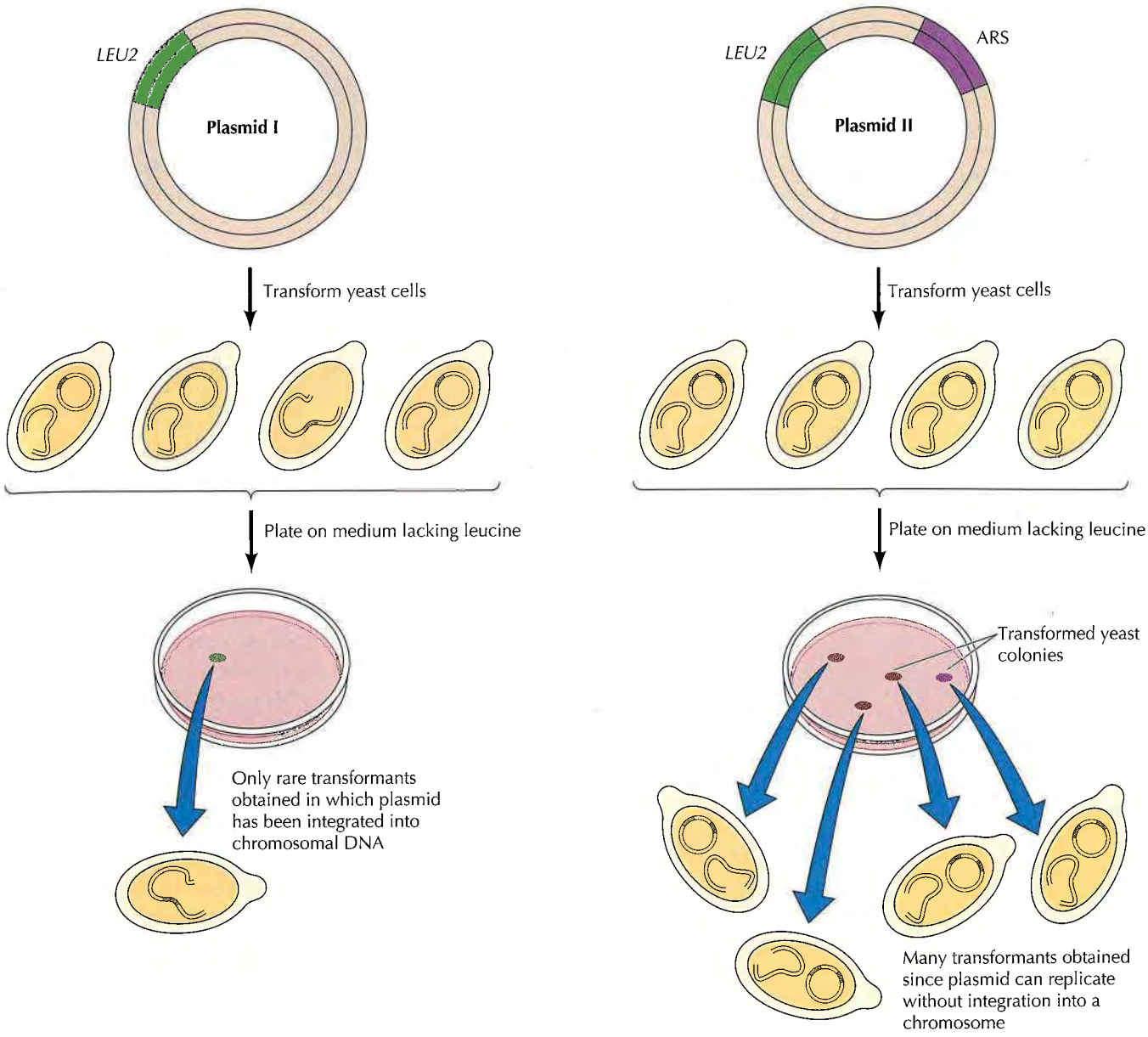
packaging of eukaryotic DNA in chromatin. Nonetheless, the genomes of mammalian cells are typically replicated within a few hours, necessitating the use of thousands of replication origins.

The presence of multiple replication origins in eukaryotic cells was first demonstrated by the exposure of cultured mammalian cells to radioactive thymidine for different time intervals, followed by autoradiography to detect newly synthesized DNA. The results of such studies indicated that DNA synthesis is initiated at multiple sites, from which it then proceeds in both directions along the chromosome (Figure 5.14). The replication origins in mammalian cells are spaced at intervals of approximately 50 to 300 kb; thus the human genome has about 30,000 origins of replication. The genomes of simpler eukaryotes also have multiple origins; for example, replication in yeasts initiates at origins separated by intervals of approximately 40 kb.

The origins of replication of eukaryotic chromosomes were first studied in the yeast *S. cerevisiae*, in which they were identified as sequences that can support the replication of plasmids in transformed cells (Figure 5.15). This has provided a functional assay for these sequences, and several such elements (called **autonomously replicating sequences,** or **ARSs**) have been isolated. Their role as origins of replication has been verified by direct biochemical analysis, not only in plasmids but also in yeast chromosomal DNA.

Functional ARS elements span about 100 base pairs, including an 11-base-pair core sequence common to many different ARSs (Figure 5.16). This core sequence is essential for ARS function and has been found to be the binding site of a protein complex (called the **origin replication complex**, or **ORC**) that is required for initiation of DNA replication at *S. cerevisiae* origins. The ORC complex appears to recruit other proteins (including DNA helicases) to the origin, leading to the initiation of replication. The mechanism of initiation of DNA replication in *S. cerevisiae* thus appears similar to that in prokaryotes and eukaryotic viruses; that is, an initiator protein specifically binds to origin sequences.

Subsequent studies have shown that the role of ORC proteins as initiators of replication is conserved in all eukaryotes, from yeasts to mammals. However, replication origins in other eukaryotes are much less well defined than the ARS elements of *S. cerevisiae*. In the fission yeast *S. pombe*, origin sequences are spread over about 1 kb of DNA. The *S. pombe* origins lack the clearly defined ORC binding site of the *S cerevisiae* ARS elements, but they contain repeats of AT-rich sequences that appear to serve as binding sites



for the *S. pombe* ORC complex. A *Drosophila* replication origin has been found to span over 2 kb of DNA and to contain several ORC binding sites, but these sequences have not been defined. In mammals, some origins have been localized to a few kb of DNA. In other cases, however, replication may initiate at multiple origins within large "initiation zones" spanning 10 to 50 kb. It thus appears that the sequences that define replication origins vary widely among eukaryotes, although the role of ORC proteins as initiators of replication is highly conserved.

Figure 5.15 Identification of origins of replication in yeast
Both plasmids I and II contain a selectable marker gene (*LEU2*) that allows transformed cells to grow on medium lacking leucine. Only plasmid II, however, contains an origin of replication (ARS). The transformation of yeasts with plasmid I yields only rare transformants in which the plasmid has integrated into chromosomal DNA. Plasmid II, however, is able to replicate without integration into a yeast chromosome (autonomous replication), so many more transformants result from its introduction into yeast cells.

Telomeres and Telomerase: Replicating the Ends of Chromosomes

Because DNA polymerases extend primers only in the 5′ to 3′ direction, they are unable to copy the extreme 5′ ends of linear DNA molecules. Consequently, special mechanisms are required to replicate the terminal sequences of the linear chromosomes of eukaryotic cells. These sequences (**telomeres**) consist of tandem repeats of simple-sequence DNA (see Chapter 4). They are replicated by the action of a unique enzyme called **telo-**

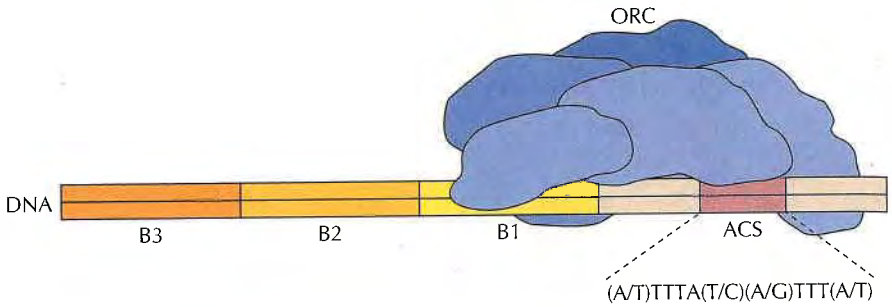


Figure 5.16 A yeast ARS element
The element contains an 11-base-pair ARS consensus sequence (ACS), which is the specific binding site of the origin replication complex (ORC). Three additional elements (B1, B2, and B3) are individually not essential but together contribute to ARS function.

merase, which is able to maintain telomeres by catalyzing their synthesis in the absence of a DNA template.

Telomerase is a **reverse transcriptase**, one of a class of DNA polymerases, first discovered in retroviruses (see Chapter 3), that synthesize DNA from an RNA template. Importantly, telomerase carries its own template RNA, which is complementary to the telomere repeat sequences, as part of the enzyme complex. The use of this RNA as a template allows telomerase to generate multiple copies of the telomeric repeat sequences, thereby maintaining telomeres in the absence of a conventional DNA template to direct their synthesis.

The mechanism of telomerase action was initially elucidated in 1985 by Carol Greider and Elizabeth Blackburn during studies of the protozoan *Tetrahymena* (Figure 5.17). The *Tetrahymena* telomerase is complexed to a 159-nucleotide-long RNA that includes the sequence 3′–AACCCCAAC–5′. This sequence is complementary to the *Tetrahymena* telomeric repeat (5′–TTGGGG–3′) and serves as the template for the synthesis of telomeric DNA. The use of this RNA as a template allows telomerase to extend the 3′ end of chromosomal DNA by one repeat unit beyond its original length. The complementary strand can then be synthesized by the polymerase α-primase complex using conventional RNA priming. Removal of the RNA primer leaves an overhanging 3′ end of chromosomal DNA, which can form loops at the ends of eukaryotic chromosomes (see Figure 4.22).

Telomerase has been identified in a variety of eukaryotes, and genes encoding telomerase RNAs have been cloned from *Tetrahymena*, yeasts, mice, and humans. In each case, the telomerase RNA contains sequences complementary to the telomeric repeat sequence of that organism (see Table 4.4). Moreover, the introduction of mutant telomerase RNA genes into yeasts has been shown to result in corresponding alterations of the chromosomal telomeric repeat sequences, directly demonstrating the function of telomerase in maintaining the termini of eukaryotic chromosomes.

DNA Repair

DNA, like any other molecule, can undergo a variety of chemical reactions. Because DNA uniquely serves as a permanent copy of the cell genome, however, changes in its structure are of much greater consequence than are alterations in other cell components, such as RNAs or proteins. Mutations

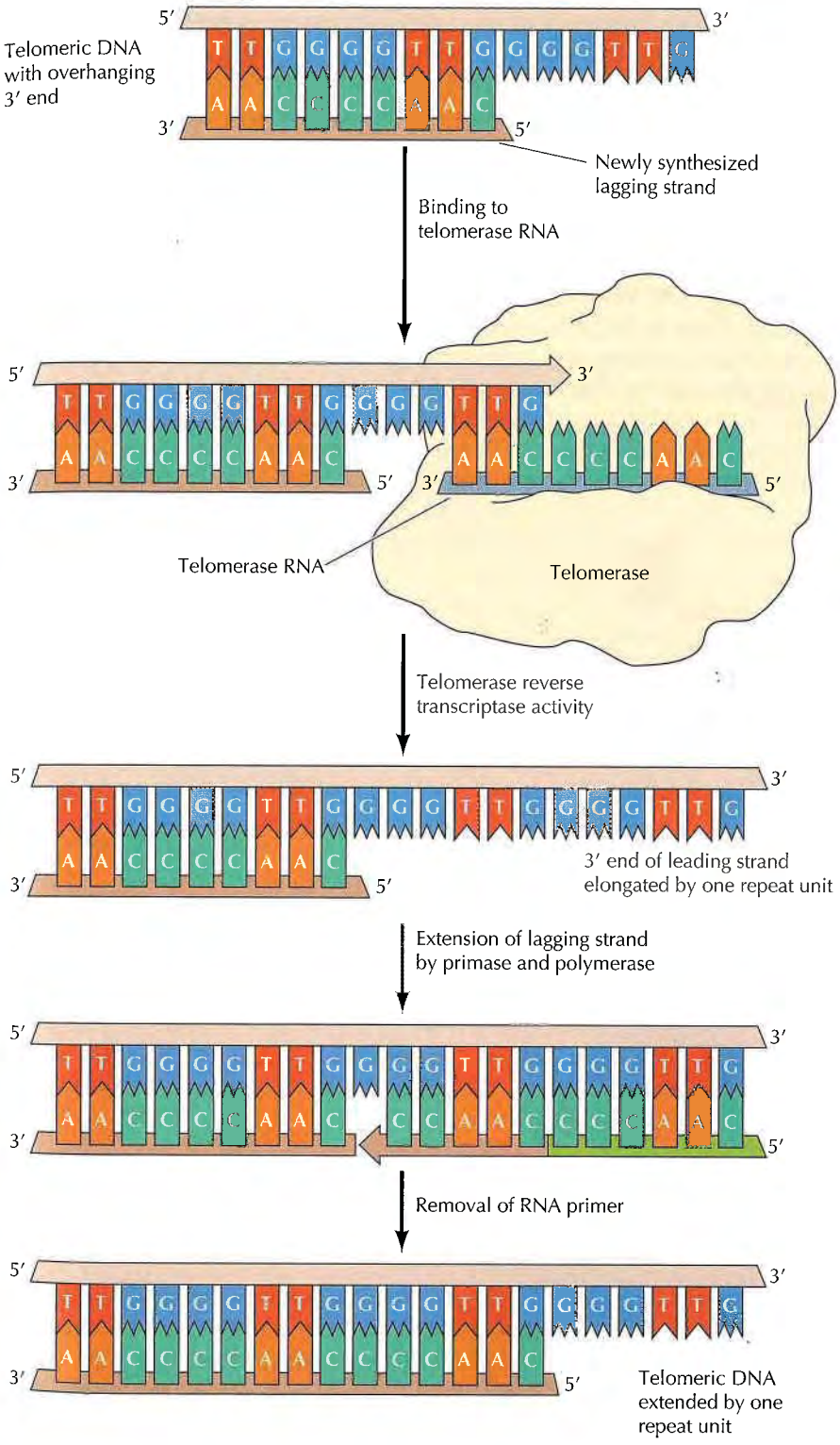


Figure 5.17 Action of telomerase
Telomeric DNA is a simple repeat sequence with an overhanging 3′ end on the newly synthesized leading strand. Telomerase carries its own RNA molecule, which is complementary to telomeric DNA, as part of the enzyme complex. The overhanging end of telomeric DNA binds to the telomerase RNA, which then serves as a template for extension of the leading strand by one repeat unit. The lagging strand of telomeric DNA can then be elongated by conventional RNA priming and DNA polymerase activity.

can result from the incorporation of incorrect bases during DNA replication. In addition, various chemical changes occur in DNA either spontaneously (Figure 5.18) or as a result of exposure to chemicals or radiation (Figure 5.19). Such damage to DNA can block replication or transcription, and can

result in a high frequency of mutations—consequences that are unacceptable from the standpoint of cell reproduction. To maintain the integrity of their genomes, cells have therefore had to evolve mechanisms to repair damaged DNA. These mechanisms of DNA repair can be divided into two general classes: (1) direct reversal of the chemical reaction responsible for DNA damage, and (2) removal of the damaged bases followed by their replacement with newly synthesized DNA. Where DNA repair fails, additional mechanisms have evolved to enable cells to cope with the damage.

Direct Reversal of DNA Damage

Most damage to DNA is repaired by removal of the damaged bases followed by resynthesis of the excised region. Some lesions in DNA, however, can be repaired by direct reversal of the damage, which may be a more efficient way of dealing with specific types of DNA damage that occur frequently. Only a few types of DNA damage are repaired in this way, particularly pyrimidine dimers resulting from exposure to ultraviolet (UV) light

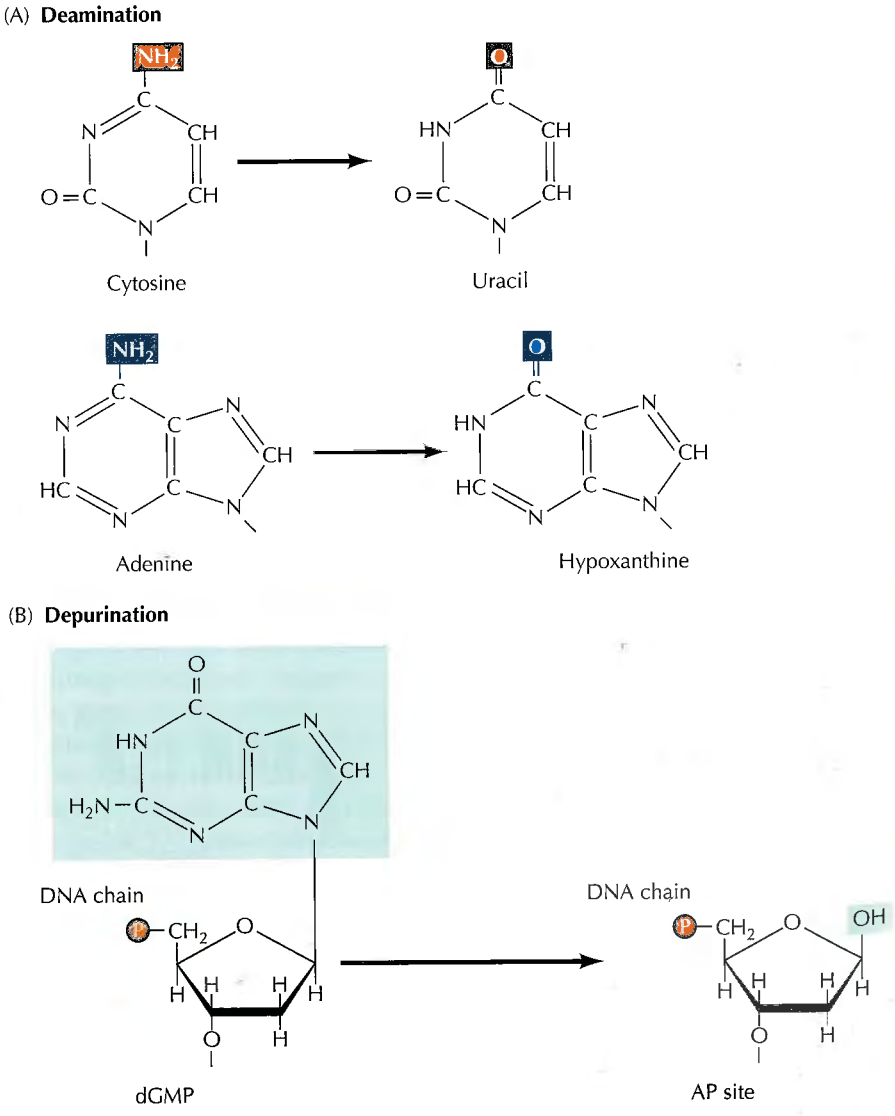
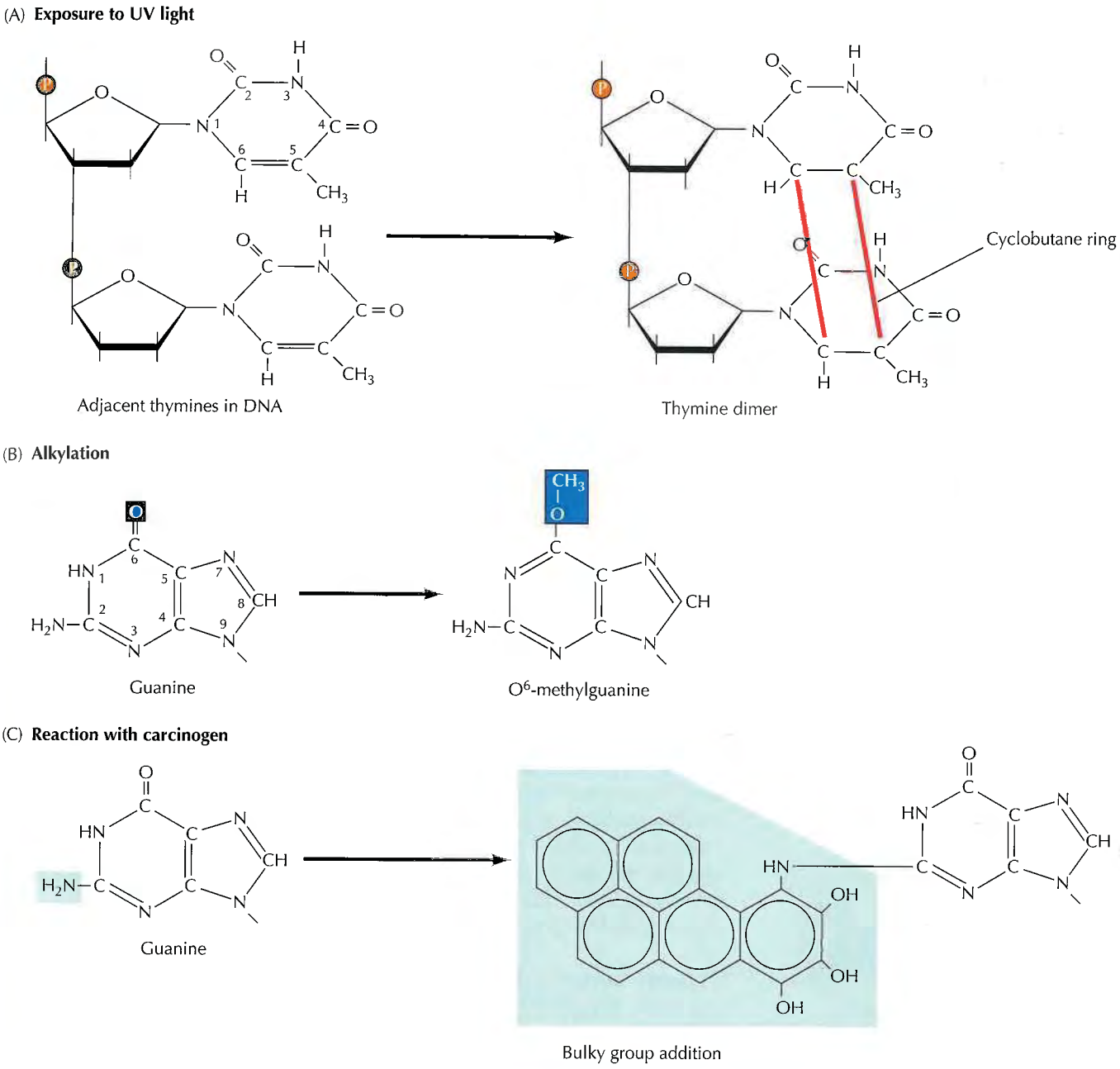


Figure 5.18 Spontaneous damage to DNA
There are two major forms of spontaneous DNA damage: (A) deamination of adenine, cytosine, and guanine, and (B) depurination (loss of purine bases) resulting from cleavage of the bond between the purine bases and deoxyribose, leaving an apurinic (AP) site in DNA. dGMP = deoxyguanosine monophosphate.



and alkylated guanine residues that have been modified by the addition of methyl or ethyl groups at the O^6 position of the purine ring.

UV light is one of the major sources of damage to DNA and is also the most thoroughly studied form of DNA damage in terms of repair mechanisms. Its importance is illustrated by the fact that exposure to solar UV irradiation is the cause of almost all skin cancer in humans. The major type of damage induced by UV light is the formation of **pyrimidine dimers,** in which adjacent pyrimidines on the same strand of DNA are joined by the formation of a cyclobutane ring resulting from saturation of the double bonds between carbons 5 and 6 (see Figure 5.19A). The formation of such dimers distorts the structure of the DNA chain and blocks transcription or replication past the site of damage, so their repair is

Figure 5.19 Examples of DNA damage induced by radiation and chemicals
(A) UV light induces the formation of pyrimidine dimers, in which two adjacent pyrimidines (e.g., thymines) are joined by a cyclobutane ring structure. (B) Alkylation is the addition of methyl or ethyl groups to various positions on the DNA bases. In this example, alkylation of the O^6 position of guanine results in formation of O^6-methylguanine. (C) Many carcinogens (e.g., benzo-(*a*)pyrene) react with DNA bases, resulting in the addition of large bulky chemical groups to the DNA molecule.

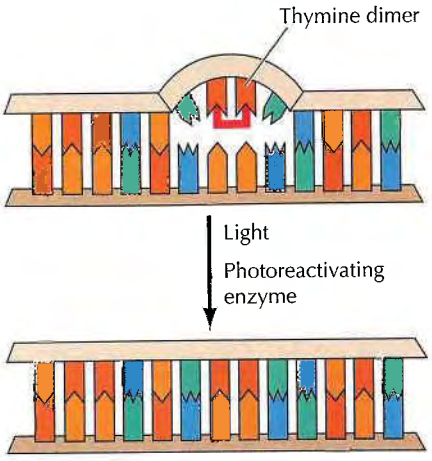


Figure 5.20 Direct repair of thymine dimers
UV-induced thymine dimers can be repaired by photoreactivation, in which energy from visible light is used to split the bonds forming the cyclobutane ring.

closely correlated with the ability of cells to survive UV irradiation. One mechanism of repairing UV-induced pyrimidine dimers is direct reversal of the dimerization reaction. The process is called **photoreactivation** because energy derived from visible light is utilized to break the cyclobutane ring structure (Figure 5.20). The original pyrimidine bases remain in DNA, now restored to their normal state. As might be expected from the fact that solar UV irradiation is a major source of DNA damage for diverse cell types, the repair of pyrimidine dimers by photoreactivation is common to a variety of prokaryotic and eukaryotic cells, including *E. coli*, yeasts, and some species of plants and animals. Curiously, however, photoreactivation is not universal; many species (including humans) lack this mechanism of DNA repair.

Another form of direct repair deals with damage resulting from the reaction between alkylating agents and DNA. Alkylating agents are reactive compounds that can transfer methyl or ethyl groups to a DNA base, thereby chemically modifying the base (see Figure 5.19B). A particularly important type of damage is methylation of the O^6 position of guanine, because the product, O^6-methylguanine, forms complementary base pairs with thymine instead of cytosine. This lesion can be repaired by an enzyme (called O^6-methylguanine methyltransferase) that transfers the methyl group from O^6-methylguanine to a cysteine residue in its active site (Figure 5.21). The potentially mutagenic chemical modification is thus removed, and the original guanine is restored. Enzymes that catalyze this direct repair reaction are widespread in both prokaryotes and eukaryotes, including humans.

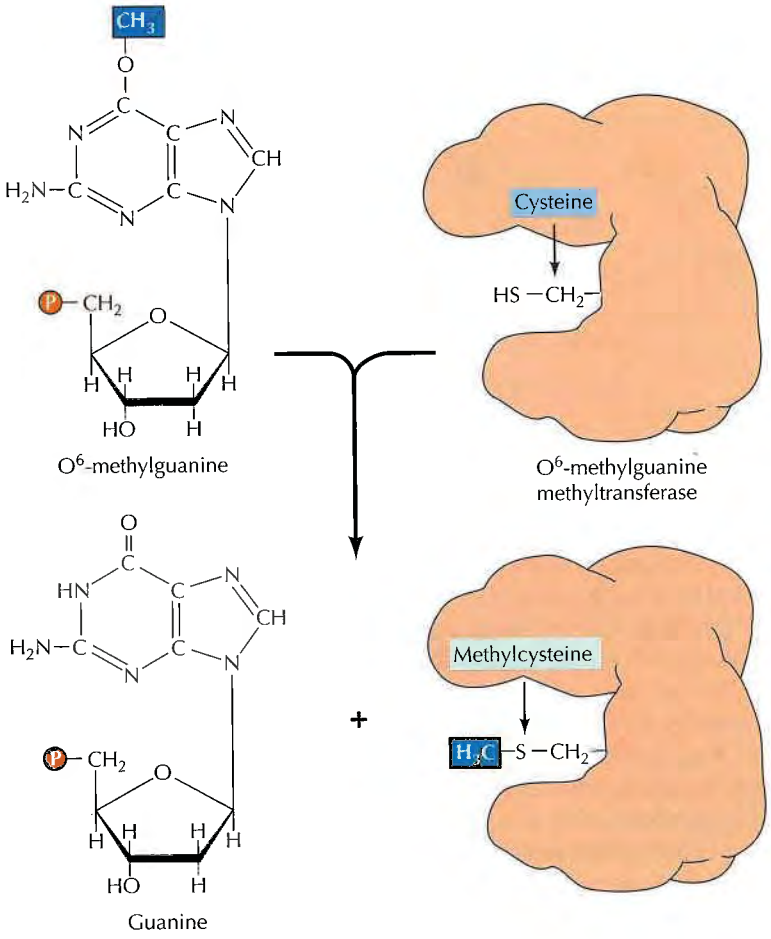


Figure 5.21 Repair of O^6-methylguanine
O^6-methylguanine methyltransferase transfers the methyl group from O^6-methylguanine to a cysteine residue in the enzyme's active site.

Excision Repair

Although direct repair is an efficient way of dealing with particular types of DNA damage, excision repair is a more general means of repairing a wide variety of chemical alterations to DNA. Consequently, the various types of excision repair are the most important DNA repair mechanisms in both prokaryotic and eukaryotic cells. In excision repair, the damaged DNA is recognized and removed, either as free bases or as nucleotides. The resulting gap is then filled in by synthesis of a new DNA strand, using the undamaged complementary strand as a template. Three types of excision repair—base-excision, nucleotide-excision, and mismatch repair—enable cells to cope with a variety of different kinds of DNA damage.

The repair of uracil-containing DNA is a good example of **base-excision repair**, in which single damaged bases are recognized and removed from the DNA molecule (Figure 5.22). Uracil can arise in DNA by two mechanisms: (1) Uracil (as dUTP [deoxyuridine triphosphate]) is occasionally incorporated in place of thymine during DNA synthesis, and (2) uracil can be formed in DNA by the deamination of cytosine (see Figure 5.18A). The second mechanism is of much greater biological significance because it alters the normal pattern of complementary base pairing and thus represents a mutagenic event. The excision of uracil in DNA is catalyzed by **DNA glycosylase**, an enzyme that cleaves the bond linking the base (uracil) to the deoxyribose of the DNA backbone. This reaction yields free uracil and an apyrimidinic site—a sugar with no base attached. DNA glycosylases also recognize and remove other abnormal bases, including hypoxanthine formed by the deamination of adenine, pyrimidine dimers, alkylated purines other than O^6-alkylguanine, and bases damaged by oxidation or ionizing radiation.

The result of DNA glycosylase action is the formation of an apyridiminic or apurinic site (generally called an AP site) in DNA. Similar AP sites are formed as the result of the spontaneous loss of purine bases (see Figure 5.18B), which occurs at a significant rate under normal cellular conditions. For example, each cell in the human body is estimated to lose several thousand purine bases daily. These sites are repaired by **AP endonuclease**, which cleaves adjacent to the AP site (see Figure 5.22). The remaining deoxyribose moiety is then removed, and the resulting single-base gap is filled by DNA polymerase and ligase.

Whereas DNA glycosylases recognize only specific forms of damaged bases, other excision repair systems recognize a wide variety of damaged bases that distort the DNA molecule, including UV-induced pyrimidine dimers and bulky groups added to DNA bases as a result of the reaction of many carcinogens with DNA (see Figure 5.19C). This widespread form of DNA repair is known as **nucleotide-excision repair**, because the damaged bases (e.g., a thymine dimer) are removed as part of an oligonucleotide containing the lesion (Figure 5.23).

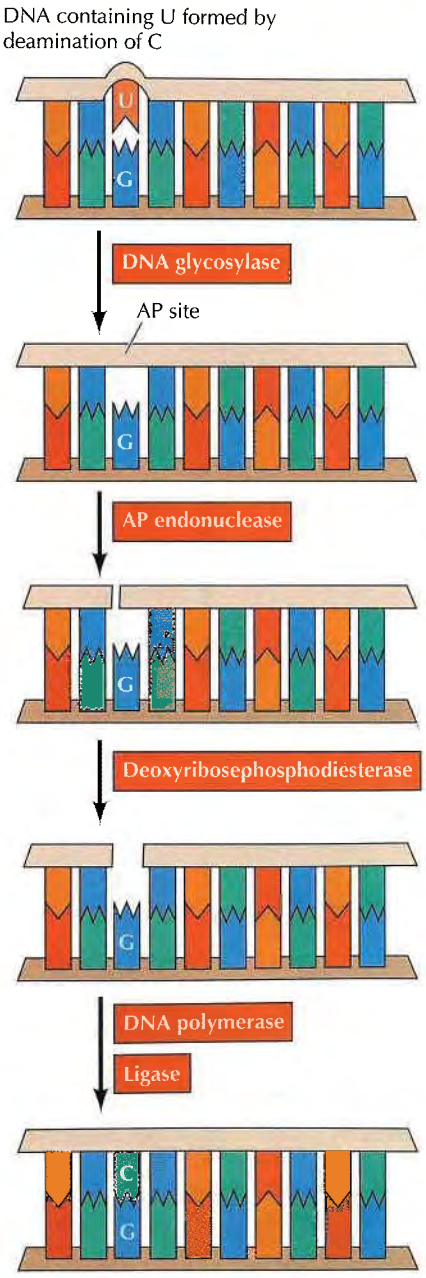


Figure 5.22 Base-excision repair
In this example, uracil (U) has been formed by deamination of cytosine (C) and is therefore opposite a guanine (G) in the complementary strand of DNA. The bond between uracil and the deoxyribose is cleaved by a DNA glycosylase, leaving a sugar with no base attached in the DNA (an AP site). This site is recognized by AP endonuclease, which cleaves the DNA chain. The remaining deoxyribose is removed by deoxyribosephosphodiesterase. The resulting gap is then filled by DNA polymerase and sealed by ligase, leading to incorporation of the correct base (C) opposite the G.

In *E. coli*, nucleotide-excision repair is catalyzed by the products of three genes (*uvrA*, *B*, and *C*) that were identified because mutations at these loci result in extreme sensitivity to UV light. The protein UvrA recognizes damaged DNA and recruits UvrB and UvrC to the site of the lesion. UvrB and UvrC then cleave on the 3′ and 5′ sides of the damaged site, respectively, thus excising an oligonucleotide consisting of 12 or 13 bases. The UvrABC complex is frequently called an **excinuclease**, a name that reflects its ability to directly *excise* an oligonucleotide. The action of a helicase is then required to remove the damage-containing oligonucleotide from the double-stranded DNA molecule, and the resulting gap is filled by DNA polymerase I and sealed by ligase.

Nucleotide-excision repair systems have also been studied extensively in eukaryotes, including yeasts, rodents, and humans. In yeasts, as in *E. coli*, several genes involved in DNA repair (called *RAD* genes for *rad*iation sensitivity) have been identified by the isolation of mutants with increased sensitivity to UV light. In humans, DNA repair genes have been identified

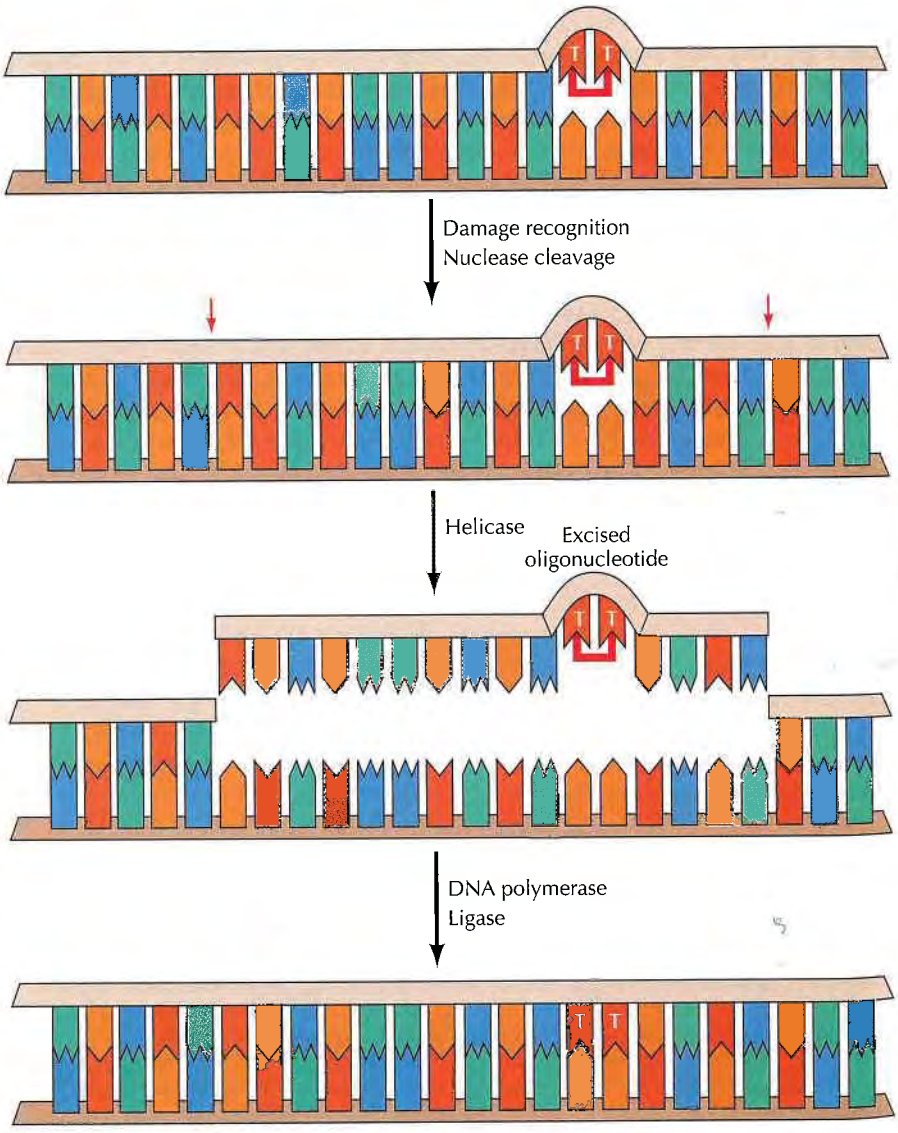


Figure 5.23 Nucleotide-excision repair of thymine dimers
Damaged DNA is recognized and then cleaved on both sides of a thymine dimer by 3′ and 5′ nucleases. Unwinding by a helicase results in excision of an oligonucleotide containing the damaged bases. The resulting gap is then filled by DNA polymerase and sealed by ligase.

largely by studies of individuals suffering from inherited diseases resulting from deficiencies in the ability to repair DNA damage. The most extensively studied of these diseases is xeroderma pigmentosum (XP), a rare genetic disorder that affects approximately one in 250,000 people. Individuals with this disease are extremely sensitive to UV light and develop multiple skin cancers on the regions of their bodies that are exposed to sunlight. In 1968 James Cleaver made the key discovery that cultured cells from XP patients were deficient in the ability to carry out nucleotide-excision repair. This observation not only provided the first link between DNA repair and cancer, but also suggested the use of XP cells as an experimental system to identify human DNA repair genes. The identification of human DNA repair genes has been accomplished by studies not only of XP cells, but also of two other human diseases resulting from DNA repair defects (Cockayne's syndrome and trichothiodystrophy) and of UV-sensitive mutants of rodent cell lines. The availability of mammalian cells with defects in DNA repair has allowed the cloning of repair genes based on the ability of wild-type alleles to restore normal UV sensitivity to mutant cells in gene transfer assays, thereby opening the door to experimental analysis of nucleotide-excision repair in mammalian systems.

Molecular cloning has identified seven different repair genes (designated *XPA* through *XPG*) that are mutated in cases of xeroderma pigmentosum, as well as genes that are mutated in Cockayne's syndrome, trichothiodystrophy, and UV-sensitive mutants of rodent cells. The proteins encoded by these mammalian DNA repair genes are closely related to proteins encoded by yeast *RAD* genes, indicating that nucleotide-excision repair is highly conserved throughout eukaryotes. With cloned yeast and mammalian repair genes available, it has been possible to purify their encoded proteins and develop *in vitro* systems to study their roles in the repair process (Figure 5.24). The initial step in excision repair in mammalian cells involves recognition of disrupted base pairing by a complex consisting of XPC and a protein called hHR23B, which is a homolog of the yeast Rad23 protein. This is followed by recruitment of the XPB, XPD, and XPG proteins to the damaged DNA. The XPB and XPD proteins are components of a multisubunit transcription factor (called TFIIH) required to initiate the transcription of eukaryotic genes (see Chapter 6); they act as helicases to unwind approximately 30 base pairs of DNA around the site of damage. The XPA protein then acts to confirm the damage, and recruits XPF as a heterodimer with ERCC1 (a repair protein identified in UV-sensitive rodent cells) to the repair complex. XPF/ERCC1 and XPG are endonucleases, which cleave DNA on the 5′ and 3′ sides of the damaged site, respectively. This cleavage excises an oligonucleotide consisting of approximately 30 bases. The resulting gap is then filled by DNA polymerase δ or ε (in association with RFC and PCNA) and sealed by ligase.

Whereas the XPC/hHR23B complex can recognize damaged DNA throughout the genome, an alternative form of nucleotide-excision repair,

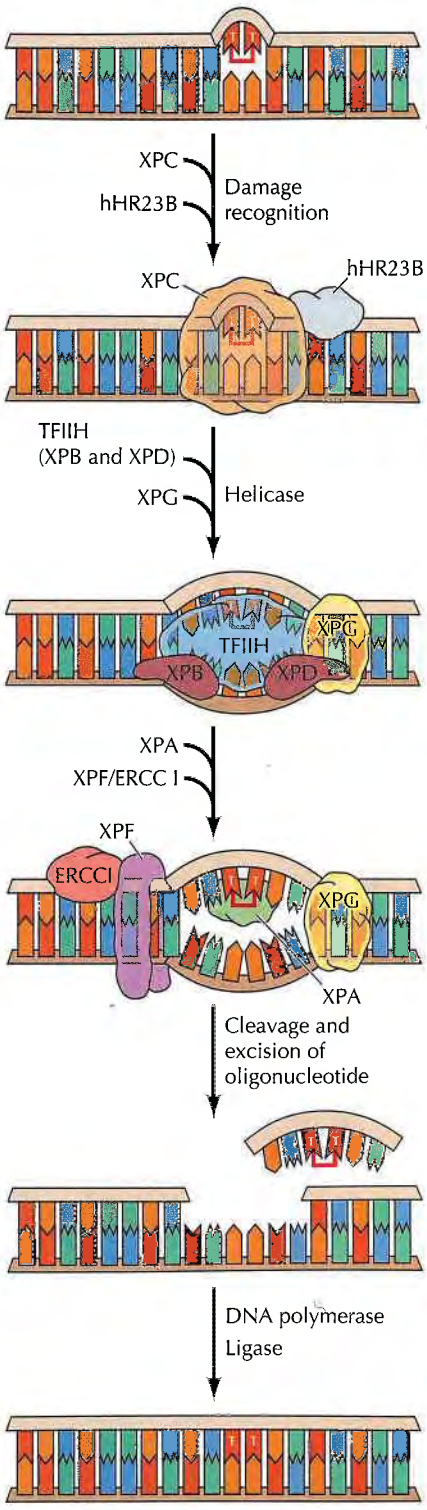


Figure 5.24 Nucleotide-excision repair in mammalian cells
DNA damage (e.g., a thymine dimer) is recognized by the XPC/hHR23B complex. The transcription factor TFIIH, which contains the XPB and XPD helicases, and XPG are then recruited to the damaged DNA. Following unwinding of the DNA by XPB and XPD, the damage is confirmed by XPA and the XPF/ERCC1 complex is recruited. The DNA is then cleaved by the XPF/ERCC1 and XPG endonucleases, excising the damaged oligonucleotide. The resulting gap is filled by DNA polymerase and sealed by ligase.

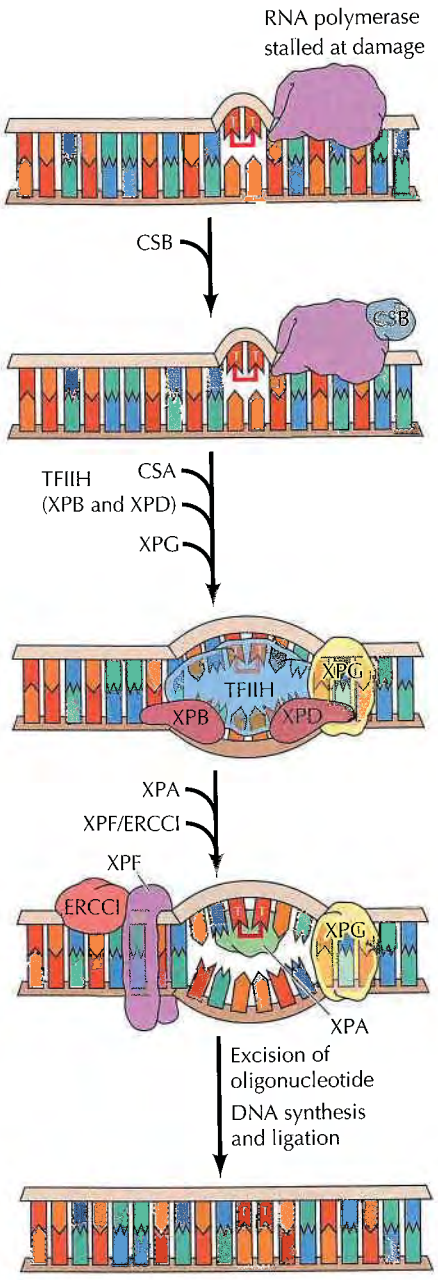


Figure 5.25 Transcription-coupled repair in mammalian cells
RNA polymerase stalls at a lesion in the DNA strand being transcribed. The stalled RNA polymerase is recognized by the transcription-repair coupling proteins CSA and CSB, which recruit TFIIH and XPG to the damaged DNA. Repair then proceeds by the general nucleotide-excision repair pathway (see Figure 5.24).

called **transcription-coupled repair**, is specifically dedicated to repairing damage within actively transcribed genes. A connection between transcription and repair was first suggested by experiments showing that transcribed strands of DNA are repaired more rapidly than nontranscribed strands in both *E. coli* and mammalian cells. Since DNA damage blocks transcription, this transcription-repair coupling is thought to be advantageous by allowing the cell to preferentially repair damage to genes that are actively expressed. In *E. coli*, the mechanism of transcription-repair coupling involves recognition of RNA polymerase stalled at a lesion in the DNA strand being transcribed. The stalled RNA polymerase is recognized by a protein called transcription-repair coupling factor, which displaces RNA polymerase and recruits the UvrABC excinuclease to the site of damage.

In mammalian cells, transcription-coupled repair involves recognition of stalled RNA polymerase by the CSA and CSB proteins, which are encoded by genes responsible for Cockayne's syndrome (Figure 5.25). In contrast to patients with xeroderma pigmentosum, patients with Cockayne's syndrome are specifically defective in transcription-coupled repair, consistent with the role of CSA and CSB as transcription-repair coupling factors. CSA and CSB act analogously to the XPC/hHR23B complex in recruiting XPB, XPD, and XPG to the damaged site. This is followed by recruitment of XPA and the XPF/ERCC1 complex, and excision of the damaged oligonucleotide. Transcription-coupled repair thus proceeds similarly to general nucleotide-excision repair, except for the initial recognition of stalled RNA polymerase by CSA and CSB rather than direct recognition of DNA damage by the XPC/hHR23B complex.

A third excision repair system recognizes mismatched bases that are incorporated during DNA replication. Many such mismatched bases are removed by the proofreading activity of DNA polymerase. The ones that are missed are subject to later correction by the **mismatch repair** system, which scans newly replicated DNA. If a mismatch is found, the enzymes of this repair system are able to identify and excise the mismatched base specifically from the newly replicated DNA strand, allowing the error to be corrected and the original sequence restored.

In *E. coli*, the ability of the mismatch repair system to distinguish between parental DNA and newly synthesized DNA is based on the fact that DNA of this bacterium is modified by the methylation of adenine residues within the sequence GATC to form 6-methyladenine (Figure 5.26). Since methylation occurs after replication, newly synthesized DNA strands are not methylated and thus can be specifically recognized by the mismatch repair enzymes. Mismatch repair is initiated by the protein MutS, which recognizes the mismatch and forms a complex with two other proteins called MutL and MutH. The MutH endonuclease then cleaves the unmethylated DNA strand at a GATC sequence. MutL and MutS then act together with an exonuclease and a helicase to excise the DNA between the strand break and the mismatch, with the resulting gap being filled by DNA polymerase and ligase.

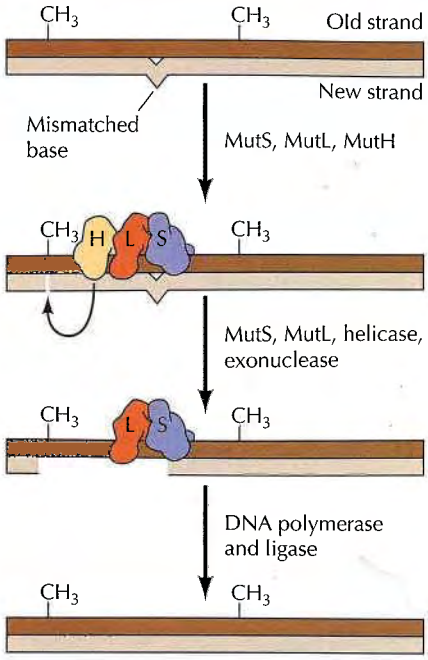


Figure 5.26 Mismatch repair in *E. coli*
The mismatch repair system detects and excises mismatched bases in newly replicated DNA, which is distinguished from the parental strand because it has not yet been methylated. MutS binds to the mismatched base, followed by MutL. The binding of MutL activates MutH, which cleaves the unmodified strand opposite a site of methylation. MutS and MutL, together with a helicase and an exonuclease, then excise the portion of the unmodified strand that contains the mismatch. The gap is then filled by DNA polymerase and sealed by ligase.

Eukaryotes have a similar mismatch repair system, although the mechanism by which eukaryotic cells identify newly replicated DNA differs from that used by *E. coli*. In mammalian cells, it appears that the strand-specificity of mismatch repair is not determined by DNA methylation. Instead, the presence of single-strand breaks (which would be present in newly replicated DNA) or associations of the eukaryotic homologs of MutS and MutL with the replication machinery may specify the strand to be repaired. MutS and MutL homologs then bind to the mismatched base and direct excision of the DNA between a strand break and the mismatch, as in *E. coli*. The importance of this repair system is dramatically illustrated by the fact that mutations in the human homologs of *MutS* and *MutL* are responsible for a common type of inherited colon cancer (hereditary nonpolyposis colorectal cancer, or HNPCC). HNPCC is one of the most common inherited diseases; it affects as many as one in 200 people and is responsible for about 15% of all colorectal cancers in this country. The relationship between HNPCC and defects in mismatch repair was discovered in 1993, when two groups of researchers cloned the human homolog of *MutS* and found that mutations in this gene were responsible for about half of all HNPCC cases. Subsequent studies have shown that most of the remaining cases of HNPCC are caused by mutations in one of three human genes that are homologs of *MutL*. Defects in these genes appear to result in a high frequency of mutations in other cell genes, with a correspondingly high likelihood that some of these mutations will eventually lead to the development of cancer.

Error-Prone Repair

The direct reversal and excision repair systems act to correct DNA damage before replication, so that replicative DNA synthesis can proceed using an undamaged DNA strand as a template. Should these systems fail, however, the cell has alternative mechanisms for dealing with damaged DNA at the replication fork. Pyrimidine dimers and many other types of lesions cannot be copied by the normal action of DNA polymerases, so replication is blocked at the sites of such damage. However, cells also possess several specialized DNA polymerases that are capable of replicating across a site of DNA damage. The replication of damaged DNA by these specialized polymerases can lead to the frequent incorporation of incorrect bases, so this form of dealing with DNA damage is called **error-prone repair.**

The first error-prone DNA polymerase was discovered in *E. coli* in 1999. This enzyme, called polymerase V, is induced in response to extensive UV irradiation and can synthesize a new DNA strand across from a thymine dimer (Figure 5.27). Two other *E. coli* DNA polymerases, polymerases II and IV, are similarly induced by DNA damage and function in error-prone repair. Eukaryotic cells also contain multiple error-prone DNA polymerases, with nine such enzymes having been identified to date in humans.

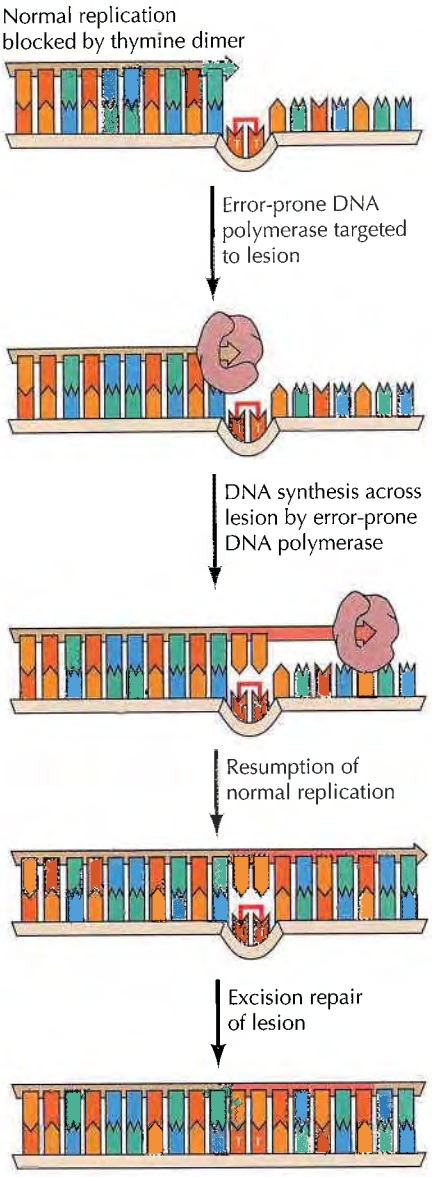


Figure 5.27 Error-prone repair
Normal replication is blocked by a thymine dimer, but error-prone DNA polymerases such as polymerase V (pol V) recognize and continue DNA synthesis across the lesion. Replication can then be resumed by the normal replicative DNA polymerase, and the thymine dimer subsequently removed by nucleotide-excision repair. DNA synthesized by the error-prone polymerase contains a high frequency of incorrect bases.

All of these error-prone DNA polymerases exhibit low fidelity when copying undamaged DNA, with error rates ranging from 100 to 10,000 times higher than the error rates of the normal replicative DNA polymerases (e.g., polymerase III in *E. coli* or polymerases δ and ε in eukaryotes). In addition, the error-prone polymerases lack the 3′→ 5′ proofreading activity that is characteristic of normal replicative DNA polymerases (see Figure 5.12).

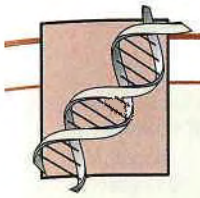
Importantly, however, the error-prone polymerases are specialized for inserting the correct base opposite specific lesions in damaged DNA, and are therefore able to accurately synthesize a new strand using some forms of damaged DNA as template. For example, *E. coli* polymerase V specifically recognizes thymine dimers and correctly inserts AA on the opposite strand. On the other hand, polymerase V makes a high frequency of errors when it synthesizes a new DNA strand opposite other forms of DNA damage. Thus, these enzymes are able to specifically insert correct bases opposite some forms of DNA damage, although they are "error-prone" in inserting bases opposite other forms of damaged DNA or in the synthesis of DNA from a normal undamaged template.

Recombinational Repair

Another means of DNA repair, **recombinational repair**, relies on replacement of the damaged DNA by recombination with an undamaged molecule. This mechanism is frequently used to repair damage encountered during DNA replication, where the presence of thymine dimers or other lesions that cannot be copied by the normal replicative DNA polymerases block the progress of a replication fork. Recombinational repair depends on the fact that one strand of the parental DNA was undamaged and therefore was copied during replication to yield a normal daughter molecule, which can then be used to repair the damaged strand.

The molecular mechanisms of recombinational repair are not entirely understood and may vary between different types of cells, but an illustrative model is presented in Figure 5.28. In this example, normal replication is blocked by the presence of a thymine dimer in one strand of DNA. Downstream of the damaged site, however, replication can be initiated again by the synthesis of an Okazaki fragment and can proceed along the damaged template strand. The result is a daughter strand that has a gap opposite the site of damage to the parental strand. The undamaged parental strand, which has been replicated to yield a normal daughter molecule, can then be used to fill the gap opposite the site of damage by recombination between homologous DNA sequences (see the next section). Because the resulting gap in the previously intact parental strand is opposite an undamaged strand, it can be filled in by DNA polymerase. Although the other parent molecule still retains the original damage (e.g., a thymine dimer), the damage now lies opposite a normal strand and can be dealt with later by excision repair.

Recombinational repair also provides a major mechanism for repair of double strand breaks, which can be introduced into DNA by ionizing radiation (such as X-rays) and some chemicals (Figure 5.29). Since this type of damage affects both strands of DNA, it is particularly difficult to repair. Recombination with homologous DNA sequences on an undamaged chromosome provides a mechanism for repairing such damage and restoring the normal DNA sequence. Alternatively, double strand breaks can be repaired simply by rejoining the broken ends of a single DNA molecule, but this leads to a high frequency of errors resulting from deletion of bases



MOLECULAR MEDICINE

Colon Cancer and DNA Repair

The Disease

Cancers of the colon and rectum (colorectal cancers) are some of the most common types of cancer in Western countries, accounting for about 140,000 cancer cases per year in the United States (approximately 10% of the total cancer incidence). Most colon cancers (like other types of cancer) are not inherited diseases; that is, they are not transmitted directly from parent to offspring. However, two inherited forms of colon cancer have been described. In both of these syndromes, the inheritance of a cancer susceptibility gene results in a very high likelihood of cancer development. One inherited form of colon cancer (familial adenomatous polyposis) is extremely rare, accounting for less than 1% of total colon cancer incidence. The second inherited form of colon cancer (hereditary nonpolyposis colorectal cancer, or HNPCC) is much more common and accounts for up to 15% of all colon cancer cases. Indeed, HNPCC is one of the most common inherited diseases, affecting as many as one in 200 people. Although colon cancers are the most common manifestation of this disease, affected individuals also suffer an increased incidence of other types of cancer, including cancers of the ovary and endometrium.

Molecular and Cellular Basis

Like other cancers, colorectal cancer results from mutations in genes that regulate cell proliferation, leading to the uncontrolled growth of cancer cells. In most cases these mutations occur sporadically in somatic cells. In hereditary cancers, however, inherited germ-line mutations predispose the individual to cancer development.

A striking advance was made in 1993 with the discovery that a gene responsible for approximately 50% of HNPCC cases encodes an enzyme involved in mismatch repair of DNA; this gene is a human homolog of the *E. coli MutS* gene. Subsequent studies have shown that three other genes, responsible for most remaining cases of HNPCC, are homologs of *MutL* and thus are also involved in the mismatch repair pathway. Defects in these genes appear to result in a high frequency of mutations in other cell genes, with a correspondingly high likelihood that some of these mutations will eventually lead to the development of cancer by affecting genes that regulate cell proliferation.

Prevention and Treatment

As with other inherited diseases, identification of the genes responsible for HNPCC allows individuals at risk for this inherited cancer to be identified by genetic testing. Moreover, prenatal genetic diagnosis may be of great importance to carriers of HNPCC mutations who are planning a family. However, the potential benefits of detecting these mutations are not limited to preventing the transmission of mutant genes to the next generation; their detection may also help prevent the development of cancer in affected individuals.

In terms of disease prevention, a key characteristic of colon cancer is that it develops gradually over several years. Early diagnosis of the disease substantially improves the chances for patient survival. The initial stage of colon cancer development is the outgrowth of small benign polyps, which eventually become malignant and invade the surrounding connective tissue. Prior to the development of malignancy, however, polyps can be easily removed surgically, effectively preventing the outgrowth of a malignant tumor. Polyps and early stages of colon cancer can be detected by examination of the colon with a thin lighted tube (colonoscopy), so frequent colonoscopy of HNPCC patients may allow polyps to be removed before cancer develops. In addition, several drugs are being tested as potential inhibitors of colon cancer development, and these drugs may be of significant benefit to HNPCC patients. By allowing the timely application of such preventive measures, the identification of mutations responsible for HNPCC may make a significant contribution to disease prevention.

A colon polyp visualized by colonoscopy. (David M. Martin, M.D./SPL/Photo Researches, Inc..)

around the site of damage. It is noteworthy that the genes responsible for inherited breast cancer (*BRCA1* and *BRCA2*) encode proteins that are involved in the repair of double strand breaks by homologous recombination, suggesting that defects in this type of DNA repair can lead to the development of one of the most common cancers in women.

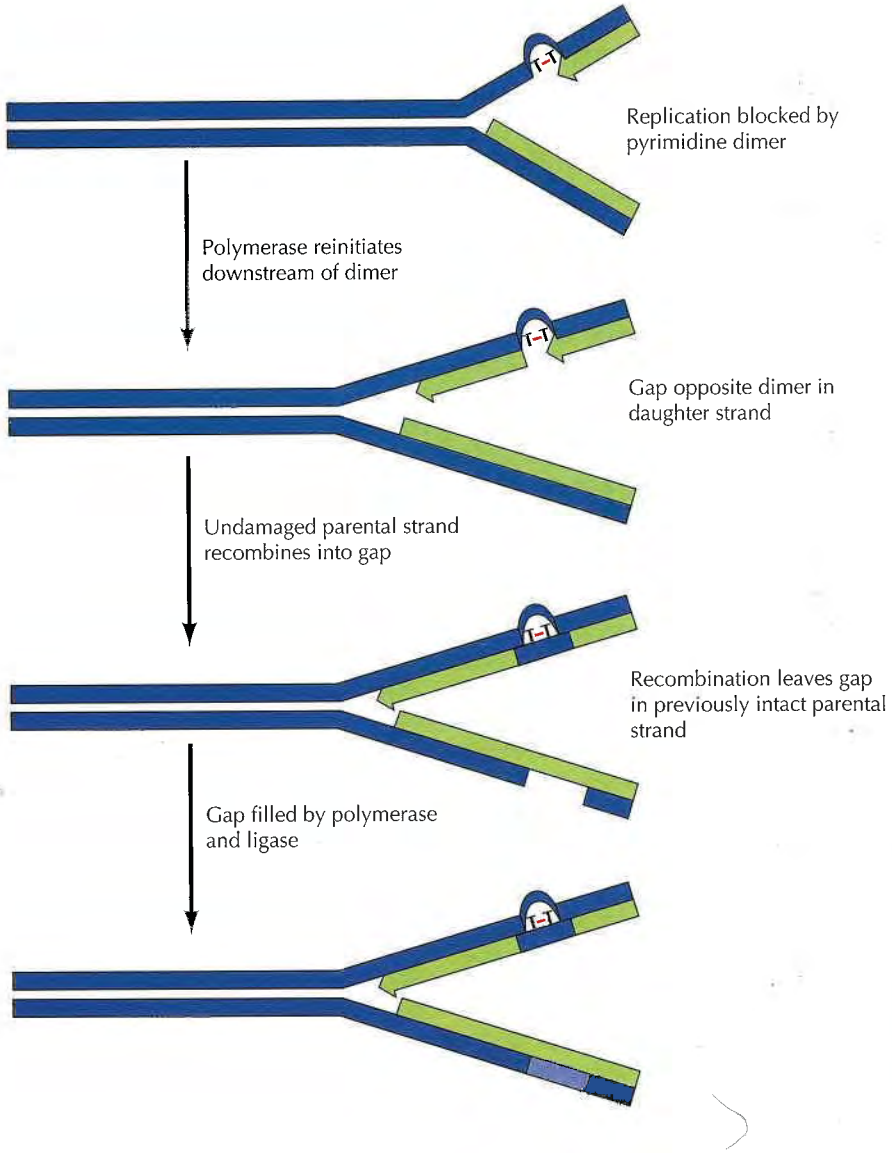


Figure 5.28 Recombinational repair The presence of a thymine dimer blocks replication, but DNA polymerase can bypass the lesion and reinitiate replication at a new site downstream of the dimer. The result is a gap opposite the dimer in the newly synthesized DNA strand. This gap is filled by recombination with the undamaged parental strand. Although this leaves a gap in the previously intact parental strand, the gap can be filled by the actions of polymerase and ligase, using the intact daughter strand as a template. Two intact DNA molecules are thus formed, and the remaining thymine dimer eventually can be removed by excision repair.

Recombination between Homologous DNA Sequences

Accurate DNA replication and repair of DNA damage are essential to maintaining genetic information and ensuring its accurate transmission from parent to offspring. As discussed in the previous section, recombination is an important mechanism for repairing damaged DNA. In addition, recombination is key to the generation of genetic diversity, which is critical from the standpoint of evolution. Genetic differences between individuals provide the essential starting material of natural selection, which allows species to evolve and adapt to changing environmental conditions. Recombination plays an central role in this process by allowing genes to be reassorted into different combinations. For example, genetic recombination results in the exchange of genes between paired homologous chromosomes during meiosis. In addition, recombination is involved in rearrangements of specific DNA sequences that alter the expression and function of some genes during development and differentiation. Thus, recombination plays important roles in the lives of individual cells and organisms, as well as contributing to the genetic diversity of the species.

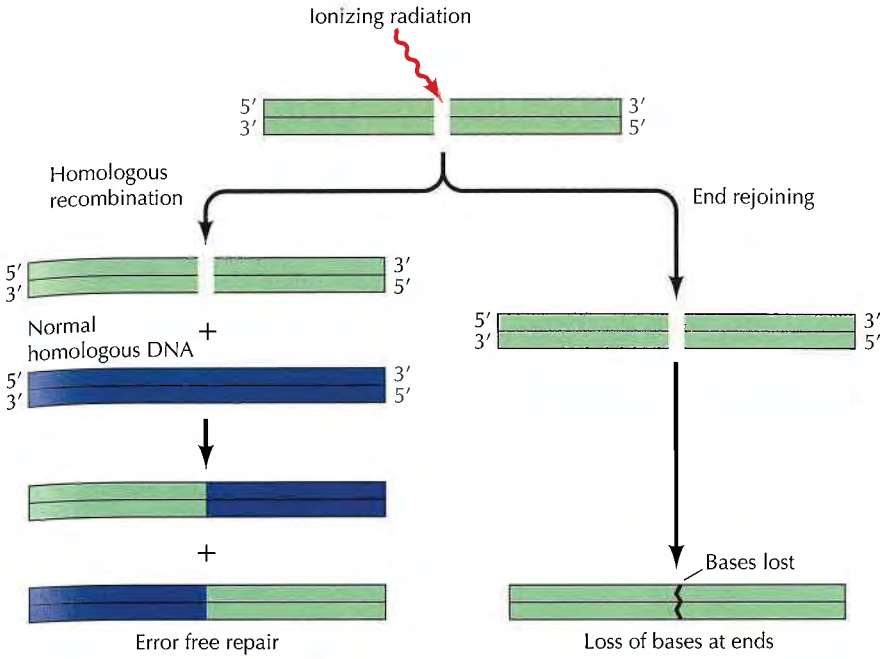


Figure 5.29 Repair of double strand breaks
Ionizing radiation and some chemicals induce double strand breaks in DNA. These breaks can be repaired by homologous recombination with a normal chromosome, leading to restoration of the original DNA sequence. Alternatively, the ends of the broken molecule can be rejoined, with the frequent loss of bases around the site of damage.

This section discusses the molecular mechanisms of recombination between DNA molecules that share extensive sequence homology. Examples include homologous recombination during DNA repair, as well as recombination between paired eukaryotic chromosomes during meiosis and recombination between bacterial chromosomes during mating. Since this type of recombination involves the exchange of genetic information between two homologous DNA molecules, it does not alter the overall arrangement of the genes on a chromosome. Other types of recombination, however, do not require extensive sequence homology and therefore can occur between unrelated DNAs. Recombination events of this type lead to gene rearrangements, which are discussed later in the chapter.

DNA Molecules Recombine by Breaking and Rejoining

Genetic recombination was first defined by studies of *Drosophila*, on the basis of the observation that genes on different copies of homologous chromosomes can reassort during meiosis. With the subsequent discovery that genes consist of DNA, two alternative models to explain recombination at the molecular level were considered (Figure 5.30). The "copy choice" model

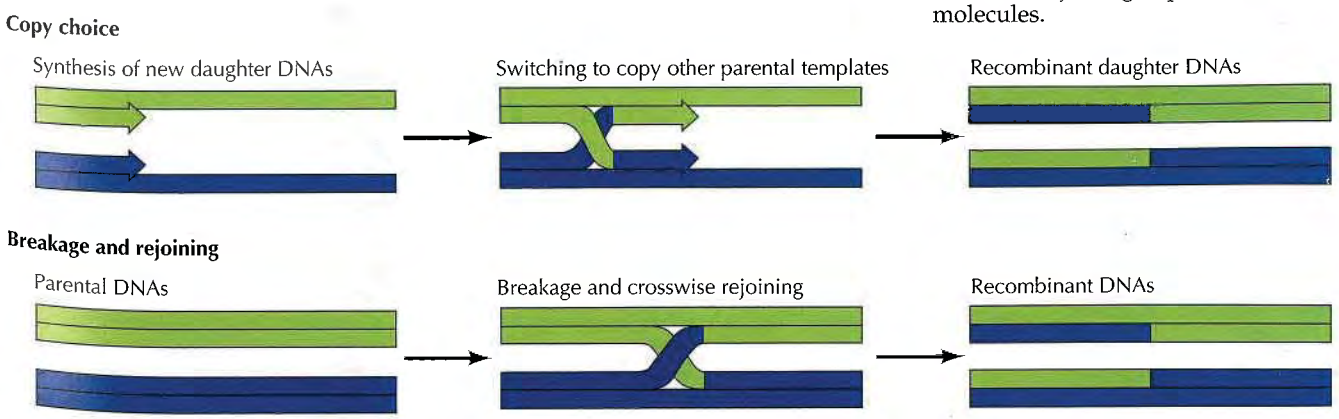


Figure 5.30 Models of recombination
In copy choice, recombination occurs during the synthesis of daughter DNA molecules. DNA replication starts with one parental DNA template and then switches to a second parental molecule, resulting in the synthesis of recombinant daughter DNAs containing sequences homologous to both parents. In breakage and rejoining, recombination occurs as a result of breakage and crosswise rejoining of parental DNA molecules.

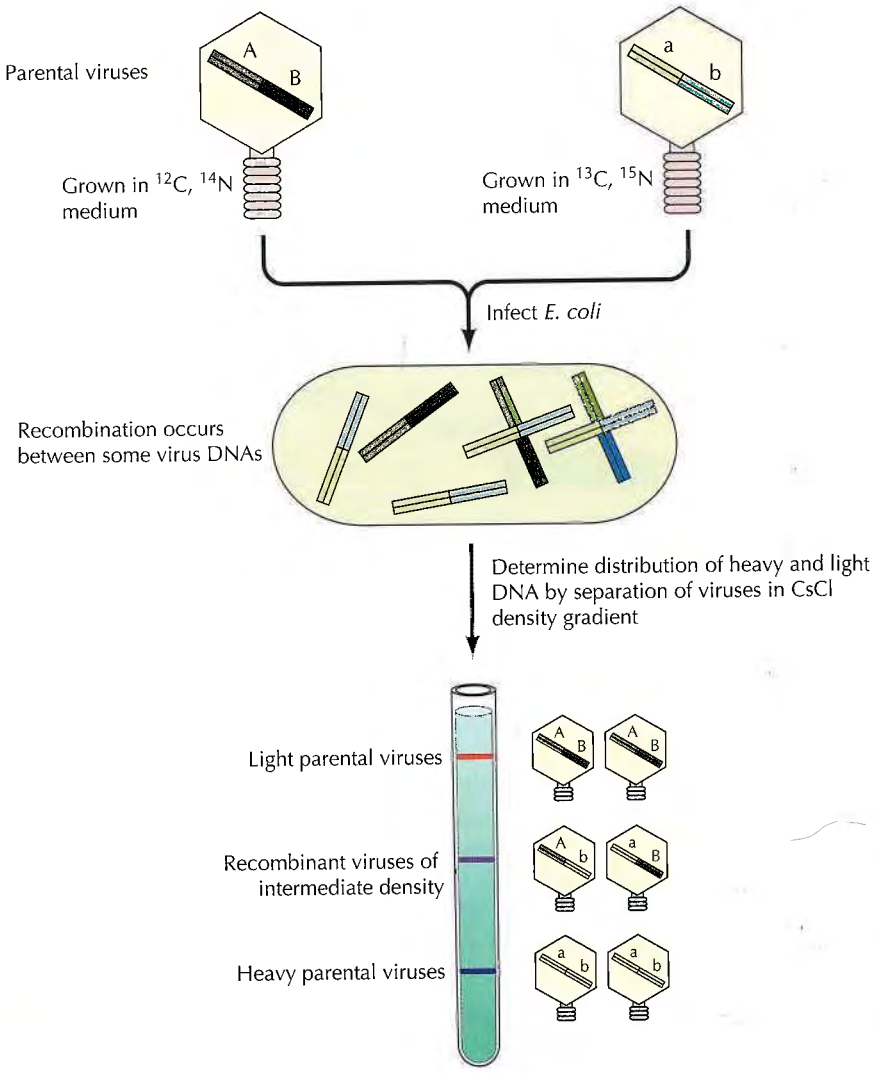
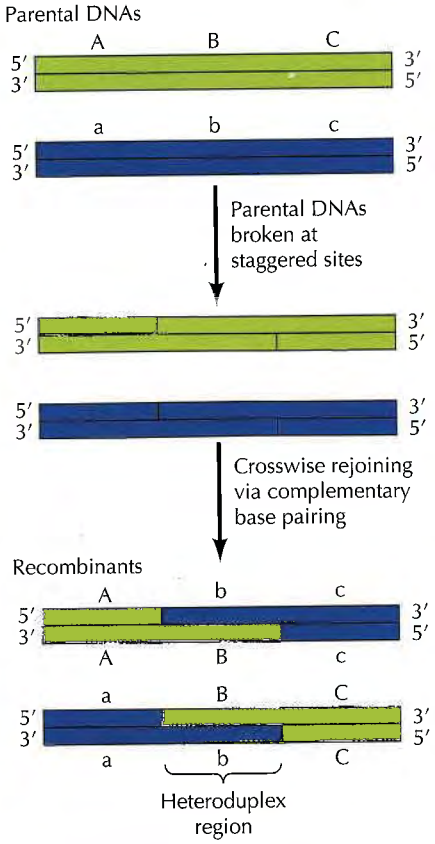


Figure 5.31 Experimental demonstration of recombination by breakage and rejoining
Genetically distinct parental viruses were grown in medium containing either light or heavy isotopes of carbon (^{12}C or ^{13}C) and nitrogen (^{14}N or ^{15}N) to density-label their DNAs. *E. coli* were infected under conditions in which replication was inhibited, and the progeny viruses were harvested and analyzed by equilibrium centrifugation in a CsCl gradient to determine the density of genetic recombinants. The recombinant viruses were found to have intermediate densities, indicating that they had acquired DNA from both parents by a breakage and rejoining mechanism.



proposed that the recombinant molecule is generated during DNA synthesis, as a result of copying first one parental DNA and then switching to copy a different template. The alternative proposal was that recombination results from the breakage and rejoining of two parental DNA molecules rather than by synthesis of new DNA.

These alternatives were first distinguished in 1961 by studies of recombination between the genomes of bacterial viruses (Figure 5.31). Infection of *E. coli* with viruses carrying different genetic markers was known to yield recombinant progeny. To determine if this recombination involved breakage and rejoining of the parental DNAs, one of the parental viruses was grown in medium containing the heavy isotopes of carbon (^{13}C) and nitrogen (^{15}N), while the other was grown in medium containing the normal

Figure 5.32 Homologous recombination by complementary base pairing
Parental DNAs are broken at staggered sites, and overlapping single-stranded regions are exchanged via base pairing with homologous sequences. The result is a heteroduplex region, in which the two DNA strands are derived from different parental molecules.

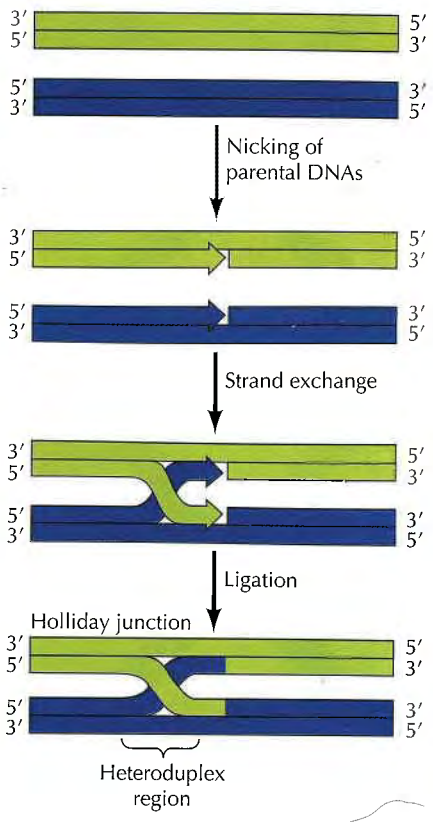


Figure 5.33 The Holliday model for homologous recombination
Single-strand nicks are introduced at the same position on both parental molecules. The nicked strands then exchange by complementary base pairing, and ligation produces a crossed-strand intermediate called a Holliday junction.

light isotopes (^{12}C and ^{14}N). The result was parental viruses having different densities, so they could be separated by equilibrium density centrifugation in a CsCl gradient. *E. coli* were then infected with these differentially labeled parental viruses under conditions in which replication was inhibited, and the progeny viruses produced were analyzed for both their density and their genetic characteristics. The important result was that genetic recombinant viruses were obtained that had intermediate densities, indicating that they had acquired DNA from both parents, as predicted by the breakage-and-rejoining, but not the copy choice, model.

Models of Homologous Recombination

The finding that recombination occurs by breakage and rejoining raises a critical question: How can two parental DNA molecules be broken at precisely the same point, so that they can rejoin without mutations resulting from the gain or loss of nucleotides at the break point? During recombination between homologous DNA molecules (**general homologous recombination**), this alignment is provided, not surprisingly, by base pairing between complementary DNA strands (Figure 5.32). Overlapping single strands are exchanged between homologous DNA molecules, leading to the formation of a heteroduplex region, in which the two strands of the recombinant double helix are derived from different parents. If the heteroduplex region contains a genetic difference, the result is a single progeny DNA molecule that contains two genetic markers. In some cases, mispaired bases in a heteroduplex may be recognized and corrected by mismatch repair systems, as discussed in preceding sections of this chapter. Genetic evidence for the formation and repair of such heteroduplex regions, obtained in studies of recombination in fungi as well as in bacteria, led to the development of a molecular model for recombination in 1964. This model, known as the **Holliday model** (after Robin Holliday), has continued to provide the basis for current thinking about recombination mechanisms, although it has been modified as new data have been obtained.

The original version of the Holliday model proposed that recombination is initiated by the introduction of nicks at the same position on the two parental DNA molecules (Figure 5.33). The nicked DNA strands partially unwind, and each invades the other molecule by pairing with the complementary unbroken strand. Ligation of the broken strands then produces a crossed-strand intermediate, known as a **Holliday junction**, that is the central intermediate in recombination. The direct demonstration of Holliday junctions by electron microscopy has provided clear support for this model of recombination (Figure 5.34).

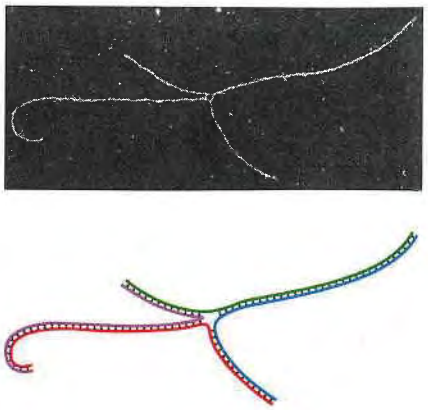


Figure 5.34 Identification of Holliday junctions by electron microscopy
Electron micrograph of a Holliday junction that was detected during recombination of plasmid DNAs in *E. coli*. An interpretive drawing of the structure is shown below. The molecule illustrates a Holliday junction in the open configuration resulting from rotation of the crossed-strand intermediate (see Figure 5.33). (Courtesy of Huntington Potter, University of South Florida, and David Dressler, University of Oxford.)

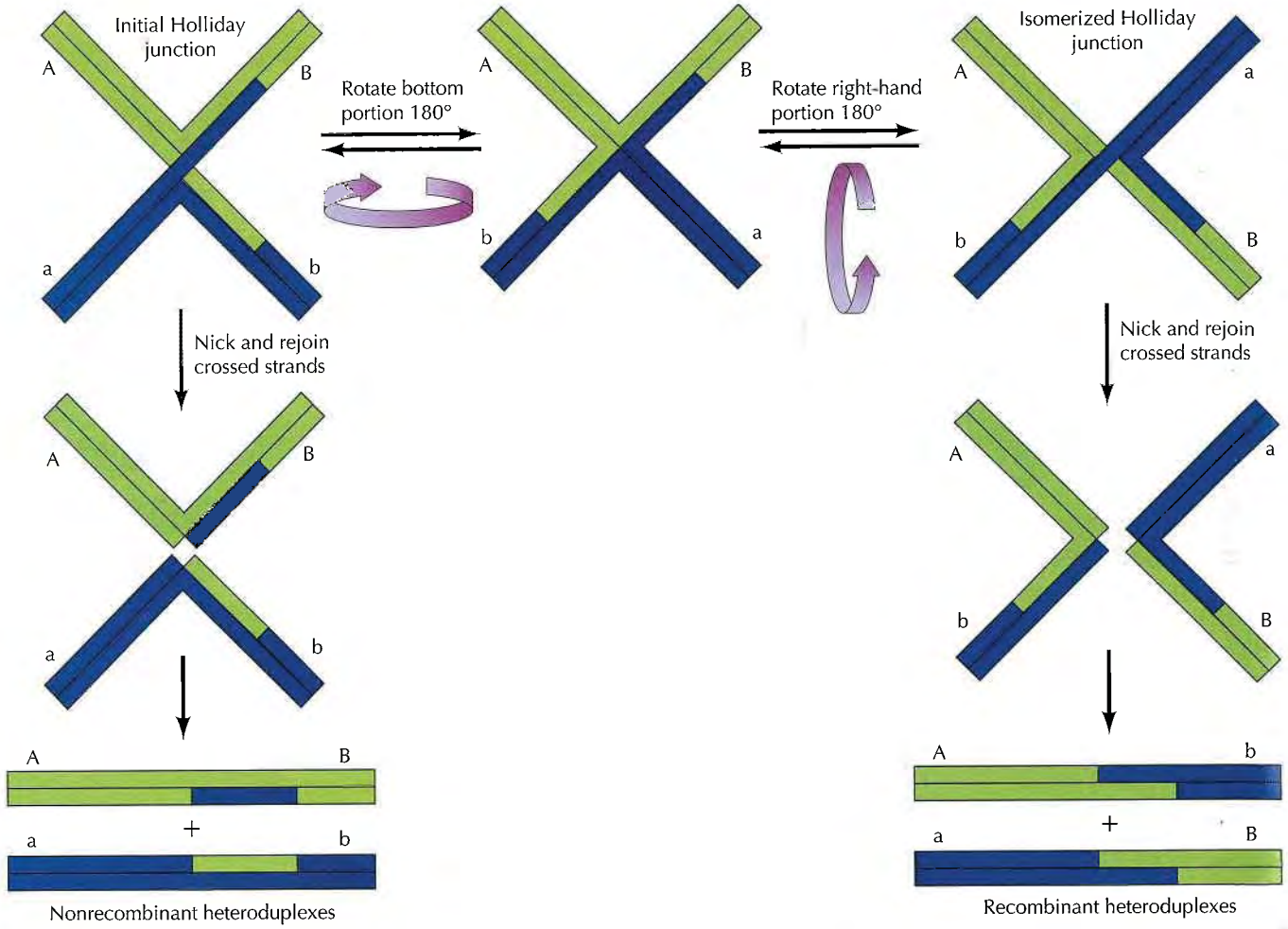


Figure 5.35 Isomerization and resolution of Holliday junctions
Holliday junctions are resolved by cutting and rejoining of the crossed strands. If the Holliday junction formed by the initial strand exchange is resolved, the resulting progeny are heteroduplexes but are not recombinant for genetic markers outside of the heteroduplex region. Two rotations of the crossed-strand molecule, however, produce an isomer in which the unbroken parental strands, rather than the initially nicked strands, are crossed. Cutting and rejoining of the crossed strands of this isomer yield progeny that are recombinant heteroduplexes.

Once a Holliday junction is formed, it can be resolved by cutting and rejoining of the crossed strands to yield recombinant molecules (Figure 5.35). This can occur in two different ways, depending on the orientation of the Holliday junction, which can readily form two different isomers. In the isomer resulting from the initial strand exchange, the crossed strands are those that were nicked at the start of the recombination process. However, simple rotation of this structure yields a different isomer in which the unbroken parental strands are crossed. Resolution of these different isomers has distinct genetic consequences. In the first case, the progeny molecules have heteroduplex regions but are nonrecombinant for DNA that flanks these regions. If isomerization occurs, however, cutting and rejoining of the crossed strands results in progeny molecules that are recombinant for DNA that flanks the heteroduplex regions. The structure of the Holliday junction thus provides the possibility of generating both recombinant and

nonrecombinant heteroduplexes, consistent with the genetic data upon which the Holliday model was based.

As initially proposed, the Holliday model failed to explain how recombination was initiated by simultaneously nicking both parental molecules at the same position. It now appears that recombination is generally initiated at double strand breaks, both during DNA repair and during recombination between homologous chromosomes during meiosis (Figure 5.36). Both strands of DNA at the double strand break are first resected by nucleases that digest DNA in the 5′ to 3′ direction, yielding single-stranded ends. These single strands then invade the other parental molecule by homologous base pairing. The gaps are then filled by repair synthesis and the strands are joined by ligation to yield a molecule with a double Holliday junction, which can be resolved to yield either recombinant or nonrecombinant heteroduplex molecules as already described.

Enzymes Involved in Homologous Recombination

Most of the enzymes currently known to be involved in recombination were first identified by analysis of recombination-defective mutants of *E. coli*. Such genetic analysis has established that recombination requires specific enzymes, in addition to proteins (such as DNA polymerase, ligase, and single-stranded DNA-binding proteins) that function in multiple aspects of DNA metabolism. The identification of genes required for efficient recombination in *E. coli* led to the isolation of their encoded proteins, which have

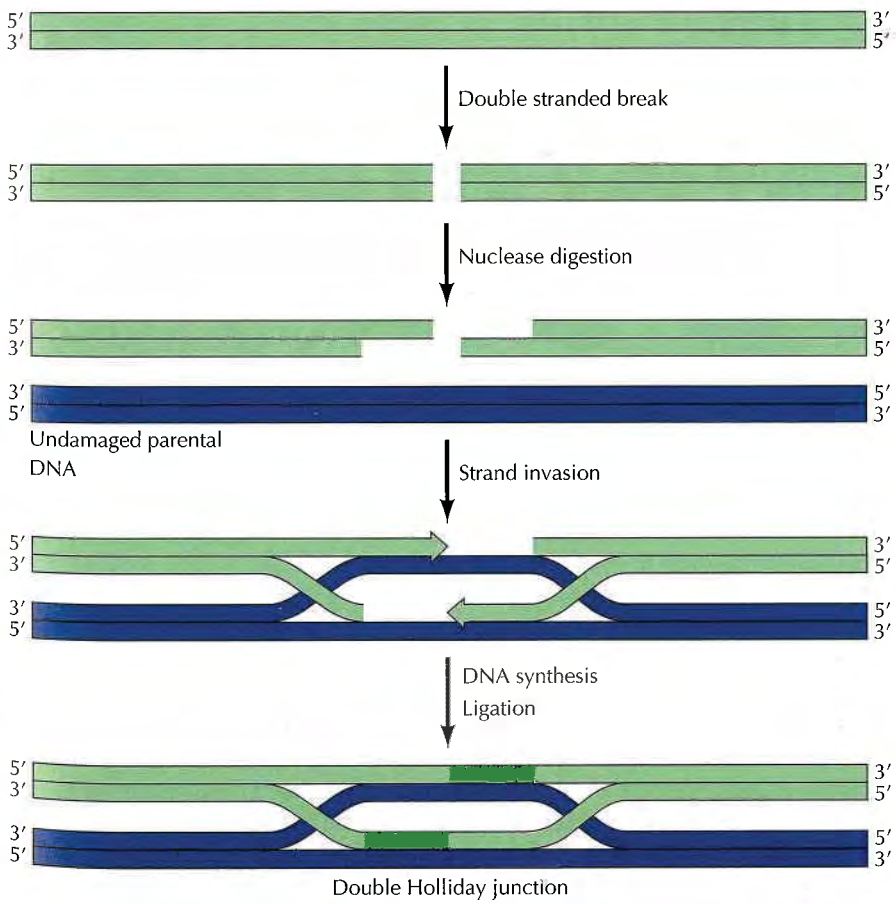


Figure 5.36 Initiation of recombination by double strand breaks Both strands of DNA at the double strand break are digested by nucleases in the 5′ to 3′ direction. The single strands then invade the other parental molecule by homologous base pairing. The gaps are then filled by repair synthesis and sealed by ligation, yielding a double Holliday junction.

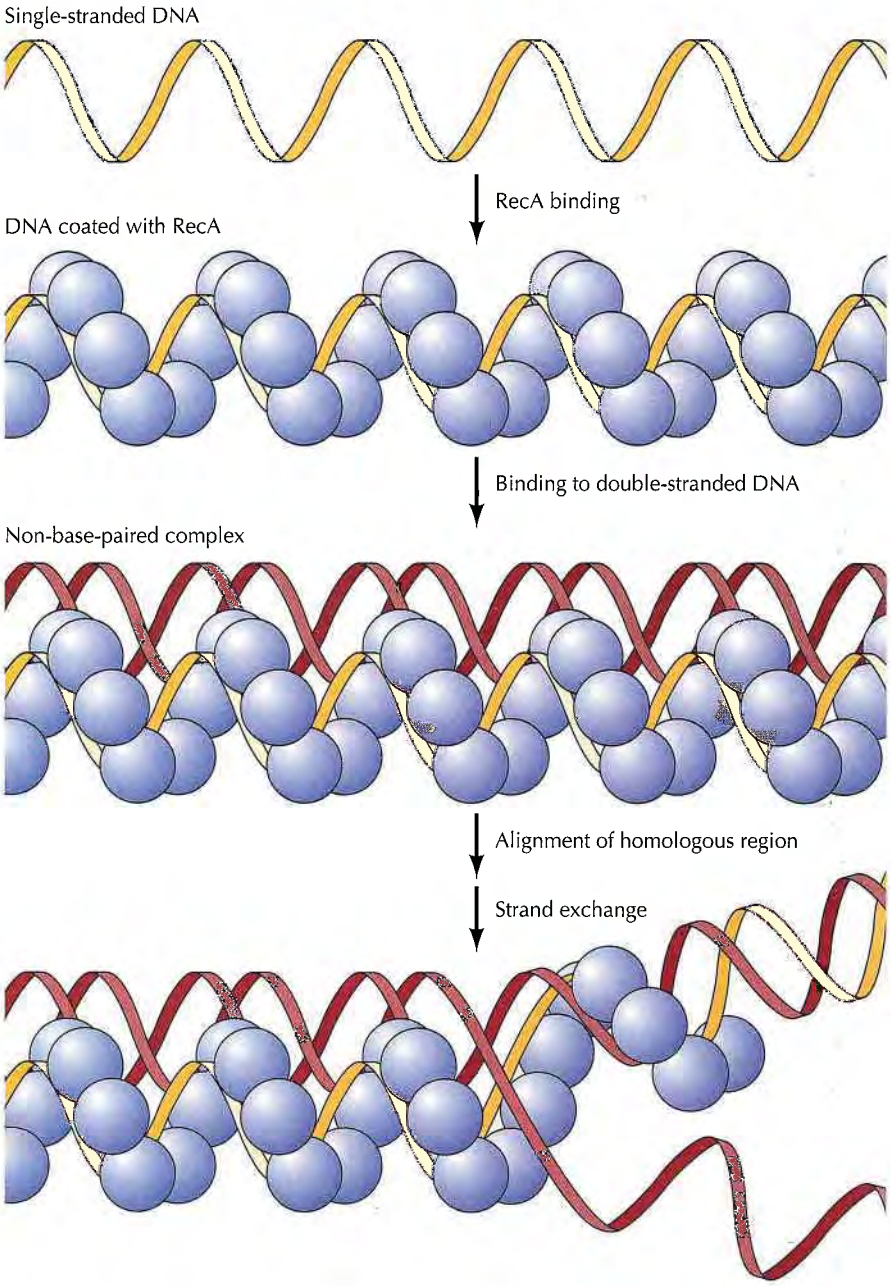


Figure 5.37 Function of the RecA protein
RecA initially binds to single-stranded DNA to form a protein-DNA filament. The RecA protein that coats the single-stranded DNA then binds to a second, double-stranded DNA molecule to form a non-base-paired complex. Complementary base pairing and strand exchange follow, forming a heteroduplex region.

been characterized by biochemical analysis in cell-free systems. These studies have elucidated the action of several enzymes in catalyzing the formation and resolution of Holliday junctions.

The central protein involved in homologous recombination is **RecA**, which promotes the exchange of strands between homologous DNAs that causes heteroduplexes to form (Figure 5.37). The action of RecA can be considered in three stages. First, the RecA protein binds to single-stranded DNA, coating the DNA to form a protein-DNA filament. Because RecA has two DNA binding sites, the RecA protein bound to single-stranded DNA is able to bind a second, double-stranded DNA molecule, forming a complex between the two DNAs. This nonspecific RecA-mediated association is followed by specific base pairing between the single-stranded DNA and its complement. The RecA protein then catalyzes strand exchange, with the single strand origi-

nally coated with RecA displacing its homologous strand to form a heteroduplex. Thus, the RecA protein is capable of catalyzing, by itself, the strand exchange reactions that are central to the formation of Holliday junctions.

In yeast, a RecA-related protein, designated RAD51, is required for genetic recombination as well as for the repair of double strand breaks. RAD51 is not only structurally similar to RecA; like RecA, it is also able to catalyze strand exchange reactions *in vitro*. Proteins related to RAD51 have been identified in complex eukaryotes, including humans, indicating that proteins related to RecA play key roles in homologous recombination in both prokaryotic and eukaryotic cells.

Once a Holliday junction is formed, a complex of three other *E. coli* proteins (RuvA, B, and C) become involved in recombination (Figure 5.38). RuvA recognizes the Holliday junction and recruits RuvB, which acts as a motor to drive migration of the site at which the DNA strands are crossed, thereby varying the extent of the heteroduplex region and the position at which the crossed strands will be cut and rejoined. RuvC then resolves Holliday junctions by cleaving the crossed DNA strands. Rejoining of the cleaved strands by ligation completes the process, yielding two recombinant molecules. Eukaryotic cells do not have homologs of the *E. coli* RuvA, B, and C proteins. Instead, the resolution of Holliday junctions in eukaryotic cells appears to be mediated by other proteins, which remain to be fully characterized.

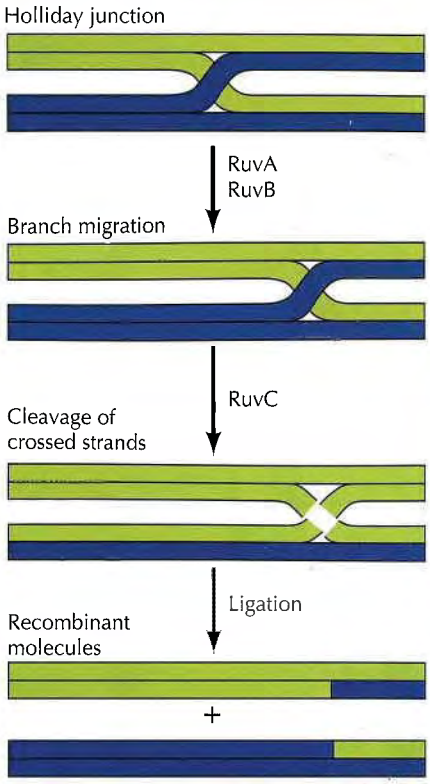


Figure 5.38 Branch migration and resolution of Holliday junctions
RuvA recognizes the Holliday junction and recruits RuvB, which catalyzes the movement of the crossed-strand site (branch migration). RuvC resolves the Holliday junctions by cleaving the crossed strands, which are then joined by ligase.

DNA Rearrangements

Homologous recombination results in the reassortment of genes between chromosome pairs without altering the arrangement of genes within the genome. In contrast, other types of recombinational events lead to rearrangements of genomic DNA. Some of these DNA rearrangements are important in controlling gene expression in specific cell types; others may play an evolutionary role by contributing to genetic diversity.

The discovery that genes can move to different chromosomal locations came from Barbara McClintock's studies of corn in the 1940s. Purely on the basis of genetic analysis, McClintock described novel genetic elements that could move to different locations in the genome and alter the expression of adjacent genes. Nearly three decades elapsed, however, before the physical basis of McClintock's work was elucidated by the discovery of transposable elements in bacteria and the notion of movable genetic elements became widely accepted by scientists. Several types of DNA rearrangements, including the transposition of elements initially described by McClintock, are now recognized in both prokaryotic and eukaryotic cells. Moreover, we now know that transposable elements constitute a large fraction of the genomes of plants and animals, including nearly half of the human genome.

Site-Specific Recombination

In contrast to general homologous recombination, which occurs at any extensive region of sequence homology, **site-specific recombination** occurs between specific DNA sequences, which are usually homologous over only a short stretch of DNA. The principal interaction in this process is mediated by proteins that recognize the specific DNA target sequences rather than by complementary base pairing.

The prototype of site-specific recombination has been provided by studies of the bacteriophage λ. When λ infects *E. coli*, it can either replicate to cause cell lysis or it can integrate into the bacterial chromosome, forming a

prophage that is then maintained as part of the *E. coli* genome (a process called **lysogeny**) (Figure 5.39). Under appropriate conditions, DNA integration can be reversed, resulting in excision of the λ DNA and initiation of lytic viral replication. Both the integration and the excision of λ DNA involve site-specific recombination between viral and host cell DNA sequences.

E. coli DNA and λ DNA recombine at specific sites, called attachment (*att*) sites. Thus, integration of λ DNA involves recombination between *att* sites of the phage (*attP*) and the bacterium (*attB*), which are about 240 and 25 nucleotides long, respectively (Figure 5.40). The process is mediated by a

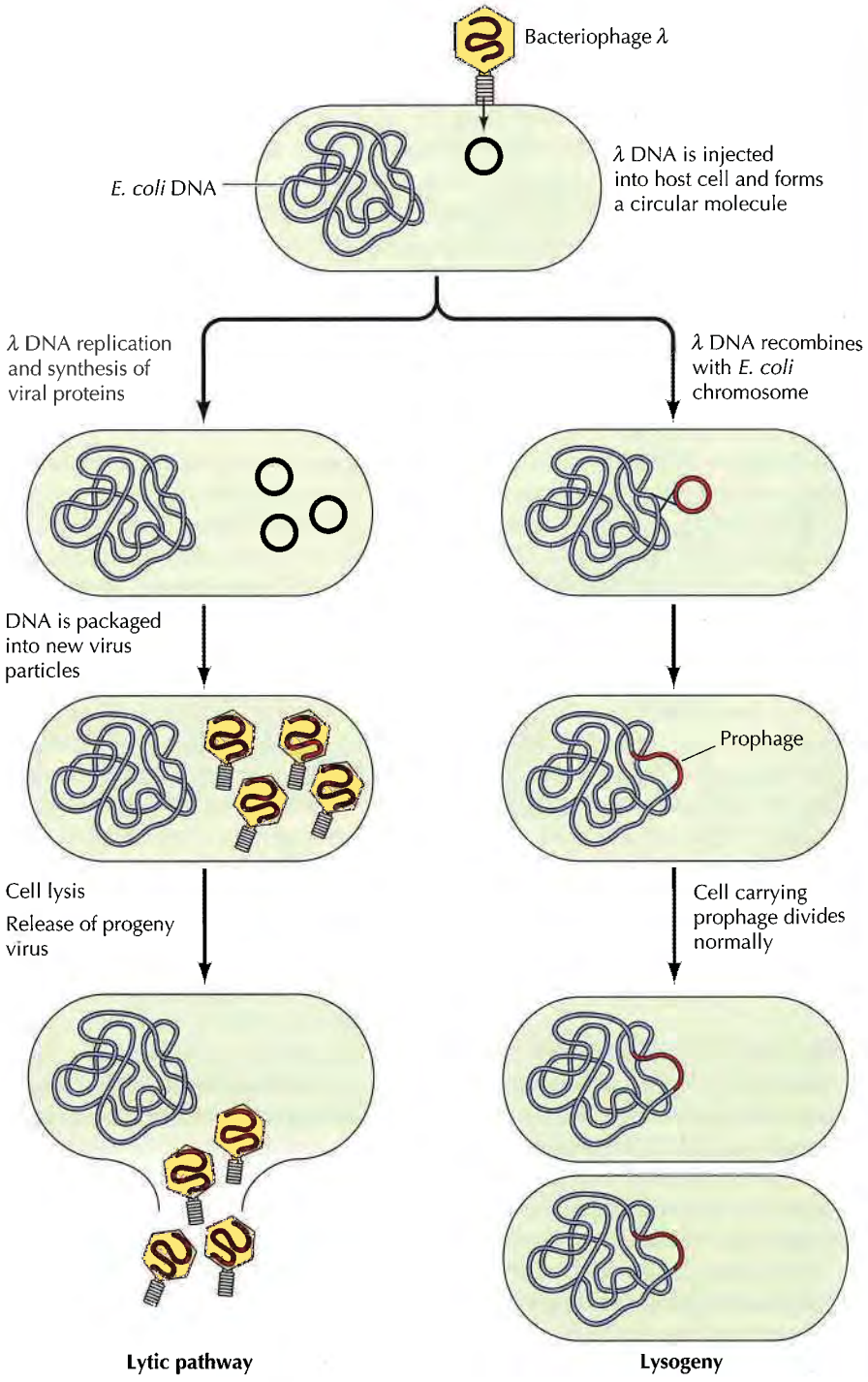


Figure 5.39 Lytic and lysogenic pathways of bacteriophage λ
Infection of *E. coli* is initiated by the injection of λ DNA, which then becomes circular within the host cell. In lytic infection, the λ DNA replicates and directs the synthesis of viral proteins. The viral DNA is then packaged into progeny virus particles, which are released upon cell lysis. In lysogenic infection, the λ DNA recombines with the host genome to form a prophage that is integrated into the *E. coli* chromosome. The integrated λ DNA does not direct the synthesis of progeny viruses, but is instead replicated along with the rest of the bacterial genome.

λ protein called integrase (Int), which specifically binds to both *attP* and *attB* sequences. Int initially binds to *attP*, forming a complex in which the *attP* DNA is wrapped around multiple copies of the Int protein. The Int-*attP* complex binds to *attB*, aligning the phage and bacterial *att* sites. The phage and bacterium then exchange strands within a 15-nucleotide core sequence shared by *attB* and *attP* (Figure 5.41). The Int protein introduces staggered cuts within the core homology region of *attB* and *attP*, catalyzes strand exchange, and then ligates the broken strands, integrating the λ DNA into the *E. coli* chromosome. The Int protein also acts in excision of the λ prophage, which is essentially the reverse of integration.

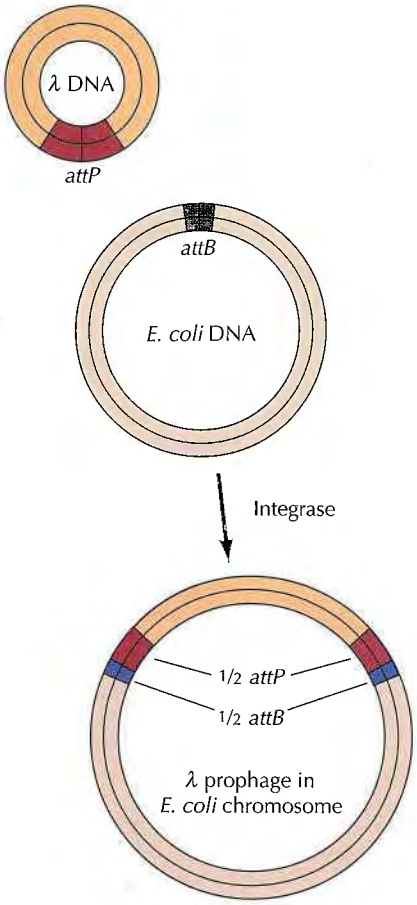


Figure 5.40 Integration of λ DNA by site-specific recombination
Integration results from recombination between specific sequences in the λ and *E. coli* genomes, called *attP* and *attB*, respectively. The process is catalyzed by a virus-encoded enzyme (integrase), which recognizes both *attP* and *attB* sequences.

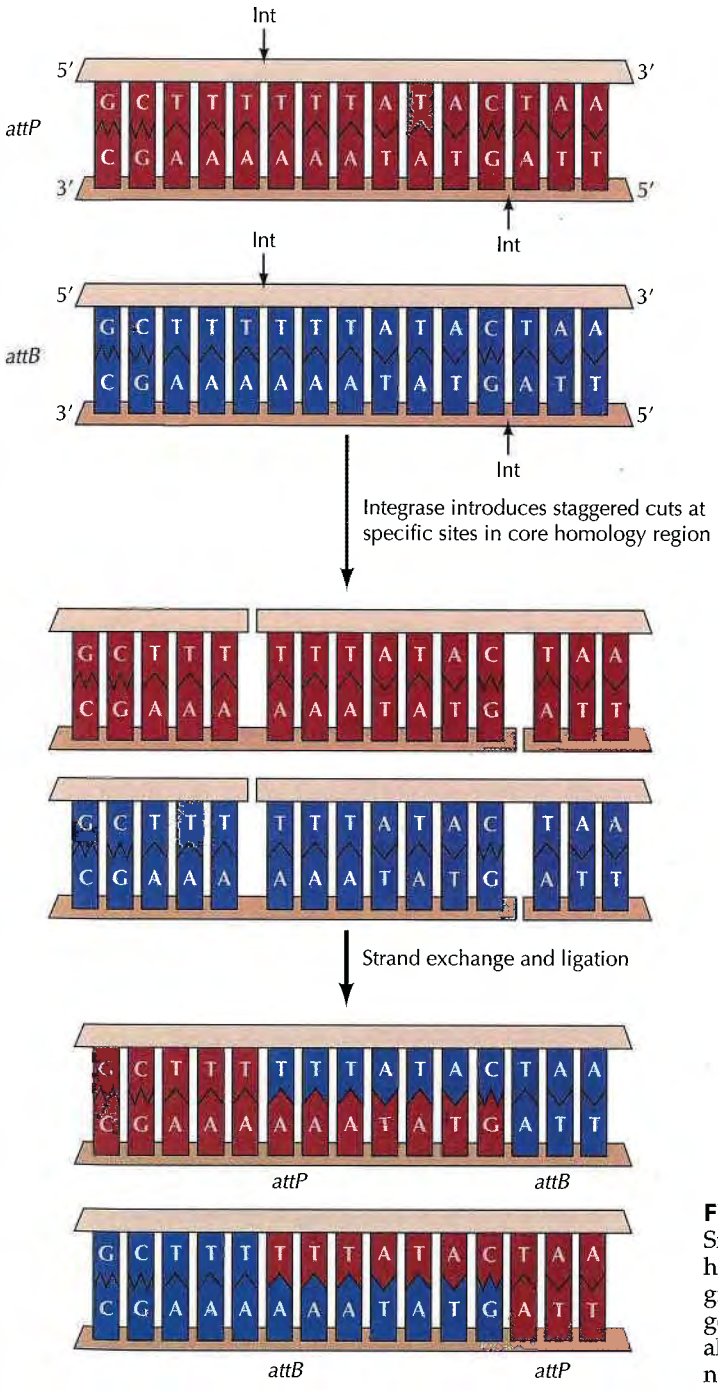


Figure 5.41 Mechanism of λ site-specific recombination
Site-specific recombination occurs within a 15-nucleotide homologous core sequence shared by *attP* and *attB*. Integrase (Int) cleaves at specific sites within this sequence to generate staggered single-stranded DNA tails. It then catalyzes strand exchange and ligation, resulting in recombination between *attP* and *attB* and integration of λ DNA.

Site-specific recombination is important not only in the interaction of viruses such as λ with their host cells, but also in programmed gene rearrangements within cell genomes. In vertebrates, site-specific recombination is critical to the development of the immune system, which recognizes foreign substances (**antigens**) and provides protection against infectious agents. There are two major classes of immune responses, which are mediated by B and T lymphocytes. B lymphocytes secrete antibodies (**immunoglobulins**) that react with soluble antigens; T lymphocytes express cell surface proteins (called **T cell receptors**) that react with antigens expressed on the surfaces of other cells. The key feature of both immunoglobulins and T cell receptors is their enormous diversity, which enables different antibody or T cell receptor molecules to recognize a vast array of foreign antigens. For example, each individual is capable of producing more than 10^{11} different antibody molecules, which is far in excess of the total number of genes in the human genome (30,000–40,000). Rather than being encoded in germ-line DNA, these diverse antibodies (and T cell receptors) are encoded by unique lymphocyte genes that are formed during development of the immune system as a result of site-specific recombination between distinct segments of immunoglobulin and T cell receptor genes.

The role of site-specific recombination in the formation of immunoglobulin genes was first demonstrated by Susumu Tonegawa in 1976. Immunoglobulins consist of pairs of identical heavy and light polypeptide chains (Figure 5.42). Both the heavy and light chains are composed of C-terminal constant regions and N-terminal variable regions. The variable regions, which have different amino acid sequences in different immunoglobulin molecules, are responsible for antigen binding, and it is the diversity of variable region amino acid sequences that allows different individual antibodies to recognize unique antigens. Although every individual is capable of producing a vast spectrum of different antibodies, each B lymphocyte produces only a single type of antibody. Tonegawa's key discovery was that each antibody is encoded by unique genes formed by site-specific recombination during B lymphocyte development. These gene rearrangements create different immunoglobulin genes in different individual B lymphocytes, so the population of approximately 10^{12} B lymphocytes in the human body includes cells capable of producing antibodies against a diverse array of foreign antigens.

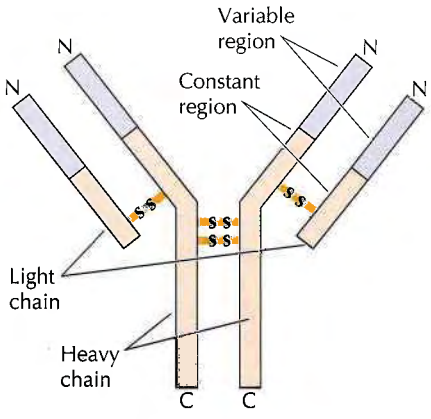


Figure 5.42 Structure of an immunoglobulin
Immunoglobulins are composed of two heavy chains and two light chains, joined by disulfide bonds. Both the heavy and the light chains consist of variable and constant regions.

The genes that encode immunoglobulin light chains consist of three regions: a V region that encodes the 95 to 96 N-terminal amino acids of the polypeptide variable region; a joining (J) region that encodes the 12 to 14 C-terminal amino acids of the polypeptide variable region; and a C region that encodes the polypeptide constant region (Figure 5.43). The major class of light-chain genes in the mouse is formed from combinations of approximately 250 V regions and four J regions with a single C region. Site-specific recombination during lymphocyte development leads to a gene rearrangement in which a single V region recombines with a single J region to generate a functional light-chain gene. Different V and J regions are rearranged in different B lymphocytes, so the possible combinations of 250 V regions with 4 J regions can generate approximately 1000 (4×250) unique light chains.

The heavy-chain genes include a fourth region (known as the diversity, or D, region), which encodes amino acids lying between V and J (Figure 5.44). Assembly of a functional heavy-chain gene requires two recombination events: A D region first recombines with a J region, and a V region then recombines with the rearranged DJ segment. In the mouse, there are about 500 heavy-chain V regions, 12 D regions, and 4 J regions, so the total num-

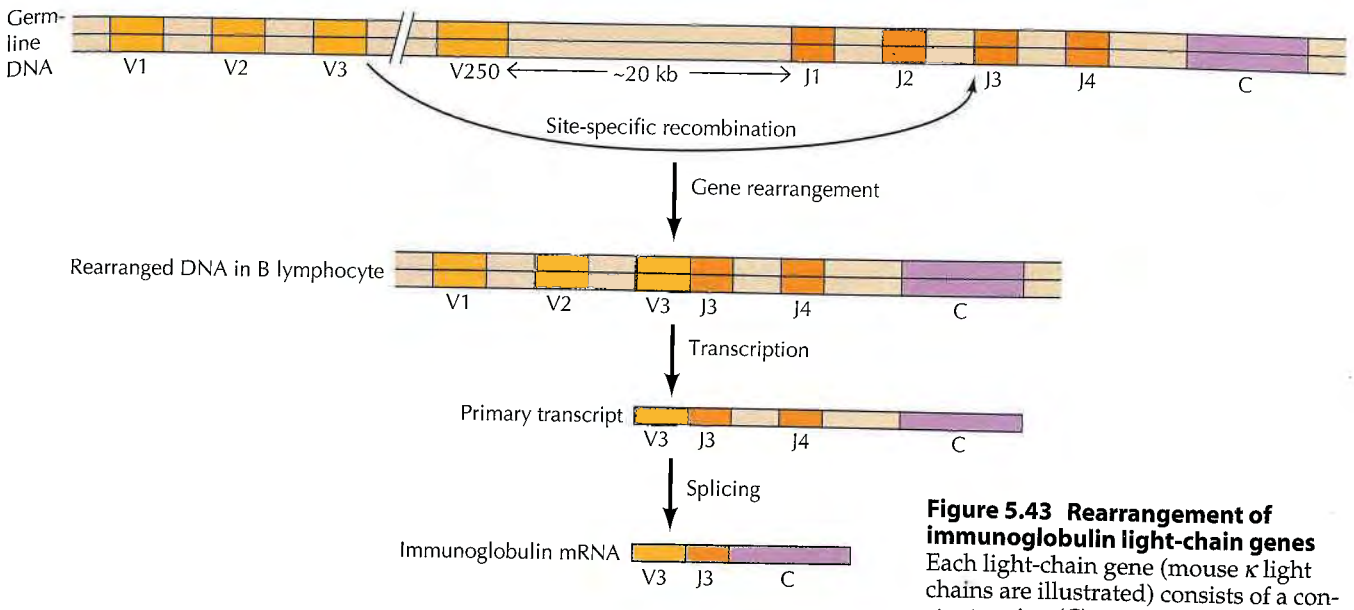


Figure 5.43 Rearrangement of immunoglobulin light-chain genes
Each light-chain gene (mouse κ light chains are illustrated) consists of a constant region (C), a joining region (J), and a variable region (V). There are approximately 250 different V regions, which are separated from J and C by about 20 kb in germ-line DNA. During the development of B lymphocytes, site-specific recombination joins one of the V regions to one of the four J regions. This rearrangement activates transcription, resulting in the formation of a primary transcript containing the rearranged VJ region together with the remaining J regions and C. The remaining unused J regions and the introns between J and C are then removed by splicing, yielding a functional mRNA.

ber of heavy chains that can be generated by the recombination events is 24,000 ($500 \times 12 \times 4$).

Combinations between the 1000 different light chains and 24,000 different heavy chains formed by site-specific recombination can generate approximately 2×10^7 different immunoglobulin molecules. This diversity is further increased because the joining of immunoglobulin gene segments often involves the loss or gain of one to several nucleotides. The mutations resulting from these deletions and insertions increase the diversity of immunoglobulin variable regions approximately a hundredfold, corresponding to the formation of about 10^5 different light chains and 2×10^6 heavy

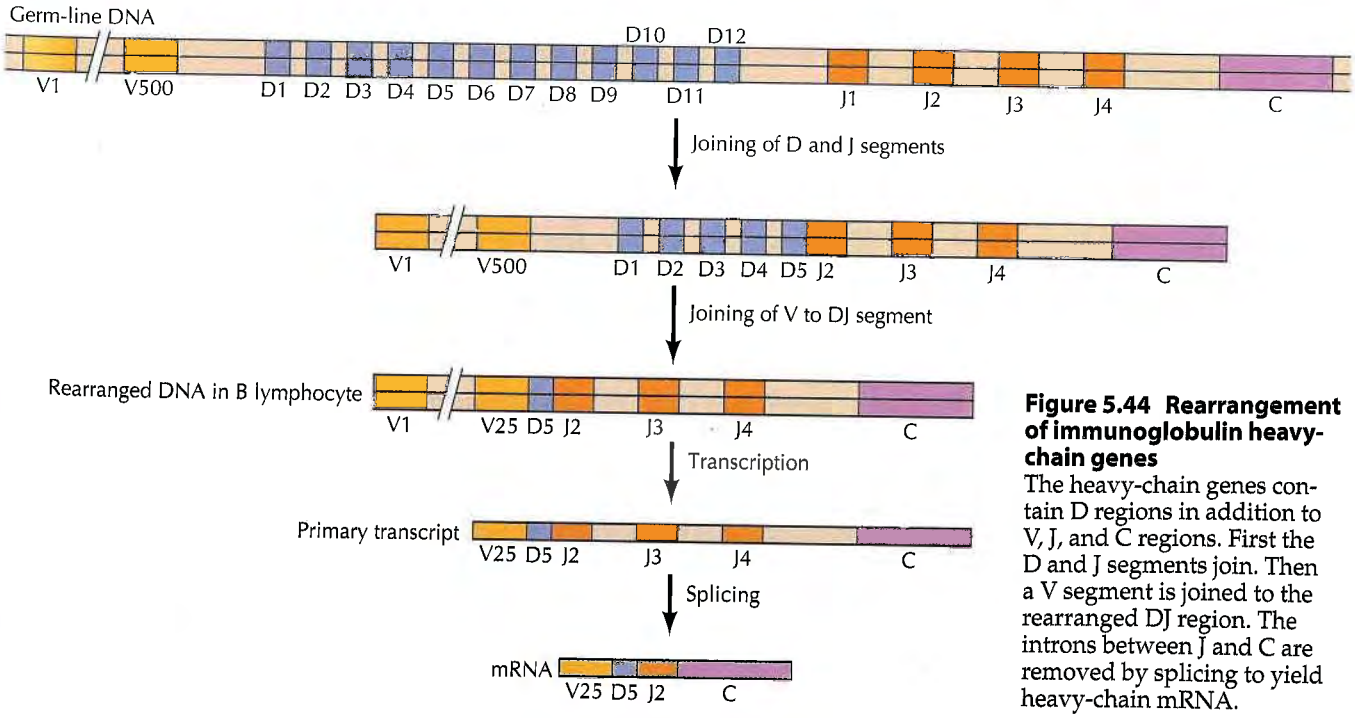


Figure 5.44 Rearrangement of immunoglobulin heavy-chain genes
The heavy-chain genes contain D regions in addition to V, J, and C regions. First the D and J segments join. Then a V segment is joined to the rearranged DJ region. The introns between J and C are removed by splicing to yield heavy-chain mRNA.

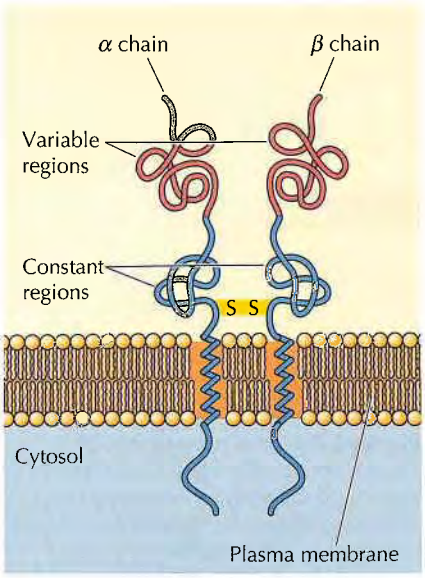


Figure 5.45 Structure of a T cell receptor
T cell receptors consist of two polypeptide chains (α and β) that span the plasma membrane and are joined by disulfide bonds. Both the α and β chains are composed of variable and constant regions.

chains, which can then combine to form more than 10^{11} distinct antibodies. Still further antibody diversity is generated after the formation of rearranged immunoglobulin genes by a process known as somatic hypermutation, which results in the introduction of frequent mutations into the variable regions of both heavy-chain and light-chain genes.

T cell receptors similarly consist of two chains (called α and β), each of which contains variable and constant regions (Figure 5.45). The genes encoding these polypeptides are generated by recombination between V and J segments (the α chain) or between V, D, and J segments (the β chain), analogous to the formation of immunoglobulin genes. Site-specific recombination between these distinct segments of DNA, in combination with mutations introduced during recombination, generates a degree of diversity in T cell receptors that is similar to that in immunoglobulins. However, T cell receptors differ from immunoglobulins in that they are not subject to the introduction of further diversity by somatic hypermutation.

VDJ recombination is mediated by a complex of two proteins, called RAG1 and RAG2, which are specifically expressed in lymphocytes. The RAG proteins recognize recombination signal (RS) sequences adjacent to the coding sequences of each gene segment, and initiate DNA rearrangements by introducing a double strand break between the RS sequences and the coding sequences (Figure 5.46). The coding ends of the gene segments are then joined to yield a rearranged immunoglobulin or T cell receptor gene, frequently with the loss or gain of nucleotides during the joining reaction. Interestingly, RAG1 is closely related to the enzymes that catalyze DNA transposition and retroviral integration, as discussed in the next section of this chapter.

Transposition via DNA Intermediates

Site-specific recombination occurs between two specific sequences that contain at least a small core of homology. In contrast, **transposition** involves the movement of sequences throughout the genome and has no requirement for sequence homology. Elements that move by transposition, such as those first described by McClintock, are called **transposable elements,** or **transposons**. They are divided into two general classes, depending on whether they transpose via DNA intermediates or via RNA intermediates. The first class of transposable elements is discussed here; transposition via RNA intermediates is considered in the next section.

The first transposons that were characterized in detail are those of bacteria, which move via DNA intermediates (Figure 5.47). The simplest of these elements are the insertion sequences, ranging in size from about 800 to 2000 nucleotides. Insertion sequences consist only of a gene for the enzyme involved in transposition (transposase) flanked by short inverted repeats, which are the sites at which transposase acts. Complex transposons consist of two insertion sequences flanking other genes, which move as a unit.

Insertion sequences move from one chromosomal site to another without replicating their DNA (Figure 5.48). Transposase introduces a staggered break in the target DNA and cleaves at the ends of the transposon inverted-repeat sequences. Although transposase acts specifically at the transposon

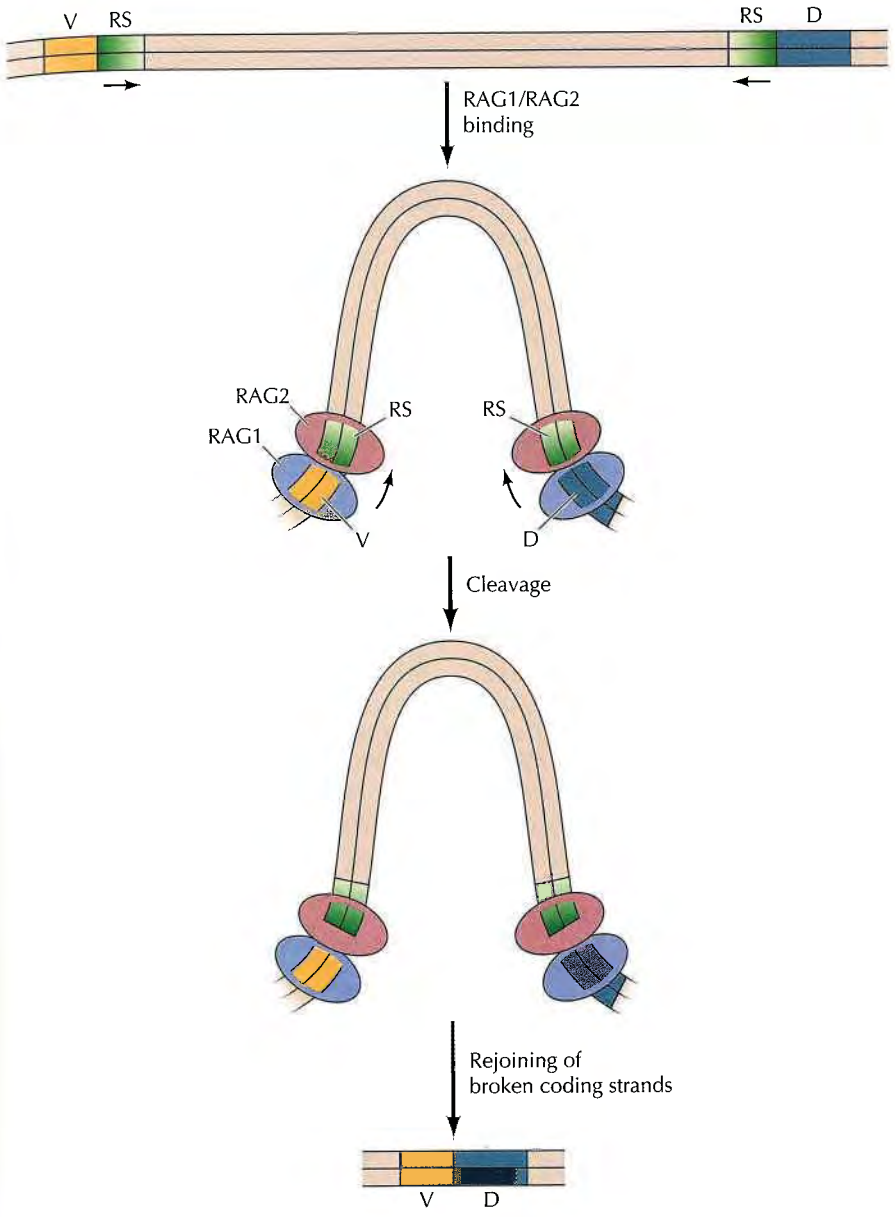


Figure 5.46 VDJ recombination
The coding segments of immunoglobulin and T cell receptor genes (e.g., a V and D segment) are flanked by short recombination signal (RS) sequences, which are in opposite orientations at the 5′ and 3′ ends of the coding sequences. The RS sequences are recognized by a complex of the lymphocyte-specific recombination proteins RAG1 and RAG2, which cleave the DNA between the coding sequence and the RS sequence. The broken coding strands are then rejoined to yield a rearranged gene segment.

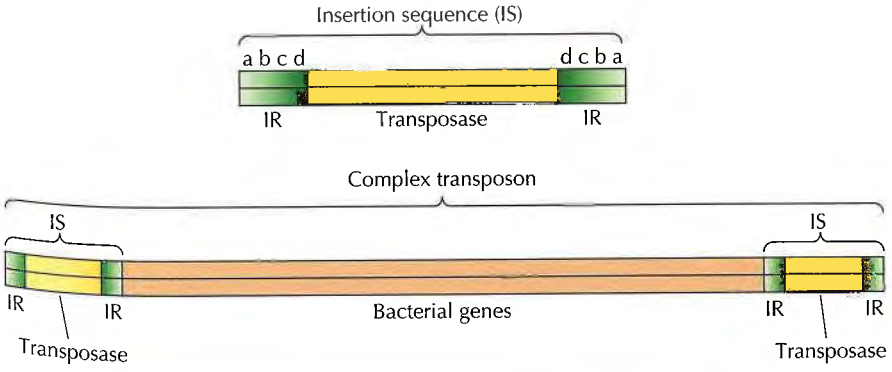


Figure 5.47 Bacterial transposons
Insertion sequences (IS) range from 800 to 2000 nucleotides and contain a gene for transposase flanked by inverted repeats (IR) of about 20 nucleotides. Complex transposons consist of two insertion sequences flanking other genes and are typically 5 to 20 kb long.

KEY EXPERIMENT

Rearrangement of Immunoglobulin Genes

Evidence for Somatic Rearrangement of Immunoglobulin Genes Coding for Variable and Constant Regions

Nobumichi Hozumi and Susumu Tonegawa
Basel Institute for Immunology, Basel, Switzerland
Proceedings of the National Academy of Sciences, USA, Volume 73, 1976, pages 3628–3632

The Context

The ability of the vertebrate immune system to recognize a seemingly infinite variety of foreign molecules implies that lymphocytes can produce a correspondingly vast array of antibodies. Since this antibody diversity is key to immune recognition, understanding the mechanism by which an apparently unlimited number of distinct immunoglobulins are encoded in genomic DNA is a central issue in immunology.

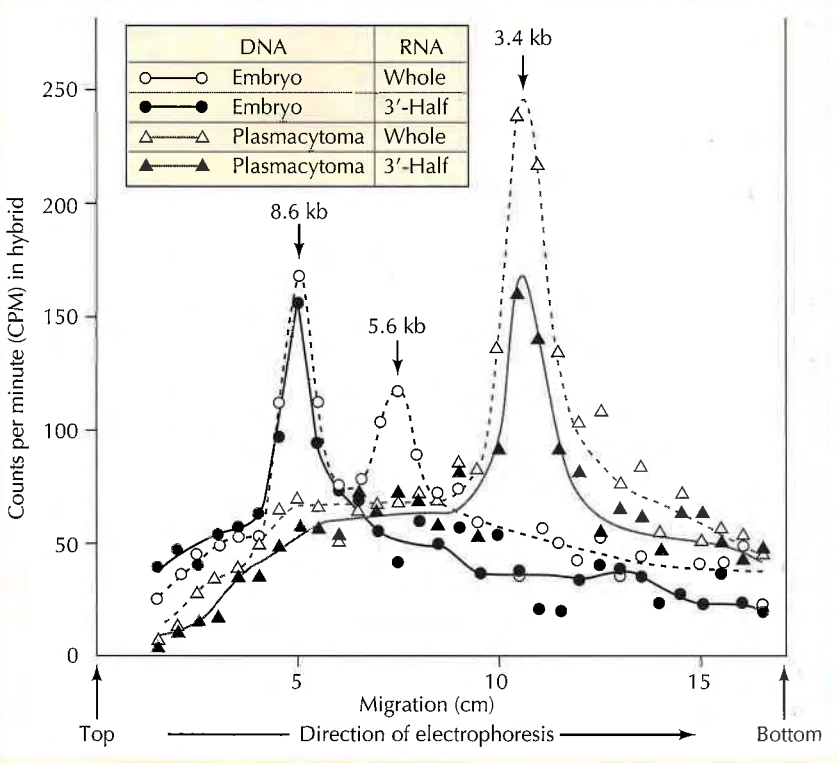
Prior to the experiments of Hozumi and Tonegawa, protein sequencing of multiple immunoglobulins had demonstrated that both heavy and light chains consist of distinct variable and constant regions. Genetic studies further indicated that mice inherit only single copies of the constant-region genes. These observations first led to the proposal that immunoglobulins are encoded by multiple variable-region genes that can associate with a single constant-region gene. The discovery of immunoglobulin gene rearrangements by Hozumi and Tonegawa provided the first direct experimental support for this hypothesis and laid the groundwork for understanding the molecular basis of antibody diversity.

The Experiments

Hozumi and Tonegawa tested the possibility that the genes encoding immunoglobulin variable and constant regions were joined at the DNA level during lymphocyte development. Their experimental approach was to use restriction endonuclease digestion to compare the organization of variable-region and constant-region sequences in DNAs extracted from mouse embryos and from cells of a mouse plasmacytoma (a B lymphocyte tumor that produces a single species of immunoglobulin).

Embryo and plasmacytoma DNAs were digested with the restriction endonuclease *Bam*HI, and DNA fragments of different sizes were separated by electrophoresis in agarose gels. The gel was then cut into slices, and DNA extracted from each slice was hybridized with radiolabeled probes that had been prepared from immunoglobulin mRNA isolated from the plasmacytoma cells. Two probes were used, corresponding either to the complete immunoglobulin mRNA or to the 3′ half of the mRNA, consisting only of constant-region sequences.

The critical result was that completely different patterns of variable-region and constant-region sequences were detected in embryo versus plasmacytoma DNAs (see figure). In embryo DNA, the complete probe hybridized to two *Bam*HI fragments of approximately 8.6 and 5.6 kb, respectively. Only the 8.6-kb fragment hybridized to the 3′ probe, suggesting that the 8.6-kb fragment contained constant-region sequences and the 5.6-kb fragment contained variable-region sequences. In striking contrast, both probes hybridized to only a single 3.4-kb fragment in plasmacytoma DNA.



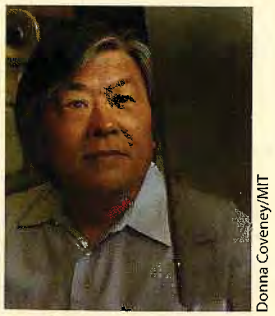
Gel electrophoresis of embryo and plasmacytoma DNAs digested with *Bam*HI and hybridized to probes corresponding to either the whole or the 3′ half of the plasmacytoma mRNA. Data are presented as the radioactivity detected in hybrid molecules with DNA from each gel slice.

The interpretation of these results was that the variable- and constant-region sequences were separated in embryo DNA but rearranged to form a single immunoglobulin gene during lymphocyte development.

The Impact

The initial results of Hozumi and Tonegawa, based on the relatively indirect approach of restriction endonuclease mapping, were confirmed and extended by the molecular cloning and sequencing of immunoglobulin genes. Such studies have now unambiguously established that these genes are generated by site-specific recombination between distinct segments of DNA in B lymphocytes. In T lymphocytes, similar DNA rearrangements are responsible for formation of the genes encoding T cell receptors. Thus, site-specific recombination and programmed gene rearrangements are central to the development of the immune system.

Further studies have shown that the variable regions of immunoglobulins and T cell receptors are generated by rearrangements of two or three distinct segments of DNA. The ability of these segments to recombine, together with a high frequency of mutations introduced at the recombination sites, is largely responsible for immunoglobulin and T cell receptor diversity.



Donna Coveney/MIT

Susumu Tonegawa

The discovery of immunoglobulin gene rearrangements thus provided the basis for understanding how the immune system can recognize and respond to a virtually unlimited range of foreign substances.

inverted-repeats, it is usually less specific with respect to the sequence of the target DNA, so it catalyzes the movement of transposons throughout the genome. Following the cleavage of transposon and target site DNAs, transposase joins the overhanging ends of the target DNA to the transposable element. The resulting gap in the target-site DNA is repaired by DNA

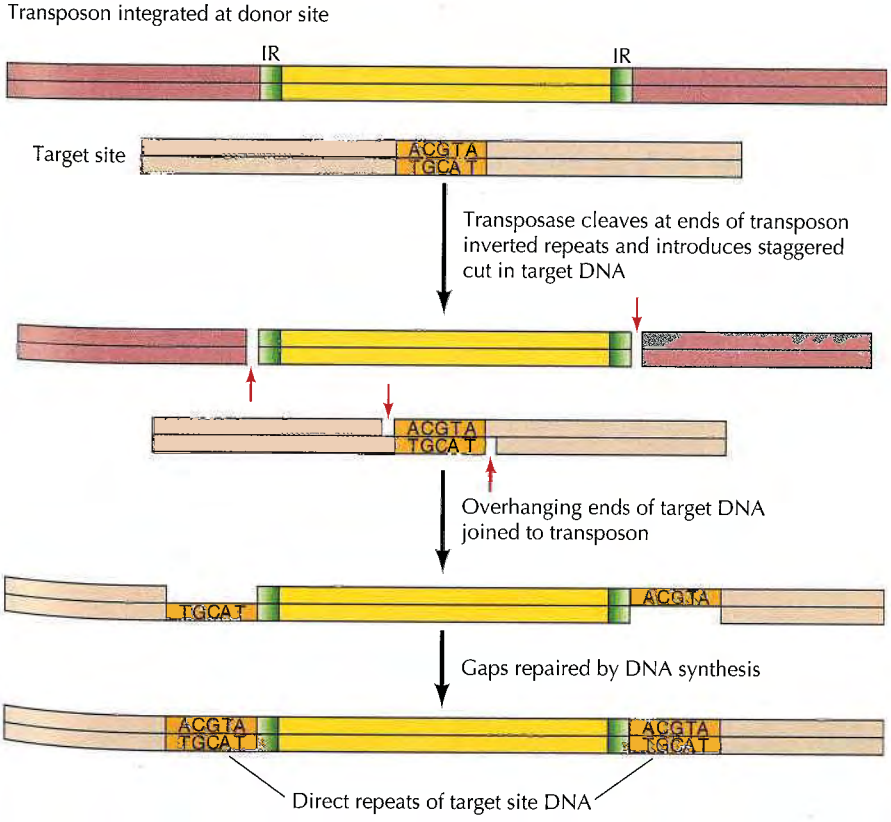


Figure 5.48 Transposition of insertion sequences
Simple transposition does not involve replication of the transposon DNA. Transposase cleaves at both ends of the transposon and introduces a staggered cut in the target DNA. The overhanging ends of target DNA are then joined to the transposon, and gaps resulting from the staggered cuts at the target site are repaired. The result is the formation of short direct repeats of target-site DNA (5 to 10 nucleotides long) flanking the integrated transposon.

synthesis, followed by ligation to the other strand of the tranposon. The result of this process is a short direct repeat of the target-site DNA on both sides of the transposable element—a hallmark of transposon integration.

This transposition mechanism causes the transposon to move from one chromosomal site to another. Other types of transposons move by a more complex mechanism, in which the transposon is replicated in concert with its integration into a new target site. This mechanism results in the integration of one copy of the transposon into a new position in the genome, while another copy remains at its original location.

Transposons that move via DNA intermediates are present in eukaryotes as well as in bacteria. For example, the human genome contains approximately 300,000 DNA transposons, which account for about 3% of human DNA. The original transposable elements described by McClintock in corn move by a nonreplicative mechanism, as do most transposable elements in other plants and animals. Like bacterial transposons, these elements move to many different target sites throughout the genome. The movement of these transposons to nonspecific sites in the genome is not likely to be useful to the cells in which it occurs, but has undoubtedly played a major role in evolution by promoting DNA rearrangements.

In yeasts and protozoans, however, transposition by a replicative mechanism is responsible for programmed DNA rearrangements that regulate gene expression. In these cases transposition is initiated by the action of a site-specific nuclease that cleaves a specific target site, at which a copy of the transposable element is then inserted. Transposable elements are thus capable not only of moving to nonspecific sites throughout the genome, but also of participating in specific gene rearrangements that result in programmed changes in gene expression.

Transposition via RNA Intermediates

Most transposons in eukaryotic cells are **retrotransposons**, which move via reverse transcription of RNA intermediates. In humans, there are almost 3 million copies of retrotransposons, accounting for more than 40% of the genome (see Table 4.1). The mechanism of transposition of these elements is similar to the replication of retroviruses, which have provided the prototype system for studying this class of movable DNA sequences.

Retroviruses contain RNA genomes in their virus particles but replicate via the synthesis of a DNA provirus, which is integrated into the chromosomal DNA of infected cells (see Figure 3.13). A DNA copy of the viral RNA is synthesized by the viral enzyme **reverse transcriptase**. The mechanism by which this occurs results in the synthesis of a DNA molecule that contains direct repeats of several hundred nucleotides at both ends (Figure 5.49). These repeated sequences, called **long terminal repeats**, or **LTRs**, arise from duplication of the sites on viral RNA at which primers bind to initiate DNA synthesis. The LTR sequences thus play central roles in reverse transcription, in addition to being involved in the integration and subsequent transcription of proviral DNA.

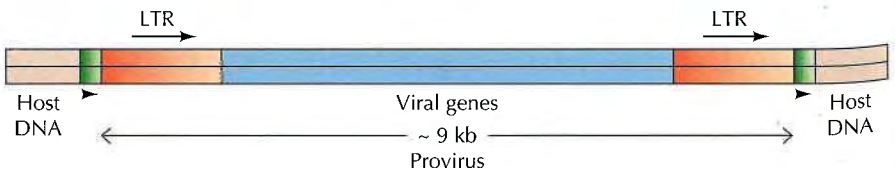


Figure 5.49 The organization of retroviral DNA
The integrated proviral DNA is flanked by long terminal repeats (LTRs), which are direct repeats of several hundred nucleotides. Viral genes, including genes for reverse transcriptase, integrase, and structural proteins of the virus particle, are located between the LTRs. The integrated provirus is flanked by short direct repeats of host DNA.

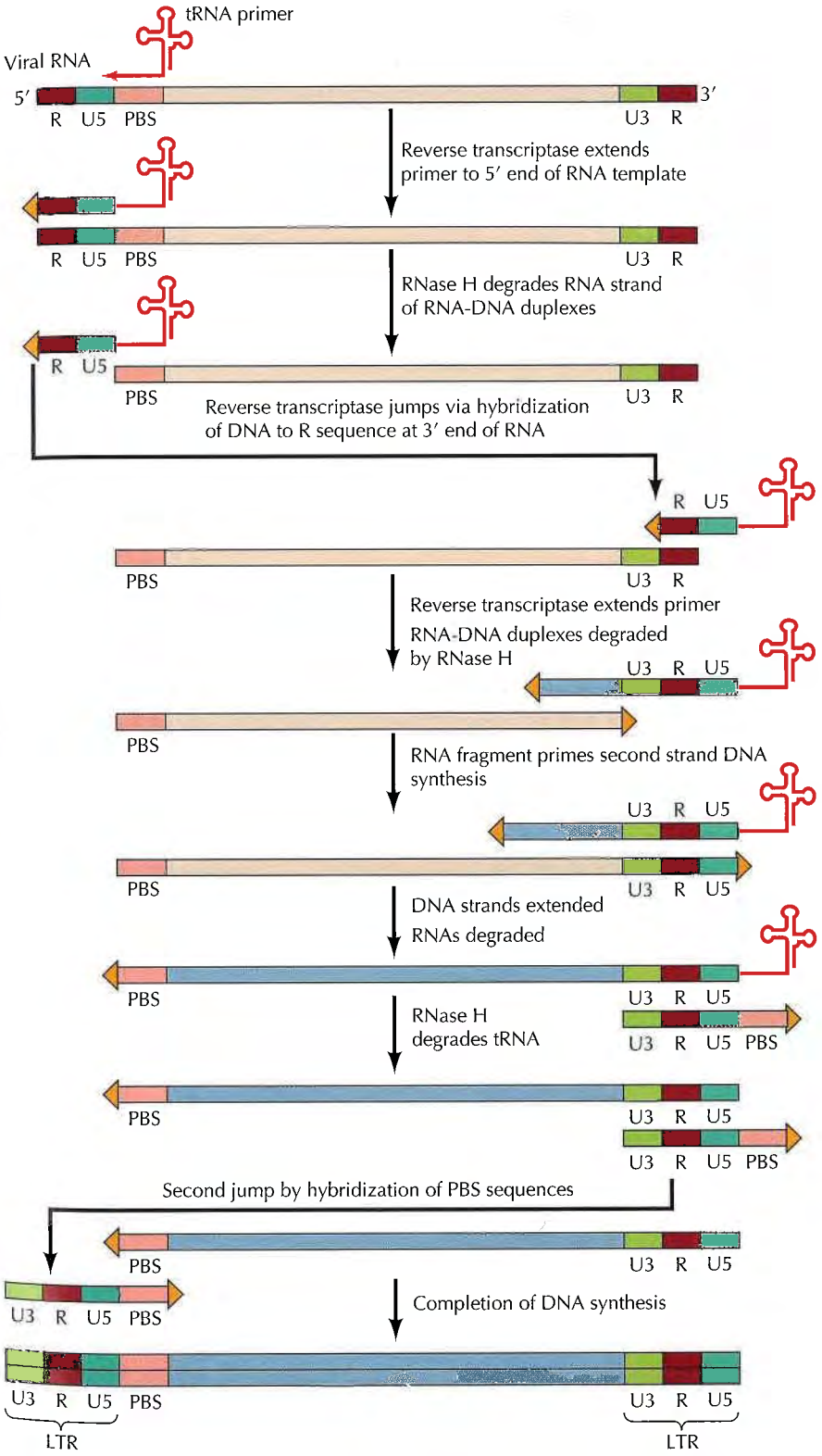


Figure 5.50 Generation of LTRs during reverse transcription
LTRs consist of three sequence elements: a short repeat sequence (R) of about 20 nucleotides that is present at both ends of the viral RNA; a sequence unique to the 5′ end of viral RNA (U5); and a sequence unique to the 3′ end of viral RNA (U3). Repeats of these sequences are generated during DNA synthesis as reverse transcriptase jumps twice between the ends of its template. Synthesis is initiated using a tRNA primer bound to a primer-binding site (PBS) adjacent to U5 at the 5′ end of the viral RNA. The polymerase copies R, and the RNA strand of the RNA-DNA hybrid is then degraded by RNase H. The polymerase then jumps to the 3′ end of the viral RNA in order to synthesize a complete DNA strand complementary to the RNA template. The polymerase jumps again during synthesis of the second strand of DNA, which is also initiated by a primer bound close to the 5′ end of its template. The result of these jumps is the formation of LTRs that contain U3-R-U5 sequences.

Like all DNA polymerases, reverse transcriptase requires a primer, which in the case of retroviruses is a tRNA molecule bound at a specific site (the primer-binding site) close to the 5′ terminus of the viral RNA (Figure 5.50). Since DNA synthesis proceeds in the 5′ to 3′ direction, only a short

piece of DNA is synthesized before reverse transcriptase reaches the end of its template. Continuation of DNA synthesis then depends on the ability of reverse transcriptase to "jump" to the 3′ end of the template RNA molecule. This is accomplished via an RNase H activity of reverse transcriptase, which degrades the RNA strand of DNA-RNA hybrids. As a result, the newly synthesized DNA is converted to a single-stranded molecule, which can hybridize to a short repeated sequence present at both the 5′ and the 3′ ends of the viral RNA. DNA synthesis can then continue, yielding a single-stranded DNA complementary to viral RNA. Synthesis of the opposite strand of DNA is initiated by a fragment of viral RNA that acts as a primer, at a site near the 3′ end of the template DNA strand. Again the result is a short piece of DNA, which includes the primer-binding site copied from the tRNA used as the initial primer for reverse transcription. The primer-binding sequence of the tRNA is then degraded by RNase H, leaving an overhanging DNA strand that again "jumps" to pair with its complementary sequence at the other end of the template. DNA synthesis can then continue once more, finally yielding a linear DNA with LTRs at both ends.

The linear viral DNA integrates into the host cell chromosome by a process that resembles the integration of DNA transposable elements. Integration is catalyzed by a viral integrase and occurs at many different target sequences in cellular DNA. The integrase cleaves two bases from the ends of viral DNA and introduces a staggered cut at the target site in cellular DNA. The overhanging ends of cellular DNA are then joined to the termini of viral DNA, and the gap is filled by DNA synthesis. The integrated provirus is therefore flanked by a direct repeat of cell sequences, similar to the repeats that flank DNA transposons.

The viral life cycle continues with transcription of the integrated provirus, which yields viral genomic RNA as well as mRNAs that direct the synthesis of viral proteins (including reverse transcriptase and integrase). The genomic RNA is then packaged into viral particles, which are released from the host cell. These progeny viruses can infect a new cell, initiating another round of DNA synthesis and integration. The net effect can be viewed as the movement of the provirus from one chromosomal site to another, via the synthesis and reverse transcription of an RNA intermediate.

Other retrotransposons differ from retroviruses in that they are not packaged into infectious particles and therefore cannot spread from one cell to another. However, these retrotransposons can move to new chromosomal sites within the same cell, via mechanisms fundamentally similar to those involved in retrovirus replication.

Some retrotransposons (called retrovirus-like elements or LTR retrotransposons) are structurally similar to retroviruses (Figure 5.51). Retrotransposons of this type account for about 8% of the human genome. They

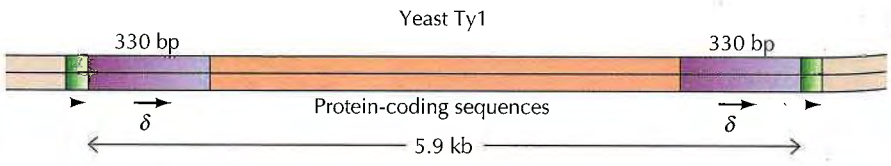


Figure 5.51 Structure of a LTR retrotransposon
The yeast Ty1 transposable element displays the same organization as a retrovirus. Protein-coding sequences, including genes for reverse transcriptase and integrase, are flanked by LTRs (called δ elements) of about 330 base pairs (bp). The integrated transposon is flanked by short direct repeats of target-site DNA.

have LTR sequences at both ends; they encode reverse transcriptase and integrase; and they transpose (like retroviruses) via transcription into RNA, synthesis of a new DNA copy by reverse transcriptase, and integration into cellular DNA.

The non-LTR retrotransposons differ from retroviruses in that they do not contain LTR sequences, although they do encode their own reverse transcriptase. In mammals, the major class of these retrotransposons consists of the highly repetitive long interspersed elements (**LINEs**), which are repeated approximately 850,000 times in the genome and account for about 21% of genomic DNA (see Chapter 4). A full-length LINE element is 6 to 7 kb long, although most members of the family are truncated at their 5′ end (Figure 5.52). At their 3′ end, LINEs have tracts of A-rich sequences thought to be derived by reverse transcription of the poly-A tails that are added to mRNAs following transcription (see Chapter 6). Like other transposable elements, LINEs are flanked by short direct repeats of the target-site DNA, indicating that integration involves staggered cuts and repair synthesis.

Since LINEs do not contain LTR sequences, the mechanism of their reverse transcription and subsequent integration into chromosomal DNA must differ from that of retroviruses and LTR-containing retrotransposons. In particular, reverse transcription is primed by a broken end of chromosomal DNA at the integration target site, resulting from cleavage of the target site DNA by a nuclease encoded by the retrotransposon (Figure 5.53). Reverse transcription then initiates within the poly-A tract at the 3′ end of the transposon RNA and continues along the molecule. The opposite strand of DNA is synthesized using the other broken end of target-site DNA as primer, resulting in simultaneous synthesis and integration of the retrotransposon DNA.

Other sequence elements, which do not encode their own reverse transcriptase, also transpose via RNA intermediates. These elements include the highly repetitive short interspersed elements (**SINEs**), of which there are approximately 1.5 million copies in mammalian genomes (see Chapter 4). The major family of these elements in humans consists of the *Alu* sequences, which are about 300 bases long. These sequences have A-rich tracts at their 3′ end and are flanked by short duplications of target-site DNA sequences, a structure similar to that of non-LTR retrotransposons (e.g., LINEs). SINEs arose by reverse transcription of small RNAs, including tRNAs and small cytoplasmic RNAs involved in protein transport. Since SINEs no longer encode functional RNA products, they represent pseudogenes that arose via RNA-mediated transposition. Pseudogenes of many protein-coding genes (called **processed pseudogenes**) have similarly arisen by reverse transcription of mRNAs (Figure 5.54). Such processed pseudogenes are readily recognized not only because they terminate in an A-rich tract but also because the

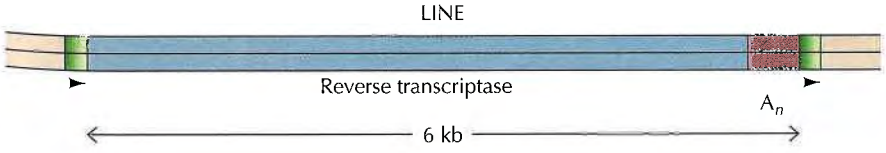


Figure 5.52 Structure of human LINEs
LINEs lack LTRs, but they do encode reverse transcriptase. They have tracts of A-rich sequences (designated A_n) at their 3′ ends, which are thought to arise from reverse transcription of poly-A tails added to the 3′ end of mRNAs. Like other transposable elements, LINEs are flanked by short direct repeats of target-site DNA.

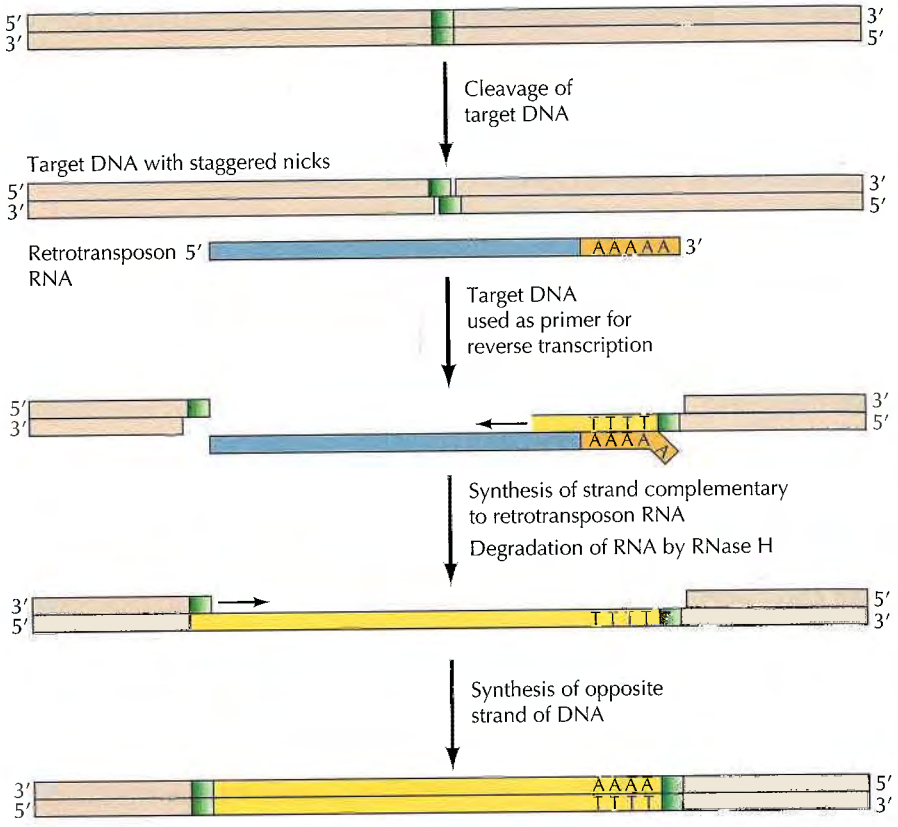


Figure 5.53 Model for reverse transcription and integration of LINEs Target site DNA is cleaved by a nuclease encoded by the retrotransposon. Reverse transcription, primed by a broken end of the target DNA, initiates within the poly-A tail at the 3′ end of retrotransposon RNA. Synthesis of the opposite strand of retrotransposon DNA is similarly primed by the other strand of DNA at the target site.

introns present in the corresponding normal gene have been removed during mRNA processing. The transposition of SINEs and of other processed pseudogenes is thought to proceed similarly to the transposition of LINEs. However, since these elements do not include genes for reverse transcriptase or a nuclease, their transposition presumably involves the action of reverse transcriptases and nucleases that are encoded elsewhere in the genome—probably by other retrotransposons, such as LINEs.

Although the highly repetitive SINEs and LINEs account for a significant fraction of genomic DNA, their transpositions to random sites in the genome are not likely to be useful for the cell in which they are located. These transposons induce mutations when they integrate at a new target site, and like mutations induced by other agents, most mutations resulting from transposon integration are expected to be harmful to the cell. Indeed, mutations resulting from the transposition of both LINEs and SINES have been associated with some cases of hemophilia, muscular dystrophy, breast cancer, and colon cancer. On the other hand, some mutations resulting from the movement of transposable elements may be beneficial, contributing in a positive way to evolution of the species. For example, some retrotransposons in mammalian genomes have been found to contain regulatory sequences that control the expression of adjacent genes.

In addition to their role as mutagens, retrotransposons have played a major role in shaping the genome by stimulating DNA rearrangements. For example, rearrangements of chromosomal DNA can result from recombination between LINEs integrated at different sites in the genome. Moreover, sequences of cellular DNA adjacent to LINEs are frequently carried along during the process of transposition. Consequently, the transposition of

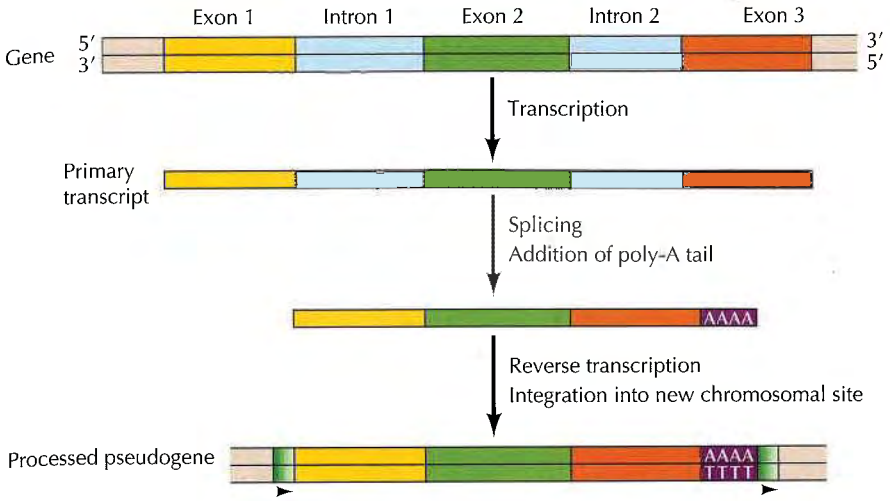


Figure 5.54 Formation of a processed pseudogene
The gene illustrated contains three exons, separated by two introns. The introns are removed from the primary transcript by splicing, and a poly-A tail is added to the 3′ end of the mRNA. Reverse transcription and integration then yield a processed pseudogene, which does not contain introns and has an A-rich tract at its 3′ end. The processed pseudogene is flanked by short direct repeats of target-site DNA that were generated during its integration.

LINEs can result in the movement of cellular DNA sequences to new genomic sites. Since LINEs can integrate into active genes, the associated transposition of cellular DNA sequences can lead to the formation of new combinations of regulatory and/or coding sequences and contribute directly to the evolution of new genes.

The vast majority of the transposable elements in the human genome are inactive, with only about 100 copies of LINEs still retaining the protein-coding sequences required for their transposition. All of the human DNA transposons and most retrotransposons thus represent evolutionary relics rather than currently functional elements. However, this is not the case in other species, including *Arabidopsis*, *C. elegans*, *Drosophila*, and mice, which have a much higher level of ongoing transposon activity. In the mouse, for example, LTR retrotransposons, LINEs, and SINEs are all active. As a consequence, it is estimated that about 10% of all mutations in mice are the result of transposons, compared to only about 1 in 600 mutations in humans. There is thus a dramatic and intriguing difference in transposon activity between mice and humans, the explanation for which remains to be determined.

Gene Amplification

The DNA rearrangements that have been discussed so far alter the position of a DNA sequence within the genome. **Gene amplification** may be viewed as a different type of alteration in genome structure; it increases the number of copies of a gene within a cell. Gene amplification results from repeated rounds of DNA replication, yielding multiple copies of a particular region (Figure 5.55). The amplified DNA sequences can be found either as free extrachromosomal molecules or as tandem arrays of sequences within a chromosome. In either case, the result is increased expression of the amplified gene, simply because more copies of the gene are available to be transcribed.

In some cases, gene amplification is responsible for developmentally programmed increases in gene expression. The prototypical example is amplification of the ribosomal RNA genes in amphibian oocytes (eggs). Eggs are extremely large cells, with correspondingly high requirements for protein synthesis. Amphibian oocytes in particular are about a million times larger in volume than typical somatic cells and must support large amounts of protein synthesis during early development. This requires increased synthesis of ribosomal RNAs, which is accomplished in part by amplification

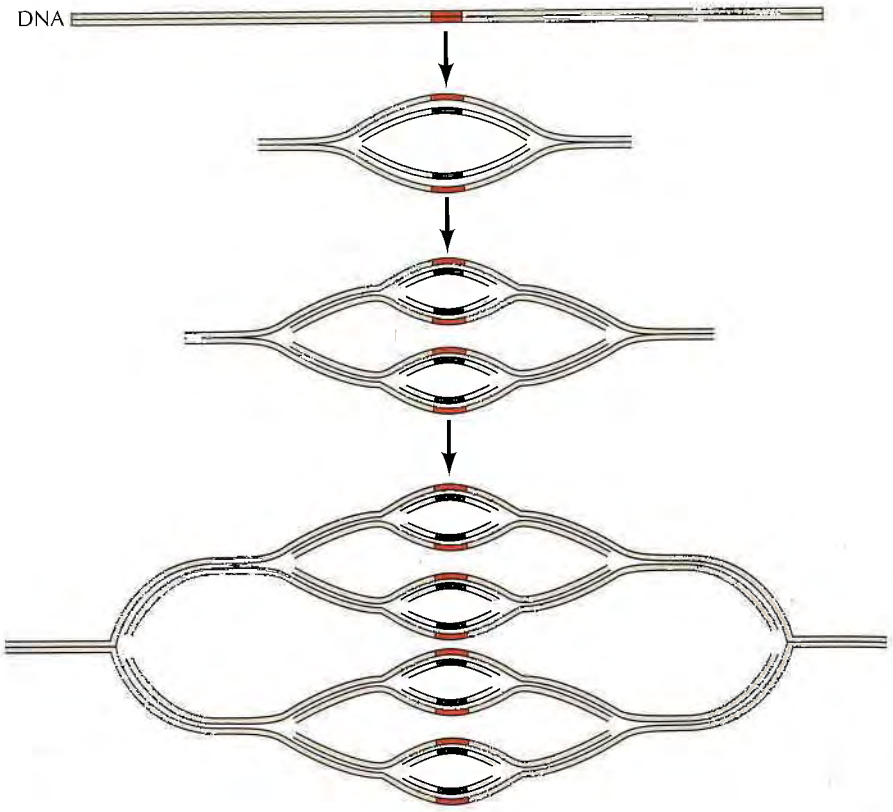


Figure 5.55 DNA amplification
Repeated rounds of DNA replication yield multiple copies of a particular chromosomal region.

of the ribosomal RNA genes. As discussed in Chapter 4, there are already several hundred copies of ribosomal RNA genes per genome, so that enough ribosomal RNA can be produced to meet the needs of somatic cells. In amphibian eggs, these genes are amplified an additional 2000-fold, to approximately 1 million copies per oocyte. Another example of programmed gene amplification occurs in *Drosophila*, where the genes that encode eggshell proteins (chorion genes) are amplified in ovarian cells to support the requirement for large amounts of these proteins. Like other programmed gene rearrangements, however, gene amplification is a relatively infrequent event that occurs in highly specialized cell types; it is not a common mechanism of gene regulation.

Gene amplification also occurs as an abnormal event in cancer cells, where it results in the increased expression of genes that contribute to uncontrolled cell growth. Such gene amplification was first recognized in cancer cells that had become resistant to methotrexate, a drug commonly used in cancer chemotherapy. Methotrexate inhibits the enzyme dihydrofolate reductase, which is involved in the synthesis of dTTP and is therefore required for DNA synthesis. Resistance to methotrexate frequently develops by amplification of the dihydrofolate reductase gene, leading to increased production of the enzyme and consequently the loss of effective inhibition by methotrexate. In addition, gene amplification in cancer cells frequently results in the increased expression of genes that drive cell proliferation (oncogenes) and thereby directly contributes to tumor development (see Chapter 15). For example, amplification of the oncogene *erb*B-2 is frequently involved in human breast cancers. Thus, as with other types of DNA rearrangements, gene amplification can have either beneficial or deleterious consequences for the cell or organism in which it occurs.

SUMMARY

DNA REPLICATION

DNA Polymerases: Different DNA polymerases play distinct roles in DNA replication and repair in both prokaryotic and eukaryotic cells. All known DNA polymerases synthesize DNA only in the 5′ to 3′ direction by the addition of dNTPs to a preformed primer strand of DNA.

DNA polymerase, mutagen

The Replication Fork: Parental strands of DNA separate and serve as templates for the synthesis of two new strands at the replication fork. One new DNA strand (the leading strand) is synthesized in a continuous manner; the other strand (the lagging strand) is formed by the joining of small fragments of DNA that are synthesized backward with respect to the overall direction of replication. DNA polymerases and various other proteins act in a coordinated manner to synthesize both leading and lagging strands of DNA.

replication fork, Okazaki fragment, DNA ligase, leading strand, lagging strand, primase, RNase H, exonuclease, helicase, single-stranded DNA-binding protein, topoisomerase

The Fidelity of Replication: DNA polymerases increase the accuracy of replication both by selecting the correct base for insertion and by proofreading newly synthesized DNA to eliminate mismatched bases.

proofreading

Origins and the Initiation of Replication: DNA replication starts at specific origins of replication, which contain binding sites for proteins that initiate the process.

origin of replication, autonomously replicating sequence (ARS), origin replication complex (ORC)

Telomeres and Telomerase: Replicating the Ends of Chromosomes: Telomeric repeat sequences at the ends of chromosomes are replicated by the action of a reverse transcriptase (telomerase) that carries its own template RNA.

telomere, telomerase, reverse transcriptase

DNA REPAIR

Direct Reversal of DNA Damage: A few types of common DNA lesions, such as pyrimidine dimers and alkylated guanine residues, are repaired by direct reversal of the damage.

pyrimidine dimer, photoreactivation

Excision Repair: Most types of DNA damage are repaired by excision of the damaged DNA. The resulting gap is filled by newly synthesized DNA, using the undamaged complementary strand as a template. In base-excision repair, specific types of single damaged bases are removed from the DNA molecule. In contrast, nucleotide excision repair systems recognize a wide variety of lesions that distort the structure of DNA and remove the damaged bases as part of an oligonucleotide. A third excision repair system specifically removes mismatched bases from newly synthesized DNA strands.

base-excision repair, DNA glycosylase, AP endonuclease, nucleotide-excision repair, excinuclease, transcription-coupled repair, mismatch repair

Error-Prone Repair: Specialized DNA polymerases are capable of replicating DNA across from a site of DNA damage, although the action of these polymerases results in a high frequency of incorporation of incorrect bases.

error-prone repair

Recombinational Repair: Damaged DNA can be replaced by recombination with an undamaged molecule. This mechanism plays an important role in repairing damage encountered during DNA replication as well as in the repair of double strand breaks.

recombinational repair

RECOMBINATION BETWEEN HOMOLOGOUS DNA SEQUENCES

DNA Molecules Recombine by Breaking and Rejoining: The molecular mechanism of recombination involves the breaking and rejoining of parental DNA molecules.

general homologous recombination, Holliday model, Holliday junction

Models of Homologous Recombination: Alignment between homologous DNA molecules is provided by complementary base pairing. Nicked strands of parental DNA invade the other parental molecule, yielding a crossed-strand intermediate known as a Holliday junction. Recombinant molecules are then formed by cleavage and rejoining of the crossed strands.

RecA

Enzymes Involved in Homologous Recombination: The central enzyme of homologous recombination is RecA, which catalyzes the exchange of strands between homologous DNAs. Other enzymes nick and unwind parental DNAs and resolve Holliday junctions.

DNA REARRANGEMENTS

site-specific recombination, lysogeny, antigen, immunoglobulin, T cell receptor

Site-Specific Recombination: Site-specific recombination takes place between specific DNA sequences that are recognized by proteins that mediate the process. In vertebrates, site-specific recombination plays a critical role in generating immunoglobulin and T cell receptor genes during development of the immune system.

transposition, transposable element, transposon

Transposition via DNA Intermediates: Most DNA transposons move throughout the genome with no requirement for specific DNA sequences at their sites of insertion. In yeasts and protozoans, however, the transposition of some DNA sequences to specific target sites results in programmed DNA rearrangements that regulate gene expression.

retrotransposon, retrovirus, reverse transcriptase, long terminal repeat (LTR), LINE, SINE, processed pseudogene

Transposition via RNA Intermediates: Most transposons in eukaryotic cells move by reverse transcription of RNA intermediates, similar to the replication of retroviruses. These retrotransposons include the highly repeated LINE and SINE sequences of mammalian genomes.

gene amplification

Gene Amplification: Gene amplification results from repeated replication of a chromosomal region. In some cases, gene amplification provides a mechanism for increasing gene expression during development. Gene amplification also frequently occurs in cancer cells, where it can result in the elevated expression of genes that contribute to uncontrolled cell proliferation.

Questions

1. Discuss the roles played by the different DNA polymerases in *E. coli*.

2. What is an Okazaki fragment? How do they become a continuous strand?

3. Compare the action of topoisomerases I and II.

4. How would you test a sequence of DNA in a yeast cell to see whether it contains an origin of replication or autonomously replicating sequence (ARS)?

5. Why does the human genome contain thousands of origins of replication but the *E. coli* genome contain only one?

6. How does DNA synthesis begin at an origin of replication?

7. How does telemerase extend the ends of chromosomes to compensate for the inability of DNA polymerase to complete DNA replication at chromosome ends?

8. Explain the process of nucleotide excision repair.

9. How can a cell repair a double strand break in its DNA?

10. How is breast cancer related to a DNA repair mechanism?

11. Patients with xeroderma pigmentosum suffer an extremely high incidence of skin cancer but have not been found to have correspondingly high incidences of cancers of internal organs (e.g., colon cancer). What might this suggest about the kinds of DNA damage responsible for most internal cancers?

12. What phenotype would you predict for a mutant mouse lacking one of the genes required for site-specific recombination in lymphocytes?

13. Many of the drugs in clinical use and under evaluation for the treatment of AIDS are inhibitors of the HIV reverse transcriptase. What reverse transcriptases in human cells might also be inhibited by these drugs? What would be the consequences of inhibiting these enzymes?

References and Further Reading

DNA Replication

Baker, T. A. and S. P. Bell. 1998. Polymerases and the replisome: Machines within machines. *Cell* 92: 296–305. [R]

Bell, S. P. 2002. The origin replication complex: from simple origins to complex functions. *Genes Dev.* 16: 659–672. [R]

Bell, S. P. and A. Dutta. 2002. DNA replication in eukaryotic cells. *Ann. Rev. Biochem.* 71: 333–374. [R]

Benkovic, S. J., A. M. Valentine and F. Salinas. 2001. Replisome-mediated DNA replication. *Ann. Rev. Biochem.* 70: 181–208. [R]

Blackburn, E. H. 1992. Telomerases. *Ann. Rev. Biochem.* 61: 113–129. [R]

Cairns, J. 1963. The chromosome of *Escherichia coli*. *Cold Spring Harbor Symp. Quant. Biol.* 28: 43–46. [P]

Champoux, J. J. 2001. DNA topoisomerases: structure, function, and mechanism. *Ann. Rev. Biochem.* 70: 369–413. [R]

Ellison, V. and B. Stillman. 2001. Opening of the clamp: an intimate view of an ATP-driven biological machine. *Cell* 106: 655–660. [R]

Frick, D. N. and C. C. Richardson. 2001. DNA primases. *Ann. Rev. Biochem.* 70: 39–80. [R]

Gilbert, D. M. 2001. Making sense of eukaryotic DNA replication origins. *Science* 294: 96–100. [R]

Goodman, M. F. 1997. Hydrogen bonding revisited: geometric selection as a principal determinant of DNA replication fidelity. *Proc. Natl. Acad. Sci. USA* 94: 10493–10495. [R]

Huberman, J. A. and A. D. Riggs. 1968. On the mechanism of DNA replication in mammalian chromosomes. *J. Mol. Biol.* 32: 327–341. [P]

Hubscher, U., G. Maga and S. Spadari. 2002. Eukaryotic DNA polymerases. *Ann. Rev. Biochem.* 71: 133–163. [R]

Kornberg, A., I. R. Lehman, M. J. Bessman and E. S. Simms. 1956. Enzymic synthesis of deoxyribonucleic acid. *Biochim. Biophys. Acta* 21: 197–198. [P]

Kunkel, T. A. and K. Bebenek. 2000. DNA replication fidelity. *Ann. Rev. Biochem.* 69: 497–529. [R]

Lohman, T. M. and K. P. Bjornson. 1996. Mechanisms of helicase-catalyzed DNA unwinding. *Ann. Rev. Biochem.* 65: 169–214. [R]

McEachern, M. J., A Krauskopf and E. H. Blackburn. 2000. Telomeres and their control. *Ann. Rev. Genet.* 34: 331–358. [R]

Ogawa, T. and T. Okazaki. 1980. Discontinuous DNA replication. *Ann. Rev. Biochem.* 49: 421–457. [R]

Stinthcomb, D. T., K. Struhl and R. W. Davis. 1979. Isolation and characterization of a yeast chromosomal replicator. *Nature* 282: 39–43. [P]

Waga, S. and B. Stillman. 1994. Anatomy of a DNA replication fork revealed by reconstitution of SV40 DNA replication *in vitro*. *Nature* 369: 207–212. [P]

Waga, S. and B. Stillman. 1998. The DNA replication fork in eukaryotic cells. *Ann. Rev. Biochem.* 67: 721–751. [R]

West, S. C. 1996. DNA helicases: New breeds of translocating motors and molecular pumps. *Cell* 86: 177–180. [R]

Wold, M. S. 1997. Replication protein A: A heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Ann. Rev. Biochem.* 66: 61–92. [R]

Zakian, V. A. 1995. Telomeres: Beginning to understand the end. *Science* 270: 1601–1607. [R]

DNA Repair

Batty, D. P. and R. D. Wood. 2000.Damage recognition in nucleotide excision repair of DNA. *Gene* 241: 193–204. [R]

Cleaver, J. E. 1968. Defective repair replication of DNA in xeroderma pigmentosum. *Nature* 218: 652–656. [P]

Cox, M. M. 2001. Recombinational DNA repair of damaged replication forks in *Escherichia coli*: questions. *Ann. Rev. Genet.* 35: 53–82. [R]

De Laat, W. L., N. G. J. Jaspers and J. H. J. Hoeijmakers. 1999. Molecular mechanism of nucleotide excision repair. *Genes Dev.* 13: 768–785. [R]

Fishel, R., M. K. Lescoe, M. R. S. Rao, N. G. Copeland, N. A. Jenkins, J. Garber, M. Kane and R. Kolodner. 1993. The human mutator gene homolog *MSH2* and its association with hereditary nonpolyposis colon cancer. *Cell* 75: 1027–1038. [P]

Friedberg, E. C., R. Wagner and M. Radman. 2002. Specialized DNA polymerases, cellular survival, and the genesis of mutations. *Science* 296: 1627–1630. [R]

Friedberg, E. C., G. C. Walker and W. Siede. 1995. *DNA Repair and Mutagenesis*. Washington, D.C.: ASM Press.

Goodman, M. F. 2002. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Ann. Rev. Biochem.* 71: 17–50. [R]

Harfe, B. D. and S. Jinks-Robertson. 2000. DNA mismatch repair and genetic instability. *Ann. Rev. Genet.* 34: 359–399. [R]

Hoeijmakers, J. H. J. 2001. Genome maintenance mechanisms for preventing cancer. *Nature* 411: 366–374. [R]

Khanna, K. K. and S. P. Jackson. 2001. DNA double strand breaks: signaling, repair and the cancer connection. *Nature Genetics* 27: 247–254. [R]

Leach, F. S. and 34 others. 1993. Mutations of a *mutS* homolog in hereditary nonpolyposis colorectal cancer. *Cell* 75: 1215–1225. [P]

Livneh, Z. 2001. DNA damage control by novel DNA polymerases: translesion replication and mutagenesis. *J. Biol. Chem.* 276: 25639–25642. [R]

Modrich, P. 1997. Strand-specific mismatch repair in mammalian cells. *J. Biol. Chem.* 272: 24727–24730. [R]

Sancar, A. 1996. DNA excision repair. *Ann. Rev. Biochem.* 65: 43–81. [R]

Seeberg, E., L. Eide and M. Bjoras. 1995. The base excision repair pathway. *Trends Biochem. Sci.* 20: 391–397. [R]

Svejstrup, J. Q. 2002. Mechanisms of transcription-coupled DNA repair. *Nature Rev. Mol. Cell. Biol.* 3: 21–29. [R]

Wood, R. D. 1997. Nucleotide excision repair in mammalian cells. *J. Biol. Chem.* 272: 23465–23468. [R]

Recombination between Homologous DNA Sequences

Baumann, P. and S. C. West. 1998. Role of the human RAD51 protein in homologous recombination and double-stranded-break repair. *Trends Biochem. Sci.* 23: 247–251. [R]

DasGupta, C., A. M. Wu, R. Kahn, R. P. Cunningham and C. M. Radding. 1981. Concerted strand exchange and formation of Holliday structures by *E. coli* RecA protein. *Cell* 25: 507–516. [P]

Haber, J. E. and W.-D. Heyer. 2001. The fuss about Mus81. *Cell* 107: 551–554. [R]

Holliday, R. 1964. A mechanism for gene conversion in fungi. *Genet. Res.* 5: 282–304. [P]

Kowalczykowski, S. C. and A. K. Eggleston. 1994. Homologous pairing and DNA strand-exchange proteins. *Ann. Rev. Biochem.* 63: 991–1043. [R]

Meselson, M. and J. J. Weigle. 1961. Chromosome breakage accompanying genetic recombination in bacteriophage. *Proc. Natl. Acad. Sci. USA* 47: 857–868. [P]

Potter, H. and D. Dressler. 1976. On the mechanism of genetic recombination: Electron microscopic observation of recombination intermediates. *Proc. Natl. Acad. Sci. USA* 73: 3000–3004. [P]

Radding, C. M. 1991. Helical interactions in homologous pairing and strand exchange driven by RecA protein. *J. Biol. Chem.* 266: 5355–5358. [R]

Shinohara, A. and T. Ogawa. 1995. Homologous recombination and the roles of double strand breaks. *Trends Biochem. Sci.* 20: 387–391. [R]

Smith, G. R. 2001. Homologous recombination near and far from DNA breaks: alternative roles and contrasting views. *Ann. Rev. Genet.* 35: 243–274. [R]

Stahl, F. 1996. Meiotic recombination in yeast: Coronation of the double-strand-break repair model. *Cell* 87: 965–968. [R]

Szostak, J. W., T. L. Orr-Weaver, R. J. Rothstein and F. W. Stahl. 1983. The double-strand-break repair model for recombination. *Cell* 33: 25–35. [P]

Taylor, A. F. 1992. Movement and resolution of Holliday junctions by enzymes from *E. coli*. *Cell* 69: 1063–1065. [R]

West, S. C. 1992. Enzymes and molecular mechanisms of genetic recombination. *Ann. Rev. Biochem.* 61: 603–640. [R]

West, S. C. 1997. Processing of recombination intermediates by the RuvABC proteins. *Ann. Rev. Genet.* 31: 213–244. [R]

DNA Rearrangements

Bassing, C. H., W. Swat and F. W. Alt. 2002. The mechanism and regulation of chromosomal V(D)J recombination. *Cell* 109: S45–S55. [R]

Boeke, J. D., D. J. Garfinkel, C. A. Styles and G. R. Fink. 1985. Ty elements transpose through an RNA intermediate. *Cell* 40: 491–500. [P]

Craig, N. L. 1995. Unity in transposition reactions. *Science* 270: 253–254. [R]

Craig, N. L. 1997. Target site selection in transposition. *Ann. Rev. Biochem.* 66: 437–474. [R]

Davis, M. M. 1990. T cell receptor gene diversity and selection. *Ann. Rev. Biochem.* 59: 475–496. [R]

Fedoroff, N. V. 2000. Transposons and genome evolution in plants. *Proc. Natl. Acad. Sci. USA* 97: 7002–7007. [R]

Fedoroff, N. and D. Botstein. 1992. *The Dynamic Genome: Barbara McClintock's Ideas in the Century of Genetics.* Plainview, N.Y.: Cold Spring Harbor Laboratory Press.

Finnegan, D. J. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet.* 5: 103–107. [R]

Gilboa, E., S. W. Mitra, S. Goff and D. Baltimore. 1979. A detailed model of reverse transcription and tests of crucial aspects. *Cell* 18: 93–100. [P]

Haren, L., B. Ton-Hoang and M. Chandler. 1999. Integrating DNA: transposases and retroviral integrases. *Ann. Rev. Microbiol.* 53: 245–281. [R]

Hozumi, N. and S. Tonegawa. 1976. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc. Natl. Acad. Sci. USA* 73: 3628–3632. [P]

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921. [P]

Kidwell, M. G. and D. Lisch. 1997. Transposable elements as sources of variation in animals and plants. *Proc. Natl. Acad. Sci. USA* 94: 7704–7711. [R]

Landy, A. 1989. Dynamic, structural, and regulatory aspects of λ site-specific recombination. *Ann. Rev. Biochem.* 58: 913–949. [R]

Lewis, S. and M. Gellert. 1989. The mechanism of antigen receptor gene assembly. *Cell* 59: 585–588. [R]

McClintock, B. 1956. Controlling elements and the gene. *Cold Spring Harbor Symp. Quant. Biol.* 21: 197–216. [P]

Moran, J. V., R. J. DeBerardinis and H. H. Kazazian Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* 283: 1530–1534. [P]

Ostertag, E. M. and H. H. Kazazian Jr. 2001. Biology of mammalian L1 retrotransposons. *Ann. Rev. Genet.* 35: 501–538. [R]

Stark, G. R., M. Debatisse, E. Giulotto and G. M. Wahl. 1989. Recent progress in understanding mechanisms of mammalian DNA amplification. *Cell* 57: 901–908. [R]

Stark, G. R. and G. M. Wahl. 1984. Gene amplification. *Ann. Rev. Biochem.* 53: 447–491. [R]

Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302: 575–581. [R]

Venter, J. C. and 273 others. 2001. The sequence of the human genome. *Science* 291: 1304–1351. [P]

Weiner, A. M., P. L. Deininger and A. Efstratiadis. 1986. Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Ann. Rev. Biochem.* 55: 631–661. [R]

Chapter 6 RNA Synthesis and Processing

CHAPTERS 4 AND 5 DISCUSSED THE ORGANIZATION and maintenance of genomic DNA, which can be viewed as the set of genetic instructions governing all cellular activities. These instructions are implemented via the synthesis of RNAs and proteins. Importantly, the behavior of a cell is determined not only by what genes it inherits, but also by which of those genes are expressed at any given time. Regulation of gene expression allows cells to adapt to changes in their environments and is responsible for the distinct activities of the multiple differentiated cell types that make up complex plants and animals. Muscle cells and liver cells, for example, contain the same genes; the functions of these cells are determined not by differences in their genomes, but by regulated patterns of gene expression that govern development and differentiation.

The first step in expression of a gene, the transcription of DNA into RNA, is the primary level at which gene expression is regulated in both prokaryotic and eukaryotic cells. RNAs in eukaryotic cells are then modified in various ways—for example, introns are removed by splicing—to convert the primary transcript into its functional form. Different types of RNA play distinct roles in cells: Messenger RNAs (mRNAs) serve as templates for protein synthesis; ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) function in mRNA translation. Still other small RNAs function in gene regulation, mRNA splicing, rRNA processing, and protein sorting in eukaryotes. Transcription and RNA processing are discussed in this chapter. The final step in gene expression, the translation of mRNA to protein, is the subject of Chapter 7.

Transcription in Prokaryotes

As in most areas of molecular biology, studies of *E. coli* have provided the model for subsequent investigations of transcription in eukaryotic cells. As reviewed in Chapter 3, mRNA was discovered first in *E. coli*. *E. coli* was also the first organism from which RNA polymerase was purified and studied. The basic mechanisms by which transcription is regulated were likewise elucidated by pioneering experiments in *E. coli*, in which regulated gene expression allows the cell to respond to variations in the environment, such as changes in the availability of nutrients. An understanding of transcription in *E. coli* has thus provided the

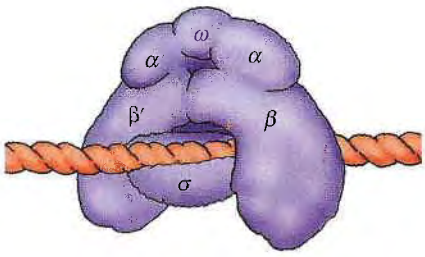


Figure 6.1 *E. coli* RNA polymerase
The complete enzyme consists of six subunits: two α, one β, one β', one ω, and one σ. The σ subunit is relatively weakly bound and can be dissociated from the other five subunits, which constitute the core polymerase.

foundation for studies of the far more complex mechanisms that regulate gene expression in eukaryotic cells.

RNA Polymerase and Transcription

The principal enzyme responsible for RNA synthesis is **RNA polymerase**, which catalyzes the polymerization of ribonucleoside 5′-triphosphates (NTPs) as directed by a DNA template. The synthesis of RNA is similar to that of DNA, and like DNA polymerase, RNA polymerase catalyzes the growth of RNA chains always in the 5′ to 3′ direction. Unlike DNA polymerase, however, RNA polymerase does not require a preformed primer to initiate the synthesis of RNA. Instead, transcription initiates *de novo* at specific sites at the beginning of genes. The initiation process is particularly important because this is a major step at which transcription is regulated.

RNA polymerase, like DNA polymerase, is a complex enzyme made up of multiple polypeptide chains. The intact bacterial enzyme consists of five different types of subunits, called α, β, β', ω, and σ (Figure 6.1). The σ subunit is relatively weakly bound and can be separated from the other subunits, yielding a core polymerase consisting of two α, one β, one β' and one ω subunits. The core polymerase is fully capable of catalyzing the polymerization of NTPs into RNA, indicating that σ is not required for the basic catalytic activity of the enzyme. However, the core polymerase does not bind specifically to the DNA sequences that signal the normal initiation of transcription; therefore, the σ subunit is required to identify the correct sites for transcription initiation. The selection of these sites is a critical element of transcription because synthesis of a functional RNA must start at the beginning of a gene.

The DNA sequence to which RNA polymerase binds to initiate transcription of a gene is called the **promoter**. The DNA sequences involved in promoter function were first identified by comparisons of the nucleotide sequences of a series of different genes isolated from *E. coli*. These comparisons revealed that the region upstream of the transcription initiation site contains two sets of sequences that are similar in a variety of genes. These common sequences encompass six nucleotides each, and are located approximately 10 and 35 base pairs upstream of the transcription start site (Figure 6.2). They are called the –10 and –35 elements, denoting their position relative to the transcription initiation site, which is defined as the +1 position. The sequences at the –10 and –35 positions in different promoters are not identical, but they are all similar enough to establish consensus sequences—the bases most frequently found at each position.

Several types of experimental evidence support the functional importance of the –10 and –35 promoter elements. First, genes with promoters that differ from the consensus sequences are transcribed less efficiently than genes whose promoters match the consensus sequences more closely. Second, mutations introduced in either the –35 or –10 consensus sequences have strong effects on promoter function. Third, the sites at which RNA

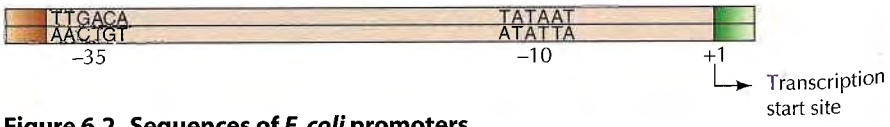


Figure 6.2 Sequences of *E. coli* promoters
E. coli promoters are characterized by two sets of sequences located 10 and 35 base pairs upstream of the transcription start site (+1). The consensus sequences shown correspond to the bases most frequently found in different promoters.

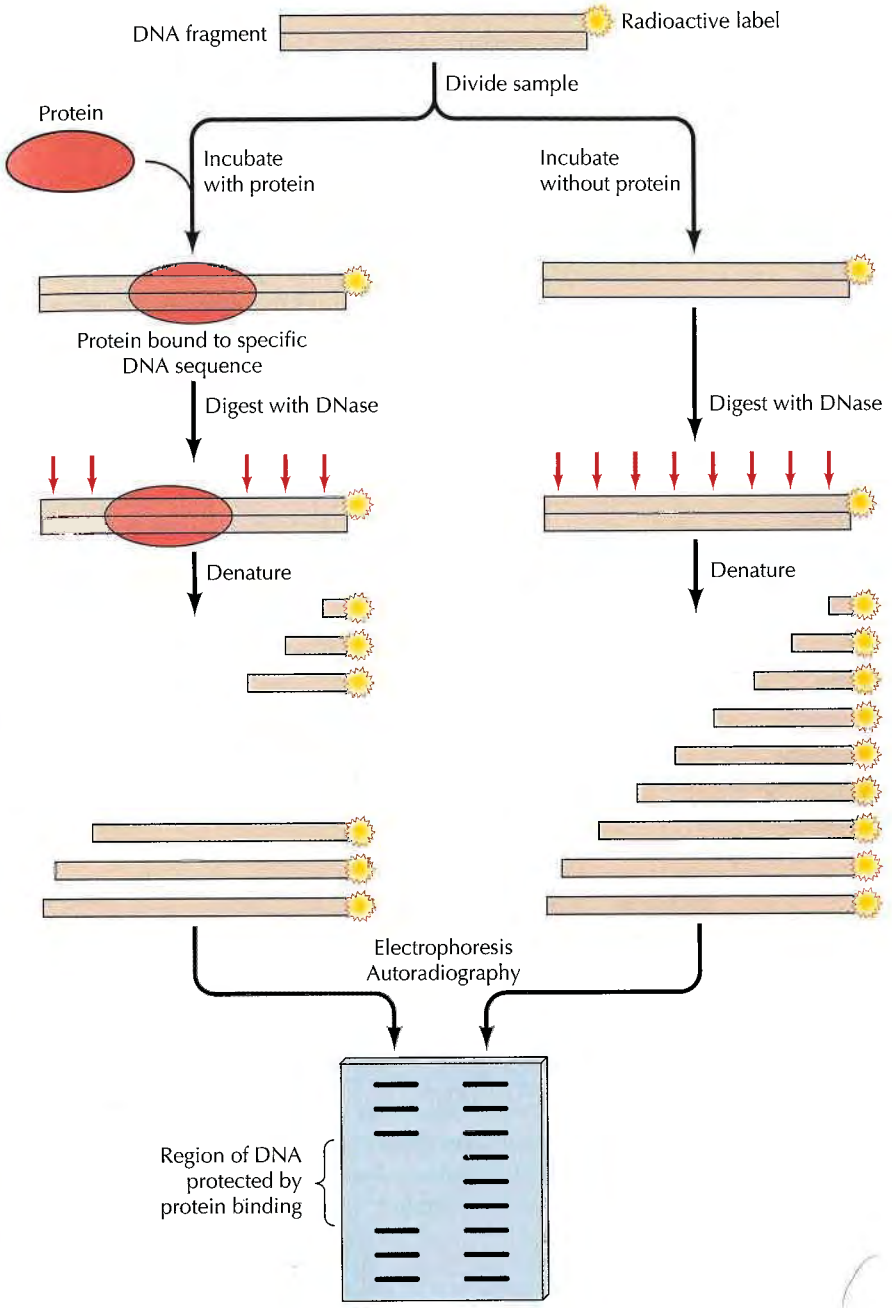


Figure 6.3 DNA footprinting
A sample containing fragments of DNA radiolabeled at one end is divided into two, and one half of the sample is incubated with a protein that binds to a specific DNA sequence within the fragment. Both samples are then digested with DNase, under conditions such that the DNase introduces an average of one cut per molecule. The region of DNA bound to the protein is protected from DNase digestion. The DNA-protein complexes are then denatured, and the sizes of the radiolabeled DNA fragments produced by DNase digestion are analyzed by electrophoresis (as for DNA sequencing). Fragments of DNA resulting from DNase cleavage within the region protected by protein binding are missing from the sample of DNA that was incubated with protein.

polymerase binds to promoters have been directly identified by **footprinting** experiments, which are widely used to determine the sites at which proteins bind to DNA (Figure 6.3). In experiments of this type, a DNA fragment is radiolabeled at one end. The labeled DNA is incubated with the protein of interest (e.g., RNA polymerase) and then subjected to partial digestion with DNase. The principle of the method is that the regions of DNA to which the protein binds are protected from DNase digestion. These regions can therefore be identified by comparison of the digestion products of the protein-bound DNA with those resulting from identical DNase treatment of a parallel sample of DNA that was not incubated with protein. Variations of this basic method, which employ chemical reagents to modify and cleave DNA at particular nucleotides, can be used to identify the specific DNA

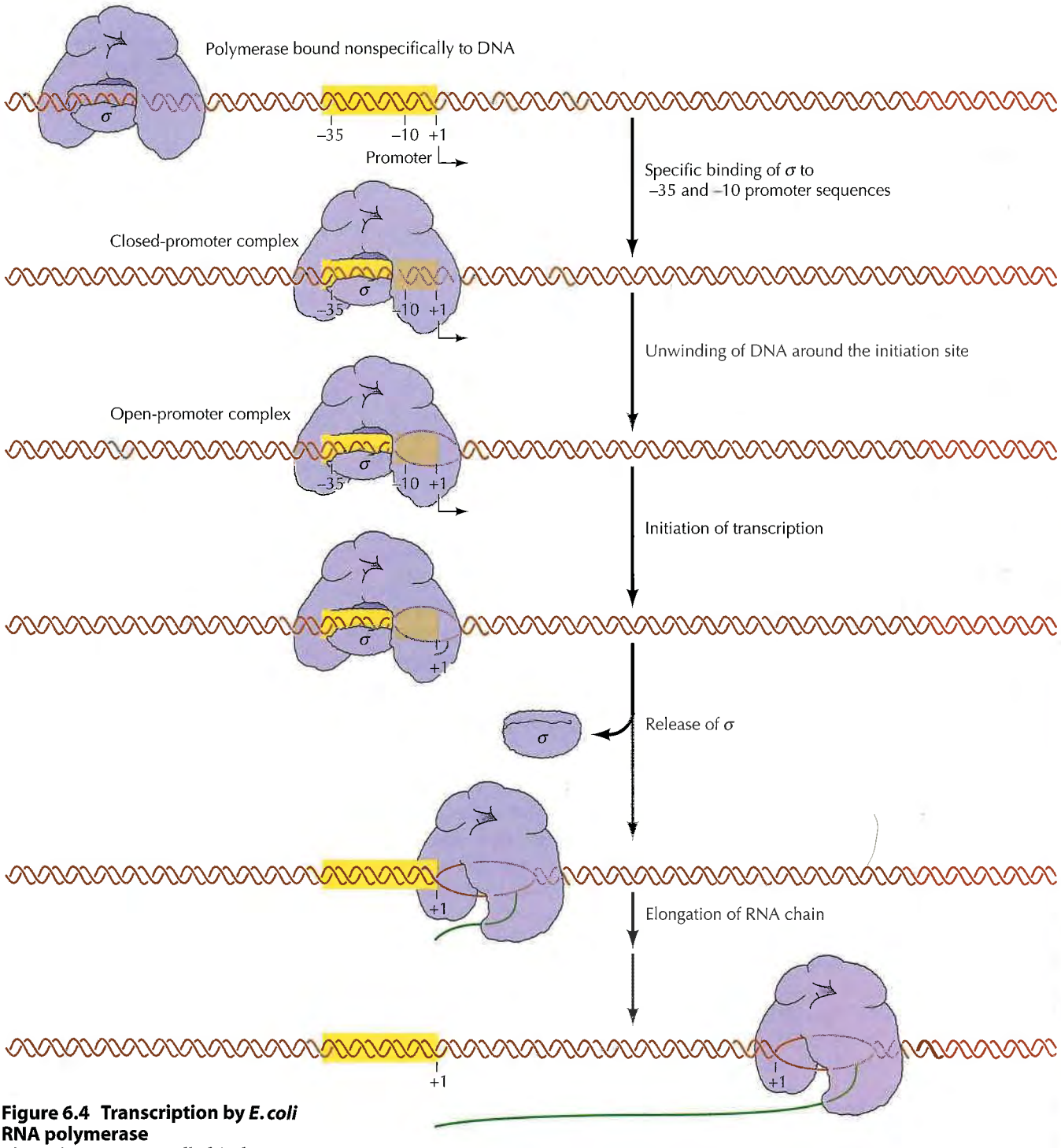


Figure 6.4 Transcription by *E. coli* RNA polymerase
The polymerase initially binds nonspecifically to DNA and migrates along the molecule until the σ subunit binds to the –35 and –10 promoter elements, forming a closed-promoter complex. The polymerase then unwinds DNA around the initiation site, and transcription is initiated by the polymerization of free NTPs. The σ subunit then dissociates from the core polymerase, which migrates along the DNA and elongates the growing RNA chain.

bases that are in contact with protein. Such footprinting analysis has shown that RNA polymerase generally binds to promoters over approximately a 60-base-pair region, extending from –40 to +20 (i.e., from 40 nucleotides upstream to 20 nucleotides downstream of the transcription start site). The σ subunit binds specifically to sequences in both the –35 and –10 promoter regions, substantiating the importance of these sequences in promoter function. In addition, some *E. coli* promoters have a third sequence, located upstream of the –35 region, that serves as a specific binding site for the RNA polymerase α subunit.

In the absence of σ, RNA polymerase binds nonspecifically to DNA with low affinity. The role of σ is to direct the polymerase to promoters by binding specifically to both the –35 and –10 sequences, leading to the initiation of transcription at the beginning of a gene (Figure 6.4). The initial binding between the polymerase and a promoter is referred to as a closed-promoter complex because the DNA is not unwound. The polymerase then unwinds 14 bases of DNA, from –12 to +2, to form an open-promoter complex in which single-stranded DNA is available as a template for transcription. Transcription is initiated by the joining of two free NTPs. After addition of about the first 10 nucleotides, σ is released from the polymerase, which then leaves the promoter and moves along the template DNA to continue elongation of the growing RNA chain.

During elongation, the polymerase remains associated with its template while it continues synthesis of mRNAs. As it travels, the polymerase unwinds the template DNA ahead of it and rewinds the DNA behind it, maintaining an unwound region of about 15 base pairs in the region of transcription. High resolution structural analysis of bacterial RNA polymerase indicates that the β and β' subunits form a crab claw-like structure that grips the DNA template (Figure 6.5). An internal channel between the β and β' subunits accommodates approximately 20 base pairs of DNA and contains the polymerase active site.

RNA synthesis continues until the polymerase encounters a termination signal, at which point transcription stops, the RNA is released from the polymerase, and the enzyme dissociates from its DNA template. The simplest and most common type of termination signal in *E. coli* consists of a symmetrical inverted repeat of a GC-rich sequence followed by four or more A residues (Figure 6.6). Transcription of the GC-rich inverted repeat results in the formation of a segment of RNA that can form a stable stem-loop structure by complementary base pairing. The formation of such a self-complementary structure in the RNA disrupts its association with the DNA template and terminates transcription. Because hydrogen bonding between A and U is weaker than that between G and C, the presence of A residues downstream of the inverted repeat sequences is thought to facilitate the dissociation of the RNA from its template. Other types of transcription termination signals, in both prokaryotic and eukaryotic cells, depend on the binding of proteins that terminate transcription to specific DNA sequences, rather than on the formation of a stem-loop structure in the RNA.

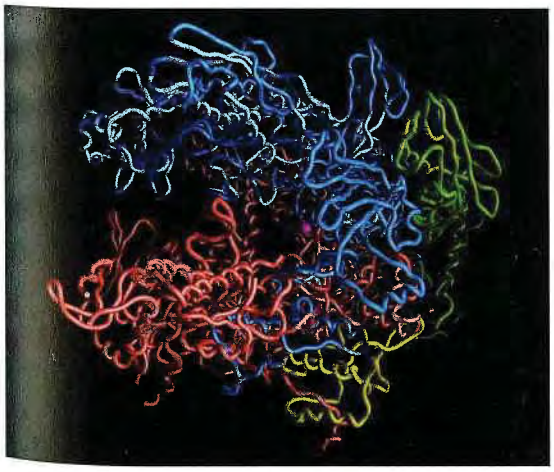


Figure 6.5 Structure of bacterial RNA polymerase
The α subunits of the polymerase are colored dark green and light green, β blue, β' pink, and ω yellow. (Courtesy of Seth Darst, Rockefeller University.)

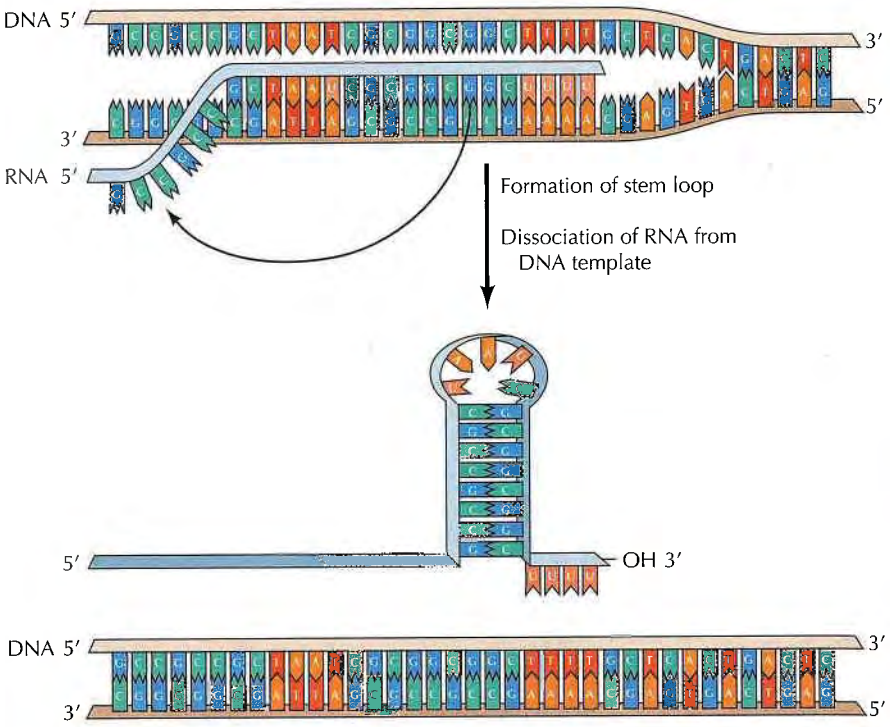


Figure 6.6 Transcription termination The termination of transcription is signaled by a GC-rich inverted repeat followed by four A residues. The inverted repeat forms a stable stem-loop structure in the RNA, causing the RNA to dissociate from the DNA template.

Repressors and Negative Control of Transcription

The pioneering studies of gene regulation in *E. coli* were carried out by François Jacob and Jacques Monod in the 1950s. These investigators and their colleagues analyzed the expression of enzymes involved in the metabolism of lactose, which can be used as a source of carbon and energy via cleavage to glucose and galactose (Figure 6.7). The enzyme that catalyzes the cleavage of lactose (β-galactosidase) and other enzymes involved in lactose metabolism are expressed only when lactose is available for use by the bacteria. Otherwise, the cell is able to economize by not investing energy in the synthesis of unnecessary RNAs and proteins. Thus, lactose induces the synthesis of enzymes involved in its own metabolism. In addition to requiring β-galactosidase, lactose metabolism involves the products of two other closely linked genes: lactose permease, which transports lactose into the cell, and a transacetylase, which is thought to inactivate toxic thiogalactosides that are transported into the cell along with lactose by the permease. On the basis of purely genetic experiments, Jacob and Monod deduced the mechanism by which the expression of these genes was regulated, thereby formulating a model that remains fundamental to our understanding of transcriptional regulation.

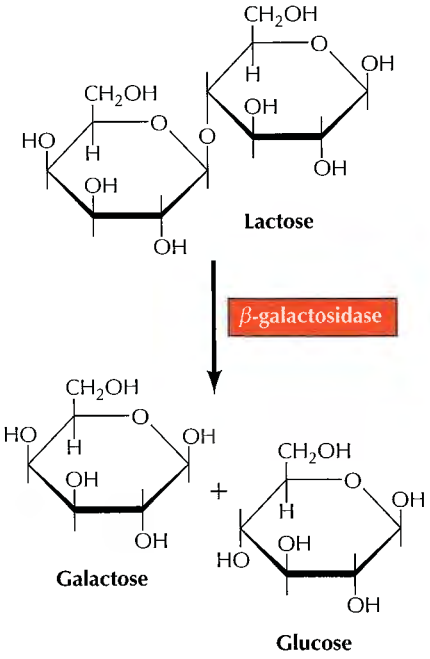


Figure 6.7 Metabolism of lactose β-galactosidase catalyzes the hydrolysis of lactose to glucose and galactose.

The starting point in this analysis was the isolation of mutants that were defective in regulation of the genes involved in lactose utilization. These mutants were of two types: constitutive mutants, which expressed all three genes even when lactose was not available, and noninducible mutants, which failed to express the genes even in the presence of lactose. Genetic mapping localized these regulatory mutants to two distinct loci, called *o* and *i*, with *o* located immediately upstream of the structural gene for β-galactosidase. Mutations affecting *o* resulted in constitutive expression; mutants of *i* were either constitutive or noninducible.

The function of these regulatory genes was probed by experiments in which two strains of bacteria were mated, resulting in diploid cells contain-

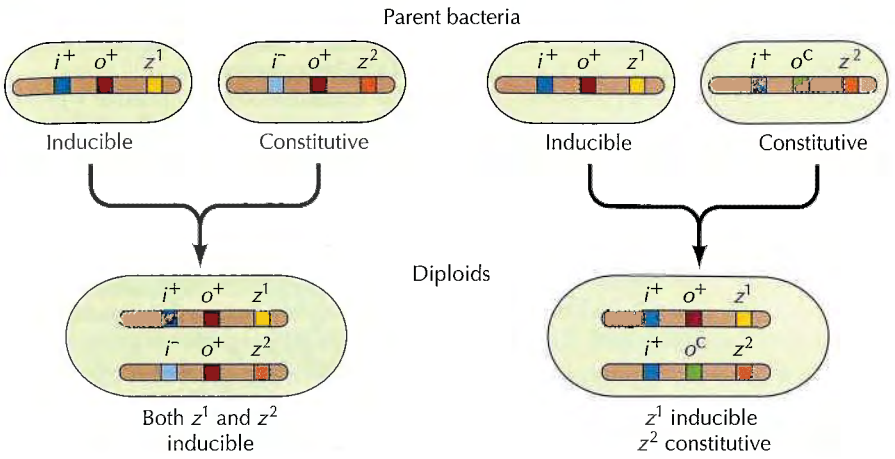


Figure 6.8 Regulation of β-galactosidase in diploid *E. coli*
The mating of two bacterial strains results in diploid cells that contain genes from both parents. In these examples, it is assumed that the genes encoding β-galactosidase (the *z* genes) can be distinguished on the basis of structural gene mutations, designated z^1 and z^2. In an i^+/i^- diploid (left), both structural genes are inducible; therefore, i^+ is dominant over i^- and affects expression of *z* genes on both chromosomes. In contrast, in an o^c/o^+ diploid (right), the *z* gene linked to o^c is constitutively expressed, whereas that linked to o^+ is inducible. Therefore, *o* affects expression of only the adjacent *z* gene on the same chromosome.

ing genes derived from both parents (Figure 6.8). Analysis of gene expression in such diploid bacteria provided critical insights by defining which alleles of these regulatory genes are dominant and which recessive. For example, when bacteria containing a normal *i* gene (i^+) were mated with bacteria carrying an *i* gene mutation resulting in constitutive expression (an i^- mutation), the resulting diploid bacteria displayed normal inducibility; therefore, the normal i^+ gene was dominant over the i^- mutant. In contrast, matings between normal bacteria and bacteria with an o^c mutation (constitutive expression) yielded diploids with the constitutive expression phenotype, indicating that o^c is dominant over o^+. Additional experiments in which mutations in *o* and *i* were combined with different mutations in the structural genes showed that *o* affects the expression of only the genes to which it is physically linked, whereas *i* affects the expression of genes on both chromosome copies in diploid bacteria. Thus, in an o^c/o^+ cell, only the structural genes that are linked to o^c are constitutively expressed. In contrast, in an i^+/i^- cell, structural genes on both chromosomes are regulated normally. These results led to the conclusion that *o* represents a region of DNA that controls the transcription of adjacent genes, whereas the *i* gene encodes a regulatory factor (e.g., a protein) that can diffuse throughout the cell and control genes on both chromosomes.

The model of gene regulation developed on the basis of these experiments is illustrated in Figure 6.9. The genes encoding β-galactosidase, permease, and transacetylase are expressed as a single unit, called an **operon**. Transcription of the operon is controlled by *o* (the **operator**), which is adjacent to the transcription initiation site. The *i* gene encodes a protein that regulates transcription by binding to the operator. Since i^- mutants (which result in constitutive gene expression) are recessive, it was concluded that these mutants failed to make a functional gene product. This result implies that the normal *i* gene product is a **repressor**, which blocks transcription when bound to *o*. The addition of lactose leads to induction of the operon because lactose binds to the repressor, thereby preventing it from binding to the operator DNA. In noninducible *i* mutants (which are dominant over i^+), the repressor fails to bind lactose, so expression of the operon cannot be induced.

The model neatly fits the results of the genetic experiments from which it was derived. In i^- cells, the repressor is not made, so the *lac* operon is constitutively expressed. Diploid i^+/i^- cells are normally inducible, since the functional repressor is encoded by the i^+ allele. Finally, in o^c mutants a functional operator has been lost and the repressor cannot be bound. Conse-

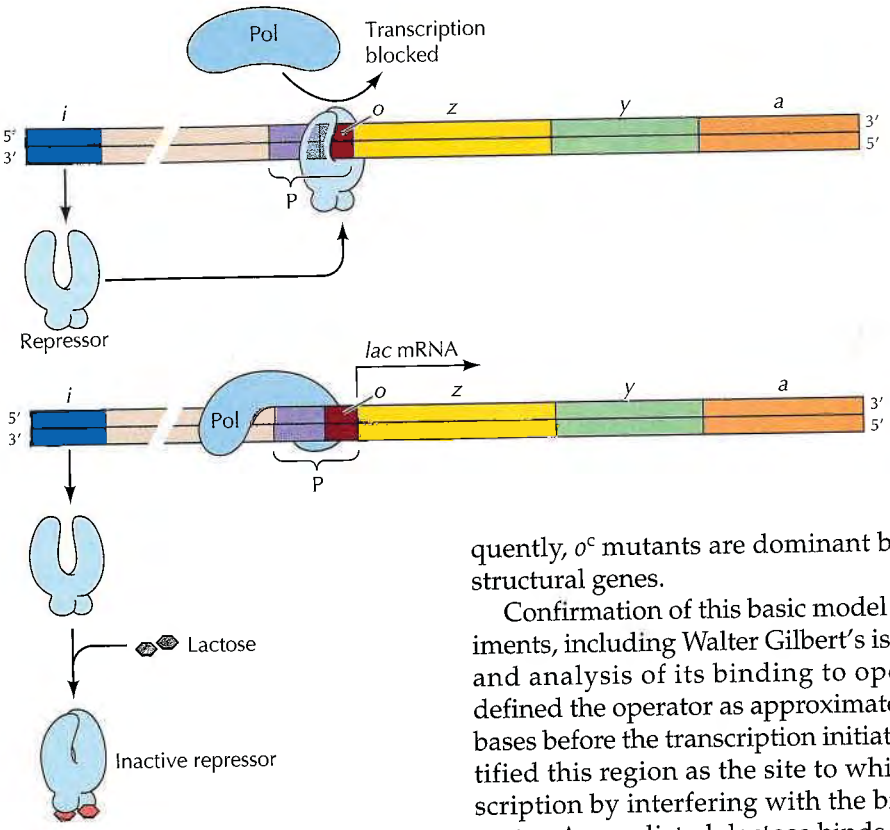


Figure 6.9 Negative control of the *lac* operon
The *i* gene encodes a repressor which, in the absence of lactose (top), binds to the operator (*o*) and interferes with the binding of RNA polymerase to the promoter, blocking transcription of the three structural genes (*z*, β-galactosidase; *y*, permease; and *a*, transacetylase). Lactose induces expression of the operon by binding to the repressor (bottom), which prevents the repressor from binding to the operator. P = promoter; Pol = polymerase.

quently, o^c mutants are dominant but affect the expression only of linked structural genes.

Confirmation of this basic model has since come from a variety of experiments, including Walter Gilbert's isolation, in the 1960s, of the *lac* repressor and analysis of its binding to operator DNA. Molecular analysis has defined the operator as approximately 20 base pairs of DNA, starting a few bases before the transcription initiation site. Footprinting analysis has identified this region as the site to which the repressor binds, blocking transcription by interfering with the binding of RNA polymerase to the promoter. As predicted, lactose binds to the repressor, which then no longer binds to operator DNA. Also as predicted, o^c mutations alter sequences within the operator, thereby preventing repressor binding and resulting in constitutive gene expression.

The central principle of gene regulation exemplified by the lactose operon is that control of transcription is mediated by the interaction of regulatory proteins with specific DNA sequences. This general mode of regulation is broadly applicable to both prokaryotic and eukaryotic cells. Regulatory sequences like the operator are called ***cis*-acting control elements**, because they affect the expression of only linked genes on the same DNA molecule. On the other hand, proteins like the repressor are called ***trans*-acting factors** because they can affect the expression of genes located on other chromosomes within the cell. The *lac* operon is an example of negative control because binding of the repressor blocks transcription. This, however, is not always the case; many *trans*-acting factors are activators rather than inhibitors of transcription.

Positive Control of Transcription

The best-studied example of positive control in *E. coli* is the effect of glucose on the expression of genes that encode enzymes involved in the breakdown (catabolism) of other sugars (including lactose) that provide alternative sources of carbon and energy. Glucose is preferentially utilized, so as long as glucose is available, enzymes involved in catabolism of alternative energy sources are not expressed. For example, if *E. coli* are grown in medium containing both glucose and lactose, the *lac* operon is not induced and only glucose is used by the bacteria. Thus, glucose represses the *lac* operon even in the presence of the normal inducer (lactose).

Glucose repression (generally called catabolite repression) is now known to be mediated by a positive control system, which is coupled to lev-

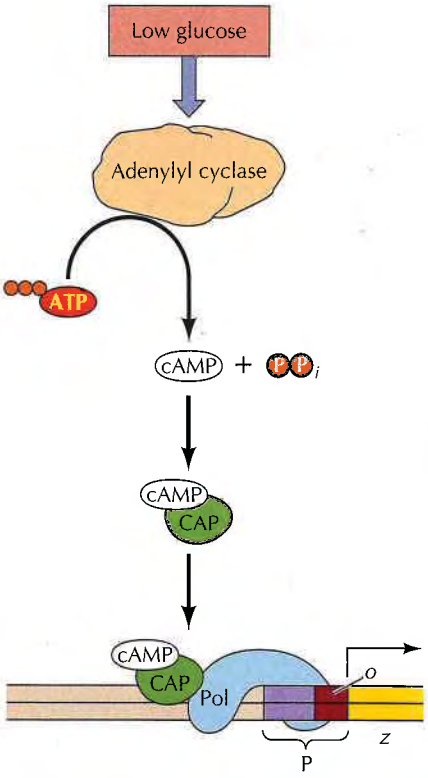


Figure 6.10 Positive control of the *lac* operon by glucose
Low levels of glucose activate adenylyl cyclase, which converts ATP to cyclic AMP (cAMP). Cyclic AMP then binds to the catabolite activator protein (CAP) and stimulates its binding to regulatory sequences of various operons concerned with the metabolism of alternative sugars, such as lactose. CAP interacts with the α subunit of RNA polymerase to facilitate the binding of polymerase to the promoter.

els of cyclic AMP (cAMP) (Figure 6.10). In bacteria, the enzyme adenylyl cyclase, which converts ATP to cAMP, is regulated such that levels of cAMP increase when glucose levels drop. cAMP then binds to a transcriptional regulatory protein called catabolite activator protein (CAP). The binding of cAMP stimulates the binding of CAP to its target DNA sequences, which in the *lac* operon are located approximately 60 bases upstream of the transcription start site. CAP then interacts with the α subunit of RNA polymerase, facilitating the binding of polymerase to the promoter and activating transcription.

Eukaryotic RNA Polymerases and General Transcription Factors

Although transcription proceeds by the same fundamental mechanisms in all cells, it is considerably more complex in eukaryotic cells than in bacteria. This is reflected in two distinct differences between the prokaryotic and eukaryotic systems. First, whereas all genes are transcribed by a single RNA polymerase in bacteria, eukaryotic cells contain multiple different RNA polymerases that transcribe distinct classes of genes. Second, rather than binding directly to promoter sequences, eukaryotic RNA polymerases need to interact with a variety of additional proteins to specifically initiate transcription. This increased complexity of eukaryotic transcription presumably facilitates the sophisticated regulation of gene expression needed to direct the activities of the many different cell types of multicellular organisms.

Eukaryotic RNA Polymerases

Eukaryotic cells contain three distinct nuclear RNA polymerases that transcribe different classes of genes (Table 6.1). Protein-coding genes are transcribed by RNA polymerase II to yield mRNAs; ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are transcribed by RNA polymerases I and III. RNA polymerase I is specifically devoted to transcription of the three largest species of rRNAs, which are designated 28S, 18S, and 5.8S according to their rates of sedimentation during velocity centrifugation. RNA polymerase III transcribes the genes for tRNAs and for the smallest species of ribosomal RNA (5S rRNA). Some of the small RNAs involved in splicing and protein transport (snRNAs and scRNAs) are also transcribed by RNA polymerase III, while others are polymerase II transcripts. In addition, separate RNA polymerases (which are similar to bacterial RNA polymerases) are found in chloroplasts and mitochondria, where they specifically transcribe the DNAs of those organelles.

All three of the nuclear RNA polymerases are complex enzymes, consisting of 12 to 17 different subunits each. Although they recognize different promoters and transcribe distinct classes of genes, they share several features in common with each other as well as with bacterial RNA polymerase. In particular, all three eukaryotic RNA polymerases contain nine conserved subunits, five of which are related to the α, β, β', and ω subunits of bacterial

TABLE 6.1 Classes of Genes Transcribed by Eukaryotic RNA Polymerases

Type of RNA synthesized	RNA polymerase
Nuclear genes	
mRNA	II
tRNA	III
rRNA	
5.8S, 18S, 28S	I
5S	III
snRNA and scRNA	II and III[a]
Mitochondrial genes	Mitochondrial[b]
Chloroplast genes	Chloroplast[b]

[a] Some small nuclear (sn) and small cytoplasmic (sc) RNAs are transcribed by polymerase II and others by polymerase III.
[b] The mitochondrial and chloroplast RNA polymerases are similar to bacterial enzymes.

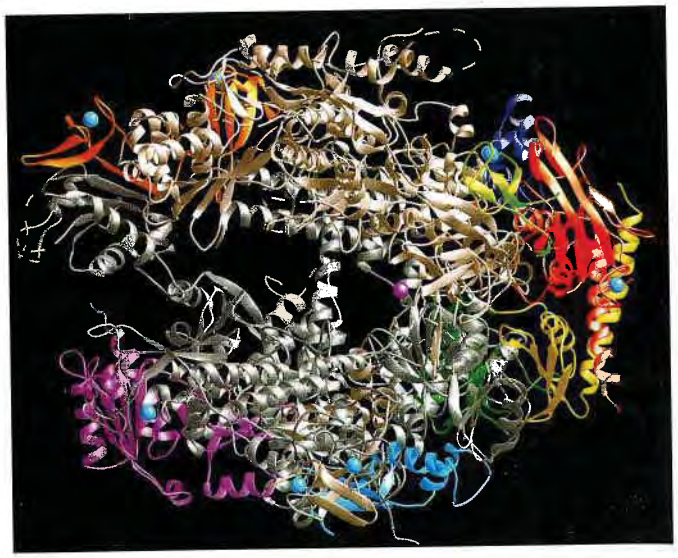


Figure 6.11 Structure of yeast RNA polymerase II Individual subunits are distinguished by colors. (From P. D. Kramer et al., 2001. *Science* 292: 1863.)

RNA polymerase. The recent determination of the structure of yeast RNA polymerase II by X-ray crystallography has further revealed that the architecture of this eukaryotic RNA polymerase is strikingly similar to that of the bacterial enzyme (Figure 6.11), suggesting that all RNA polymerases utilize fundamentally conserved mechanisms to transcribe DNA.

General Transcription Factors and Initiation of Transcription by RNA Polymerase II

Because RNA polymerase II is responsible for the synthesis of mRNA from protein-coding genes, it has been the focus of most studies of transcription in eukaryotes. Early attempts at studying this enzyme indicated that its activity is different from that of prokaryotic RNA polymerase. The accurate transcription of bacterial genes that can be accomplished *in vitro* simply by the addition of purified RNA polymerase to DNA containing a promoter is not possible in eukaryotic systems. The basis of this difference was elucidated in 1979, when Robert Roeder and his colleagues discovered that RNA polymerase II is able to initiate transcription only if additional proteins are added to the reaction. Thus, transcription in the eukaryotic system appeared to require distinct initiation factors that (in contrast to bacterial σ factors) were not associated with the polymerase.

Biochemical fractionation of nuclear extracts subsequently led to the identification of specific proteins (called **transcription factors**) that are required for RNA polymerase II to initiate transcription. Indeed, the identification and characterization of these factors represents a major part of ongoing efforts to understand transcription in eukaryotic cells. Two general types of transcription factors have been defined. **General transcription factors** are involved in transcription from all polymerase II promoters and therefore constitute part of the basic transcription machinery. Additional gene-specific transcription factors (discussed later in the chapter) bind to DNA sequences that control the expression of individual genes and are thus responsible for regulating gene expression. It is estimated that about 5% of the genes in the human genome encode transcription factors, emphasizing the importance of these proteins.

Five general transcription factors are required for initiation of transcription by RNA polymerase II in reconstituted *in vitro* systems (Figure 6.12). The promoters of many genes transcribed by polymerase II contain a sequence similar to TATAA 25 to 30 nucleotides upstream of the transcrip-

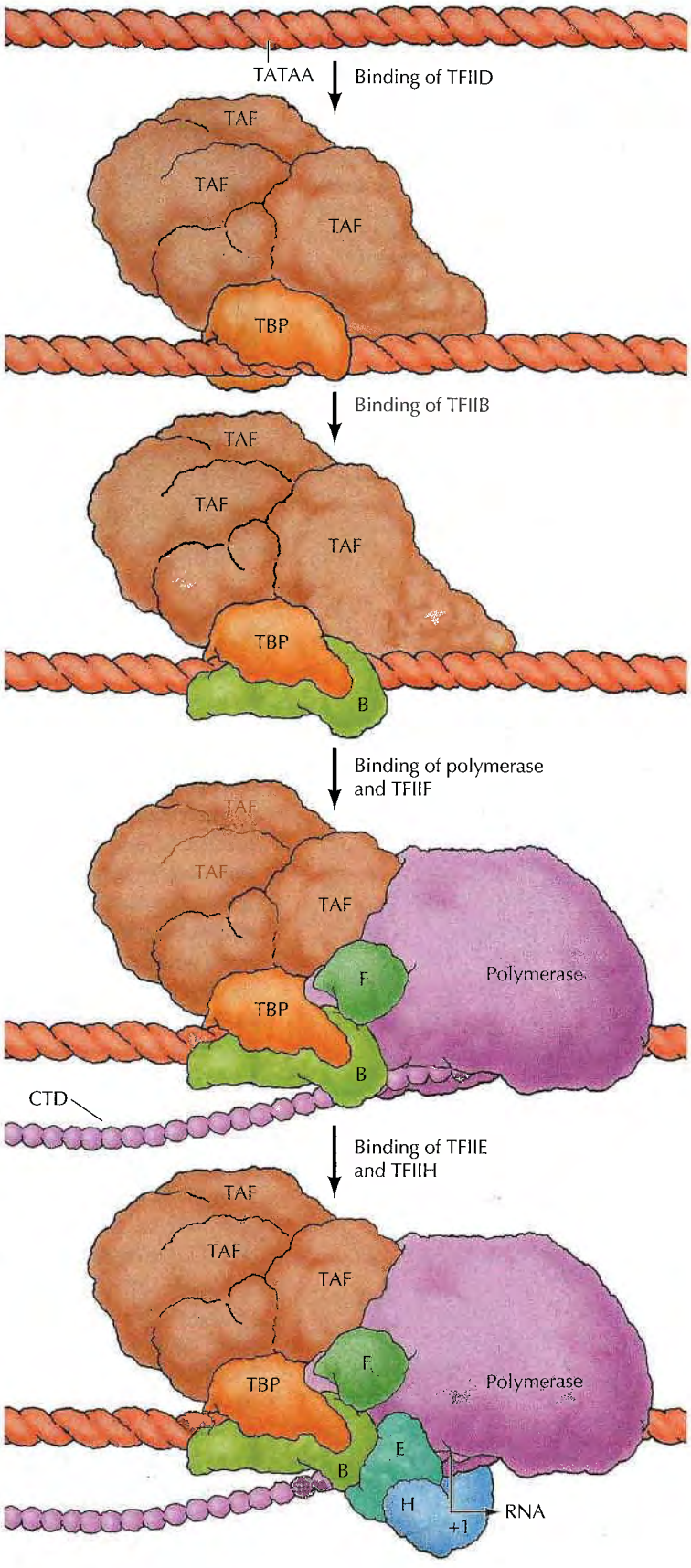


Figure 6.12 Formation of a polymerase II transcription complex
Many polymerase II promoters have a TATA box (consensus sequence TATAA) 25 to 30 nucleotides upstream of the transcription start site. This sequence is recognized by transcription factor TFIID, which consists of the TATA-binding protein (TBP) and TBP-associated factors (TAFs). TFIIB(B) then binds to TBP, followed by binding of the polymerase in association with TFIIF(F). Finally, TFIIE(E) and TFIIH(H) associate with the complex.

tion start site. This sequence (called the **TATA box**) resembles the –10 sequence element of bacterial promoters, and the results of introducing mutations into TATAA sequences have demonstrated their role in the initiation of transcription. The first step in formation of a transcription complex is the binding of a general transcription factor called TFIID to the TATA box (*TF* indicates *t*ranscription *f*actor; *II* indicates polymerase *II*). TFIID is itself composed of multiple subunits, including the **TATA-binding protein (TBP)**, which binds specifically to the TATAA consensus sequence, and approximately 10 other polypeptides, called **TBP-associated factors (TAFs)**. The binding of TFIID is followed by recruitment of a second general transcription factor (TFIIB), which binds to TBP as well as to DNA sequences that are present upstream of the TATA box in some promoters (Figure 6.13). TFIIB in turn serves as a bridge to RNA polymerase II, which binds to the TBP-TFIIB complex in association with a third factor, TFIIF.

Following recruitment of RNA polymerase II to the promoter, the binding of two additional factors (TFIIE and TFIIH) is required for initiation of transcription. TFIIH is a multisubunit factor that appears to play at least two important roles. First, two subunits of TFIIH are helicases, which unwind DNA around the initiation site. (These subunits of TFIIH are the XPB and XPD proteins which are also required for nucleotide excision repair, as discussed in Chapter 5.) Another subunit of TFIIH is a protein kinase that phosphorylates repeated sequences present in the C-terminal domain of the largest subunit of RNA polymerase II. The polymerase II C-terminal domain (or CTD) consists of tandem repeats (27 repeats in yeast and 52 in humans) of 7 amino acids with the consensus sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser. Phosphorylation of these amino acids

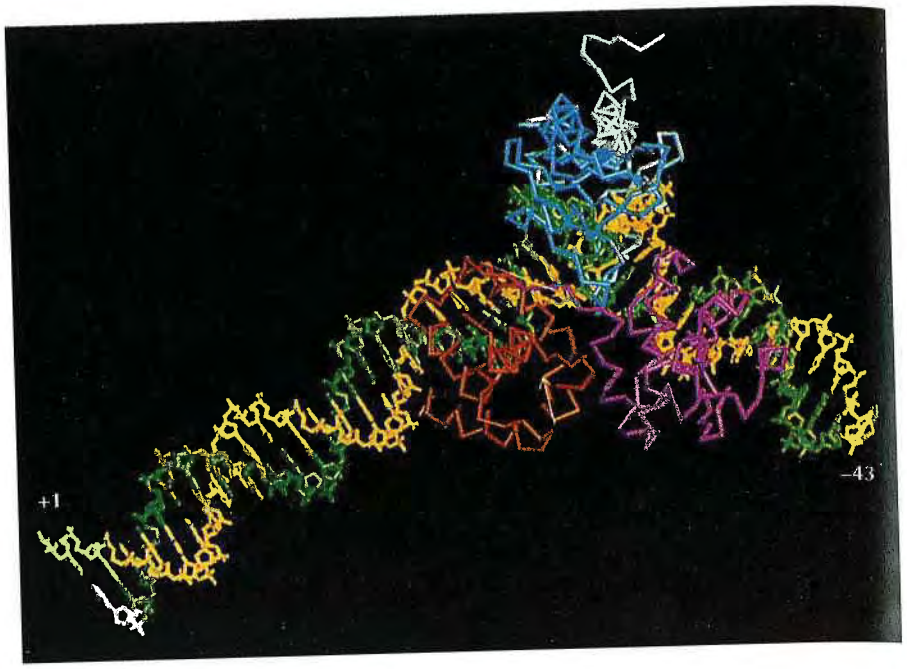


Figure 6.13 Model of the TBP-TFIIB complex bound to DNA
The DNA is shown as a stick figure consisting of yellow and green strands, with the site of transcription initiation designated +1. TBP consists of two repeats, colored light blue and dark blue. TFIIB repeats are colored orange and magenta. Note that TBP binding bends the DNA by approximately 110°. (From D. B. Nikolov et al., 1995. *Nature* 377: 119.)

releases the polymerase from its association with the preinitiation complex, and leads to the recruitment of other proteins that allow the polymerase to initiate transcription and begin synthesis of a growing mRNA chain.

In addition to a TATA box, the promoters of many genes transcribed by RNA polymerase II contain a second important sequence element (an initiator, or Inr, sequence) that spans the transcription start site. Moreover, some RNA polymerase II promoters contain only an Inr element, with no TATA box. Many promoters that lack a TATA box but contain an Inr element also contain an additional downstream promoter element (DPE), located approximately 30 base pairs downstream of the transcription start site, which functions cooperatively with the Inr sequence. Initiation at these promoters still requires TFIID (and TBP), even though TBP obviously does not recognize these promoters by binding directly to the TATA sequence. Instead, other subunits of TFIID (TAFs) appear to bind to the Inr and DPE sequences. The binding of TAFs to these elements recruits TBP to the promoter, and TFIIB, polymerase II, and additional transcription factors then assemble as already described. TBP thus plays a central role in initiating polymerase II transcription, even on promoters that lack a TATA box.

Although the sequential recruitment of five general transcription factors and RNA polymerase II described here represents the minimal system required for transcription *in vitro*, additional factors are needed within the cell. These factors include a **Mediator** protein complex that allows the polymerase to respond to the gene-specific transcription factors that regulate gene expression. At least a fraction of RNA polymerase II in both yeast and mammalian cells is present in the form of large complexes (called RNA polymerase II holoenzymes) in which the polymerase is associated with Mediator proteins, as well as with a subset of the general transcription factors, including TFIIB, TFIIE, TFIIF, TFIIH. The Mediator proteins are associated with the C-terminal domain of polymerase II, and are released from the polymerase following assembly of the preinitiation complex and phosphorylation of the polymerase C-terminal domain (Figure 6.14). The phosphorylated CTD then binds other proteins that facilitate transcriptional elongation and function in mRNA processing, as discussed later in this chapter.

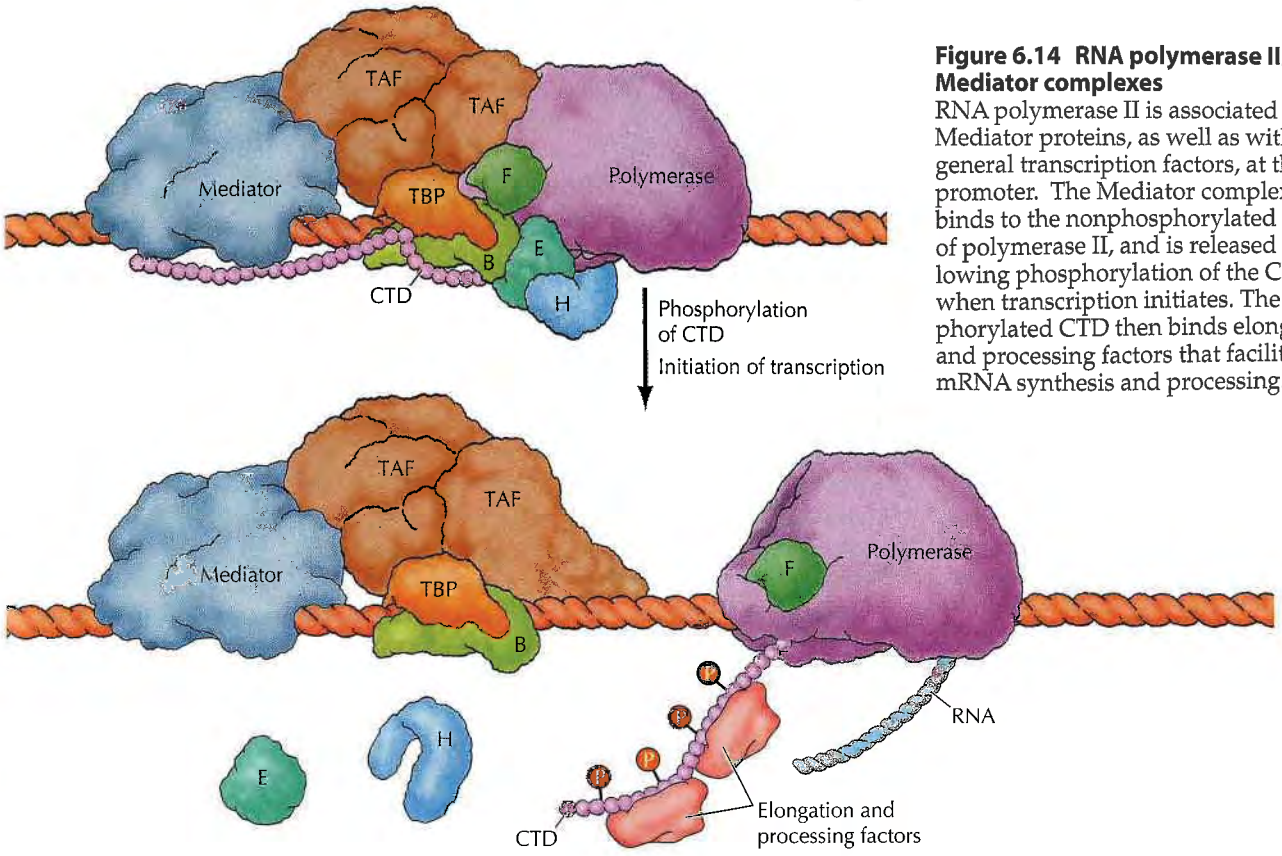


Figure 6.14 RNA polymerase II/ Mediator complexes
RNA polymerase II is associated with Mediator proteins, as well as with the general transcription factors, at the promoter. The Mediator complex binds to the nonphosphorylated CTD of polymerase II, and is released following phosphorylation of the CTD when transcription initiates. The phosphorylated CTD then binds elongation and processing factors that facilitate mRNA synthesis and processing.

Transcription by RNA Polymerases I and III

As previously discussed, distinct RNA polymerases are responsible for the transcription of genes encoding ribosomal and transfer RNAs in eukaryotic cells. All three RNA polymerases, however, require additional transcription factors to associate with appropriate promoter sequences. Furthermore, although the three different polymerases in eukaryotic cells recognize distinct types of promoters, a common transcription factor—the TATA-binding protein (TBP)—appears to be required for initiation of transcription by all three enzymes.

RNA polymerase I is devoted solely to the transcription of ribosomal RNA genes, which are present in tandem repeats. Transcription of these genes yields a large 45S pre-rRNA, which is then processed to yield the 28S, 18S, and 5.8S rRNAs (Figure 6.15). The promoter of ribosomal RNA genes

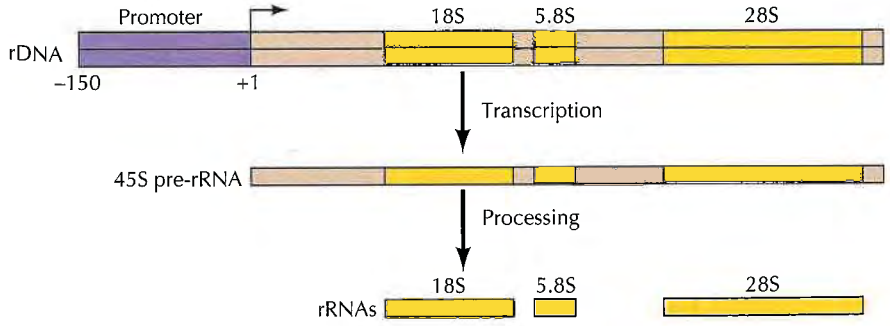


Figure 6.15 The ribosomal RNA gene
The ribosomal DNA (rDNA) is transcribed to yield a large RNA molecule (45S pre-rRNA), which is then cleaved into 28S, 18S, and 5.8S rRNAs.

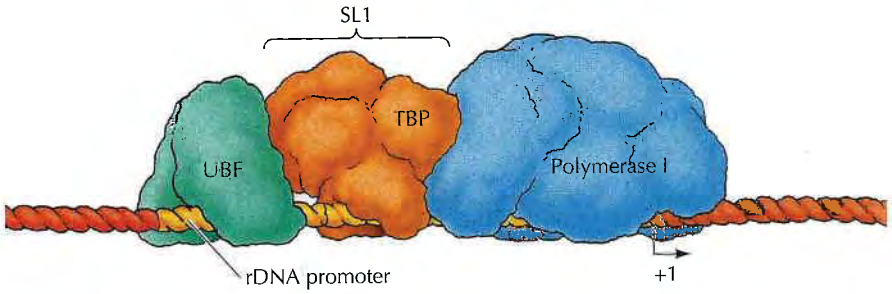


Figure 6.16 Initiation of rDNA transcription
Two transcription factors, UBF and SL1, bind cooperatively to the rDNA promoter and recruit RNA polymerase I to form an initiation complex. One subunit of SL1 is the TATA-binding protein (TBP).

spans about 150 base pairs just upstream of the transcription initiation site. These promoter sequences are recognized by two transcription factors, UBF (upstream binding factor) and SL1 (selectivity factor 1), which bind cooperatively to the promoter and then recruit polymerase I to form an initiation complex (Figure 6.16). The SL1 transcription factor is composed of four protein subunits, one of which is TBP. The role of TBP has been demonstrated directly by the finding that yeasts carrying mutations in TBP are defective not only for transcription by polymerase II, but also for transcription by polymerases I and III. Thus, TBP is a common transcription factor required by all three classes of eukaryotic RNA polymerases. Since the promoter for ribosomal RNA genes does not contain a TATA box, TBP does not bind to specific promoter sequences. Instead, the association of TBP with ribosomal RNA genes is mediated by the binding of other proteins in the SL1 complex to the promoter, a situation similar to the association of TBP with the Inr sequences of polymerase II genes that lack TATA boxes.

The genes for tRNAs, 5S rRNA, and some of the small RNAs involved in splicing and protein transport are transcribed by polymerase III. These genes are transcribed from three distinct classes of promoters, two of which lie within, rather than upstream of, the transcribed sequence (Figure 6.17). The most thoroughly studied of the genes transcribed by polymerase III are the 5S rRNA genes of *Xenopus*. TFIIIA (which is the first transcription factor to have been purified) initiates assembly of a transcription complex by binding to specific DNA sequences in the 5S rRNA promoter. This binding is followed by the sequential binding of TFIIIC, TFIIIB, and the polymerase. The promoters for the tRNA genes differ from the 5S rRNA promoter in that they do not contain the DNA sequence recognized by TFIIIA. Instead, TFIIIC binds directly to the promoter of tRNA genes, serving to recruit TFIIIB and polymerase to form a transcription complex. Promoters of the third class of genes transcribed by polymerase III, including genes encoding some of the small nuclear RNAs involved in splicing, are located upstream of the transcription start site. These promoters contain a TATA box (like promoters for polymerase II genes) as well as a binding site for another factor, called SNAP. SNAP and TFIIIB bind cooperatively to these promoters, with TFIIIB binding directly to the TATA box. This is mediated by the TATA-binding protein, TBP, which is one of the subunits of TFIIIB. As in the case of the promoters of other RNA polymerase III genes, TFIIIB then recruits the polymerase to the transcription complex.

Regulation of Transcription in Eukaryotes

Although the control of gene expression is far more complex in eukaryotes than in bacteria, the same basic principles apply. The expression of eukaryotic genes is controlled primarily at the level of initiation of transcription, although in many cases transcription is also regulated during elongation. As in bacteria, transcription in eukaryotic cells is controlled by proteins that

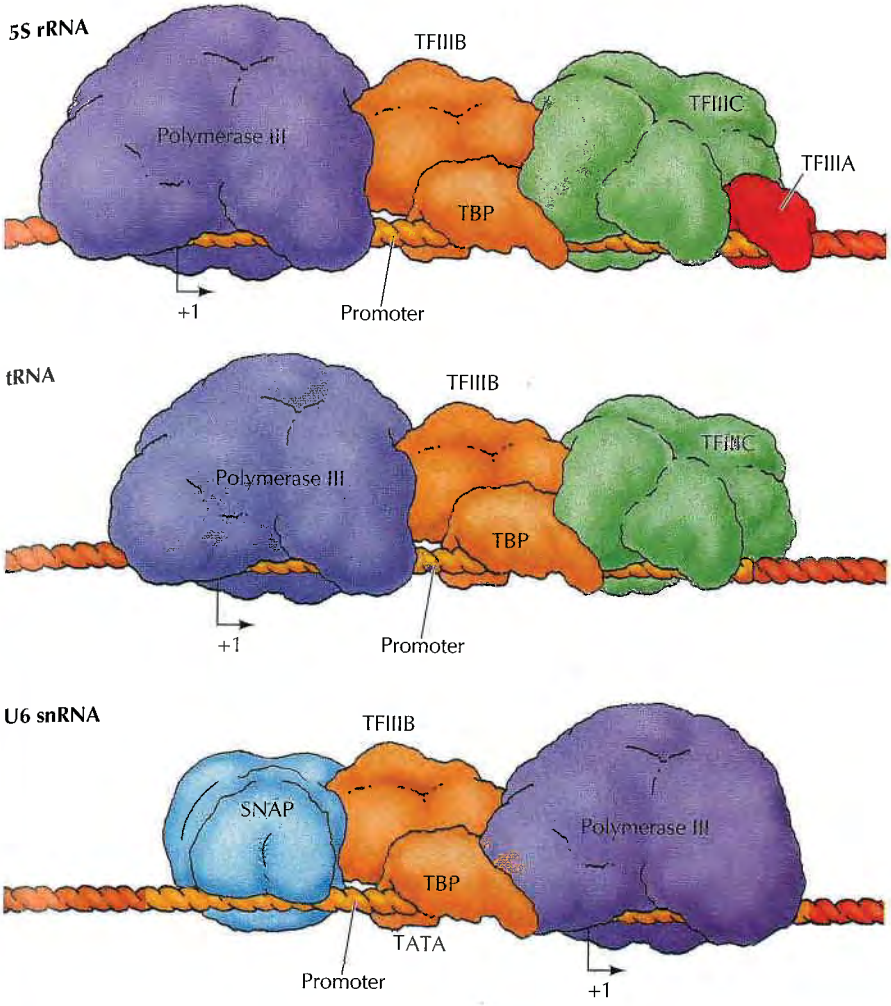


Figure 6.17 Transcription of polymerase III genes
Genes transcribed by polymerase III are expressed from three types of promoters. The promoters of 5S rRNA and tRNA genes are downstream of the transcription initiation site. Transcription of the 5S rRNA gene is initiated by the binding of TFIIIA, followed by the binding of TFIIIC, TFIIIB, and the polymerase. The tRNA promoters do not contain a binding site for TFIIIA, and TFIIIA is not required for their transcription. Instead, TFIIIC initiates the transcription of tRNA genes by binding to promoter sequences, followed by the association of TFIIIB and polymerase. The promoter of the U6 snRNA gene is upstream of the transcription start site and contains a TATA box, which is recognized by the TATA-binding protein (TBP) subunit of TFIIIB, in cooperation with another factor called SNAP.

bind to specific regulatory sequences and modulate the activity of RNA polymerase. An important difference between transcriptional regulation in prokaryotes and eukaryotes cells, however, results from the packaging of eukaryotic DNA into chromatin, which limits its availability as a template for transcription. As a result, modifications of chromatin structure play key roles in the control of transcription in eukaryotic cells.

cis-*Acting Regulatory Sequences: Promoters and Enhancers*

As already discussed, transcription in bacteria is regulated by the binding of proteins to *cis*-acting sequences (e.g., the *lac* operator) that control the transcription of adjacent genes. Similar *cis*-acting sequences regulate the expression of eukaryotic genes. These sequences have been identified in mammalian cells largely by the use of gene transfer assays to study the activity of suspected regulatory regions of cloned genes (Figure 6.18). The eukaryotic regulatory sequences are usually ligated to a reporter gene that encodes an easily detectable enzyme. The expression of the reporter gene following its transfer into cultured cells then provides a sensitive assay for the ability of the cloned regulatory sequences to direct transcription. Biologically active regulatory regions can thus be identified, and *in vitro* mutagenesis can be used to determine the roles of specific sequences within the region.

Genes transcribed by RNA polymerase II have core promoter elements, including the TATA box and the Inr sequence, that serve as specific binding sites for general transcription factors. Other *cis*-acting sequences serve as

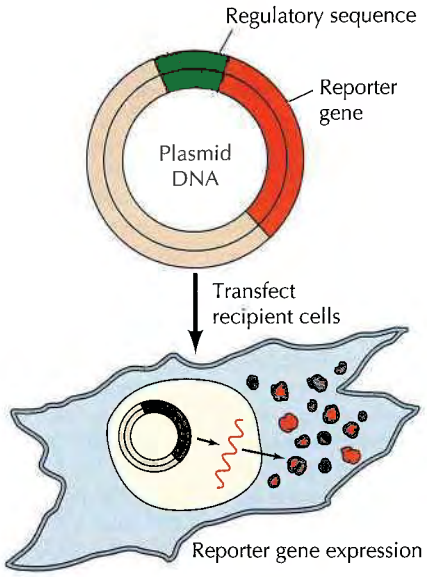


Figure 6.18 Identification of eukaryotic regulatory sequences
The regulatory sequence of a cloned eukaryotic gene is ligated to a reporter gene that encodes an easily detectable enzyme. The resulting plasmid is then introduced into cultured recipient cells by transfection. An active regulatory sequence directs transcription of the reporter gene, expression of which is then detected in the transfected cells.

binding sites for a wide variety of regulatory factors that control the expression of individual genes. These *cis*-acting regulatory sequences are frequently, though not always, located upstream of the TATA box. For example, two regulatory sequences that are found in many eukaryotic genes were identified by studies of the promoter of the herpes simplex virus gene that encodes thymidine kinase (Figure 6.19). Both of these sequences are located within 100 base pairs upstream of the TATA box: Their consensus sequences are CCAAT and GGGCGG (called a GC box). Specific proteins that bind to these sequences and stimulate transcription have since been identified.

In contrast to the relatively simple organization of CCAAT and GC boxes in the herpes thymidine kinase promoter, many genes in mammalian cells are controlled by regulatory sequences located farther away (sometimes more than 10 kilobases) from the transcription start site. These sequences, called **enhancers**, were first identified during studies of the promoter of another virus, SV40 (Figure 6.20). In addition to a TATA box and a set of six GC boxes, two 72-base-pair repeats located farther upstream are required for efficient transcription from this promoter. These sequences were found to stimulate transcription from other promoters as well as from that of SV40, and, surprisingly, their activity depended on neither their distance nor their orientation with respect to the transcription initiation site (Figure 6.21). They could stimulate transcription when placed either upstream or downstream of the promoter, in either a forward or backward orientation.

The ability of enhancers to function even when separated by long distances from transcription initiation sites at first suggested that they work by mechanisms different from those of promoters. However, this has turned out not to be the case: Enhancers, like promoters, function by binding transcription factors that then regulate RNA polymerase. This is possible because of DNA looping, which allows a transcription factor bound to a distant enhancer to interact with proteins associated with RNA polymerase at the promoter (Figure 6.22). Transcription factors bound to distant enhancers can thus work by the same mechanisms as those bound adjacent to promoters, so there is no fundamental difference between the actions of enhancers and those of *cis*-acting regulatory sequences adjacent to transcription start sites. Interestingly, although enhancers were first identified in mammalian cells, they have subsequently been found in bacteria—an unusual instance in which studies of eukaryotes served as a model for the simpler prokaryotic systems.

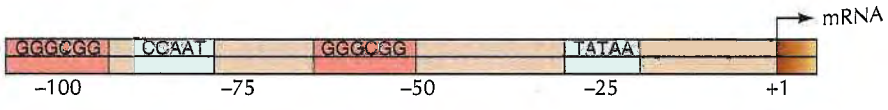


Figure 6.19 A eukaryotic promoter
The promoter of the thymidine kinase gene of herpes simplex virus contains three sequence elements upstream of the TATA box that are required for efficient transcription: a CCAAT box and two GC boxes (consensus sequence GGGCGG).

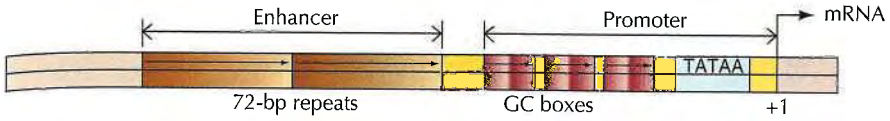


Figure 6.20 The SV40 enhancer
The SV40 promoter for early gene expression contains a TATA box and six GC boxes arranged in three sets of repeated sequences. In addition, efficient transcription requires an upstream enhancer consisting of two 72-base-pair (bp) repeats.

The binding of specific transcriptional regulatory proteins to enhancers is responsible for the control of gene expression during development and differentiation, as well as during the response of cells to hormones and growth factors. One of the most thoroughly studied mammalian enhancers controls the transcription of immunoglobulin genes in B lymphocytes. Gene transfer experiments have established that the immunoglobulin enhancer is active in lymphocytes, but not in other types of cells. Thus, this regulatory sequence is at least partly responsible for tissue-specific expression of the immunoglobulin genes in the appropriate differentiated cell type.

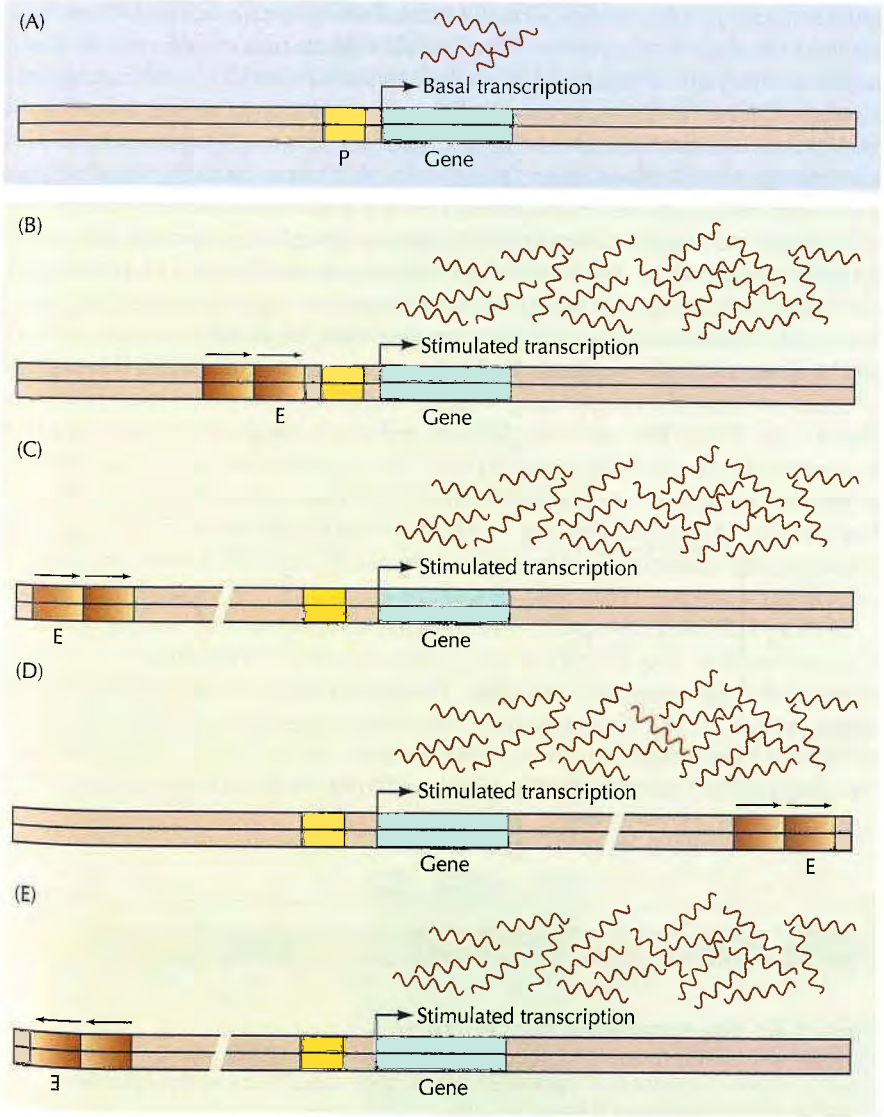


Figure 6.21 Action of enhancers
Without an enhancer, the gene is transcribed at a low basal level (A). Addition of an enhancer, E—for example, the SV40 72-base-pair repeats—stimulates transcription. The enhancer is active not only when placed just upstream of the promoter (B), but also when inserted up to several kilobases either upstream or downstream from the transcription start site (C and D). In addition, enhancers are active in either the forward or backward orientation (E).

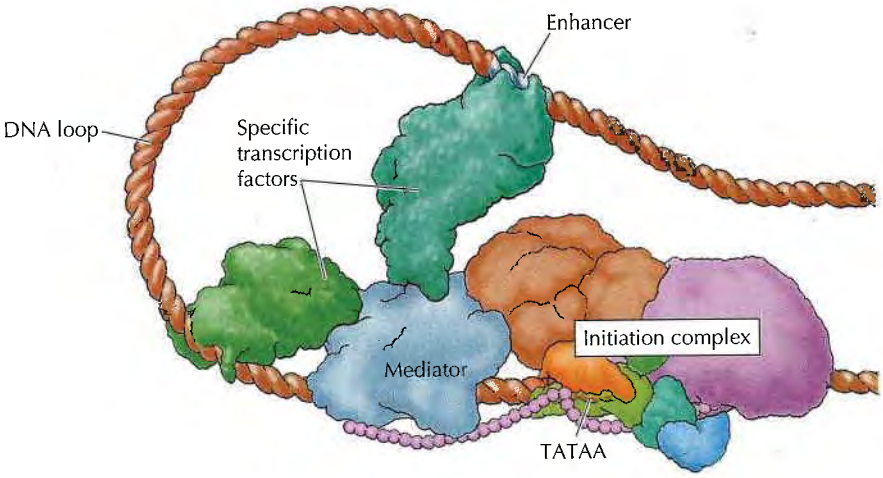


Figure 6.22 DNA looping
Transcription factors bound at distant enhancers are able to interact with the RNA polymerase II/Mediator complex or general transcription factors at the promoter because the intervening DNA can form loops. There is therefore no fundamental difference between the action of transcription factors bound to DNA just upstream of the promoter and to distant enhancers.

An important aspect of enhancers is that they usually contain multiple functional sequence elements that bind different transcriptional regulatory proteins. These proteins work together to regulate gene expression. The immunoglobulin heavy-chain enhancer, for example, spans approximately 200 base pairs and contains at least nine distinct sequence elements that serve as protein-binding sites (Figure 6.23). Mutation of any one of these sequences reduces but does not abolish enhancer activity, indicating that the functions of individual proteins that bind to the enhancer are at least partly redundant. Many of the individual sequence elements of the immunoglobulin enhancer by themselves stimulate transcription in non-lymphoid cells. The restricted activity of the intact enhancer in B lymphocytes therefore does not result from the tissue-specific function of each of its components. Instead, tissue-specific expression results from the combination of the individual sequence elements that make up the complete enhancer. These elements include some *cis*-acting regulatory sequences that bind transcriptional activators that are expressed specifically in B lymphocytes, as well as other regulatory sequences that bind repressors in nonlymphoid cells. Thus, the immunoglobulin enhancer contains negative regulatory elements that inhibit transcription in inappropriate cell types, as well as positive regulatory elements that activate transcription in B lymphocytes. The overall activity of the enhancer is greater than the sum of its parts, reflecting the combined action of the proteins associated with each of its individual sequence elements.

Although DNA looping allows enhancers to act at a considerable distance from promoters, the activity of any given enhancer is specific for the promoter of its appropriate target gene. This specificity is maintained by **insulators**, which divide chromosomes into independent domains and prevent enhancers from acting on promoters located in an adjacent domain. Insulators also prevent the chromatin structure of one domain from spreading to

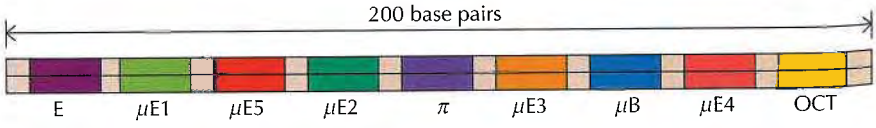


Figure 6.23 The immunoglobulin enhancer
The immunoglobulin heavy-chain enhancer spans about 200 bases and contains nine functional sequence elements (E, μE1–5, π, μB, and OCT), which together stimulate transcription in B lymphocytes.

its neighbors, thereby maintaining independently regulated regions of the genome.

Transcriptional Regulatory Proteins

The isolation of a variety of transcriptional regulatory proteins has been based on their specific binding to promoter or enhancer sequences. Protein binding to these DNA sequences is commonly analyzed by two types of experiments. The first, footprinting, was described earlier in connection with the binding of RNA polymerase to prokaryotic promoters (see Figure 6.3). The second approach is the **electrophoretic-mobility shift assay**, in which a radiolabeled DNA fragment is incubated with a protein preparation and then subjected to electrophoresis through a nondenaturing gel (Figure 6.24). Protein binding is detected as a decrease in the electrophoretic mobility of the DNA fragment, since its migration through the gel is slowed by the bound protein. The combined use of footprinting and electrophoretic-mobility shift assays has led to the correlation of protein-binding sites with the regulatory elements of enhancers and promoters, indicating that these sequences generally constitute the recognition sites of specific DNA-binding proteins.

One of the prototypes of eukaryotic transcription factors was initially identified by Robert Tjian and his colleagues during studies of the transcription of SV40 DNA. This factor (called Sp1, for specificity protein 1) was found to stimulate transcription from the SV40 promoter, but not from several other promoters, in cell-free extracts. Then, stimulation of transcription by Sp1 was found to depend on the presence of the GC boxes in the SV40 promoter: If these sequences were deleted, stimulation by Sp1 was abolished. Moreover, footprinting experiments established that Sp1 binds specifically to the GC box sequences. Taken together, these results indicate that the GC box represents a specific binding site for a transcriptional activator—Sp1. Similar experiments have established that many other transcriptional regulatory sequences, including the CCAAT sequence and the

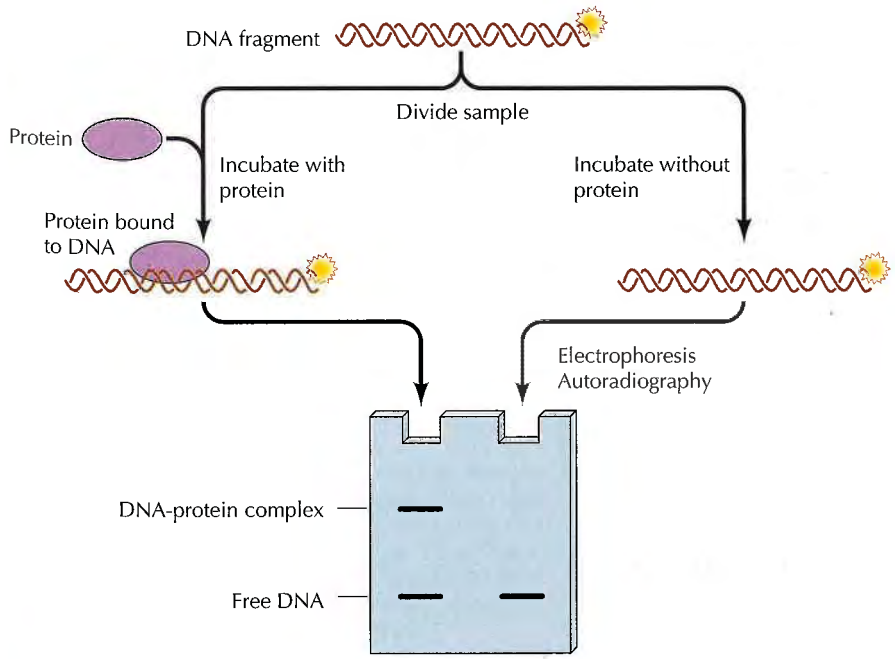


Figure 6.24 Electrophoretic-mobility shift assay
A sample containing radiolabeled fragments of DNA is divided into two, and one half of the sample is incubated with a protein that binds to a specific DNA sequence. Samples are then analyzed by electrophoresis in a nondenaturing gel so that the protein remains bound to DNA. Protein binding is detected by the slower migration of DNA-protein complexes compared to that of free DNA. Only a fraction of the DNA in the sample is actually bound to protein, so both DNA-protein complexes and free DNA are detected following incubation of the DNA with protein.

TABLE 6.2 Examples of Transcription Factors and Their DNA-Binding Sites

Transcription factor	Consensus binding site
Specificity protein 1 (Sp1)	GGGCGG
CCAAT/Enhancer binding protein (C/EBP)	CCAAT
Activator protein 1 (AP1)	TGACTCA
Octamer binding proteins (OCT-1 and OCT-2)	ATGCAAAT
E-box binding proteins (E12, E47, E2-2)	CANNTG[a]

[a] N stands for any nucleotide.

various sequence elements of the immunoglobulin enhancer, also represent recognition sites for sequence-specific DNA-binding proteins (Table 6.2).

The specific binding of Sp1 to the GC box not only established the action of Sp1 as a sequence-specific transcription factor; it also suggested a general approach to the purification of transcription factors. The isolation of these proteins initially presented a formidable challenge because they are present in very small quantities (e.g., only 0.001% of total cell protein) that are difficult to purify by conventional biochemical techniques. This problem was overcome in the purification of Sp1 by **DNA-affinity chromatography** (Figure 6.25). Multiple copies of oligonucleotides corresponding to the GC box sequence were bound to a solid support, and cell extracts were passed through the oligonucleotide column. Because Sp1 bound to the GC box

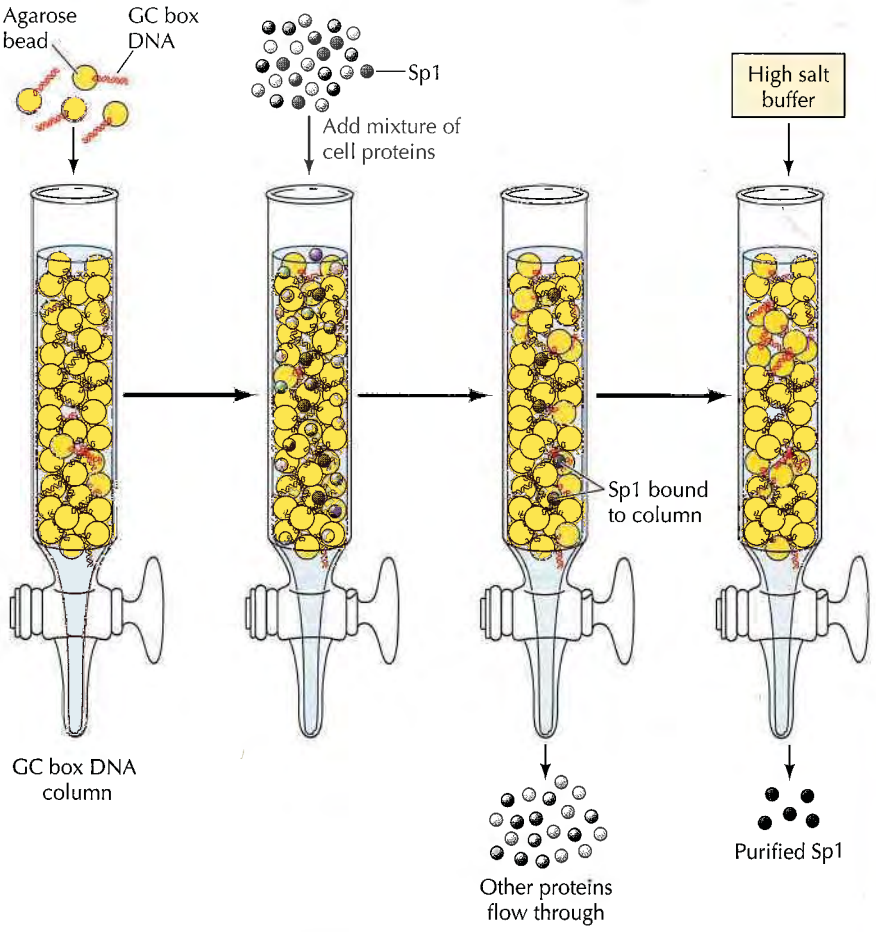


Figure 6.25 Purification of Sp1 by DNA-affinity chromatography
A double-stranded oligonucleotide containing repeated GC box sequences is bound to agarose beads, which are poured into a column. A mixture of cell proteins containing Sp1 is then applied to the column; because Sp1 specifically binds to the GC box oligonucleotide, it is retained on the column while other proteins flow through. Washing the column with high salt buffer then dissociates Sp1 from the GC box DNA, yielding purified Sp1.

KEY EXPERIMENT

Isolation of a Eukaryotic Transcription Factor

Affinity Purification of Sequence-Specific DNA-Binding Proteins

James T. Kadonaga and Robert Tjian
University of California, Berkeley

Proceedings of the National Academy of Sciences, USA, 1986, Volume 83, pages 5889–5893

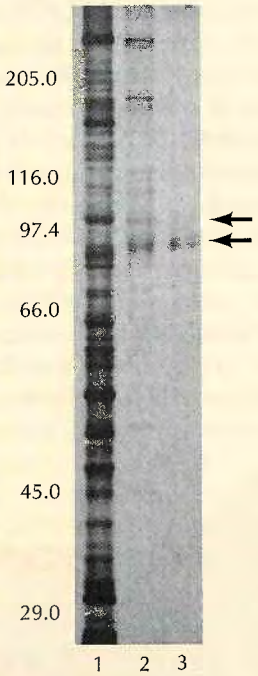
Robert Tjian

The Context

Starting with studies of the *lac* operon by Jacob and Monod, it became clear that transcription is regulated by proteins that bind to specific DNA sequences. One of the prototype systems for studies of gene expression in eukaryotic cells was the monkey virus SV40, in which several regulatory DNA sequences were identified in the early 1980s. In 1983 William Dynan and Robert Tjian first demonstrated that one of these sequence elements (the GC box) is the specific binding site of a protein detectable in nuclear extracts of human cells. This protein (called Sp1 for specificity protein 1) not only binds to the GC box sequence; it also stimulates transcription *in vitro*, demonstrating that it is a sequence-specific transcriptional activator.

To study the mechanism of Sp1 action, it then became necessary to obtain the transcription factor in pure form and eventually to clone the *Sp1* gene. The isolation of pure Sp1 thus became a high priority, but it also posed a daunting technical challenge.

Sp1 and other transcription factors appeared to represent only about 0.001% of total cell protein, so they could not be purified by conventional biochemical techniques. James Kadonaga and Robert Tjian solved this problem by developing a method of DNA-affinity chromatography that led to the purification not only of Sp1 but also of many other eukaryotic transcription factors, thereby opening the door to molecular analysis of transcriptional regulation in eukaryotic cells.



Purification of Sp1. Gel electrophoresis of proteins initially present in the crude nuclear extract (lane 1) and of proteins obtained after either one or two sequential cycles of DNA-affinity chromatography (lanes 2 and 3, respectively). The sizes of marker proteins (in kilodaltons) are indicated to the left of the gel, and the Sp1 polypeptides are indicated by arrows.

The Experiments

The DNA-affinity chromatography method developed by Kadonaga and Tjian exploited the specific high-affinity binding of Sp1 to the GC box sequence, GGGCGG. Synthetic oligonucleotides containing multiple copies of this sequence were coupled to solid beads, and a crude nuclear extract was passed through a column consisting of beads linked to GC box DNA. The beads were then washed to remove proteins that had failed to bind specifically to the oligonucleotides. Finally, the beads were washed with a high salt buffer (0.5 *M* KCl), which disrupted the binding of Sp1 to DNA, thereby releasing Sp1 from the column.

Gel electrophoresis demonstrated that the crude nuclear extract initially applied to the column was a complex mixture of proteins (see figure). In contrast, approximately 90% of the protein recovered after two cycles of DNA-affinity chromatography corresponded to only two polypeptides, which were identified as Sp1 by DNA binding and by their activity in *in vitro* transcription assays. Thus, Sp1 had been successfully purified by DNA-affinity chromatography.

The Impact

In their 1986 paper, Kadonaga and Tjian stated that the DNA-affinity chromatography technique "should be generally applicable for the purification of other sequence-specific DNA binding proteins." This prediction has been amply verified; many eukaryotic transcription factors have been purified by this method. The genes that encode still other transcription factors have been isolated by an alternative approach (developed independently in 1988 in the laboratories of Phillip Sharp and Steven McKnight) in which cDNA expression libraries are screened with oligonucleotide probes to detect recombinant proteins that bind specifically to the desired DNA sequences. The ability to isolate sequence-specific DNA-binding proteins by these methods has led to detailed characterization of the structure and function of a wide variety of transcriptional regulatory proteins, providing the basis for our current understanding of gene expression in eukaryotic cells.

with high affinity, it was specifically retained on the column while other proteins were not. Highly purified Sp1 could thus be obtained and used for further studies, including partial determination of its amino acid sequence, which in turn led to cloning of the gene for Sp1.

The general method of DNA-affinity chromatography, first optimized for the purification of Sp1, has been used successfully to isolate a wide variety of sequence-specific DNA-binding proteins from eukaryotic cells. Genes encoding other transcription factors have been isolated by screening cDNA expression libraries to identify recombinant proteins that bind to specific DNA sequences. The cloning and sequencing of transcription factor cDNAs has led to the accumulation of a great deal of information on the structure and function of these critical regulatory proteins.

Structure and Function of Transcriptional Activators

Because transcription factors are central to the regulation of gene expression, understanding the mechanisms of their action is a major area of ongoing research in cell and molecular biology. The most thoroughly studied of these proteins are **transcriptional activators**, which, like Sp1, bind to regulatory DNA sequences and stimulate transcription. In general, these factors consist of two domains: One region of the protein specifically binds DNA; the other stimulates transcription by interacting with other proteins, including components of the transcriptional machinery (Figure 6.26). Transcriptional activators appear to be modular proteins, in the sense that the DNA binding and activation domains of different factors can frequently be interchanged using recombinant DNA techniques. Such manipulations result in hybrid transcription factors, which activate transcription by binding to promoter or enhancer sequences determined by the specificity of their DNA-binding domains. It therefore appears that the basic function of the DNA-binding domain is to anchor the transcription factor to the proper site on DNA; the activation domain then independently stimulates transcription through protein-protein interactions.

Many different transcription factors have now been identified in eukaryotic cells, as might be expected, given the intricacies of tissue-specific and inducible gene expression in complex multicellular organisms. Molecular characterization has revealed that the DNA-binding domains of many of these proteins are related to one another (Figure 6.27). **Zinc finger domains** contain repeats of cysteine and histidine residues that bind zinc ions and fold into looped structures ("fingers") that bind DNA. These domains were

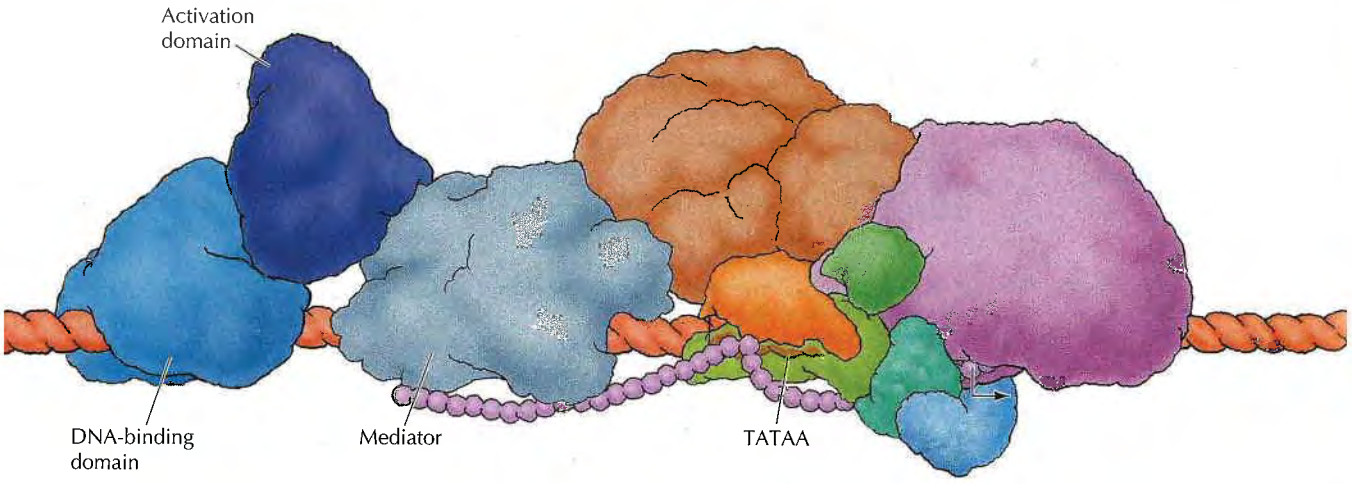


Figure 6.26 Structure of transcriptional activators
Transcriptional activators consist of two independent domains. The DNA-binding domain recognizes a specific DNA sequence, and the activation domain interacts with other components of the transcriptional machinery.

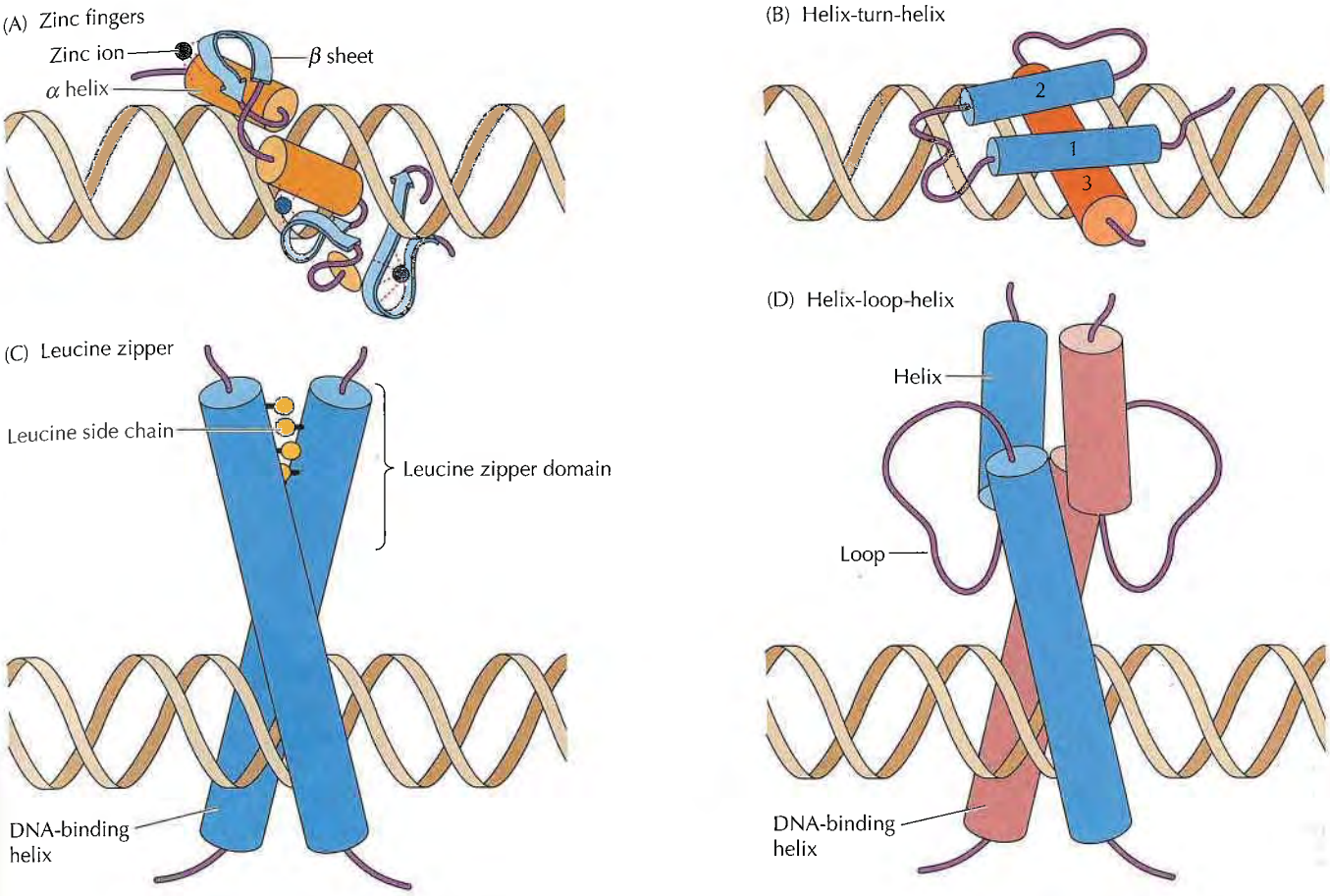


Figure 6.27 Families of DNA-binding domains
(A) Zinc finger domains consist of loops in which an α helix and a β sheet coordinately bind a zinc ion. (B) Helix-turn-helix domains consist of three (or in some cases four) helical regions. One helix (helix 3) makes most of the contacts with DNA, while helices 1 and 2 lie on top and stabilize the interaction. (C) The DNA-binding domains of leucine zipper proteins are formed from two distinct polypeptide chains. Interactions between the hydrophobic side chains of leucine residues exposed on one side of a helical region (the leucine zipper) are responsible for dimerization. Immediately following the leucine zipper is a DNA-binding helix, which is rich in basic amino acids. (D) Helix-loop-helix domains are similar to leucine zippers, except that the dimerization domains of these proteins each consist of two helical regions separated by a loop.

initially identified in the polymerase III transcription factor TFIIIA but are also common among transcription factors that regulate polymerase II promoters, including Sp1. Other examples of transcription factors that contain zinc finger domains are the **steroid hormone receptors**, which regulate gene transcription in response to hormones such as estrogen and testosterone.

The **helix-turn-helix** motif was first recognized in prokaryotic DNA-binding proteins, including the *E. coli* catabolite activator protein (CAP). In these proteins, one helix makes most of the contacts with DNA, while the other helices lie across the complex to stabilize the interaction. In eukaryotic cells, helix-turn-helix proteins include the **homeodomain** proteins, which play critical roles in the regulation of gene expression during embryonic development. The genes encoding these proteins were first discovered as developmental mutants in *Drosophila*. Some of the earliest recognized *Drosophila* mutants (termed homeotic mutants in 1894) resulted in the development of flies in which one body part was transformed into another. For example, in the homeotic mutant called *Antennapedia*, legs rather than antennae grow out of the head of the fly (Figure 6.28). Genetic analysis of these mutants, pioneered by Ed Lewis in the 1940s, has shown that *Drosophila* contains nine homeotic genes, each of which specifies the identity of a different body segment. Molecular cloning and analysis of these genes then indicated that they contain conserved sequences of 180 base pairs (called **homeoboxes**) that encode DNA-binding domains (homeodomains) of transcription factors. A wide variety of the additional homeodomain proteins have since been identified in fungi, plants, and other animals, including humans. Verte-

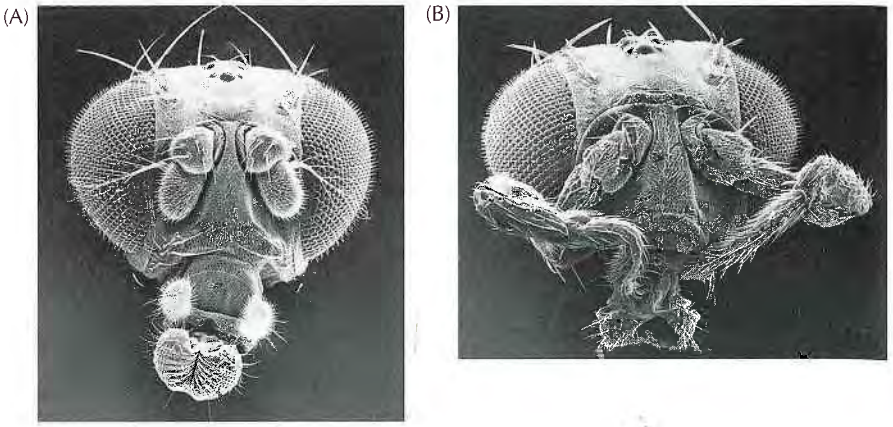


Figure 6.28 The *Antennapedia* mutation
Antennapedia mutant flies have legs growing out of their heads in place of antennae. (A) Head of a normal fly. (B) Head of an *Antennapedia* mutant. (Courtesy of F. Rudolf Turner, Indiana University.)

brate homeobox genes are strikingly similar to their *Drosophila* counterparts in both structure and function, demonstrating the highly conserved roles of these transcription factors in animal development.

Two other families of DNA-binding proteins, **leucine zipper** and **helix-loop-helix** proteins, contain DNA-binding domains formed by dimerization of two polypeptide chains. The leucine zipper contains four or five leucine residues spaced at intervals of seven amino acids, resulting in their hydrophobic side chains being exposed at one side of a helical region. This region serves as the dimerization domain for the two protein subunits, which are held together by hydrophobic interactions between the leucine side chains. Immediately following the leucine zipper is a region rich in positively charged amino acids (lysine and arginine) that binds DNA. The helix-loop-helix proteins are similar in structure, except that their dimerization domains are each formed by two helical regions separated by a loop. An important feature of both leucine zipper and helix-loop-helix transcription factors is that different members of these families can dimerize with each other. Thus, the combination of distinct protein subunits can form an expanded array of factors that can differ both in DNA sequence recognition and in transcription-stimulating activities. Both leucine zipper and helix-loop-helix proteins play important roles in regulating tissue-specific and inducible gene expression, and the formation of dimers between different members of these families is a critical aspect of the control of their function.

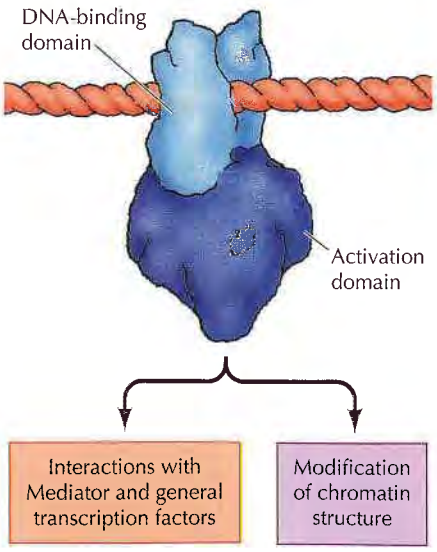


Figure 6.29 Action of transcriptional activators
Eukaryotic activators stimulate transcription by two mechanisms. They interact with Mediator proteins and general transcription factors to facilitate the assembly of a transcription complex, and they interact with coactivators that facilitate transcription by modifying chromatin structure.

The activation domains of transcription factors are not as well characterized as their DNA-binding domains. Some, called acidic activation domains, are rich in negatively charged residues (aspartate and glutamate); others are rich in proline or glutamine residues. The activation domains of eukaryotic transcription factors stimulate transcription by two distinct mechanisms (Figure 6.29). First, they interact with Mediator proteins and general transcription factors, such as TFIIB or TFIID, to recruit RNA polymerase and facilitate the assembly of a transcription complex on the promoter, similar to transcriptional activators in bacteria (see Figure 6.10). In addition, eukaryotic transcription factors interact with a variety of **coactivators** that stimulate transcription by modifying chromatin structure, as discussed later in this chapter.

Eukaryotic Repressors

Gene expression in eukaryotic cells is regulated by repressors as well as by transcriptional activators. Like their prokaryotic counterparts, eukaryotic

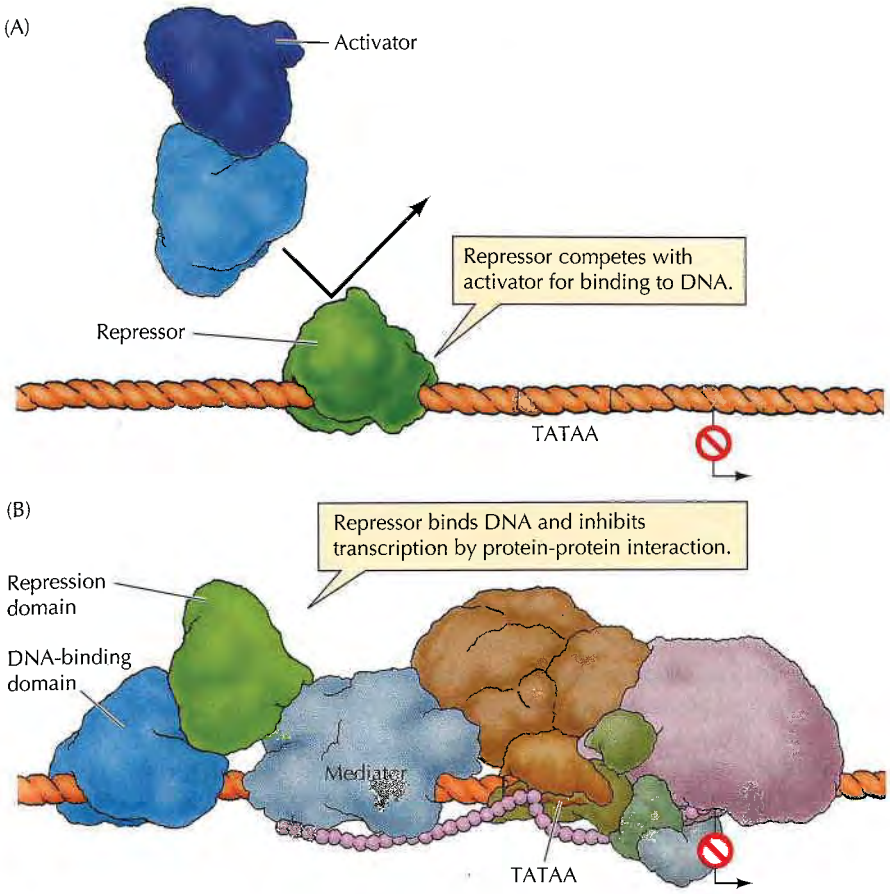


Figure 6.30 Action of eukaryotic repressors
(A) Some repressors block the binding of activators to regulatory sequences. (B) Other repressors have active repression domains that inhibit transcription by interactions with Mediator proteins or general transcription factors, as well as with corepressors that act to modify chromatin structure.

repressors bind to specific DNA sequences and inhibit transcription. In some cases, eukaryotic repressors simply interfere with the binding of other transcription factors to DNA (Figure 6.30A). For example, the binding of a repressor near the transcription start site can block the interaction of RNA polymerase or general transcription factors with the promoter, which is similar to the action of repressors in bacteria. Other repressors compete with activators for binding to specific regulatory sequences. Some such repressors contain the same DNA-binding domain as the activator but lack its activation domain. As a result, their binding to a promoter or enhancer blocks the binding of the activator, thereby inhibiting transcription.

In contrast to repressors that simply interfere with activator binding, many repressors (called active repressors) contain specific functional domains that inhibit transcription via protein-protein interactions (Figure 6.30B). The first such active repressor was described in 1990 during studies of a gene called *Krüppel*, which is involved in embryonic development in *Drosophila*. Molecular analysis of the Krüppel protein demonstrated that it contains a discrete repression domain, which is linked to a zinc finger DNA-binding domain. The Krüppel repression domain could be interchanged with distinct DNA-binding domains of other transcription factors.

These hybrid molecules also repressed transcription, indicating that the Krüppel repression domain inhibits transcription via protein-protein interactions, irrespective of its site of binding to DNA.

Many active repressors have since been found to play key roles in the regulation of transcription in animal cells, in many cases serving as critical regulators of cell growth and differentiation. As with transcriptional activators, several distinct types of repression domains have been identified. For example, the repression domain of Krüppel is rich in alanine residues, whereas other repression domains are rich in proline or acidic residues. The functional targets of repressors are also diverse, repressors can inhibit transcription by interacting with specific activator proteins, with Mediator proteins or general transcription factors, and with **corepressors** that act by modifying chromatin structure.

The regulation of transcription by repressors as well as by activators considerably extends the range of mechanisms that control the expression of eukaryotic genes. One important role of repressors may be to inhibit the expression of tissue-specific genes in inappropriate cell types. For example, as noted earlier, a repressor-binding site in the immunoglobulin enhancer is thought to contribute to its tissue-specific expression by suppressing transcription in nonlymphoid cell types. Other repressors play key roles in the control of cell proliferation and differentiation in response to hormones and growth factors (see Chapters 13 and 14).

Relationship of Chromatin Structure to Transcription

As noted in the preceding discussion, both activators and repressors regulate transcription in eukaryotes not only by interacting with other components of the transcriptional machinery, but also by inducing changes in the structure of chromatin. Rather than being present within the nucleus as naked DNA, the DNA of all eukaryotic cells is tightly bound to histones. The basic structural unit of chromatin is the nucleosome, which consists of 146 base pairs of DNA wrapped around two molecules each of histones H2A, H2B, H3, and H4, with one molecule of histone H1 bound to the DNA as it enters the nucleosome core particle (see Figure 4.12). The chromatin is then further condensed by being coiled into higher-order structures organized into large loops of DNA. This packaging of eukaryotic DNA in chromatin clearly has important consequences in terms of its availability as a template for transcription, so chromatin structure is a critical aspect of gene expression in eukaryotic cells.

Actively transcribed genes are found in relatively decondensed chromatin, probably corresponding to the 30-nm chromatin fibers discussed in Chapter 4 (see Figure 4.13). For example, microscopic visualization of the polytene chromosomes of *Drosophila* indicates that regions of the genome that are actively engaged in RNA synthesis correspond to decondensed chromosome regions (Figure 6.31). Nonetheless, actively transcribed genes remain bound to histones and packaged in nucleosomes, so transcription factors and RNA polymerase are still faced with the problem of interacting with chromatin rather than with naked DNA. The tight winding of DNA around the nucleosome core particle is a major obstacle to transcription, affecting both the ability of transcription factors to bind DNA and the ability of RNA polymerase to transcribe through a chromatin template.

Several modifications are characteristic of transcriptionally active chromatin, including modifications of histones, rearrangements of nucleosomes, and the association of two nonhistone chromosomal proteins, called **HMGN proteins,** with the nucleosomes of actively transcribed genes. The

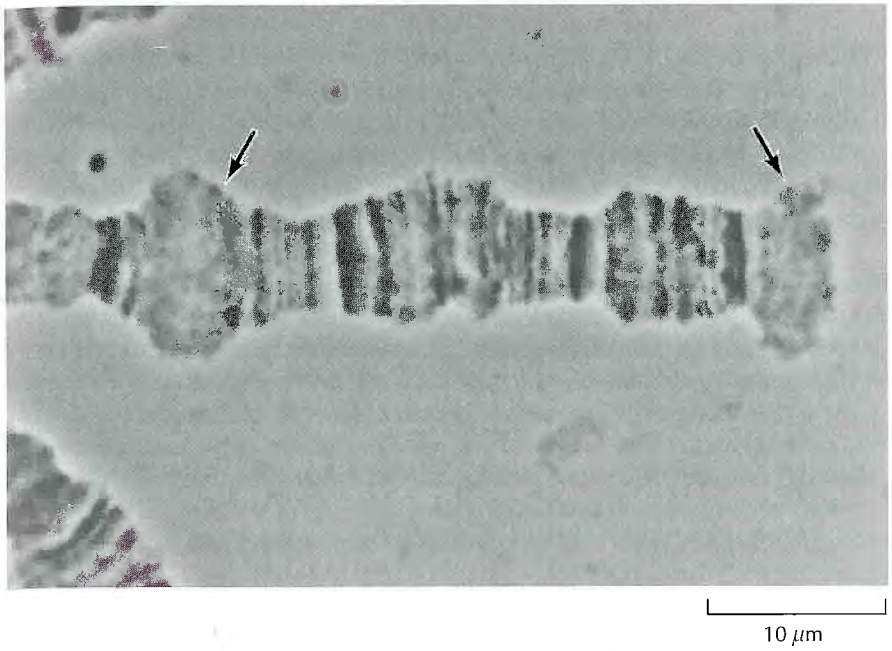


Figure 6.31 Decondensed chromosome regions in *Drosophila*
A light micrograph showing decondensed regions of polytene chromosomes (arrows), which are active in RNA synthesis. (Courtesy of Joseph Gall, Carnegie Institute.)

binding sites of the HMGN proteins on nucleosomes overlap the binding site of histone H1, and it appears that HMGN proteins stimulate transcription by altering the interaction of histone H1 with nucleosomes to maintain a decondensed chromatin structure.

Histone acetylation has been correlated with transcriptionally active chromatin in a wide variety of cell types (Figure 6.32). The core histones (H2A, H2B, H3 and H4) have two domains: a histone fold domain, which is involved in interactions with other histones and in wrapping DNA around the nucleosome core particle, and an amino-terminal tail domain, which extends outside of the nucleosome. The amino-terminal tail domains are rich

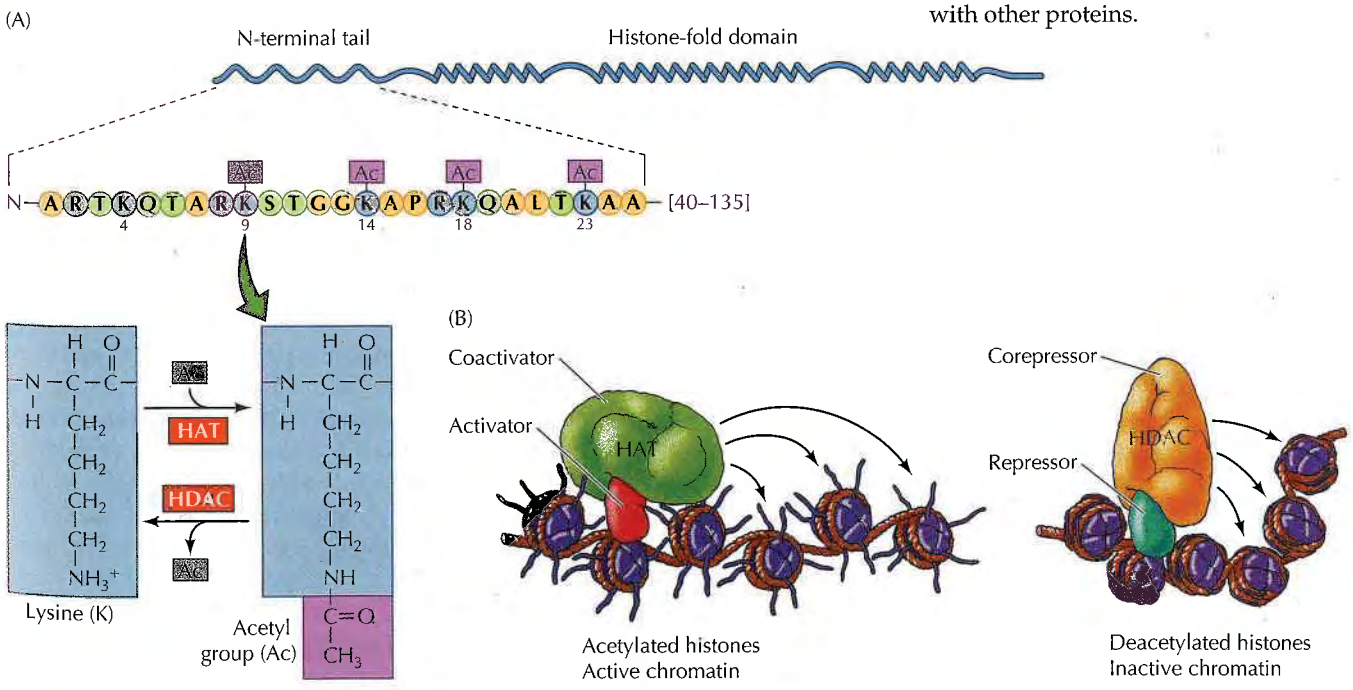


Figure 6.32 Histone acetylation
(A) The core histones have histone-fold domains, which interact with other histones and with DNA in the nucleosome, and N-terminal tails, which extend outside of the nucleosome. The N-terminal tails of the core histones (e.g., H3) are modified by the addition of acetyl groups (Ac) to the side chains of specific lysine residues. (B) Transcriptional activators and repressors are associated with coactivators and corepressors, which have histone acetyltransferase (HAT) and histone deacetylase (HDAC) activities, respectively. Histone acetylation is characteristic of actively transcribed chromatin and may weaken the binding of histones to DNA or alter their interactions with other proteins.

in lysine and can be modified by acetylation at specific lysine residues. Acetylation reduces the net positive charge of the histones, and may weaken their binding to DNA as well as altering their interactions with other proteins. Moreover, acetylation of histones has been shown to facilitate the binding of transcription factors to nucleosomal DNA, indicating that histone acetylation increases the accessibility of chromatin to DNA-binding proteins.

Studies from two groups of researchers in 1996 provided direct links between histone acetylation and transcriptional regulation by demonstrating that transcriptional activators and repressors are associated with histone acetyltransferases and deacetylases, respectively. This association was first revealed by cloning a gene encoding a histone acetyltransferase from *Tetrahymena*. Unexpectedly, the sequence of this histone acetyltransferase was closely related to a previously known yeast transcriptional coactivator called Gcn5p, which stimulates transcription in association with several different sequence-specific transcriptional activators. Further experiments revealed that Gcn5p itself has histone acetyltransferase activity, suggesting that transcriptional activation results directly from histone acetylation. These results have been extended by demonstrations that histone acetyltransferases are also associated with a number of mammalian transcriptional coactivators, as well as with the general transcription factor TFIID. Conversely, many transcriptional corepressors in both yeast and mammalian cells function as histone deacetylases, which remove the acetyl groups from histone tails. Histone acetylation is thus targeted directly by both transcriptional activators and repressors, indicating that it plays a key role in regulation of eukaryotic gene expression.

Histones are modified not only by acetylation, but also by phosphorylation of serine residues, methylation of lysine and arginine residues, and addition of ubiquitin (a small peptide discussed in Chapter 7) to lysine residues. Like acetylation, these modifications occur at specific amino acid residues in the histone tails and are associated with changes in transcriptional activity (Figure 6.33). In addition to affecting chromatin structure, it has been proposed that specific histone modifications affect gene expression by providing binding sites for other transcriptional regulatory proteins. According to this hypothesis, combinations of specific histone modifications constitute a "**histone code**" that regulates gene expression by recruiting other regulatory proteins to the chromatin template. For example, transcriptionally active chromatin is associated with several specific modifications of histone H3, including methylation of lysine-4, phosphorylation of serine-10 and acetylation of lysine-9 and lysine-14. In contrast, methylation of lysine-9 is associated with repression, and the enzyme that catalyzes methylation of H3 lysine-9 is recruited to target genes by corepressors. The methylated H3 lysine-9 residues have further been shown to serve as binding sites for proteins that induce chromatin condensation, directly linking this histone modification to transcriptional repression.

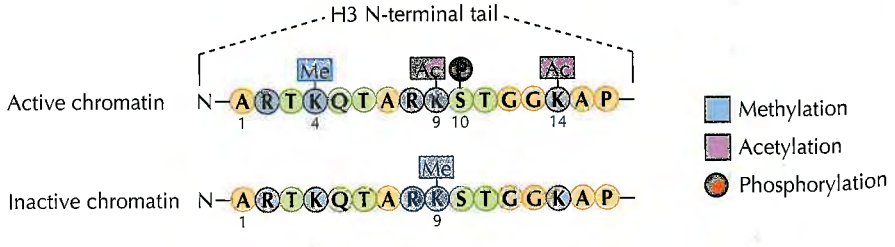


Figure 6.33 Histone methylation and phosphorylation
Transcriptional activity of chromatin is affected by methylation and phosphorylation of specific amino acid residues in histone tails, as well as by their acetylation. For example, transcriptionally active chromatin is characterized by methylation of H3 lysine-4, phosphorylation of serine-10, and acetylation of lysine-9 and lysine-14. In contrast, inactive chromatin is characterized by methylation of H3 lysine-9.

It is notable that these modifications of histone tails regulate one another, leading to the establishment of distinct patterns of histone modification that correlate with transcriptional activity. For example, phosphorylation of H3 serine-10 promotes acetylation of lysine-14 but inhibits methylation of lysine-9, leading to the establishment of a pattern of H3 modification that is characteristic of transcriptionally active chromatin. Methylation of lysine-4 also inhibits methylation of lysine-9, and vice versa, consistent with the opposing effects of methylation of these two lysine residues on transcriptional activation versus repression. The interplay between modifications of these different residues thus results in patterns of histone modification that may provide a stable regulatory code for the transcriptional activity of chromatin.

In contrast to the enzymes that regulate chromatin structure by modifying histones, **nucleosome remodeling factors** are protein complexes that alter the arrangement or structure of nucleosomes, without removing or covalently modifying the histones (Figure 6.34). One mechanism by which nucleosome remodeling factors act is to catalyze the sliding of histone octamers along the DNA molecule, thereby repositioning nucleosomes to change the accessibility of specific DNA sequences to transcription factors. Alternatively, nucleosome remodeling factors may act by inducing changes in the conformation of nucleosomes, again affecting the ability of specific DNA sequences to interact with transcriptional regulatory proteins. Like histone modifying enzymes, nucleosome remodeling factors can be recruited to DNA in association with either transcriptional activators or repressors, and can alter the arrangement of nucleosomes to either stimulate or inhibit transcription.

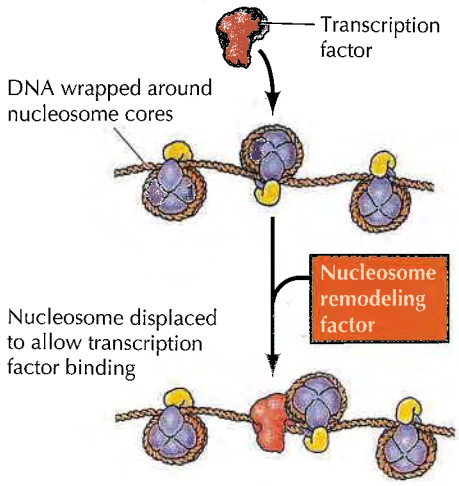


Figure 6.34 Nucleosome remodeling factors
Nucleosome remodeling factors alter the arrangement or structure of nucleosomes. For example, a nucleosome remodeling factor can facilitate the binding of transcription factors to chromatin by repositioning nucleosomes on the DNA.

The recruitment of histone modifying enzymes and nucleosome remodeling factors by transcriptional activators stimulates the initiation of transcription by altering the chromatin structure of enhancer and promoter regions. However, following the initiation of transcription, RNA polymerase is still faced with the problem of transcriptional elongation through a chromatin template. Perhaps surprisingly, the packaging of DNA in nucleosomes does not present an impassable barrier to RNA polymerase, which is able to transcribe through a nucleosome core by disrupting histone-DNA contacts. The ability of RNA polymerase to transcribe chromatin templates is facilitated by the association of HMGN proteins with the nucleosomes of actively transcribed genes, as well as by **elongation factors** that become associated with the phosphorylated C-terminal domain of RNA polymerase II when transcription is initiated (see Figure 6.14). These elongation factors recruit histone acetyltransferases, as well as acting directly to disrupt nucleosome structure during transcription.

Regulation of Transcription by Noncoding RNAs

A series of recent advances indicate that gene expression can be regulated not only by the transcriptional regulatory proteins discussed so far, but also by noncoding regulatory RNA molecules. One mode of action of noncoding regulatory RNAs is to inhibit translation by RNA interference; a phenomenon in which short double-stranded RNAs induce degradation of a homologous mRNA (see Figure 3.41). In addition, noncoding RNAs appear to play important roles in repressing transcription at some chromosomal loci by inducing histone modifications that lead to chromatin condensation and the formation of heterochromatin. Although much remains to be learned concerning their mechanism of action, noncoding RNAs clearly play important roles in regulating chromatin structure and function in eukaryotic cells.

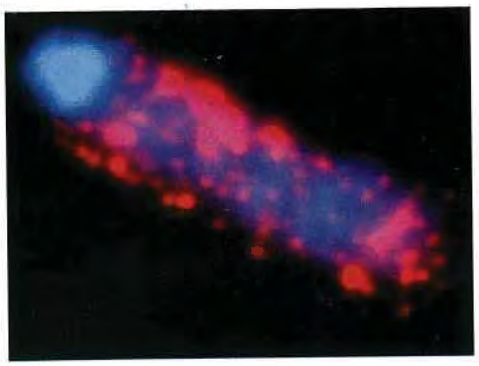


Figure 6.35 X chromosome inactivation
The inactive X chromosome (blue) is coated by *Xist* RNA (red). (From B. Panning and R. Jaenisch, 1998. *Cell* 93: 305.)

The phenomenon of **X chromosome inactivation** provides an example of the role of a noncoding RNA in regulating gene expression in mammals. In many animals, including humans, females have two X chromosomes, and males have one X and one Y chromosome. The X chromosome contains hundreds of genes that are not present on the much smaller Y chromosome (see Figure 4.29). Thus, females have twice as many copies of most X chromosome genes as males have. Despite this difference, female and male cells contain equal amounts of the proteins encoded by the majority of X chromosome genes. This results from a dosage compensation mechanism in which most of the genes on one of the two X chromosomes in female cells are inactivated by being converted to heterochromatin early in development. Consequently, only one copy of most genes located on the X chromosome are available for transcription in either female or male cells.

Although the mechanism of X chromosome inactivation is not yet fully understood, the key element appears to be a noncoding RNA transcribed from a regulatory gene, called *Xist*, on the inactive X chromosome. *Xist* RNA remains localized to the inactive X, binding to and coating this chromosome (Figure 6.35). In addition, *Xist* RNA recruits regulatory proteins that repress transcription of most genes on the inactive X. Although these proteins remain to be identified, it is clear that a principal effect of *Xist* RNA is the induction of methylation of histone H3 lysine-9, leading to chromatin condensation and conversion of the inactive X to heterochromatin.

Noncoding RNAs have also been recently shown to play a key role in transcriptional silencing and formation of heterochromatin at the centromeres of the fission yeast *Schizosaccharomyces pombe*. In this case, RNAs homologous to repeated centromeric DNA sequences act to repress transcription and induce heterochromatin formation at the centromere. As in X chromosome inactivation, methylation of histone H3 lysine-9 is an early event in formation of heterochromatin at yeast centromeres. Interestingly, the action of regulatory centromeric RNAs in *S. pombe* requires their conversion to small double-stranded RNA molecules by the cellular machinery responsible for generating the small interfering RNAs that block gene expression by targeting mRNAs for degradation.

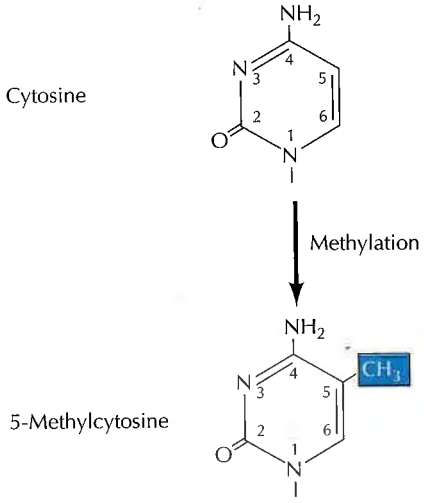


Figure 6.36 DNA methylation
A methyl group is added to the 5-carbon position of cytosine residues in DNA.

DNA Methylation

The methylation of DNA is another general mechanism by which control of transcription in vertebrates is linked to chromatin structure. Cytosine residues in vertebrate DNA can be modified by the addition of methyl groups at the 5-carbon position (Figure 6.36). DNA is methylated specifically at the C's that precede G's in the DNA chain (CpG dinucleotides). This methylation is correlated with reduced transcriptional activity of genes that contain high frequencies of CpG dinucleotides in the vicinity of their promoters. Methylation inhibits transcription of these genes by interfering with the binding of some transcriptional activators, as well as by recruiting repressors that specifically bind methylated DNA. The repressors that bind methylated DNA function as complexes with histone deacetylases, linking DNA methylation to alterations in histone acetylation and nucleosome structure.

Although DNA methylation is capable of inhibiting transcription, it generally appears that only genes that are already repressed become methylated. Rather than being the primary cause of transcriptional inactivation, DNA methylation may serve principally to stabilize and maintain gene inactivation during development. For example, genes on the inactive X

chromosome become methylated following transcriptional repression by *Xist* RNA and the methylation of histone H3 lysine-9, which may serve to target the enzymes responsible for inducing DNA methylation to inactive genes. In plants, it has also been suggested that noncoding RNAs target DNA methylation of repressed genes.

One important regulatory role of DNA methylation has been established in the phenomenon known as **genomic imprinting**, which controls the expression of some genes involved in the development of mammalian embryos. In most cases, both the paternal and maternal alleles of a gene are expressed in diploid cells. However, there are a few imprinted genes (over two dozen have been described in mice and humans) whose expression depends on whether they are inherited from the mother or from the father. In some cases, only the paternal allele of an imprinted gene is expressed, and the maternal allele is transcriptionally inactive. For other imprinted genes, the maternal allele is expressed and the paternal allele is inactive.

DNA methylation appears to play a key role in distinguishing between the paternal and maternal alleles of imprinted genes. A good example is the gene *H19*, which is transcribed only from the maternal copy (Figure 6.37). The *H19* gene is specifically methylated during the development of male, but not female, germ cells. The union of sperm and egg at fertilization therefore yields an embryo containing a methylated paternal allele and an unmethylated maternal allele of the gene. These differences in methylation are maintained following DNA replication by an enzyme that specifically methylates CpG sequences of a daughter strand that is hydrogen-bonded to a methylated parental strand (Figure 6.38). The paternal *H19* allele therefore remains methylated, and transcriptionally inactive, in embryonic cells and somatic tissues. However, the paternal *H19* allele becomes demethylated in the germ line, allowing a new pattern of methylation to be established for transmittal to the next generation.

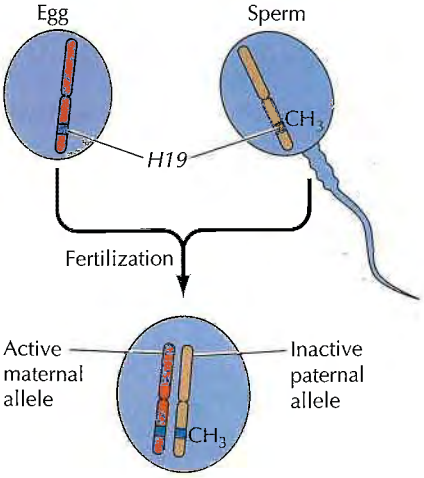


Figure 6.37 Genomic imprinting
The *H19* gene is specifically methylated during development of male germ cells. Therefore, sperm contain a methylated *H19* allele and eggs contain an unmethylated allele. Following fertilization, the methylated paternal allele remains transcriptionally inactive, and only the unmethylated maternal allele is expressed in the embryo.

RNA Processing and Turnover

Although transcription is the first and most highly regulated step in gene expression, it is usually only the beginning of the series of events required to produce a functional RNA. Most newly synthesized RNAs must be mod-

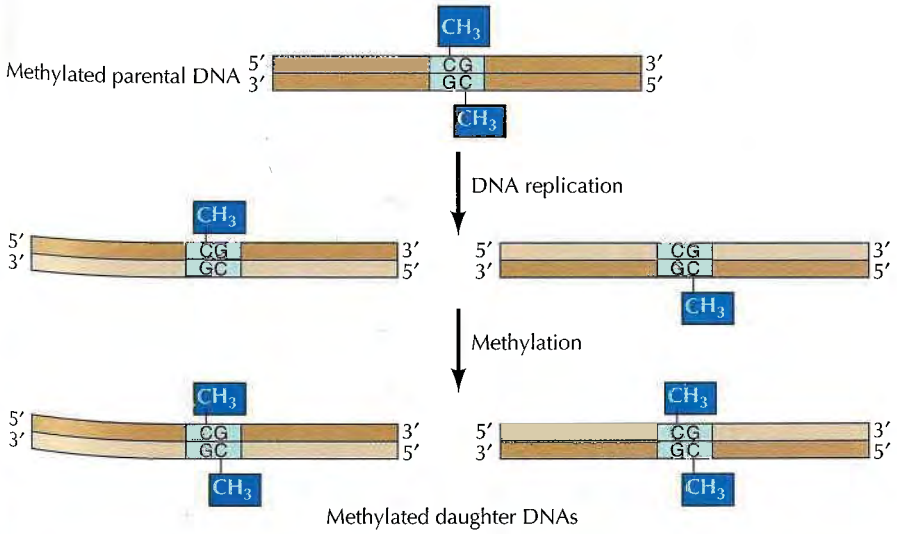


Figure 6.38 Maintenance of methylation patterns
In parental DNA, both strands are methylated at complementary CpG sequences. Following replication, only the parental strand of each daughter molecule is methylated. The newly synthesized daughter strands are then methylated by an enzyme that specifically recognizes CpG sequences opposite a methylation site.

ified in various ways to be converted to their functional forms. Bacterial mRNAs are an exception; they are used immediately as templates for protein synthesis while still being transcribed. However, the primary transcripts of both rRNAs and tRNAs must undergo a series of processing steps in prokaryotic as well as eukaryotic cells. Primary transcripts of eukaryotic mRNAs similarly undergo extensive modifications, including the removal of introns by splicing, before they are transported from the nucleus to the cytoplasm to serve as templates for protein synthesis. Regulation of these processing steps provides an additional level of control of gene expression, as does regulation of the rates at which different mRNAs are subsequently degraded within the cell.

Processing of Ribosomal and Transfer RNAs

The basic processing of ribosomal and transfer RNAs in prokaryotic and eukaryotic cells is similar, as might be expected given the fundamental roles of these RNAs in protein synthesis. As discussed previously, eukaryotes have four species of ribosomal RNAs (see Table 6.1), three of which (the 28S, 18S, and 5.8S rRNAs) are derived by cleavage of a single long precursor transcript, called a **pre-rRNA** (Figure 6.39). Prokaryotes have three ribosomal RNAs (23S, 16S, and 5S), which are equivalent to the 28S, 18S, and 5S rRNAs of eukaryotic cells and are also formed by the processing of a single pre-rRNA transcript. The only rRNA that is not processed extensively is the 5S rRNA in eukaryotes, which is transcribed from a separate gene.

Prokaryotic and eukaryotic pre-rRNAs are processed in several steps. Initial cleavages of bacterial pre-rRNA yield separate precursors for the three individual rRNAs; these are then further processed by secondary cleavages to the final products. In eukaryotic cells, pre-rRNA is first cleaved at a site adjacent to the 5.8S rRNA on its 5′ side, yielding two separate precursors that contain the 18S and the 28S + 5.8S rRNAs, respectively. Further cleavages then convert these to their final products, with the 5.8S rRNA becoming

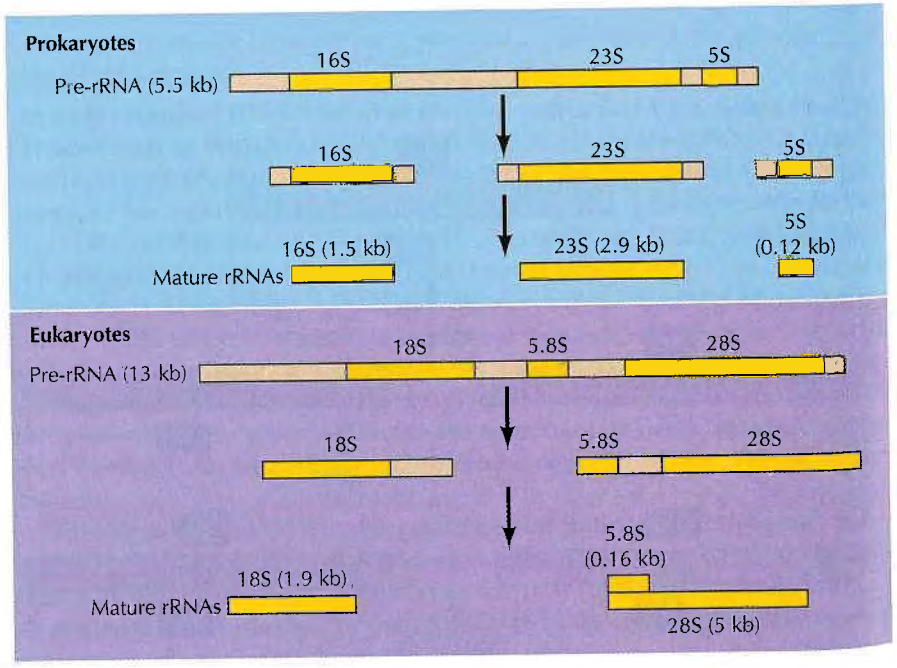


Figure 6.39 Processing of ribosomal RNAs
Prokaryotic cells contain three rRNAs (16S, 23S, and 5S), which are formed by cleavage of a pre-rRNA transcript. Eukaryotic cells (e.g., human cells) contain four rRNAs. One of these (5S rRNA) is transcribed from a separate gene; the other three (18S, 28S, and 5.8S) are derived from a common pre-rRNA. Following cleavage, the 5.8S rRNA (which is unique to eukaryotes) becomes hydrogen-bonded to 28S rRNA.

hydrogen-bonded to the 28S molecule. In addition to these cleavages, rRNA processing involves the addition of methyl groups to the bases and sugar moieties of specific nucleotides. Processing of rRNA takes place within the nucleolus of eukaryotic cells, and will be discussed in detail in Chapter 8.

Like rRNAs, tRNAs in both bacteria and eukaryotes are synthesized as longer precursor molecules (**pre-tRNAs**), some of which contain several individual tRNA sequences (Figure 6.40). In bacteria, some tRNAs are included in the pre-rRNA transcripts. The processing of the 5′ end of pre-tRNAs involves cleavage by an enzyme called **RNase P**, which is of special interest because it is a prototypical model of a reaction catalyzed by an RNA enzyme. RNase P consists of RNA and protein molecules, both of which are required for maximal activity. In 1983 Sidney Altman and his colleagues demonstrated that the isolated RNA component of RNase P is itself capable of catalyzing pre-tRNA cleavage. These experiments established that RNase P is a **ribozyme**—an enzyme in which RNA rather than protein is responsible for catalytic activity.

The 3′ end of tRNAs is generated by the action of a conventional protein RNase, but the processing of this end of the tRNA molecule also involves an unusual activity: the addition of a CCA terminus. All tRNAs have the sequence CCA at their 3′ ends. This sequence is the site of amino acid attachment, so it is required for tRNA function during protein synthesis. The CCA terminus is encoded in the DNA of some tRNA genes, but in others it is not, instead being added as an RNA processing step by an enzyme that recognizes and adds CCA to the 3′ end of all tRNAs that lack this sequence.

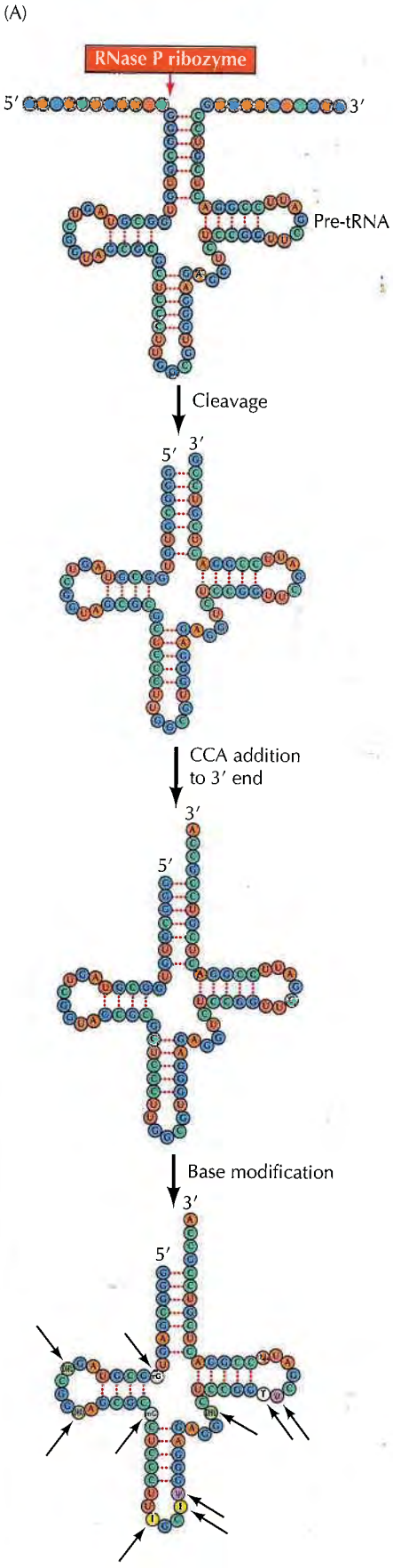
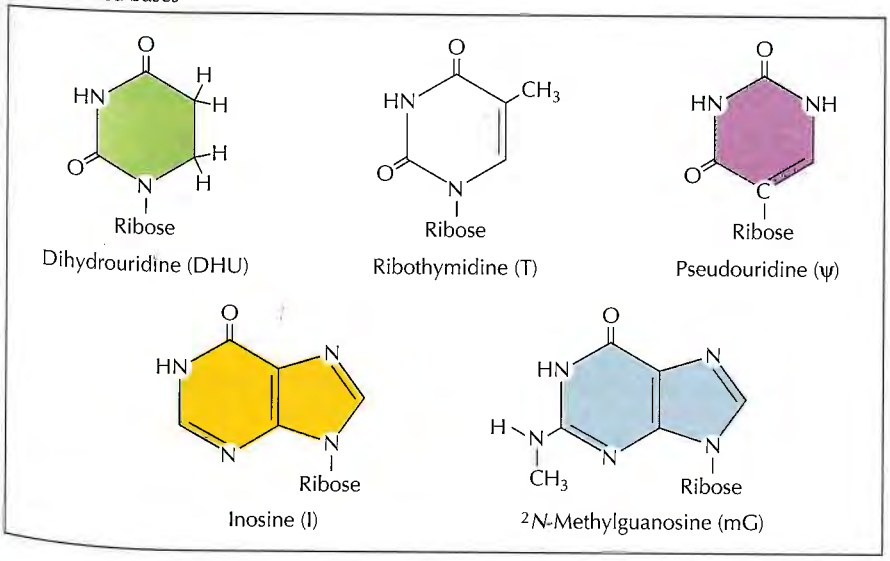


Figure 6.40 Processing of transfer RNAs
(A) Transfer RNAs are derived from pre-tRNAs, some of which contain several individual tRNA molecules. Cleavage at the 5′ end of the tRNA is catalyzed by the RNase P ribozyme; cleavage at the 3′ end is catalyzed by a conventional protein RNase. A CCA terminus is then added to the 3′ end of many tRNAs in a posttranscriptional processing step. Finally, some bases are modified at characteristic positions in the tRNA molecule. In this example, these modified nucleosides include dihydrouridine (DHU), methylguanosine (mG), inosine (I), ribothymidine (T), and pseudouridine (ψ). (B) Structure of modified bases. Ribothymidine, dihydrouridine, and pseudouridine are formed by modification of uridines in tRNA. Inosine and methylguanosine are formed by the modification of guanosines.

(B) Modified bases



Another unusual aspect of tRNA processing is the extensive modification of bases in tRNA molecules. Approximately 10% of the bases in tRNAs are altered to yield a variety of modified nucleotides at specific positions in tRNA molecules (see Figure 6.40). The functions of most of these modified bases are unknown, but some play important roles in protein synthesis by altering the base-pairing properties of the tRNA molecule (see Chapter 7).

Some pre-tRNAs, as well as pre-rRNAs in a few organisms, contain introns that are removed by splicing. In contrast to other splicing reactions, which (as discussed in the next section) involve the activities of catalytic RNAs, tRNA splicing is mediated by conventional protein enzymes. An endonuclease cleaves the pre-tRNA at the splice sites to excise the intron, followed by joining of the exons to form a mature tRNA molecule.

Processing of mRNA in Eukaryotes

In contrast to the processing of ribosomal and transfer RNAs, the processing of messenger RNAs represents a major difference between prokaryotic and eukaryotic cells. In bacteria, ribosomes have immediate access to mRNA and translation begins on the nascent mRNA chain while transcription is still in progress. In eukaryotes, mRNA synthesized in the nucleus must first be transported to the cytoplasm before it can be used as a template for protein synthesis. Moreover, the initial products of transcription in eukaryotic cells (**pre-mRNAs**) are extensively modified before export from the nucleus. The processing of mRNA includes modification of both ends of the intial transcript, as well as the removal of introns from its middle (Figure 6.41). Rather than occurring as independent events following synthesis

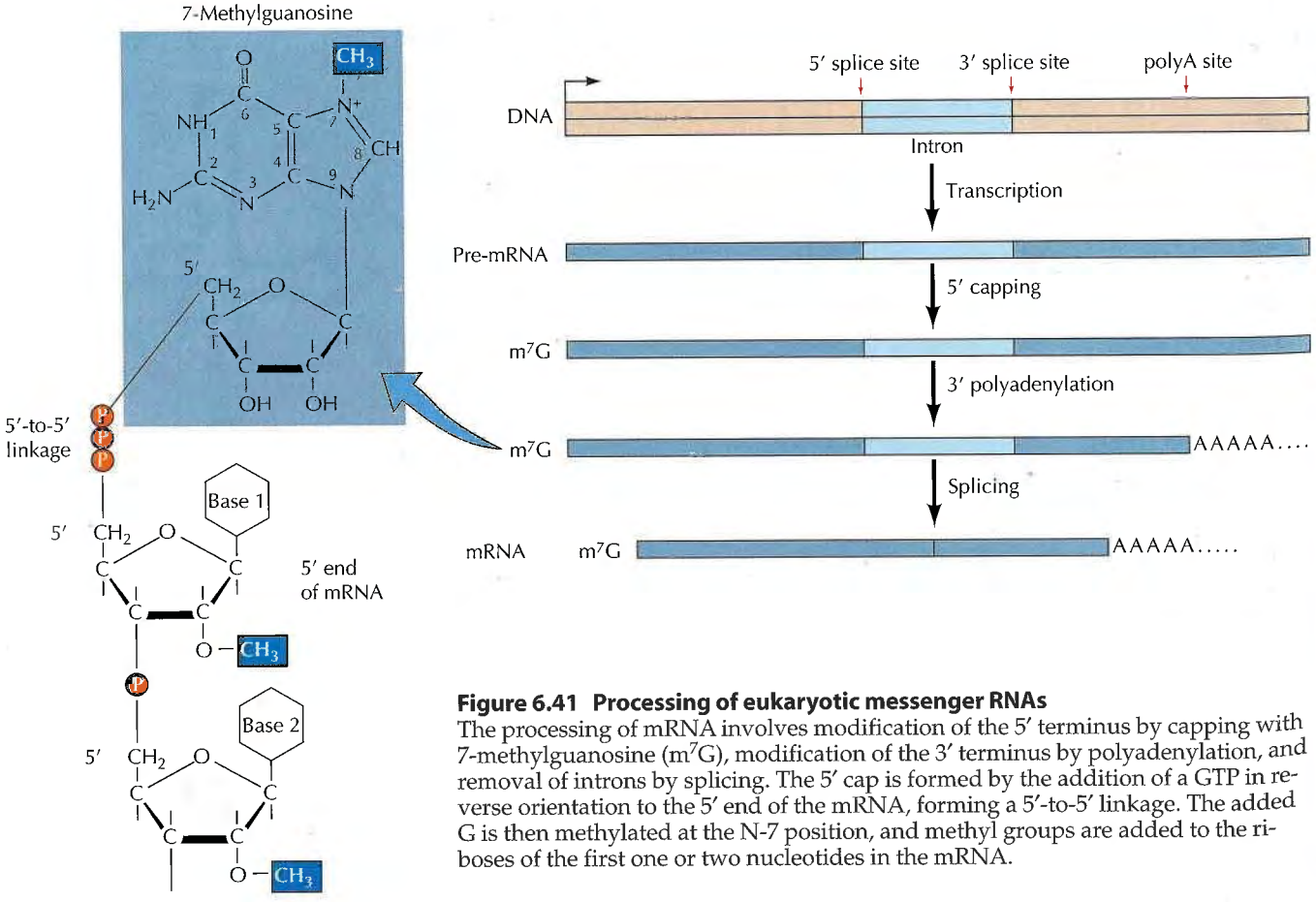


Figure 6.41 Processing of eukaryotic messenger RNAs
The processing of mRNA involves modification of the 5′ terminus by capping with 7-methylguanosine (m^7G), modification of the 3′ terminus by polyadenylation, and removal of introns by splicing. The 5′ cap is formed by the addition of a GTP in reverse orientation to the 5′ end of the mRNA, forming a 5′-to-5′ linkage. The added G is then methylated at the N-7 position, and methyl groups are added to the riboses of the first one or two nucleotides in the mRNA.

of a pre-mRNA, these processing reactions are coupled to transcription, so that mRNA synthesis and processing are closely coordinated steps in gene expression. The C-terminal domain (CTD) of RNA polymerase II plays a key role in coordinating these processes by serving as a binding site for the enzyme complexes involved in mRNA processing. The association of these processing enzymes with the CTD of polymerase II accounts for their specificity in processing mRNAs; polymerases I and III lack a CTD, so their transcripts are not processed by the same enzyme complexes.

The first step in mRNA processing is the modification of the 5′ end of the transcript by the addition of a structure called a **7-methylguanosine cap**. The enzymes responsible for capping are recruited to the phosphorylated CTD following initiation of transcription, and the cap is added after transcription of the first 20-30 nucleotides of the RNA. Capping is initiated by the addition of a GTP in reverse orientation to the 5′ terminal nucleotide of the RNA. Then methyl groups are added to this G residue and to the ribose moieties of one or two 5′ nucleotides of the RNA chain. The 5′ cap stabilizes the RNA, as well as aligning eukaryotic mRNAs on the ribosome during translation (see Chapter 7).

The 3′ end of most eukaryotic mRNAs is defined not by termination of transcription, but by cleavage of the primary transcript and addition of a **poly-A tail**—a processing reaction called **polyadenylation** (Figure 6.42). The signals for polyadenylation include a highly conserved hexanucleotide (AAUAAA in mammalian cells), which is located 10 to 30 nucleotides upstream of the site of polyadenylation, and a G-U rich downstream sequence element. In addition, some genes have a U-rich sequence element upstream of the AAUAAA. These sequences are recognized by a complex of proteins, including an endonuclease that cleaves the RNA chain and a separate poly-A polymerase that adds a poly-A tail of about 200 nucleotides to the transcript. These processing enzymes are associated with the phosphorylated CTD of RNA polymerase II, and may travel with the polymerase all the way from the transcription initiation site. Cleavage and polyadenylation signal the termination of transcription, which usually occurs several hundred nucleotides downstream of the site of poly-A addition.

Almost all mRNAs in eukaryotes are polyadenylated, and poly-A tails have been shown to regulate both translation and mRNA stability. In addition, polyadenylation plays an important regulatory role in early development, where changes in the length of poly-A tails control mRNA translation. For example, many mRNAs are stored in unfertilized eggs in an untranslated form with short poly-A tails (usually 30 to 50 nucleotides

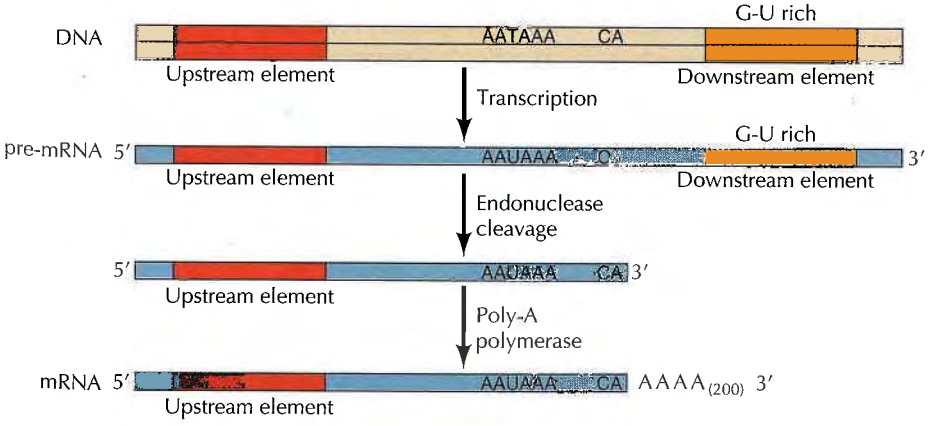


Figure 6.42 Formation of the 3′ ends of eukaryotic mRNAs
Polyadenylation signals in mammalian cells consist of the hexanucleotide AAUAAA in addition to upstream and downstream (G-U rich) elements. An endonuclease cleaves the pre-mRNA 10 to 30 nucleotides downstream of the AAUAAA, usually at a CA sequence. Poly-A polymerase then adds a poly-A tail consisting of about 200 A's to the 3′ end of the RNA.

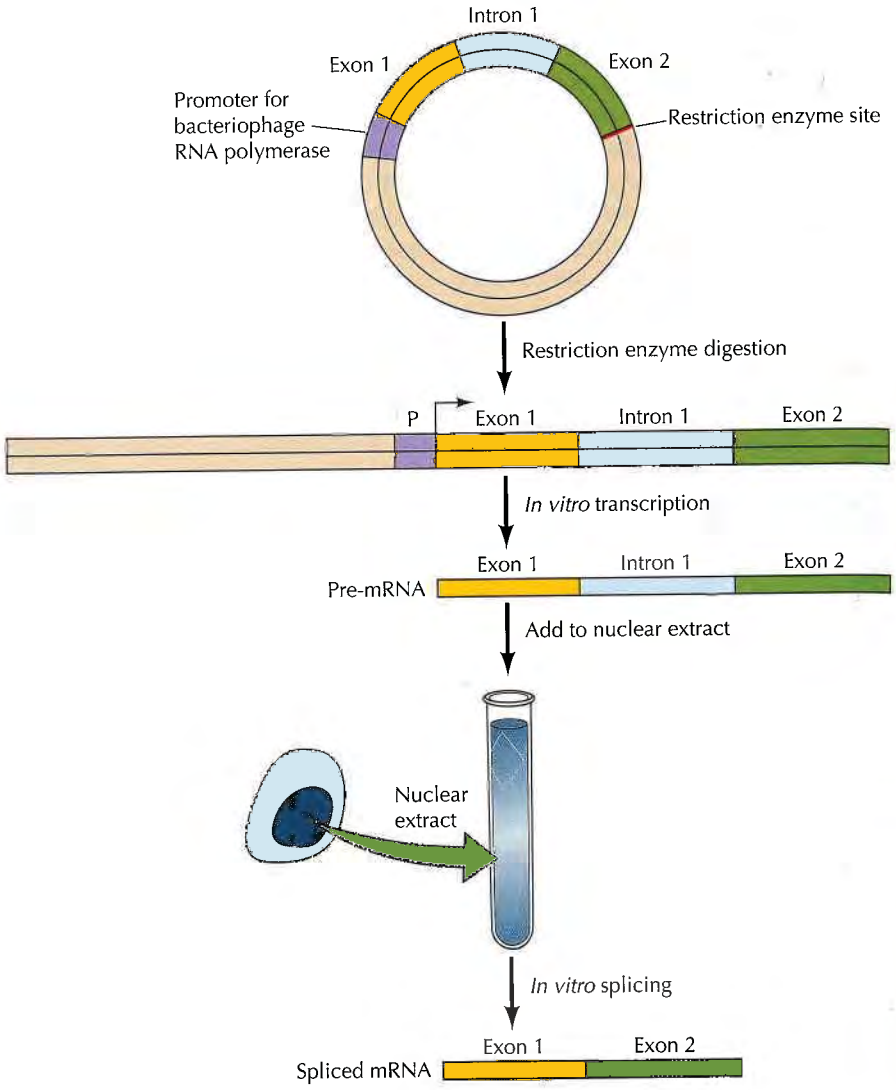


Figure 6.43 *In vitro* splicing
A gene containing an intron is cloned downstream of a promoter (P) recognized by a bacteriophage RNA polymerase. The plasmid is digested with a restriction enzyme that cleaves at the 3′ end of the inserted gene to yield a linear DNA molecule. This DNA is then transcribed *in vitro* with the bacteriophage polymerase to produce pre-mRNA. Splicing reactions can then be studied *in vitro* by addition of this pre-mRNA to nuclear extracts of mammalian cells.

long). Fertilization stimulates the lengthening of the poly-A tails of these stored mRNAs, which in turn activates their translation and the synthesis of proteins required for early embryonic development.

The most striking modification of pre-mRNAs is the removal of introns by splicing. As discussed in Chapter 4, the coding sequences of most eukaryotic genes are interrupted by noncoding sequences (introns) that are precisely excised from the mature mRNA. In mammals, most genes contain multiple introns, which typically account for about ten times more pre-mRNA sequences than the exons do. The unexpected discovery of introns in 1977 generated an active research effort directed toward understanding the mechanism of splicing, which had to be highly specific to yield functional mRNAs. Further studies of splicing have not only illuminated new mechanisms of gene regulation; they have also revealed novel catalytic activities of RNA molecules.

Splicing Mechanisms

The key to understanding pre-mRNA splicing was the development of *in vitro* systems that efficiently carried out the splicing reaction (Figure 6.43). Pre-mRNAs were synthesized *in vitro* by the cloning of structural genes (with their introns) adjacent to promoters for bacteriophage RNA poly-

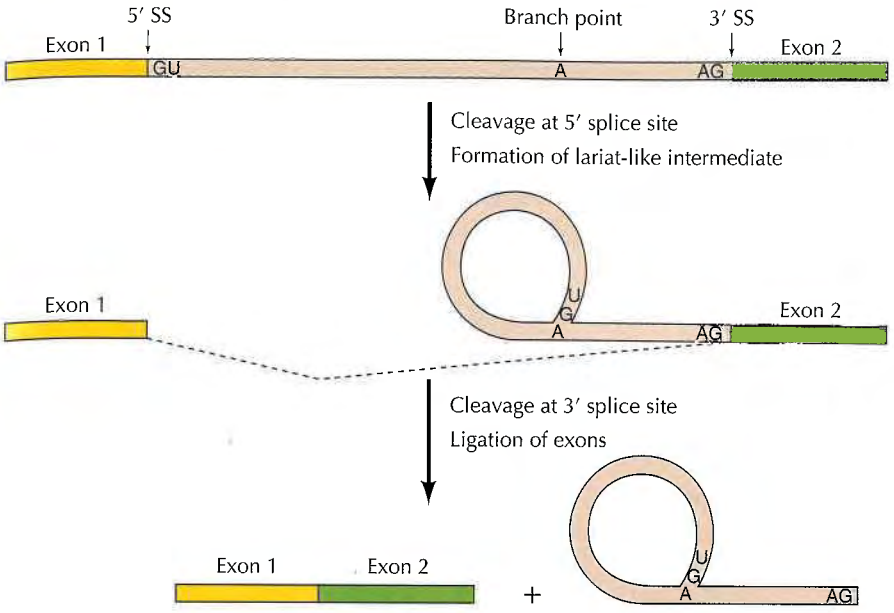


Figure 6.44 Splicing of pre-mRNA
The splicing reaction proceeds in two steps. The first step involves cleavage at the 5′ splice site (SS) and joining of the 5′ end of the intron to an A within the intron (the branch point). This reaction yields a lariat-like intermediate, in which the intron forms a loop. The second step is cleavage at the 3′ splice site and simultaneous ligation of the exons, resulting in excision of the intron as a lariat-like structure.

merases, which could readily be isolated in large quantities. Transcription of these plasmids could then be used to prepare large amounts of pre-mRNAs that, when added to nuclear extracts of mammalian cells, were found to be correctly spliced. As with transcription, the use of such *in vitro* systems has allowed splicing to be analyzed in much greater detail than would have been possible in intact cells.

Analysis of the reaction products and intermediates formed *in vitro* revealed that pre-mRNA splicing proceeds in two steps (Figure 6.44). First, the pre-mRNA is cleaved at the 5′ splice site, and the 5′ end of the intron is joined to an adenine nucleotide within the intron (near its 3′ end). In this step an unusual bond forms between the 5′ end of the intron and the 2′ hydroxyl group of the adenine nucleotide. The resulting intermediate is a lariat-like structure, in which the intron forms a loop. The second step in splicing then proceeds with simultaneous cleavage at the 3′ splice site and ligation of the two exons. The intron is thus excised as a lariat-like structure, which is then linearized and degraded within the nucleus of intact cells.

These reactions define three critical sequence elements of pre-mRNAs: sequences at the 5′ splice site, sequences at the 3′ splice site, and sequences within the intron at the branch point (the point at which the 5′ end of the intron becomes ligated to form the lariat-like structure) (see Figure 6.44). Pre-mRNAs contain similar consensus sequences at each of these positions, allowing the splicing apparatus to recognize pre-mRNAs and carry out the cleavage and ligation reactions involved in the splicing process.

Biochemical analysis of nuclear extracts has revealed that splicing takes place in large complexes, called **spliceosomes**, composed of proteins and RNAs. The RNA components of the spliceosome are five types of **small nuclear RNAs (snRNAs)** called U1, U2, U4, U5, and U6. These snRNAs, which range in size from approximately 50 to nearly 200 nucleotides, are

KEY EXPERIMENT

The Discovery of snRNPs

Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus

Michael R. Lerner and Joan A. Steitz
Yale University, New Haven, Connecticut

Proceedings of the National Academy of Sciences, USA, 1979, Volume 76, pages 5495-5499

Joan Steitz

The Context

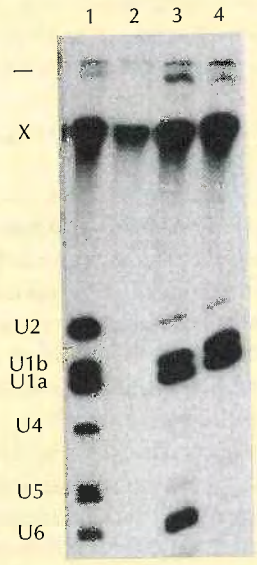
The discovery of introns in 1977 implied that a totally unanticipated processing reaction was required to produce mRNA in eukaryotic cells. Introns had to be precisely excised from pre-mRNA, followed by the joining of exons to yield a mature mRNA molecule. Given the unexpected nature of pre-mRNA splicing, understanding the mechanism of the splicing reaction captivated the attention of many molecular biologists. One of the major steps in elucidating this mechanism was the discovery of snRNPs and their involvement in pre-mRNA splicing.

Small nuclear RNAs were first identified in eukaryotic cells in the late 1960s. However, the function of snRNAs remained unknown. In this 1979 paper, Michael Lerner and Joan Steitz demonstrated that the most abundant snRNAs were present as RNA-protein complexes called snRNPs. In addition, they provided the first suggestion that these RNA-protein complexes might function in pre-mRNA splicing. This identification of snRNPs led to a variety of experiments that confirmed their roles and elucidated the mechanism by which pre-mRNA splicing takes place.

The Experiments

The identification of snRNPs was based on the use of antisera from patients with systemic lupus erythematosus, an autoimmune disease in which patients produce antibodies against their own normal cell constituents. Many of the antibodies produced by systemic lupus erythematosus patients are directed against components of the nucleus, including DNA, RNA, and histones. The discovery of snRNPs arose from studies in which Lerner and Steitz sought to characterize two antigens, called ribonucleoprotein (RNP) and Sm, that were recognized by antibodies from systemic lupus erythematosus patients. Indirect data suggested that RNP consisted of both protein and RNA, as its name implies, but neither RNP nor Sm had been characterized at the molecular level.

To identify possible RNA components of the RNP and Sm antigens, nuclear RNAs of mouse cells were radiolabeled with ^{32}P and immunoprecipitated with antisera from different systemic lupus erythematosus patients (see Figure 3.30). Six specific species of snRNAs were found to be selectively immunoprecipitated by antisera from different patients, but not by serum from a normal control patient (see figure). Anti-Sm serum immunoprecipitated all six of these snRNAs, which were designated U1a, U1b, U2, U4, U5, and U6. Anti-RNP serum immunoprecipitated only U1a and U1b, and serum from a third patient (which had been characterized as mostly anti-RNP) immunoprecipitated U1a, U1b, and U6. The immunoprecipitated snRNAs were further characterized by sequence analysis, which demonstrated that U1a, U1b, and U2 were identical to the most abundant snRNAs previously reported in mammalian nuclei, with U1a and U1b representing sequence variants of a single species of U1 snRNA present in human cells. In contrast, the U4, U5, and U6 snRNAs were newly identified by Lerner and Steitz in these experiments.



Immunoprecipitation of snRNAs with antisera from systemic lupus erythematosus patients. Lane 1, anti-Sm; lane 2,normal control serum; lane 3, antiserum recognizing primarily the RNP antigen; lane 4, anti-RNP. Note that a nonspecific RNA designated X is present in all immunoprecipitates, including the control.

Importantly, the immunoprecipitation of these snRNAs demonstrated that they were components of RNA-protein complexes. The anti-Sm serum, which immunoprecipitated all six of the snRNAs, had previously been shown to be directed against a protein antigen. Similarly, protein was known to be required for antigen recognition by anti-RNP serum. More-

over, Lerner and Steitz showed that none of the snRNAs could be immunoprecipitated if protein was first removed by extraction of the RNAs with phenol. Further analysis of cells in which proteins had been radiolabeled with ^{35}S-methionine identified seven prominent nuclear proteins that were immunoprecipitated along with the snRNAs by anti-Sm and anti-RNP sera. These data therefore indicated that each of the six snRNAs was present in an snRNP complex with specific nuclear proteins.

The Impact

The finding that snRNAs were components of snRNPs that were recognized by specific antisera opened a new approach to studying snRNA function. Lerner and Steitz noted that a "most intriguing" possible role for snRNAs might be in pre-mRNA splicing, and pointed out that sequences near the 5′ terminus of U1 snRNA were complementary to splice sites.

Steitz and her colleagues then proceeded with a series of experiments that established the critical involvement of snRNPs in splicing. These studies included more extensive sequence analysis that demonstrated the complementarity of conserved 5′ sequences of U1 snRNA to the consensus sequences of 5′ splice sites, suggesting that U1 functioned in 5′ splice site recognition. In addition, antisera against snRNPs were used to demonstrate that U1 was required for pre-mRNA splicing both in isolated nuclei and in *in vitro* splicing extracts. Further studies have gone on to show that the snRNAs themselves play critical roles not only in the identification of splice sites, but also as catalysts of the splicing reaction. The initial discovery that snRNAs were components of snRNPs that could be recognized by specific antisera thus opened the door to understanding the mechanism of pre-mRNA splicing.

complexed with six to ten protein molecules to form **small nuclear ribonucleoprotein particles (snRNPs)**, which play central roles in the splicing process. The U1, U2, and U5 snRNPs each contain a single snRNA molecule, whereas U4 and U6 snRNAs are complexed to each other in a single snRNP.

The first step in spliceosome assembly is the binding of U1 snRNP to the 5′ splice site of pre-mRNA (Figure 6.45). This recognition of 5′ splice sites involves base pairing between the 5′ splice site consensus sequence and a complementary sequence at the 5′ end of U1 snRNA (Figure 6.46). U2 snRNP then binds to the branch point, by similar complementary base pairing between U2 snRNA and branch point sequences. A preformed complex consisting of U4/U6 and U5 snRNPs is then incorporated into the spliceosome, with U5 binding to sequences upstream of the 5′ splice site. The splicing reaction is then accompanied by rearrangements of the snRNAs. Prior to the first reaction step (formation of the lariat-like intermediate, see Figure 6.44), U6 dissociates from U4 and displaces U1 at the 5′ splice site. U5 then binds to sequences at the 3′ splice site, followed by excision of the intron and ligation of the exons.

Not only do the snRNAs recognize consensus sequences at the branch points and splice sites of pre-mRNAs; they also catalyze the splicing reaction directly. The catalytic role of RNAs in splicing was demonstrated by the discovery that some RNAs are capable of **self-splicing**; that is, they can catalyze the removal of their own introns in the absence of other protein or RNA factors. Self-splicing was first described by Tom Cech and his colleagues during studies of the 28S rRNA of the protozoan *Tetrahymena*. This RNA contains an intron of approximately 400 bases that is precisely removed following incubation of the pre-rRNA in the absence of added proteins. Further studies have revealed that splicing is catalyzed by the intron, which acts as a ribozyme to direct its own excision from the pre-rRNA molecule. The discovery of self-splicing of *Tetrahymena* rRNA, together with the studies of RNase P already discussed, provided the first demonstrations of the catalytic activity of RNA.

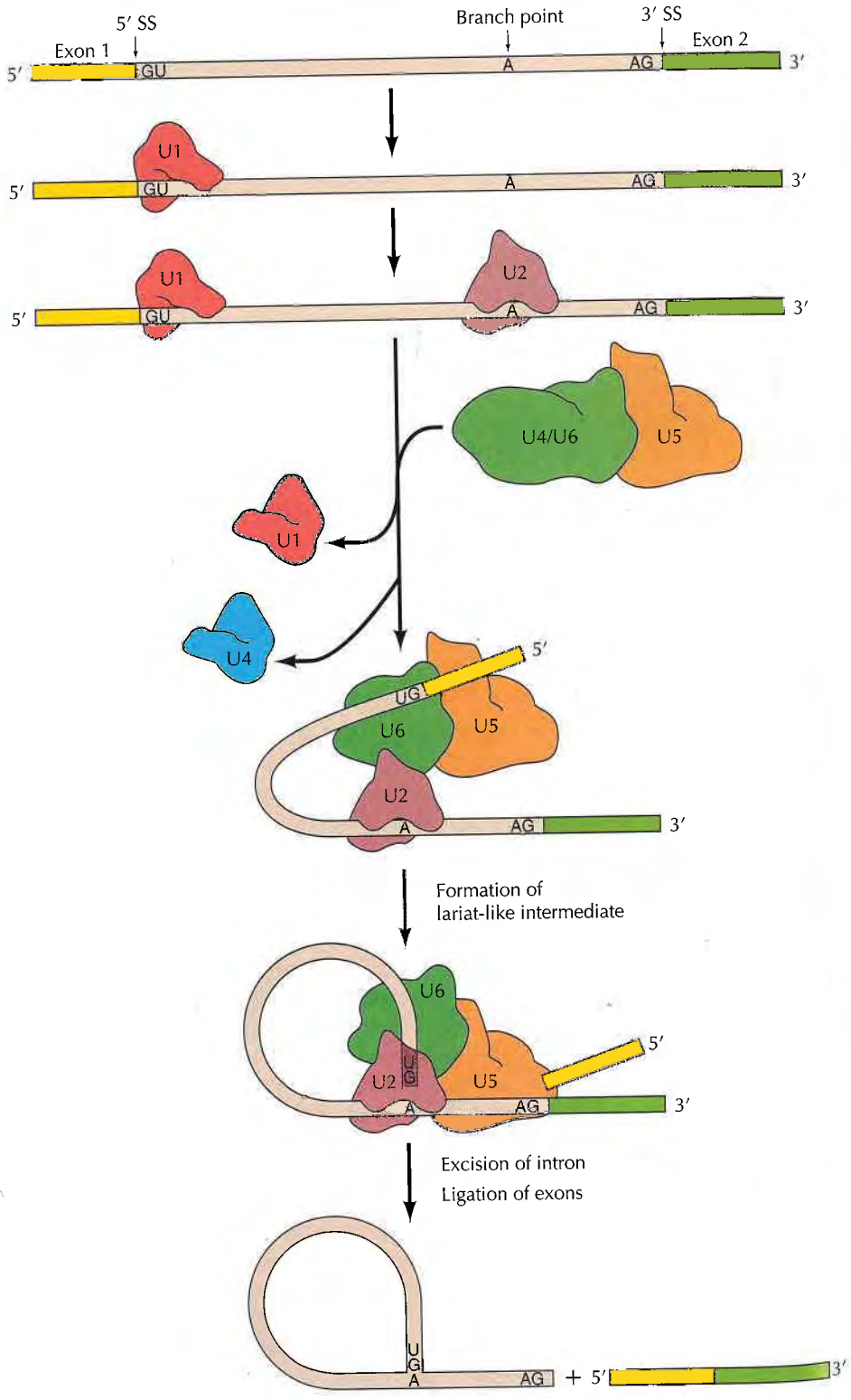


Figure 6.45 Assembly of the spliceosome
The first step in spliceosome assembly is the binding of U1 snRNP to the 5′ splice site (SS), followed by the binding of U2 snRNP to the branch point. A preformed complex consisting of U4/U6 and U5 snRNPs then enters the spliceosome. U5 binds to sequences upstream of the 5′ splice site, and U6 dissociates from U4 and displaces U1 prior to formation of the lariat-like intermediate. U5 then binds to the 3′ splice site, followed by excision of the intron and ligation of the exons.

Additional studies have revealed self-splicing RNAs in mitochondria, chloroplasts, and bacteria. These self-splicing RNAs are divided into two classes on the basis of their reaction mechanisms (Figure 6.47). The first step in splicing for group I introns (e.g., *Tetrahymena* pre-rRNA) is cleavage at the 5′ splice site mediated by a guanosine cofactor. The 3′ end of the free

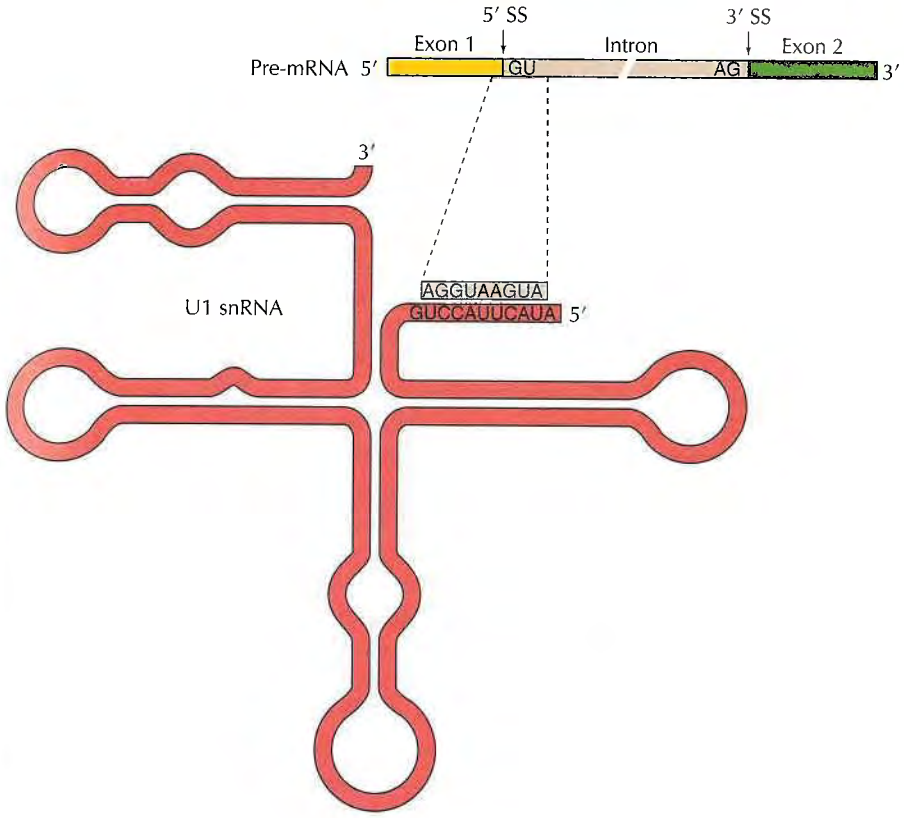
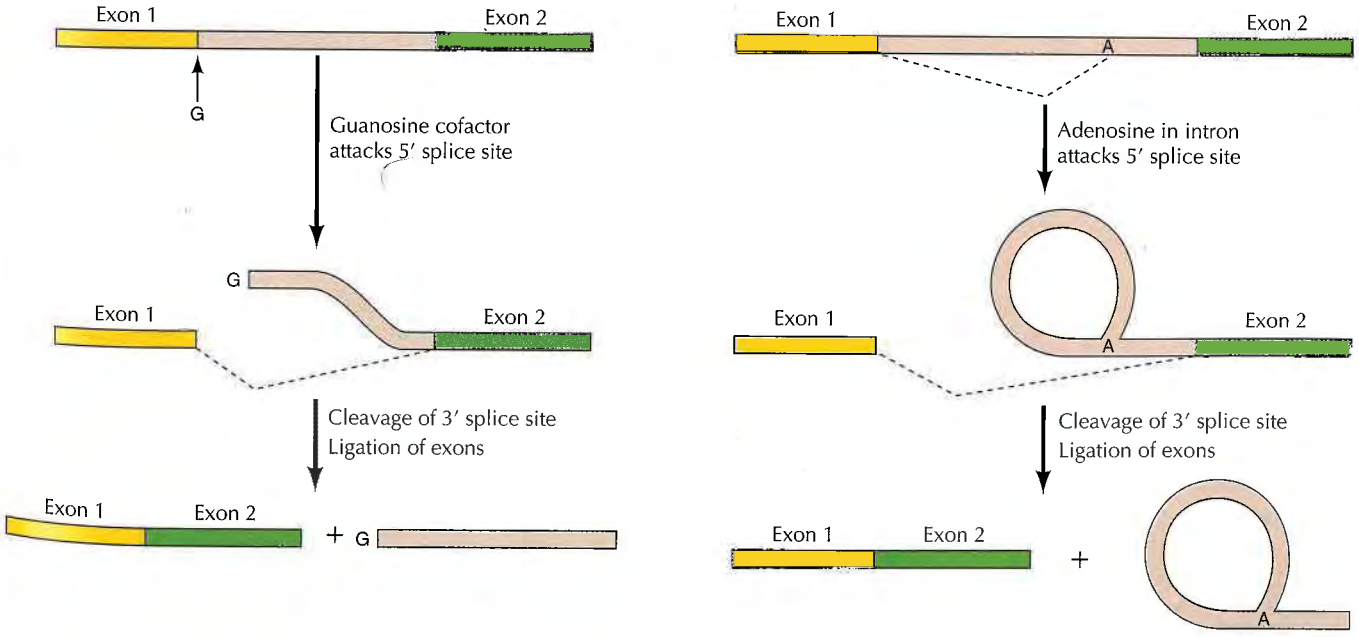


Figure 6.46 Binding of U1 snRNA to the 5′ splice site
The 5′ terminus of U1 snRNA binds to consensus sequences at 5′ splice sites by complementary base pairing.

exon then reacts with the 3′ splice site to excise the intron as a linear RNA. In contrast, the self-splicing reactions of group II introns (e.g., some mitochondrial pre-mRNAs) closely resemble those characteristic of nuclear pre-mRNA splicing, in which cleavage of the 5′ splice site results from attack by an adenosine nucleotide in the intron. As with pre-mRNA splicing, the result is a lariat-like intermediate, which is then excised.

Figure 6.47 Self-splicing introns
Group I and group II self-splicing introns are distinguished by their reaction mechanisms. In group I introns, the first step in splicing is cleavage of the 5′ splice site by reaction with a guanosine cofactor. The result is a linear intermediate with a G added to the 5′ end of the intron. In group II introns (as in pre-mRNA splicing), the first step is cleavage of the 5′ splice site by reaction with an A within the intron, forming a lariat-like intermediate. In both cases, the second step is simultaneous cleavage of the 3′ splice site and ligation of the exons.



The similarity between spliceosome-mediated pre-mRNA splicing and self-splicing of group II introns strongly suggested that the active catalytic components of the spliceosome were RNAs rather than proteins. In particular, these similarities suggested that pre-mRNA splicing was catalyzed by the snRNAs of the spliceosome. Continuing studies of pre-mRNA splicing have provided clear support for this view, including the demonstration that U2 and U6 snRNAs, in the absence of proteins, can catalyze the first step in pre-mRNA splicing. Pre-mRNA splicing is thus considered to be an RNA-based reaction, catalyzed by spliceosome snRNAs acting analogously to group II self-splicing introns. Within the cell, protein components of the snRNPs are also required, however, and participate in both assembly of the spliceosome and the splicing reaction.

A number of protein splicing factors that are not snRNP components also play critical roles in spliceosome assembly, particularly in identification of the correct splice sites in pre-mRNAs (Figure 6.48). Mammalian pre-mRNAs typically contain multiple short exons (an average of 150 nucleotides in humans) separated by much larger introns (average of 3,500 nucleotides). Introns frequently contain many sequences that resemble splice sites, so the splicing machinery must be able to identify the appropriate 5′ and 3′ splice sites at intron/exon boundaries to produce a functional mRNA. Splicing factors serve to direct spliceosomes to the correct splice sites by binding to specific RNA sequences within exons and then recruiting U1 and U2 snRNPs to the appropriate sites on pre-mRNA by protein-protein interactions. In addition, splicing factors couple splicing to transcription by associating with the phosphorylated CTD of RNA polymerase II. This anchoring of the splicing machinery to RNA polymerase is thought to be important in ensuring that exons are joined in the correct order as the pre-mRNA is synthesized.

Alternative Splicing

The central role of splicing in the processing of pre-mRNA opens the possibility of regulation of gene expression by control of the splicing machinery. Since most pre-mRNAs contain multiple introns, different mRNAs can be produced from the same gene by different combinations of 5′ and 3′ splice sites. The possibility of joining exons in varied combinations provides a novel means of controlling gene expression by generating multiple mRNAs (and therefore multiple proteins) from the same pre-mRNA. This process, called **alternative splicing**, occurs frequently in genes of complex eukaryotes. For example, it is estimated that alternative splicing can result in the production of three or more mRNAs from the average mammalian gene, considerably increasing the diversity of proteins that can be encoded by the estimated 30,000–40,000 genes in mammalian genomes. Because patterns of alternative splicing can vary in different tissues, alternative splicing pro-

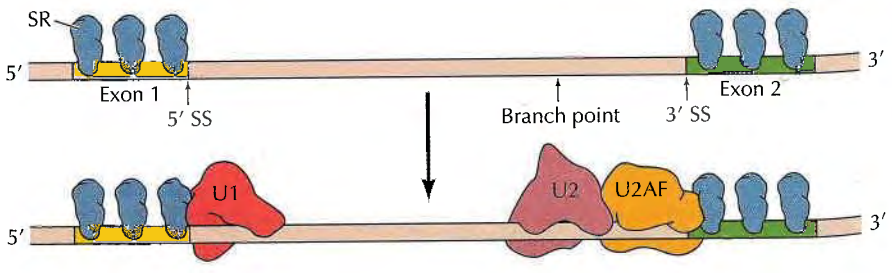


Figure 6.48 Role of splicing factors in spliceosome assembly
Splicing factors (SR proteins) bind to specific sequences within exons. The SR proteins recruit U1 snRNP to the 5′ splice site and an additional splicing factor (U2AF) to the 3′ splice site. U2AF then recruits U2 snRNP to the branchpoint.

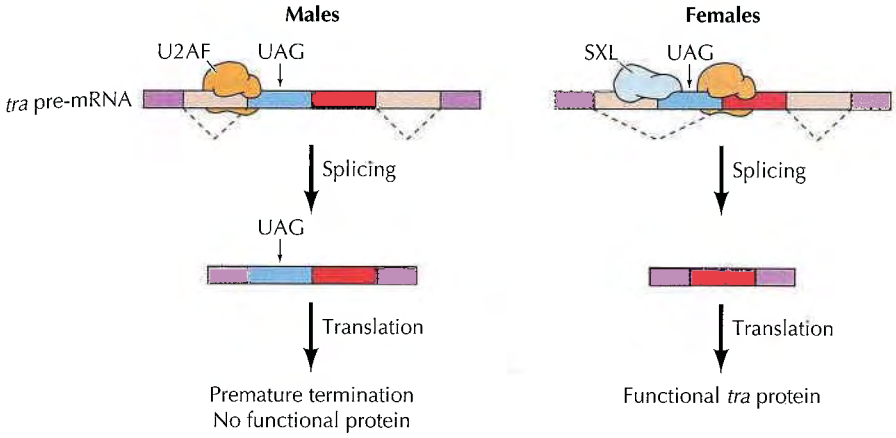


Figure 6.49 Alternative splicing in *Drosophila* sex determination
Alternative splicing of *transformer (tra)* mRNA is regulated by the SXL protein, which is only expressed in female flies. In males, the first exon of *tra* mRNA is joined to a 3′ splice site that yields a 2nd exon containing a translation termination codon, so no *tra* protein is expressed. In females, the binding of SXL protein blocks the binding of U2AF to this 3′ splice site, resulting in the use of an alternative site further downstream in exon 2. This alternative 3′ splice site is downstream of the translation termination codon, so the mRNA expressed in females directs the synthesis of functional *tra* protein.

vides an important mechanism for tissue-specific and developmental regulation of gene expression.

One well-studied example of tissue-specific alternative splicing is provided by sex determination in *Drosophila*, where alternative splicing of the same pre-mRNA determines whether a fly is male or female (Figure 6.49). Alternative splicing of the pre-mRNA of a gene called *transformer* is controlled by a protein (SXL) that is only expressed in female flies. The *transformer* pre-mRNA has three exons, but a different second exon is incorporated into the mRNA as a result of using alternate 3′ splice sites in the two different sexes. In males, exon 1 is joined to the most upstream of these 3′ splice sites, which is selected by the binding of the U2AF splicing factor to sequences in exon 2. In females, the SXL protein binds to this site in exon 2, blocking the binding of U2AF. Consequently, the upstream 3′ splice site is skipped in females, and exon 1 is instead joined to an alternate 3′ splice site that is further downstream. The exon 2 sequences included in the male *transformer* mRNA contain a translation termination codon, so no protein is produced. This termination codon is not included in the female mRNA, so female flies express functional *transformer* protein, which acts a key regulator of sex determination.

The alternative splicing of *transformer* illustrates the action of a repressor (the SXL protein) that functions by blocking the binding of a splicing factor (U2AF). In other cases, alternative splicing is controlled by activators that recruit splicing factors to splice sites that would otherwise not be recognized. Multiple mechanisms can thus regulate alternative splicing, and variations in alternative splicing make a major contribution to the diversity of proteins expressed during development and differentiation.

RNA Editing

RNA editing refers to RNA processing events (other than splicing) that alter the protein-coding sequences of some mRNAs. This unexpected form of RNA processing was first discovered in mitochondrial mRNAs of trypanosomes, in which U residues are added and deleted at multiple sites along the molecule. More recently, editing has also been described in mitochondrial mRNAs of other organisms, chloroplast mRNAs of higher plants, and nuclear mRNAs of some mammalian genes.

Editing in mammalian nuclear mRNAs, as well as in mitochondrial and chloroplast RNAs of higher plants, involves single base changes as a result of base modification reactions, similar to those involved in tRNA processing. In mammalian cells, RNA editing reactions include the deamination of

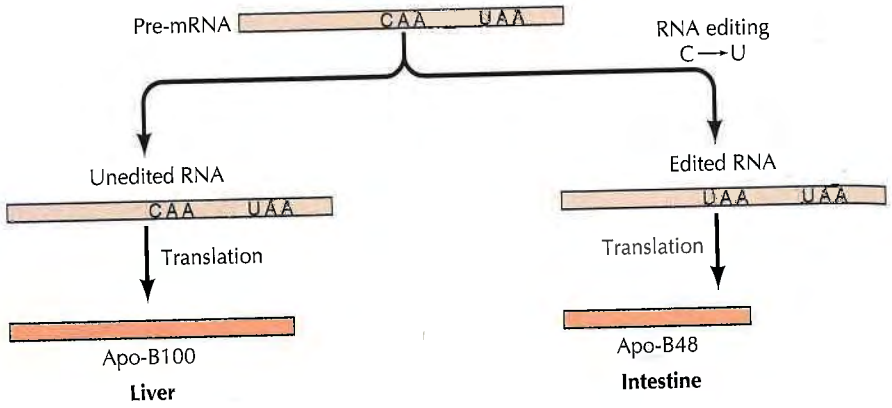


Figure 6.50 Editing of apolipoprotein B mRNA
In human liver, unedited mRNA is translated to yield a 4536-amino-acid protein called Apo-B100. In human intestine, however, the mRNA is edited by a base modification that changes a specific C to a U. This modification changes the codon for glutamine (CAA) to a termination codon (UAA), resulting in synthesis of a shorter protein (Apo-B48, consisting of only 2152 amino acids).

cytosine to uridine and of adenosine to inosine. One of the best-studied examples is editing of the mRNA for apolipoprotein B, which transports lipids in the blood. In this case, tissue-specific RNA editing results in two different forms of apolipoprotein B (Figure 6.50). In humans, Apo-B100 (4536 amino acids) is synthesized in the liver by translation of the unedited mRNA. However, a shorter protein (Apo-B48, 2152 amino acids) is synthesized in the intestine as a result of translation of an edited mRNA in which a C has been changed to a U by deamination. This alteration changes the codon for glutamine (CAA) in the unedited mRNA to a translation termination codon (UAA) in the edited mRNA, resulting in synthesis of the shorter Apo-B protein. Tissue-specific editing of Apo-B mRNA thus results in the expression of structurally and functionally different proteins in liver and intestine. The full-length Apo-B100 produced by the liver transports lipids in the circulation; Apo-B48 functions in the absorption of dietary lipids by the intestine.

RNA editing by the deamination of adenosine to inosine is the most common form of nuclear RNA editing in mammals. This form of editing plays an important role in the nervous system, where A-to-I editing results in single amino acid changes in the receptors for some signaling molecules on the surface of neurons. The importance of this editing reaction has been clearly demonstrated using homologous recombination to inactivate the gene encoding the enzyme responsible for A-to-I editing in mice (see Chapter 3). Mice lacking this enzyme die at a young age after suffering repeated epileptic seizures as a result of dysfunction of the improperly edited receptors.

RNA Degradation

The processing steps discussed in the previous section result in the formation of mature mRNAs, which are then transported to the cytoplasm and function to direct protein synthesis. However, most of the sequences transcribed into pre-mRNA are instead degraded within the nucleus. Over 90% of pre-mRNA sequences are introns, which are degraded within the nucleus following their excision by splicing. This is carried out by an enzyme that recognizes the unique 2′-5′ bond formed at the branchpoint, as well as by enzymes that recognize either the 5′ or 3′ ends of RNA molecules and catalyze degradation of the RNA in either direction. The 5′ and 3′ ends of processed mRNAs are protected from this degradation machinery by capping and polyadenylation, respectively, while the unprotected ends of introns are recognized and degraded.

In addition to degrading introns, cells possess a quality-control system (called **nonsense-mediated mRNA decay**) that leads to the degradation of

mRNAs that lack complete open-reading frames. This eliminates defective mRNA molecules and prevents the synthesis of abnormal truncated proteins. In yeast, nonsense-mediated mRNA decay takes place in the cytoplasm and is triggered when a premature termination codon is encountered by a ribosome during protein synthesis. In mammals, however, at least some nonsense-mediated mRNA decay takes place within the nucleus. The mechanism by which termination codons are recognized within the nucleus of mammalian cells is not yet understood, although some recent studies have suggested that ribosomes within the nucleus could be involved in recognizing and even translating nuclear mRNAs.

What may be considered the final aspect of the processing of an RNA molecule is its eventual degradation in the cytoplasm. Since the intracellular level of any RNA is determined by a balance between synthesis and degradation, the rate at which individual RNAs are degraded is another level at which gene expression can be controlled. Both ribosomal and transfer RNAs are very stable, and this stability largely accounts for the high levels of these RNAs (greater than 90% of all RNA) in both prokaryotic and eukaryotic cells. In contrast, bacterial mRNAs are rapidly degraded, usually having half-lives of only 2 to 3 minutes. This rapid turnover of bacterial mRNAs allows the cell to respond quickly to alterations in its environment, such as changes in the availability of nutrients required for growth. In eukaryotic cells, however, different mRNAs are degraded at different rates, providing an additional parameter to the regulation of eukaryotic gene expression.

The cytoplasmic degradation of most eukaryotic mRNAs is initiated by shortening of their poly-A tails. Then follows removal of the 5′ cap and degradation of the RNA by nucleases acting from both ends. The half-lives of mRNAs in mammalian cells vary from less than 30 minutes to approximately 20 hours. The unstable mRNAs frequently code for regulatory proteins, including certain transcription factors, whose levels within the cell vary rapidly in response to environmental stimuli. These mRNAs often contain specific AU-rich sequences near their 3′ ends that appear to signal rapid degradation by promoting deadenylation.

The stability of some mRNAs can also be regulated in response to extracellular signals. A good example is provided by the mRNA that encodes transferrin receptor—a cell surface protein involved in the uptake of iron by mammalian cells. The amount of transferrin receptor within cells is regulated by the availability of iron, largely as a result of modulation of the stability of its mRNA (Figure 6.51). In the presence of adequate amounts of iron, transferrin receptor mRNA is rapidly degraded as a result of specific nuclease cleavage at a sequence near its 3′ end. If an adequate supply of iron is not available, however, the mRNA is stabilized, resulting in increased synthesis of transferrin receptor and more iron uptake by the cell.

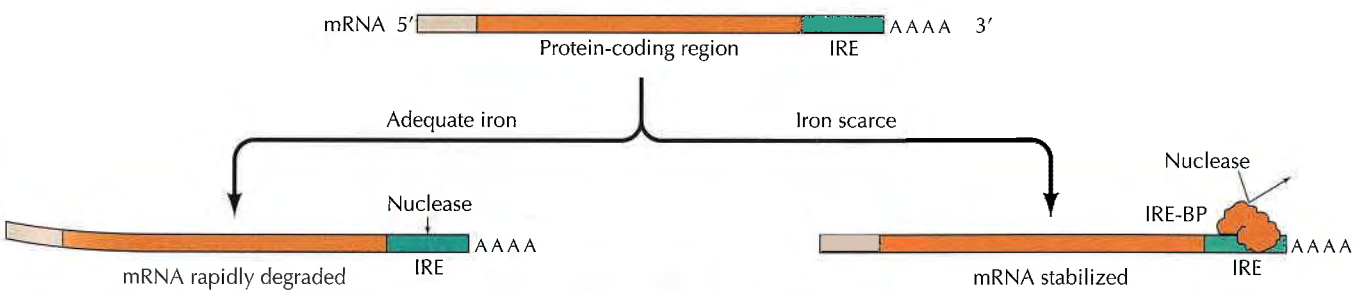


Figure 6.51 Regulation of transferrin receptor mRNA stability
The levels of transferrin receptor mRNA are regulated by the availability of iron. If the supply of iron is adequate, the mRNA is rapidly degraded as a result of nuclease cleavage near the 3′ end. If iron is scarce, however, a regulatory protein (called the iron response element–binding protein, or IRE-BP) binds to a sequence near the 3′ end of the mRNA (the iron response element, or IRE), protecting the mRNA from nuclease cleavage.

This regulation is mediated by a protein that binds to specific sequences (called the iron response element, or IRE) near the 3′ end of transferrin receptor mRNA and protects the mRNA from cleavage. The binding of this regulatory protein to the IRE is in turn controlled by the levels of iron within the cell: If iron is scarce, the protein binds to the IRE and protects transferrin receptor mRNA from degradation. Similar changes in the stability of other mRNAs are involved in the regulation of gene expression by certain hormones. Thus, although transcription remains the primary level at which gene expression is regulated, variations in the rate of mRNA degradation also play an important role in controlling steady-state levels of mRNAs within the cell.

KEY TERMS

RNA polymerase, promoter, footprinting

operon, operator, repressor, *cis*-acting control element, *trans*-acting factor

transcription factor, general transcription factor, TATA box, TATA-binding protein (TBP), TBP-associated factor (TAF), Mediator

enhancer, insulator

SUMMARY

TRANSCRIPTION IN PROKARYOTES

RNA Polymerase and Transcription: *E. coli* RNA polymerase consists of $\alpha, \beta, \beta', \omega$,and σ subunits. Transcription is initiated by the binding of σ to promoter sequences. After synthesis of about the first ten nucleotides of RNA, the core polymerase dissociates from σ and travels along the template DNA as it elongates the RNA chain. Transcription then continues until the polymerase encounters a termination signal.

Repressors and Negative Control of Transcription: The prototype model for gene regulation in bacteria is the *lac* operon, which is regulated by the binding of a repressor to specific DNA sequences within the promoter.

Positive Control of Transcription: Some bacterial genes are regulated by transcriptional activators rather than repressors.

EUKARYOTIC RNA POLYMERASES AND GENERAL TRANSCRIPTION FACTORS

Eukaryotic RNA Polymerases: Eukaryotic cells contain three distinct nuclear RNA polymerases that transcribe genes encoding mRNAs (polymerase II), rRNAs (polymerases I and III), and tRNAs (polymerase III).

General Transcription Factors and Initiation of Transcription by RNA Polymerase II: Eukaryotic RNA polymerases do not bind directly to promoter sequences; they require additional proteins (general transcription factors) to initiate transcription. The promoter sequences of many polymerase II genes are recognized by the TATA-binding protein, which recruits additional transcription factors and RNA polymerase to the promoter.

Transcription by RNA Polymerases I and III: RNA polymerases I and III also require additional transcription factors to bind to the promoters of rRNA, tRNA, and some snRNA genes.

REGULATION OF TRANSCRIPTION IN EUKARYOTES

cis-*Acting Regulatory Sequences: Promoters and Enhancers:* Transcription of eukaryotic genes is controlled by proteins that bind to regulatory sequences, which can be located up to several kilobases away from the transcription start site. Enhancers typically contain binding sites for multiple proteins that work together to regulate gene expression.

Transcriptional Regulatory Proteins: Many eukaryotic transcription factors have been isolated on the basis of their binding to specific DNA sequences.

electrophoretic-mobility shift assay, DNA-affinity chromatography

Structure and Function of Transcriptional Activators: Transcriptional activators are modular proteins, consisting of distinct DNA-binding and activation domains. DNA-binding domains mediate association with specific regulatory sequences; activation domains stimulate transcription by interacting with Mediator proteins and general transcription factors, as well as with coactivators that modify chromatin structure.

transcriptional activator, zinc finger domain, steroid hormone receptor, helix-turn-helix, homeodomain, homeobox, leucine zipper, helix-loop-helix, coactivator

Eukaryotic Repressors: Gene expression in eukaryotic cells is regulated by repressors as well as by activators. Some repressors interfere with the binding of activators or general transcription factors to DNA. Other repressors contain discrete repression domains that inhibit transcription by interacting with either general transcription factors, transcriptional activators, or corepressors that affect chromatin structure.

corepressor

Relationship of Chromatin Structure to Transcription: The packaging of DNA in nucleosomes presents an impediment to transcription in eukaryotic cells. Modification of histones by acetylation increases the accessibility of nucleosomal DNA to transcription factors, and this modification of chromatin is tightly linked to transcriptional regulation. Enzymes that catalyze histone acetylation are associated with transcriptional activators, whereas histone deacetylases are associated with repressors. Histones are also modified by phosphorylation and methylation, and specific modifications of histones affect gene expression by serving as binding sites for other regulatory proteins. In addition, nucleosome remodeling factors facilitate the binding of transcription factors to DNA by altering the arrangement or structures of nucleosomes. RNA polymerase is then able to transcribe through nucleosomes by disrupting histone-DNA contacts. Transcriptional elongation is facilitated by the nonhistone HMGN chromosomal proteins, and by elongation factors that recruit histone acetyltransferases as well as acting directly to disrupt nucleosome structure.

HMGN proteins, histone acetylation, histone code, nucleosome remodeling factor, elongation factor

Regulation of Transcription by Noncoding RNAs: Transcription can be regulated by noncoding RNAs, as well as by regulatory proteins. X chromosome inactivation provides an example of gene regulation by a noncoding RNA in mammals.

X chromosome inactivation

DNA Methylation: Methylation of cytosine residues can inhibit the transcription of vertebrate genes. Regulation of gene expression by methylation plays an important role in genomic imprinting, which controls the transcription of some genes involved in mammalian development.

genomic imprinting

RNA PROCESSING AND TURNOVER

Processing of Ribosomal and Transfer RNAs: Ribosomal and transfer RNAs are derived by cleavage of long primary transcripts in both prokaryotic and eukaryotic cells. Methyl groups are added to rRNAs, and various bases are modified in tRNAs.

pre-rRNA, pre-tRNA, RNase P, ribozyme

Processing of mRNA in Eukaryotes: Eukaryotic pre-mRNAs are modified by the addition of 7–methylguanosine caps and 3′ poly-A tails, in addition to the removal of introns by splicing.

pre-mRNA, 7-methylguanosine cap, poly-A tail, polyadenylation

spliceosome, small nuclear RNA (snRNA), snRNP, self-splicing

Splicing Mechanisms: Splicing of nuclear pre-mRNAs takes place in large complexes, called spliceosomes, composed of proteins and small nuclear RNAs (snRNAs). The snRNAs recognize sequences at the splice sites of pre-mRNAs and catalyze the splicing reaction. Some mitochondrial, chloroplast, and bacterial RNAs undergo self-splicing, in which the splicing reaction is catalyzed by intron sequences.

alternative splicing

Alternative Splicing: Exons can be joined in various combinations as a result of alternative splicing, which provides an important mechanism for tissue-specific control of gene expression in complex eukaryotes.

RNA editing

RNA Editing: Some mRNAs are modified by processing events that alter their protein-coding sequences. Editing of mitochondrial mRNAs in some protozoans involves the addition and deletion of U residues at multiple sites in the molecule. Other forms of RNA editing in plant and mammalian cells involve the modification of specific bases.

nonsense-mediated mRNA decay

RNA Degradation: Introns are degraded within the nucleus, and abnormal mRNAs lacking complete open-reading frames are eliminated by nonsense-mediated mRNA decay. Functional mRNAs in eukaryotic cells are degraded at different rates, providing an additional mechanism for control of gene expression. In some cases, rates of mRNA degradation are regulated by extracellular signals.

Questions

1. How does lactose induce the expression of the proteins that are needed for *E. coli* to take up and metabolize lactose?

2. The consensus sequence of the *E. coli* –10 promoter element is TATAAT. You are comparing two promoters that have –10 element sequences of TATGAT and CATGAT, respectively. Which would you expect to be transcribed more efficiently?

3. You are working with two strains of *E. coli*. One contains a wild-type β-galactosidase gene and an i^- mutation; the other contains a temperature-sensitive β-galactosidase gene and an o^c mutation. After mating these strains, you assay for the production of β-galactosidase at both permissive and nonpermissive temperatures in the absence of lactose. What do you expect to find?

4. How does DNA footprinting show where a protein binds to a specific DNA sequence?

5. What is the role of sigma (σ) factors in bacterial RNA synthesis?

6. How is an *E. coli* mRNA terminated?

7. Eukaryotic cells have three distinct RNA polymerases. Which RNAs does each transcribe?

8. You are comparing the requirements for *in vitro* basal transcription of two polymerase II genes, one containing a TATA box and the other containing only an Inr sequence. Does transcription from these promoters require TBP or TFIID?

9. How do enhancers differ from promoters as *cis*-acting regulatory sequences in eukaryotes?

10. You are studying the enhancer of a gene that normally is expressed only in neurons. Constructs in which this enhancer is linked to a reporter gene are expressed in neuronal cells but not in fibroblasts. However, if you mutate a specific sequence element within the enhancer, you find expression in both fibroblasts and neuronal cells. What type of regulatory protein would you expect to bind to that enhancer element?

11. A transcription factor is found to activate transcription by binding to different DNA sequences in muscle cells and liver cells. How might alternative splicing be involved in determining this tissue specificity of activator function?

12. What are the functions of insulators?

13. Explain the mechanism of X chromosome inactivation in human females.

References and Further Reading

Transcription in Prokaryotes

Busby, S. and R. H. Ebright. 1994. Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell* 79: 743–746. [R]

Darst, S. A. 2001. Bacterial RNA polymerase. *Curr. Opin. Struc. Biol.* 11: 155–162. [R]

Gilbert, W. and B. Muller-Hill. 1966. Isolation of the *lac* repressor. *Proc. Natl. Acad. Sci. USA* 56: 1891–1899. [P]

Hochschild, A. and S. L. Dove. 1998. Protein-protein contacts that activate and repress prokaryotic transcription. *Cell* 92: 597–600. [R]

Jacob, F. and J. Monod. 1961. Genetic and regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3: 318–356. [P]

Lewis, M., G. Chang, N. C. Horton, M. A. Kirchner, H. C. Pace, M. A. Schumacher, R. G. Brennan and P. Lu. 1996. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* 271: 1247–1254. [P]

Murakami, K. S., S. Masuda and S. A. Darst. 2002. Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 Å resolution. *Science* 296: 1280–1284. [P]

Ptashne, M. and A. Gann. 2002. *Genes and Signals*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Uptain, S. M., C. M. Kane and M. J. Chamberlin. 1997. Basic mechanisms of transcript elongation and its regulation. *Ann. Rev. Biochem.* 66: 117–172. [R]

Vassylyev, D. G., S. Sekine, O. Laptenko, J. Lee, M. N. Vassylyeva, S. Borukhov and S. Yokoyama. 2002. Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution. *Nature* 417: 712–719. [P]

Zhang, G., E. A. Campbell, E. A., L. Minakhin, C. Richter, K. Severinov and S. A. Darst. 1999. Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. *Cell* 98: 811–824. [P]

Eukaryotic RNA Polymerases and General Transcription Factors

Butler, J. E. F. and J. T. Kadonaga. 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* 16: 2583–2592. [R]

Conaway, J. W., A. Shilatifard, A. Dvir and R. C. Conaway. 2000. Control of elongation by RNA polymerase II. *Trends Biochem. Sci.* 25: 375–380. [R]

Cramer, P., D. A. Bushnell, J. Fu, A. L. Gnatt, B. Maier-Davis, N. E. Thompson, R. R. Burgess, A. M. Edwards, P. R. David and R. D. Kornberg. 2000. Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* 288: 640–649. [P]

Cramer, P., D. A. Bushnell and R. D. Kornberg. 2001. Structural basis of transcription: RNA polymerase II at 2.8 Å resolution. *Science* 292: 1863–1876. [P]

Ebright, R. H. 2000. RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J. Mol. Biol.* 304: 687–698. [R]

Lee, T. I. and R. A. Young. 2000. Transcription of eukaryotic protein-coding genes. *Ann. Rev. Genet.* 34: 77–137. [R]

Matsui, T., J. Segall, P. A. Weil and R. G. Roeder. 1980. Multiple factors are required for accurate initiation of transcription by purified RNA polymerase II. *J. Biol. Chem.* 255: 11992–11996. [P]

Myers, L. C. and R. D. Kornberg. 2000. Mediator of transcriptional regulation. *Ann. Rev.Biochem.* 69: 729–749. [R]

Naar, A. M., B. D. Lemon and R. Tjian. 2001. Transcriptional coactivator complexes. *Ann. Rev. Biochem.* 70: 475–501. [R]

Nikolov, D. B., H. Chen, E. D. Halay, A. A. Usheva, K. Hisatake, D. K. Lee, R. G. Roeder and S. K. Burley. 1995. Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature* 377: 119–128. [P]

Orphanides, G. and D. Reinberg. 2002. A unified theory of gene expression. *Cell* 108: 439–451. [R]

Schramm, L. and N. Hernandez. 2002. Recruitment of RNA polymerase III to its target promoters. *Genes Dev.* 16: 2593–2620. [R]

Weil, P. A., D. S. Luse, J. Segall and R. G. Roeder. 1979. Selective and accurate transcription at the Ad2 major late promoter in a soluble system dependent on purified RNA polymerase II and DNA. *Cell* 18: 469–484. [P]

Woychik, N. A. and M. Hampsey. 2002. The RNA polymerase II machinery: structure illuminates function. *Cell* 108: 453–463. [R]

Regulation of Transcription in Eukaryotes

Atchison, M. L. 1988. Enhancers: Mechanisms of action and cell specificity. *Ann. Rev. Cell. Biol.* 4: 127–153. [R]

Banerji, J., S. Rusconi and W. Schaffner. 1981. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27: 299–308. [P]

Berger, S. L. 2002. Histone modifications in transcriptional regulation. *Curr. Opin. Genet. Dev.* 12: 142–148. [R]

Bird, A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev.* 16: 6–21. [R]

Brent, R. and M. Ptashne. 1985. A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell* 43: 729–736. [P]

Brownell, J. E., J. Zhou, T. Ranalli, R. Kobayashi, D. G. Edmondson, S. Y. Roth and C. D. Allis. 1996. *Tetrahymena* histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell* 84: 843–851. [P]

Bustin, M. 2001. Chromatin unfolding and activation by HMGN chromosomal proteins. *Trends Biochem. Sci.* 26: 431–437. [R]

Cohen, D. E. and J. T. Lee. 2002. X-chromosome inactivation and the search for chromosome-wide silencers. *Curr. Opin. Genet. Dev.* 12: 219–224. [R]

Conaway, J. W., A. Shilatifard, A. Dvir and R. C. Conaway. 2000. Control of elongation by RNA polymerase II. *Trends Biochem. Sci.* 25: 375–380. [R]

Courey, A. J. and S. Jia. 2001. Transcriptional repression: the long and the short of it. *Genes Dev.* 15: 2786–2796. [R]

Dynan, W. S. and R. Tjian. 1983. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* 35: 79–87. [P]

Ferguson-Smith, A. C. and M. A. Surani. 2001. Imprinting and the epigenetic asymmetry between parental genomes. *Science* 293: 1086–1089. [R]

Hanna-Rose, W. and U. Hansen. 1996. Active repression mechanisms of eukaryotic transcription repressors. *Trends Genet.* 12: 229–234. [R]

Horn, P. J. and C. L. Peterson. 2002. Chromatin higher order folding: wrapping up transcription. *Science* 297: 1824–1827. [R]

Jenuwein, T. and C. D. Allis. 2001. Translating the histone code. *Science* 293: 1074–1080. [R]

Kadonaga, J. T. 1998. Eukaryotic transcription: An interlaced network of transcription factors and chromatin-modifying machines. *Cell* 92: 307–313. [R]

Kadonaga, J. T. and R. Tjian. 1986. Affinity purification of sequence-specific DNA binding proteins. *Proc. Natl. Acad. Sci. USA* 83: 5889–5893. [P]

Lee, T. I. and R. A. Young. 2000. Transcription of eukaryotic protein-coding genes. *Ann. Rev. Genet.* 34: 77–137. [R]

Licht, J. D., M. J. Grossel, J. Figge and U. M. Hansen. 1990. *Drosophila Krüppel* protein is a transcriptional repressor. *Nature* 346: 76–79. [P]

Matzke, M., A. J. M. Matzke and J. M. Kooter. 2001. RNA: guiding gene silencing. *Science* 293: 1080–1083. [R]

McKenna, N. J. and B. W. O'Malley. 2002. Combinatorial control of gene expression by nuclear receptors and coregulators. *Cell* 108: 465–474. [R]

McKnight, S. L. and R. Kingsbury. 1982. Transcriptional control signals of a eukaryotic protein-coding gene. *Science* 217: 316–324. [P]

Naar, A. M., B. D. Lemon and R. Tjian. 2001. Transcriptional coactivator complexes. *Ann. Rev. Biochem.* 70: 475–501. [R]

Narlikar, G. J., H.-Y. Fan and R. E. Kingston. 2002. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* 108: 475–487. [R]

Orphanides, G. and D. Reinberg. 2000. RNA polymerase II elongation through chromatin. *Nature* 407: 471–475. [R]

Pabo, C. O. and R. T. Sauer. 1992. Transcription factors: Structural families and principles of DNA recognition. *Ann. Rev. Biochem.* 61: 1053–1095. [R]

Panning, B. and R. Jaenisch. 1998. RNA and the epigenetic regulation of X chromosome inactivation. *Cell* 93: 305–308. [R]

Ptashne, M. and A. Gann. 2002. *Genes and Signals.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Richards, E. J. and S. C. R. Elgin. 2002. Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell* 108: 489–500. [R]

Schreiber, S. L. and B. E. Bernstein. 2002. Signaling network model of chromatin. *Cell* 111: 771–778. [R]

Staudt, L. M. and M. J. Lenardo. 1991. Immunoglobulin gene transcription. *Ann. Rev. Immunol.* 9: 373–398. [R]

Taunton, J., C. A. Hassig and S. L. Schreiber. 1996. A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science* 272: 408–411. [P]

Tilghman, S. M. 1999. The sins of the fathers and mothers: Genomic imprinting in mammalian development. *Cell* 96: 185–193. [R]

Turner, B. M. 2002. Cellular memory and the histone code. *Cell* 111: 285–291. [R]

Volpe, T. A., C. Kidner, I. M. Hall, G. Teng, S. I. S. Grewal and R. A. Martienssen. 2002. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* 297: 1833–1837. [P]

West, A. G., M. Gaszner and G. Felsenfeld. 2002. Insulators: many functions, many mechanisms. *Genes Dev.* 16: 271–288. [R]

Wolffe, A. 1998. *Chromatin: Structure and Function.* 3rd. ed. New York: Academic Press.

RNA Processing and Turnover

Abelson, J., C. R. Trotta and H. Li. 1998. tRNA splicing. *J. Biol. Chem.* 273:12685–12688. [R]

Baker, B. S. 1989. Sex in flies: the splice of life. *Nature* 340: 521–524. [R]

Blanc, V. and N. O. Davidson. 2003. C-to-U RNA editing: mechanisms leading to genetic diversity. *J. Biol. Chem.* 278: 1395–1398. [R]

Cech, T. R. 1990. Self-splicing of group I introns. *Ann. Rev. Biochem.* 59: 543–568. [R]

Dodson, R. E. and D. J. Shapiro. 2002. Regulation of pathways of mRNA destabilization and stabilization. *Prog. Nucl. Acid Res.* 72: 129–164. [R]

Frank, D. N. and N. R. Pace. 1998. Ribonuclease P: Unity and diversity in a tRNA processing ribozyme. *Ann. Rev. Biochem.* 67: 153–180. [R]

Guerrier-Takada, C., K. Gardiner, T. Marsh, N. Pace and S. Altman. 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* 35: 849–857. [P]

Hirose, Y. and J. L. Manley. 2000. RNA polymerase II and the integration of nuclear events. *Genes Dev.* 14: 1415–1429. [R]

Hopper, A. K. and E. M. Phizicky. 2003. tRNA transfers to the limelight. *Genes Dev.* 17: 162–180. [R]

Klausner, R. D., T. A. Rouault and J. B. Harford. 1993. Regulating the fate of mRNA: The control of cellular iron metabolism. *Cell* 72: 19–28. [R]

Kruger, K., P. J. Grabowski, A. Zaug, A. J. Sands, D. E. Gottschling and T. R. Cech. 1982. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell* 31: 147–157. [P]

Maas, S., A. Rich and K. Nishikura. 2003. A-to-I RNA editing: recent news and residual mysteries. *J. Biol. Chem.* 278: 1391–1394. [R]

Madison-Antenucci, S., J. Grams and S. L. Hajduk. 2002. Editing machines: the complexities of trypanosome RNA editing. *Cell* 108: 435–438. [R]

Maniatis, T. and R. Reed. 2002. An extensive network of coupling among gene expression machines. *Nature* 416: 499–506. [R]

Maniatis, T. and B. Tasic. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418: 236–243. [R]

Moore, M. J. 2002. Nuclear RNA turnover. *Cell* 108: 431–434. [R]

Padgett, R. A., M. M. Konarska, P. J. Grabowski, S. F. Hardy and P. A. Sharp. 1984. Lariat RNAs as intermediates and products in the splicing of messenger RNA precursors. *Science* 225: 898–903. [P]

Padgett, R. A., S. M. Mount, J. A. Steitz and P. A. Sharp. 1983. Splicing of messenger RNA precursors is inhibited by antisera to small nuclear ribonucleoprotein. *Cell* 35: 101–107. [P]

Proudfoot, N. J., A Furger and M. J. Dye. 2002. Integrating mRNA processing with transcription. *Cell* 108: 501–512. [R]

Ruskin, B., A. R. Krainer, T. Maniatis and M. R. Green. 1984. Excision of an intact intron as a novel lariat structure during pre-mRNA splicing *in vitro*. *Cell* 38: 317–331. [P]

Smith, C. W. J. and J. Valcarcel. 2000. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* 25: 381–388. [R]

Valadkhan, S. and J. L. Manley. 2001. Splicing-related catalysis by protein-free snRNAs. *Nature* 413: 701–7078. [P]

Van Hoof, A. and R. Parker. 2002. Messenger RNA degradation: beginning at the end. *Curr. Biol.* 12: R285–R287. [R]

Villa, T., J. A. Pleiss and C. Guthrie. 2002. Spliceosomal snRNAs: Mg^{2+}-dependent chemistry at the catalytic core? *Cell* 109: 149–152. [R]